

Syracuse University

**SURFACE**

---

Dissertations - ALL

SURFACE

---

May 2016

## FINE-GRAINED EMOTION DETECTION IN MICROBLOG TEXT

Jasy Suet Yan Liew  
*Syracuse University*

Follow this and additional works at: <https://surface.syr.edu/etd>



Part of the [Social and Behavioral Sciences Commons](#)

---

### Recommended Citation

Liew, Jasy Suet Yan, "FINE-GRAINED EMOTION DETECTION IN MICROBLOG TEXT" (2016). *Dissertations - ALL*. 440.

<https://surface.syr.edu/etd/440>

This Dissertation is brought to you for free and open access by the SURFACE at SURFACE. It has been accepted for inclusion in Dissertations - ALL by an authorized administrator of SURFACE. For more information, please contact [surface@syr.edu](mailto:surface@syr.edu).

# Abstract

Automatic emotion detection in text is concerned with using natural language processing techniques to recognize emotions expressed in written discourse. Endowing computers with the ability to recognize emotions in a particular kind of text, microblogs, has important applications in sentiment analysis and affective computing. In order to build computational models that can recognize the emotions represented in tweets we need to identify a set of suitable emotion categories. Prior work has mainly focused on building computational models for only a small set of six basic emotions (happiness, sadness, fear, anger, disgust, and surprise). This thesis describes a taxonomy of 28 emotion categories, an expansion of these six basic emotions, developed inductively from data. This set of 28 emotion categories represents a set of fine-grained emotion categories that are representative of the range of emotions expressed in tweets, microblog posts on Twitter.

The ability of humans to recognize these fine-grained emotion categories is characterized using inter-annotator reliability measures based on annotations provided by expert and novice annotators. A set of 15,553 human-annotated tweets form a gold standard corpus, EmoTweet-28. For each emotion category, we have extracted a set of linguistic cues (i.e., punctuation marks, emoticons, emojis, abbreviated forms, interjections, lemmas, hashtags and collocations) that can serve as salient indicators for that emotion category.

We evaluated the performance of automatic classification techniques on the set of 28 emotion categories through a series of experiments using several classifier and feature combinations. Our results shows that it is feasible to extend machine learning classification to fine-grained emotion detection in tweets (i.e., as many as 28 emotion categories) with results that are comparable to state-of-the-art classifiers that detect six to eight basic emotions in text. Classifiers using features extracted from the linguistic cues associated with each category equal or better the performance of conventional corpus-based and lexicon-based features for fine-grained emotion classification.

This thesis makes an important theoretical contribution in the development of a taxonomy of emotion in text. In addition, this research also makes several practical contributions, particularly in the creation of language resources (i.e., corpus and lexicon) and machine learning models for fine-grained emotion detection in text.

# FINE-GRAINED EMOTION DETECTION IN MICROBLOG TEXT

by

Jasy Liew Suet Yan

B.S. Computer Science, Universiti Sains Malaysia, 2008  
M.S. Information Management, Syracuse University, 2011

Dissertation

Submitted in partial fulfillment of the requirements for the degree of  
Doctor of Philosophy in *Information Science and Technology*

Syracuse University  
May 2016

Copyright © Jasy Liew Suet Yan 2016  
All Rights Reserved

# Acknowledgments

I would like to express my deepest appreciation to my advisor, Howard Turtle, for his thoughtful insights, his guidance, his patience and his relentless faith in me throughout this dissertation. To Elizabeth Liddy, my co-advisor, I am grateful for her encouragement, inspiration and mentorship. They are the closest confidants in my academic family and I could not imagine completing this dissertation without their immense knowledge and tireless support.

I am very fortunate to have Ping Zhang, Nancy McCracken, Jennifer Stromer-Galley, Bei Yu, Diana Inkpen and Michael Kalish serve on my committee. I thank them for carefully reviewing my dissertation, providing insightful comments and challenging me with hard questions and criticism to help widen my perspective. I would also like to acknowledge Phyllis Turtle for her editorial comments on the dissertation.

Special thanks to the students who volunteered their time to perform the annotation task and worked on building the tools to support my research: Rudy Rusli, Olivia Rhinehart, Tuo Wu, Jiaqi Li, Mohammed Hassan Aldrees, Tim Fu, Ke Ding, Shujin Cao, Yang Liu, Dane Dell, Feifei Zhang, Rucha Somani, Venkata Chadalawada, Anusha Ambati, Yiwei Jin, Yueming Sun, Ruby Cuate Ayala, Sharon Lee, Brian Dobreski and Aravind Gopalakrishnan. Their voluntary spirit, dedication and contribution to knowledge are commendable.

I am also indebted to all faculty members who have helped me grown into a confident scholar and teacher as well as to the administrative staff who have helped made my journey in the program a smooth and successful one. Many thanks to my mentors, colleagues, peers and friends for the constant boost in morale and social support. Particularly, I wish to extend my heartfelt appreciation to my cohort mates, Katie DeVries Hassman and Douglas Crescenzi. I am blessed to be given the opportunity to grow, laugh and cry with them throughout my journey in

the PhD program. Special thanks to Joy Ying Tang and Stephanie Santoso for making my life as a PhD student especially colorful and memorable.

This dissertation is made possible by the joint financial support from Universiti Sains Malaysia and the Ministry of Higher Education Malaysia. I thank them for investing in the future of the university and the country. I would also like to express my gratitude to Christine Larsen for funding part of the data collection in this research under the Liddy Fellowship.

Last but not least, I would like to thank my family especially my parents and sisters who stood by me and cheered me on. I also want to specially dedicate this dissertation to my niece and nephew, who provided me with generous doses of laughter all throughout the dissertation writing process. Most of my determination to complete the dissertation came from the hope that their generation and many others to come will also benefit from this research.

# Table of Content

<b>Abstract .....</b>	<b>i</b>
<b>Chapter 1: Introduction .....</b>	<b>1</b>
1.1    Motivation .....	1
1.2    Problem Definition.....	7
1.3    Research Goals .....	12
1.4    Research Questions .....	12
1.5    Thesis Overview .....	13
<b>Chapter 2: Literature Review .....</b>	<b>15</b>
2.1    Introduction .....	15
2.2    Emotion Theories: The Psychology of Emotion .....	16
2.2.1    The Darwinian Perspective: Emotion as Expression .....	16
2.2.2    The Jamesian Perspective: Emotion as Embodiment .....	20
2.2.3    The Cognitive Perspective: Emotion as Appraisal .....	21
2.2.4    The Social Constructivist Perspective: Emotion as Social Constructs .....	25
2.3    Models of Emotion .....	27
2.3.1    Categorical Model .....	27
2.3.2    Dimensional Model .....	28
2.4    Application of Emotion Theories in Automatic Emotion Detection in Text: A Summary	29
2.5    Distinguishing Emotion from Related Terms .....	32
2.5.1    Emotion and Subjectivity .....	33
2.5.2    Emotion and Sentiment.....	33

2.5.3	Emotion and Affect .....	35
2.5.4	Hierarchical Organization of Related Terms.....	36
2.6	Conceptualizing Emotion in Text.....	37
2.7	Approaches to Identify Emotions in Text .....	40
2.8	Linguistic Representations of Emotion .....	42
2.9	Automatic Emotion Detection in Text .....	43
2.9.1	Lexicon-based Approach .....	44
2.9.2	Learning-based Approach.....	48
2.9.2.1	Supervised Learning .....	48
2.9.2.2	Unsupervised Learning .....	59
2.9.3	Manually Constructed Rules .....	60
2.9.4	Ontology-based Approach .....	62
2.9.5	Hybrid Approach .....	64
2.10	Automatic Emotion Detection on Twitter .....	66
2.11	Conclusion.....	72
<b>Chapter 3: Methodology</b>	.....	<b>73</b>
3.1	Introduction.....	73
3.2	Data Collection .....	74
3.2.1	Random Sampling [RANDOM].....	76
3.2.2	Sampling by Topic [TOPIC].....	76
3.2.3	Sampling by User .....	77
3.3	Corpus Development .....	78
3.3.1	Phase 1: Small-scale Content Analysis.....	78



3.3.1.1	Task 1: Inductive Coding .....	79
3.3.1.2	Task 2: Card Sorting .....	84
3.3.1.3	Task 3: Emotion Word Rating .....	86
3.3.2	Phase 2: Large-scale Content Analysis.....	95
3.4	Phase 3: Machine Learning Experiments.....	102
3.4.1	Task 1: Classifier-related Experiments.....	104
3.4.2	Task 2: Feature-related Experiments .....	109
3.4.2.1	Corpus-based Features .....	110
3.4.2.2	Lexicon-based Features .....	110
3.4.2.3	Cue-based Features .....	111
3.4.2.4	Cross-Group Combinations.....	113
3.4.3	Task 3: Sample-related Experiments .....	113
3.5	Conclusion .....	114
<b>Chapter 4: Conceptual Analysis .....</b>		<b>115</b>
4.1	Class Distinctiveness .....	116
4.1.1	Emotion Categories: Set-48.....	116
4.1.2	Emotion Categories: Set-28.....	125
4.2	Level of Agreement.....	131
4.2.1	Overall Inter-annotator Agreement.....	131
4.2.2	Category Level Agreement .....	137
4.3	Class Intuitiveness .....	139
4.4	Summary: Human Recognition of Emotion Categories in Tweets .....	142
4.5	EmoTweet-28: Corpus Characteristics.....	144

4.5.1	Emotion Distributions .....	144
4.5.2	Multiple Emotions in a Tweet .....	147
4.5.3	Emotion Expressions and Descriptions in Tweets .....	149
4.6	Emotion Cues .....	151
4.6.1	Emotion Cue Characteristics.....	151
4.7	Linguistic Analysis.....	153
4.7.1	Lexical Diversity and Density .....	154
4.7.2	Lexical Uniqueness.....	156
4.7.3	Lexical Indicators .....	159
4.7.3.1	Symbols.....	160
4.7.3.2	Lemmas.....	163
4.7.3.3	Hashtags .....	168
4.7.3.4	Collocations .....	172
4.7.3.5	Part-of-speech (POS) Tags.....	180
4.8	Summary: Salient Linguistic Cues for the Emotion Categories.....	182
4.9	Conclusion.....	183
<b>Chapter 5: Machine Learning Results .....</b>		<b>185</b>
5.1	Definition of Terms and Evaluation Metrics .....	186
5.2	Task 1: Classifier-related Experiments.....	187
5.2.1	Comparison with Baseline.....	188
5.2.2	Tiered Model Results .....	190
5.2.2.1	Levels of Granularity .....	190
5.2.2.2	Flat versus Hierarchical Classification.....	192

5.2.3	Ensemble Methods .....	193
5.2.4	Summary: Classifier-related Experiments .....	197
5.3	Task 2: Feature-related Experiments .....	197
5.3.1	Comparison of P1 and P2 Data .....	198
5.3.2	Comparison across Features .....	202
5.3.2.1	Term Weights .....	202
5.3.2.2	One-Size-Fits-All Model .....	203
5.3.2.3	Custom Model.....	214
5.3.3	Summary: Feature-related Experiments.....	219
5.4	Task 3: Sample-related Experiments .....	219
5.4.1	Class Imbalance Strategies .....	219
5.4.2	Effects of Sampling Strategies on Classifier Performance .....	224
5.4.3	Summary: Sample-related Experiments.....	227
5.5	Discussion .....	227
5.5.1	Comparing Machine Learning Classification Performance .....	227
5.5.2	Comparing Human and Machine Classification Performance.....	230
5.6	Conclusion.....	234
<b>Chapter 6: Conclusion and Future Work.....</b>		<b>237</b>
6.1	Contributions.....	237
6.1.1	Theoretical.....	237
6.1.2	Language Resources.....	238
6.1.3	Machine Learning .....	239
6.2	Limitations .....	240
6.3	Challenges and Future Work .....	242

6.4 Conclusion.....	247
<b>Appendix A.....</b>	<b>249</b>
<b>Appendix B.....</b>	<b>252</b>
<b>Appendix C.....</b>	<b>253</b>
<b>Appendix D.....</b>	<b>254</b>
<b>Appendix E.....</b>	<b>273</b>
<b>References .....</b>	<b>279</b>

# List of Tables

Table 2.1: Emotion and its associated action tendency and function .....	17
Table 2.2: Emotion and its core relational theme .....	22
Table 2.3: Summary of commonly-used emotion theories and models among sentiment analysis researchers .....	31
Table 2.4: Comparison between dictionary and common literature definitions of sentiment, affect, and emotion .....	32
Table 2.5: Examples of explicit emotion cues .....	42
Table 2.6: Descriptions and examples of implicit emotion cues .....	43
Table 2.7: List of popular emotion/affect lexicons .....	45
Table 2.8: Examples of emotion corpora .....	49
Table 2.9: Feature sets for supervised machine learning .....	53
Table 3.1: Distribution of tweets for 4 samples .....	75
Table 3.2: Description of topics included in TOPIC .....	76
Table 3.3: Classification schemes for four facets of emotion .....	79
Table 3.4: Demographic information of annotators in Phase 1 .....	81
Table 3.5: Number of categories proposed by each card sorting team.....	85
Table 3.6: Mapping between 50 emotion words and ANEW words .....	89
Table 3.7: Mapping between the final set of 28 emotion categories to the original set of 48.....	94
Table 3.8: Batches of annotation contributed by AMT workers and volunteers.....	101
Table 3.10: Micro-averaged F1 of BayesNet, SMO, J48 and KNN for the three experimental setups in Task 1 .....	106
Table 3.11: Frequency counts based on best performing classifier (F1) for each emotion category.....	107
Table 3.12: Feature groups and the description of features .....	109

Table 4.1: Emotion tags associated with each of the 17 positive emotion categories .....	117
Table 4.2: Emotion tags associated with each of the 10 neutral emotion categories .....	117
Table 4.3: Emotion tags associated with each of the 21 negative emotion categories.....	118
Table 4.4: Distinctiveness of 48 emotion categories based on $\kappa$ .....	119
Table 4.5: General description of 28 emotion categories.....	127
Table 4.6: Distinctiveness of 28 emotion categories based on $\kappa$ .....	128
Table 4.7: Emotion categories with high, medium and low levels of class distinctiveness .....	130
Table 4.8: Inter-annotator agreement statistics for emotion/non-emotion, valence, arousal, emotion category, and emotion cue .....	133
Table 4.9: Agreement per round for 28 emotion categories among expert annotators.....	134
Table 4.10: Inter-annotator reliability statistics from related work on emotion annotation .....	136
Table 4.11: Proportion of full, partial and no agreement for 28 emotion categories among three annotators.....	137
Table 4.12: Proportion of full agreement (FA) for 28 emotion categories (Cat) .....	138
Table 4.13: Emotion categories with high, medium and low levels of full agreement.....	139
Table 4.14: Percent matches and deviation between annotator labels and gold labels .....	139
Table 4.15: Mean pairwise agreement between annotator and gold labels per emotion category .....	141
Table 4.16: Emotion categories with high, medium and low levels of intuitiveness.....	142
Table 4.17: Triangulation of measures to determine the emotion categories humans can detect in microblog text.....	143
Table 4.18: Word composition of the four samples in the corpus .....	144
Table 4.19: Distribution of emotional and non-emotional tweets.....	145
Table 4.20: Distribution of tweets based on emotion valence .....	145
Table 4.21: Frequency distribution of all emotion categories in the corpus.....	147
Table 4.22: Distribution of tweets containing single and multiple emotion categories .....	148

Table 4.23: Emotion cue and segment statistics .....	151
Table 4.24: Composition of token types .....	152
Table 4.25: Composition of POS tags .....	153
Table 4.26: Lexical composition for each emotion category .....	155
Table 4.27: Symbols associated with each emotion category .....	162
Table 4.28: Frequent canonical form of words and abbreviations for each emotion category ..	165
Table 4.29: Primary and secondary indicators of each emotion category .....	166
Table 4.30: Hashtags associated with each emotion category .....	169
Table 4.31: Top ranking bigrams and trigrams based on frequency and PMI .....	174
Table 4.32: Common collocated words or phrases associated with each emotion category ....	177
Table 4.33: Negated phrases used to affirm and negate emotive meaning .....	180
Table 4.34: POS tag composition based on content words in each emotion category .....	181
Table 5.1: Comparison between basic models and baselines for emotion classification.....	189
Table 5.2: Accuracy (A), precision (P), recall (R) and F1 across classification schemes with different levels of granularity .....	191
Table 5.3: Overall precision, recall and F1 using boosting, bagging, voting and stacking .....	194
Table 5.4: F1 of each emotion category based on boosting and bagging .....	195
Table 5.5: F1 of each emotion category based on voting and stacking .....	196
Table 5.6: Precision of basic SMO and BayesNet classifiers across P1, P2 and P1+P2 .....	199
Table 5.7: Recall of basic SMO and BayesNet classifiers across P1, P2 and P1+P2 .....	200
Table 5.8: F1 of basic SMO and BayesNet classifiers across P1, P2 and P1+P2 .....	201
Table 5.9: F1 of basic SMO and BayesNet classifiers based on four weighting schemes .....	202
Table 5.10: Precision, recall and F1 of classifiers based on feature sets.....	203
Table 5.11: F1 of each emotion category for SMO and BayesNet using corpus-based features .....	206
Table 5.12: F1 of each emotion category based on SMO using lexicon-based features.....	208

Table 5.13: F1 of each emotion category based on BayesNet using lexicon-based features...	209
Table 5.14: F1 of each emotion category based on SMO using cue-based features .....	210
Table 5.15: F1 of each emotion category based on BayesNet using cue-based features.....	211
Table 5.16: Comparing F1 of each emotion category in the best one-size-fits-all model to the best model per category.....	213
Table 5.17: Overall precision, recall and F1 based on custom cue features .....	215
Table 5.18: F1 of each emotion category based on SMO custom model.....	216
Table 5.19: Precision, recall and F1 for each emotion category between E2 and E6 .....	217
Table 5.20: Precision, recall and F1 for SMO classifier based downsampled data evaluated using cross validation and train/test split .....	220
Table 5.21: Precision, recall and F1 for BayesNet classifier based downsampled data evaluated using cross validation and train/test split .....	221
Table 5.22: Precision, recall and F1 of SMO and BayesNet trained with different samples .....	225
Table 5.23: F1 of SMO based on testing classifiers trained and tested based on different subsamples.....	226
Table 5.24: F1 of BayesNet based on testing classifiers trained and tested based on different subsamples.....	226
Table 5.25: Comparing F1 of each emotion category between SMO-E6 and the best performing binary classifier per category.....	228
Table 5.26: Emotion categories with high, medium and low performance .....	229
Table 5.27: Comparing human and machine annotation performance .....	233
Table 5.28: Classifier performance in the state-of-the-art of automatic emotion detection in tweets .....	235
Table C.1: Frequency distribution of emotion categories for each sample.....	253



# List of Figures

Figure 2.1: Ekman, Plutchik, and Izard's basic emotions .....	18
Figure 2.2: Global structure of emotion types.....	23
Figure 2.3: Commonly-used terms related to emotion in sentiment analysis .....	36
Figure 2.4: Concepts related to emotion in text.....	38
Figure 3.1: Overview of three-phase study.....	74
Figure 3.2: Four-step procedure in the card sorting activity .....	84
Figure 3.3: List of 48 emotion categories.....	86
Figure 3.4: Scaling coordinates for 28 affect words in two-dimensional space .....	88
Figure 3.5: Two-dimensional pleasure and arousal plot for 38 ANEW words representing 35 emotion categories based on ANEW ratings .....	90
Figure 3.6: Two-dimensional pleasure and arousal plot for 50 emotion words based on AMT ratings.....	92
Figure 3.7: Web annotation application for data collection in Phase 2.....	96
Figure 3.8: Design of the HIT for the emotion annotation task on AMT .....	97
Figure 3.9: Gold labels for valence and emotion tags obtained from Phase 1's ground truth displayed in the first five tweets in a batch .....	99
Figure 3.10: Processes for running machine learning experiments .....	103
Figure 3.11: Two Tier 2 approaches to train classifiers that distinguish only instances containing emotion: a) single multi-class classification and b) multi-class binary classification.....	108
Figure 3.12: Pseudocode on bigram extraction for custom phrase cues.....	112
Figure 4.1: Heatmap showing co-occurrence frequencies of paired annotator labels .....	122
Figure 4.2: Heatmap showing co-occurrence frequencies of paired gold and annotator labels.....	123
Figure 4.3: Four groupings of emotion categories based on class frequency and $\kappa$ .....	124
Figure 4.4: Comparing $\kappa$ for 28 emotion categories across P1, P2 and P1+P2 .....	129

Figure 4.5: Distribution of tweets based on mean arousal ratings .....	146
Figure 4.6: Frequency of emotion cue and segment length based on word count .....	152
Figure 4.7: Heatmap based on term similarity between emotion category pairs .....	157
Figure 5.1: Comparing BayesNet performance in flat and two-tiered classification.....	192
Figure 5.2: Precision, recall and F1 based on 12 iterations of downsampled training data .....	223
Figure 5.3: Bubble chart with three data dimensions (x: precision, y: recall and bubble size: percentage of full agreement) .....	230
Figure 5.4: Bubble chart with three data dimensions (x: precision, y: recall and bubble size: class size) .....	232
Figure A.1: HIT describing the emotion word rating task on AMT .....	249
Figure A.2: Instructions on the emotion word rating task .....	249
Figure A.3: Valence rating for 50 emotion words.....	250
Figure A.4: Arousal rating for 50 emotion words.....	251

# Chapter 1: Introduction

The ways that individuals write provide windows into their emotional worlds (Pennebaker, Chung, Ireland, Gonzales, & Booth, 2007). In addition to facial expression, vocal intonation, body language, and physiological response (Picard, 1998), humans can choose to convey their emotional experiences in written text (Cornelius, 1996; Fussell, 2002). Encoding emotional information in text is a common practice especially in online interactions. In the absence of non-verbal cues, writers adapt to the medium by imbuing messages with emotion cues either explicitly (e.g., emotion words or emoticons) or implicitly (e.g., metaphors or metonymy) to allow for more natural or enhanced communication (Walther, Loh, & Granka, 2005). With the prevalence of emotive content on the Web, especially on social media and microblogs, automatic emotion detection in text is attracting significant attention from researchers and businesses interested in investigating how emotions affect decision making, behaviors, and quality of life.

## 1.1 Motivation

Automatic emotion detection in text is concerned with using natural language processing techniques to recognize emotions expressed in written discourse. Endowing computers with the ability to recognize emotions in text has important applications in the field of computational linguistics. In sentiment analysis, emotion provides a promising direction for fine-grained analysis of subjective content (Aman & Szpakowicz, 2008; Chaumartin, 2007). Current sentiment analysis research operates at a coarser level than emotion detection. Sentiment analysis is mainly focused on detecting the subjectivity (objective or subjective) (Wiebe, Wilson, Bruce, Bell, & Martin, 2004) or semantic orientation (positive or negative) (Agarwal, Xie,

Vovsha, Rambow, & Passonneau, 2011; Kouloumpis, Wilson, & Moore, 2011; Pak & Paroubek, 2010; Pang, Lee, & Vaithyanathan, 2002) of a unit of text rather than a specific emotion. Often times, knowing exactly how one reacts emotionally towards a particular stimulus does matter (Mohammad, Zhu, & Martin, 2014). For example, while fear and sadness are both negative emotions, distinguishing between them can be important. In the event of a disaster, fear may be used to identify an onset of the disaster whereas sadness may be associated with later stages.

In real world business applications, automatic emotion detectors can provide richer insights into how a targeted audience feels about a product, person, event or topic. Businesses are finding innovative ways to analyze user-generated content to learn about consumer emotional reactions toward their products, services, and events. For example, automatic emotion detectors used on online product reviews can help businesses identify and track emotional reactions toward their products and services. Automatic anger detection in customer service emails can help customer service representatives quickly identify angry customers so that immediate action can be taken to reduce customer attrition. In consumer analytics, automatic emotion detectors provide businesses with non-invasive strategies to better market and advertise their offerings to their customers.

There is a growing demand to create more emotion-sensitive systems that can understand and express emotions to enhance human-computer interactions (Picard, 1998). An automatic emotion detector is a key component in building expressive conversational agents (L. Zhang, 2013), intelligent user interfaces (Liu, Lieberman, & Selker, 2003), as well as textual affect sensing and visualization systems (Garcia & Schweitzer, 2011; Kalra & Karahalios, 2005; Kennedy et al., 2011; Shaikh, Islam, & Ishizuka, 2006). Detecting emotions in text can be challenging due to the complexity of language used to express emotion. Our individual knowledge of what constitutes emotion cues in text may be limited by our cultural backgrounds. Automatic emotion detectors, on the other hand, could be trained to recognize both obvious

emotion cues that are widely used as well as the less evident cues used by different individuals, groups or cultures.

Most existing automatic emotion detectors are constructed to detect emotion based on the recognition of single words contained within an emotion lexicon. This notion that emotion is expressed using emotion words works well, to a certain extent, in formal English text (Mohammad & Turney, 2008; Strapparava & Valitutti, 2004). However, much less is known about what and how emotions are expressed in microblog text. To improve the performance of automatic emotion detectors on microblog text, this thesis focuses on the detection of emotions expressed on Twitter<sup>1</sup>. Twitter, a microblogging site with more than 100 million monthly active users<sup>2</sup>, is particularly rich with content containing how users feel about events, entities and topics shared publicly on a global scale. The text on Twitter (better known as “tweets”) can be mined to gain insights regarding users’ perceptions, behaviors, and the social interactions between different individuals and populations of interest in a non-invasive manner.

Interest in analyzing emotions on Twitter is evidenced by studies of how emotions expressed on microblogs affect stock market trends (Bollen, Mao, & Zeng, 2011), relate to fluctuations in social and economic indicators (Bollen, Pepe, & Mao, 2011), serve as a measure for the population’s level of happiness (Dodds & Danforth, 2010; Quercia, Ellis, Capra, & Crowcroft, 2012), provide situational awareness for both the authorities and the public in the event of disasters (Vo & Collier, 2013), and reflect clinical depression (Park, C. Cha, & M. Cha, 2012). With 500 million tweets being sent a day<sup>2</sup>, automatic emotion detectors would greatly augment our ability to analyze and understand emotive content. It is impossible to characterize emotions expressed in millions of tweets through human effort because it is very labor-intensive and costly.

---

<sup>1</sup> Twitter: <https://twitter.com/>

<sup>2</sup> Twitter usage and company facts: <https://about.twitter.com/company>

The quest for fine-grained emotion detection in tweets is potentially useful for applications such as:

- **Personality detection:** An individual's personality reveals key information about the person's attitude towards risks, decision making tendencies and preferences in various aspects of life. Fine-grained emotions have been shown to be significant indicators of an individual's personality (Mohammad & Kiritchenko, 2013, 2015). Therefore, fine-grained emotion detection in tweets can help businesses gain better knowledge of their customer's personality. Businesses can then provide more targeted product or service recommendations to their customers. For example, a customer who frequently expresses fear, regret and doubt is more likely to be risk averse. Being aware of such risk tendency, a financial consultant might recommend more conservative financial products for risk-averse customers. Personality detection systems are also helpful to help companies screen potential candidates to select the ones who are most suitable for a given job.
- **Public and behavioral health:** At the individual level, fine-grained emotion detection in tweets can help doctors and health professionals monitor changes in a person's emotional well-being and mental health through their interactions on social media and microblogs. Such a monitoring system might send alerts to care givers when early signs of depression are detected (e.g., high levels of sadness, regret, exhaustion and desperation are expressed by an individual over a long period of time) (De Choudhury, Counts, & Horvitz, 2013; De Choudhury, Gamon, Counts, & Horvitz, 2013). Early intervention might then be made to mitigate major psychological disorders or prevent suicidal behaviors. Microblogs like Twitter also allows researchers and practitioners to harness information about a population's emotional well-being (e.g. level of happiness or hope across populations in different communities or nations).

- Analyzing consumer attitude:** Apart from being able to detect if the consumer's sentiment or attitude towards an entity (e.g., product, service, company, etc.) is positive or negative, fine-grained emotion detection allows businesses to identify more specific emotional reactions from the consumers towards various aspects of an entity, thus assisting them in collecting feedback to refine their product or service offerings as well as in designing different strategies to handle various consumer's emotional reactions (e.g., angry versus sad consumers). In addition, gauging consumer's excitement, curiosity, confidence or doubt towards an event or marketing campaign from microblog posts can help companies collect feedback on its potential success. On the other hand, consumers are also in need of fine-grained emotion detection tools to help them mine relevant experiences from others especially when making difficult purchase decisions. For instance, a fine-grained emotion detector that is able to differentiate between expressions of sadness, anger, indifference and hope can provide consumers with a more informed view about how others feel about an item of interest.
- Market analysis and investment trend:** Increased interests have been observed in examining if public emotions expressed on microblogs such as Twitter correlate or predict stock market and other socioeconomic indicators. Existing research uses a unidimensional measure (i.e., positive or negative) to assess public emotion. Mixed results are observed using this coarse-grained measure of public emotion. The ratio of positive and negative messages on a given day generated from OpinionFinder (Wilson et al., 2005) has shown to correlate with the Consumer Confidence Index from Gallup and the Reuters/University of Michigan Surveys of Consumers over a period of time but exhibit no causal relations with the Dow Jones Industrial Average (DJIA) (Bollen et al., 2011). However, Bollen et al. (2011) and Dong, Chen, Qian, & Zhou (2015) reported that the predictive accuracy of stock market and trading volume can be significantly improved

by including certain emotions. Fine-grained emotion detection allows market analysts to study and leverage a richer set of human emotions into these predictive models.

- **Security and crises management:** Fine-grained emotion detection in tweets can be used to track public emotions for security threats, disruptions or natural disasters (e.g., terrorism, protests, earthquakes, etc.) as well as flag and monitor individuals that are threatening, abusive or risky (Karlgrén, Sahlgrén, Olsson, Espinoza, & Hamfors, 2012). In particular, it is crucial for such systems to be able to distinguish between more specific negative emotions (e.g., anger versus sadness, sympathy versus sadness or fear versus desperation).

Prior work has focused on adapting conventional theories to represent emotions expressed on Twitter and has not attempted to discover the actual range of emotions expressed in tweets or how these emotions are actually characterized in this type of text. An important starting point to build computational models that can recognize the emotions represented in tweets is the identification of a set of suitable emotion categories. Instead of borrowing a set of emotion categories from existing emotion theories in psychology, this research aims to first expose a set of categories that are representative of the emotions expressed on Twitter by analyzing the range of emotions humans can reliably detect in microblog text. Second, this research attempts to use a more systematic approach to surface pertinent linguistic cues associated with each emotion category based on cues that humans have identified as emotion expressions in text. Our findings will prove useful to improve our understanding of what and how emotion is expressed in text, as well as to advance scientific knowledge on emotion for the purpose of natural language processing. Third, we experiment with computational techniques that can allow automatic emotion detectors to recognize this representative set of emotions expressed on microblogs.



## 1.2 Problem Definition

Over a decade of active research in sentiment analysis has led to automatic emotion detectors that can be applied for large-scale analysis of emotions in different types of formal and informal English text. However, the dearth of research in understanding the richness of actual emotions that humans express and describe in text has resulted in existing automatic emotion detectors that detect only a small set of emotions (Alm, Roth, & Sproat, 2005; Aman & Szpakowicz, 2007; Liu et al., 2003). For automatic emotion detection on Twitter, the most commonly used emotion categories are adopted from Ekman's six basic emotions (happiness, sadness, fear, anger, disgust, and surprise) (Ekman, 1971) mainly because they are assumed to be universal emotions according to the emotion theories in psychology (Mohammad, 2012a; Wang, Chen, Thirunarayan, & Sheth, 2012). The current emphasis on these basic emotions poses limitations on the development of automatic emotion detectors that can capture the richness of actual human emotional expression.

First, it is unclear if these basic emotions are representative of the range of emotions humans express on Twitter. It is possible that the basic emotions framework is a poor fit or is too crude to adequately capture the range of emotions expressed in tweets. Mohammad & Kiritchenko (2014) found a few hundred emotion words being expressed explicitly using hashtags (notably used to indicate topics) on Twitter. While the basic emotions framework offers a starting point to study emotions expressed in text, it is crucial to note that the basic emotions represent only emotions recognized for their adaptive value in dealing with "fundamental life tasks" such as separation or failure, presence of threat, etc. (Ekman, 1999; Johnson-Laird & Oatley, 1992; Lazarus, 1991). The basic-emotions framework was derived from facial expressions and physiological responses and is not grounded on language theories. Humans use language to express and describe a wide range of emotions as illustrated in Example 1.1

and Example 1.2. Such nuances in tweeters' emotion language cannot be captured using the basic emotion framework.

**Example 1.1:** "Attack on Embassy in Libya is terrible tragedy. I extend my extreme & definite condolences to families of those lost. <http://t.co/H77gCiLI>" **[Sympathy]**

**Example 1.2:** "Thank you to all the Nebraskans who joined us for breakfast this morning! <http://t.co/FWYusNpf>" **[Gratitude]**

Second, many emotions that are not included as part of the basic set are either ignored or worse, force-fit into one of the available emotion categories. Example 1.1 is an obvious case of "sympathy" as the writer is expressing his or her condolences to people affected by a tragedy. Since "sympathy" is not one of the six basic emotions, Example 1.1 is most likely classified as the basic emotion "sadness". Similarly, Example 1.2 is an expression of "gratitude" that would most likely be classified as the basic emotion "happiness". The coarseness of the basic emotion taxonomy makes it more difficult for automatic emotion detectors to identify pertinent linguistic patterns for each emotion category because of the considerable amount of fuzziness and noise introduced into the corpus.

With the basic emotions accepted as the state-of-the-art, existing emotion corpora and other emotion-related language resources that serve as the basis for building and evaluating mechanisms to detect emotion in tweets are only annotated with the basic emotion categories as the finest level of granularity. For instance, Pak & Paroubek (2010) created a corpus with two emotion categories: positive and negative, while Mohammad (2012a) and Wang et al. (2012) applied Ekman's six emotions in the construction of their corpora. As a result, automatic emotion detectors developed using these resources are only able to give us a limited picture of actual human emotion expression. Complex emotions, as well as variations within each basic emotion are "virgin territories" that have not yet been explored by researchers in this area. Efforts to increase the utility of automatic emotion detectors have to start with extending

language resources to cover other emotion categories that both humans and computers can reliably detect in text.

Automatic emotion detection on Twitter presents a different set of challenges because tweets exhibit a unique set of characteristics that are not shared by other types of text. Unlike traditional text, tweets consist of short messages expressed within the limit of 140 characters. Due to the length limitation, language used to express emotions in tweets differs significantly from that found in longer documents (e.g., blogs, news, and stories). Language use on Twitter is also typically informal. It is common for abbreviations, acronyms, emoticons, unusual orthographic elements, slang, and misspellings to occur in these short messages (see Example 1.3). On top of that, retweets (i.e., propagating messages of other users), referring to @username when responding to another user's tweet, and using #hashtags to represent topics are prevalent in tweets (see Example 1.4 and Example 1.5). Even though users are restricted to post only 140 characters per tweet, it is not uncommon to find a tweet containing more than one emotion (see Example 1.6).

**Example 1.3:** "@HavokGrimey yes. Mexico.! im not from nikaragua. lma0. i swear idk were yu got that from.. im not even dark.! on the 17th(=" **[Amusement]**

**Example 1.4:** "RT @AylaBrown I love these guys with all of my heart!!! @scottbrownma @gailonthetrail <http://t.co/p20yhwD1>" **[Love]**

**Example 1.5:** "Many of our #NYC #veterans have been impacted by #Sandy, incl those in the #Manhattan VA hospital. Thx to #FortHamilton for taking them in." **[Gratitude]**

**Example 1.6:** "Having long hair is the most annoying and wonderful thing ever" **[Anger, Happiness]**

Emotion cues are not limited to only emotion words such as *happy*, *amused*, *sad*, *miserable*, *scared*, etc. Given the immense richness of English language, people use a variety of ways to express emotions. For instance, a person expressing happiness may use the

emotion word “happy” (see Example 1.7), the interjection “woop” (see Example 1.8), the emoticon “:)” (see Example 1.9), or the emoji “😄” (see Example 1.10).

**Example 1.7:** “I can now finally say I am at a place in my life where I am happy with who am and the stuff I have coming for me in the future #blessed” **[Happiness]**

**Example 1.8:** “its midnight and i am eating a lion bar woop” **[Happiness]**

**Example 1.9:** “The wait is almost over LA, will be out in just a little! 😄😄😄😄” **[Happiness]**

**Example 1.10:** “Enjoying a night of #Dexter with @DomoniqueP07 :)” **[Happiness]**

In addition to explicit expressions of emotion, users on Twitter also express their emotions in figurative forms through the use of idiomatic expressions (see Example 1.11), similes (see Example 1.12), metaphors (see Example 1.13) or other descriptors (see Example 1.14). In these figurative expressions of emotion, each word if treated individually does not directly convey any emotion. When combined together and, depending on the context of use, they act as implicit indicators of emotion. Automatic emotion detectors that rely solely on the recognition of emotion words will likely fail to recognize the emotions conveyed in these examples.

**Example 1.11:** “@ter2459 it was!!! I am still on cloud nine! I say and watched them for over two hours. I couldn't leave! They are incredible!” **[Happiness]**

**Example 1.12:** “Getting one of these bad boys in your cereal box and feeling like your day simply couldn't get any better <http://t.co/Fae9EjyN61>” **[Happiness]**

**Example 1.13:** “Loving the #IKEAHomeTour décor #ideas! Between the showroom and the catalog I am in heaven” **[Happiness]**

**Example 1.14:** “I did an adult thing by buying stylish bed sheets and not fucking it up when setting them up. \*cracks beer open\*” **[Happiness]**

The occurrence of an emotion word in a tweet does not always indicate the tweeter's emotion. The emotion word "happy" in Example 1.15 is not used to describe how the tweeter feels about the tune but is instead used to characterize the affective quality<sup>3</sup> of the tune. The tweeter is in fact expressing anger towards the "happy" tune. Similarly, #Happiness in Example 1.16 is part of a book's title so the emotion word hashtag functions as a topic more than an expression or description of an individual's emotion. The common practice of using emotion word hashtags to retrieve self-annotated examples as ground truth to build emotion classifiers, a method known as "distant supervision", (Hasan, Agu, & Rundensteiner, 2014; Mohammad, 2012a; Mohammad & Kiritchenko, 2014; Wang et al., 2012) is susceptible to this weakness.

**Example 1.15:** "@Anjjade I was at this party on the weekend, that happy tune was played endlessly, really not my stuff, it was like the cure's torture ha" **[Anger]**

**Example 1.16:** "Hear Carrie Goodwiler's audition for the audio version of my book #Happiness & Honey on #SoundCloud" **[No Emotion]**

These challenges associated with emotion expressions and descriptions in tweets remain a virgin territory that has not been thoroughly explored. Our objective is to deepen the understanding of how emotions are expressed on microblogs based on layman's conception of emotion and how automatic techniques can be used to detect those expressions. We will first inductively develop a set of emotion categories from the data. We will then manually annotate a large corpus to serve as training data for computer models. Each emotion category is defined by a set of linguistic properties. This will allow automatic emotion detectors to leverage the linguistic properties identified in our study to learn both explicit and implicit expressions of emotion.

---

<sup>3</sup> Affective quality is defined as the affective property of a stimulus that can affect how we feel about the stimulus (Russell, 2003; P. Zhang, 2013). In Example 1.15, the tune is the stimulus. The tweeter attributes a happy quality to the tune. The tweeter is not expressing happiness towards the tune.

### 1.3 Research Goals

Broadly, this thesis is intended as an exploratory analysis to discover the range of emotions humans and computers can detect in microblog text, to identify the linguistic cues that are most informative for the detection of each emotion category, and to investigate what level of performance we can expect from training human annotators as well as supervised machine learning models to classify fine-grained emotions in text. Specifically, the goals of this thesis are:

- Discover inductively from data the set of emotion categories that are representative of the emotions expressed by tweeters and emotional phenomena described in tweets (e.g., emotion of others or emotion that one should feel in a particular situation).
- Identify the linguistic characteristics of each emotion category based on the emotion cues humans use to identify emotions in tweets.
- Test the extent to which the linguistic cues pertinent to each emotion category can be leveraged to improve the performance of machine learning (ML) models used for automatic emotion detection in text.

The outcome of this research can be used to suggest ways in which the linguistic characteristics of this richer set of emotion categories might be used to advance research in sentiment analysis and the design of more emotion-sensitive systems for real world applications.

### 1.4 Research Questions

Using data from Twitter, an initial framework will be developed to study the range of emotions expressed in text. This study is guided by these research questions:

- *R1: What emotions can humans detect in microblog text? (Phase 1 & 2)*

- *R2: What salient linguistic cues are associated with each emotion? (Phase 1 & 2)*
- *R3: Do the salient cues humans associate with each emotion serve as better features for machine learning classification of emotion in text? (Phase 3)*
- *R4: How do current machine learning techniques perform on more fine-grained categories of emotion? (Phase 3)*

To answer the research questions above, the proposed research design consists of three phases: 1) small-scale content analysis for code book development and testing, 2) large-scale content analysis for gold standard data development, and 3) the design of machine learning experiments to test the effectiveness of automatic emotion detection in text. Phase 1 of the investigation focuses on discovering the set of emotion categories expressed in tweets and the emotion cues associated with each emotion using grounded theory (Corbin & Strauss, 2008). Phase 2 tests the representativeness of the emotion categories emerging from Phase 1 on a larger set of tweets and creates a large corpus of annotated data through crowdsourcing. Analysis on the emotion corpus will be used to address R1 and R2. To answer R3 and R4, human annotations from Phase 1 and Phase 2 will serve as gold standard data to build machine learning models and to evaluate their performance on these fine-grained emotion categories.

## **1.5 Thesis Overview**

This section provides an overview of how the thesis is organized. Chapter 2 surveys related work on emotion detection in text by pulling together literature from three disciplines: computational linguistics, psychology and linguistics. We discuss how emotion in text is positioned within these three research areas. Specifically, it reviews existing research on automatic emotion detection in text in the context of tweets. The chapter also defines the terms and concepts related to emotion in text and clarifies how emotion differs from closely related terms such as subjectivity, sentiment, affect and mood.

We present details of our three-phase methodology used to investigate fine-grained emotion detection in tweets in Chapter 3. Phase 1 and Phase 2 are dedicated to the development of the largest tweet corpus (EmoTweet-28) annotated with a set of fine-grained emotion categories. This corpus is then used as ground truth in our machine learning experiments in Phase 3.

Chapter 4 begins by characterizing the set of 28 fine-grained emotion categories identified from tweets. This set of emotion categories are deemed to be representative of the range of emotions expressed in tweets. We then report the level of performance we can expect from annotators in recognizing each emotion category. Based on the emotion cues highlighted by annotators as part of the annotation task, we derive linguistic patterns that can be used to distinguish an emotion category from the others.

Chapter 5 presents the outcome of our machine learning experiments based on EmoTweet-28. We apply machine learning techniques to train and evaluate classifiers in detecting 28 emotion categories. We compare a feature selection approach utilizing the emotion cues identified from our study (i.e., cue-based features) to the conventional corpus-based features (i.e., features statistically-generated from the corpus such as unigrams from the corpus) and lexicon-based features (i.e., features originating from an emotion lexicon). Our research shows that it is feasible to perform fine-grained emotion classification on tweets using an extended set of 28 emotion categories. We show that our classifiers give performance that is comparable to the current state-of-the-art using only the limited six basic emotion categories.

Finally, Chapter 6 highlights the contributions of this study, presents conclusions and reviews topics for future work.



# Chapter 2: Literature Review

## 2.1 Introduction

This chapter presents the background on emotion detection in text drawing from three research areas: computational linguistics, psychology, and linguistics. Over the years, vast amounts of research have been conducted to better define and understand the concept of emotion especially in the discipline of psychology, but very little emphasis has been placed on the systematic study of how emotion is expressed in language for the development of natural language processing systems used to automatically detect emotion in text. Each research area contributes relevant knowledge to help us better understand the theories, methods and applications related to emotion detection in text but there has been little effort to piece together related work from each individual research area.

This literature review aims to integrate related work from these three research threads not only to identify gaps in the literature, but also to provide a theoretically and methodologically sound discussion of existing work in automatic emotion detection in text. The focal point of our survey is on automatic emotion detection employed for sentiment analysis. Psychology and linguistics offer a rich discourse in emotion theories to shed light on the definition of emotion and how emotion in text is conceptualized.

The first part of this literature review explores the different conceptualizations of emotion in text, and how these notions are employed in automatic emotion detection in text. We also define terms and concepts related to emotion in text and attempt to draw the distinction between emotion and other closely-related concepts such as subjectivity, sentiment, affect and mood.

The second part then focuses on the various computational mechanisms developed to automatically detect emotion in text. The third part reviews literature specific to the detection of emotion on Twitter.

## **2.2 Emotion Theories: The Psychology of Emotion**

The discussion of what emotions can be detected in text, and how they are measured should first start with an examination of the four main theoretical perspectives on emotion drawn from the psychology literature. While research on textual emotion detection is relatively new in the boom of online user-generated content, classic work on the theory of emotion can be traced back to Darwin (1872) in his seminal publication, *The Expression of the Emotions in Man and Animals*. Thus, the Darwinian perspective represents the earliest attempt in psychology to shed light on what constitutes emotion, followed by the Jamesian, the cognitive, and the social constructivist perspectives (Cornelius, 1996). Researchers in sentiment analysis have borrowed these emotion theories and used them in various ways to operationalize the concept of emotion in text in order to inform the development of automatic emotion detectors. We first provide an overview of the four theoretical perspectives on emotion; we then review the extent to which sentiment analysis researchers have engaged with these emotion theories. The level of engagement by sentiment analysis researchers with emotion theories varies on a continuum from limited (i.e., theoretical framework is identified with no or limited application) to significant (i.e., the theoretical framework is integral to the work).

### **2.2.1 The Darwinian Perspective: Emotion as Expression**

The Darwinian perspective defines emotion as being “expressions” (Calvo & D’Mello, 2010). These emotional expressions can appear in any observable form such as facial, behavioral, and physical. The underlying tenet of Darwin’s theory of emotion is that emotional expressions are tied to biological actions that are essential for human adaptation to the natural

environment (Darwin, 1872). According to the evolutionary perspective, humans do not express emotions just for the sake of expressing emotions. Rather, emotional expressions are action tendencies that are tied to survival functions in response to an emotional stimulus (Frijda, 1987). Table 2.1 shows Frijda's mapping of emotion to its associated action tendency and function (Frijda, 1986). More importantly, the Darwinian perspective claims that there is a consistent set of patterns associated with the expression of each distinct emotion (Cornelius, 1996). This implies that there is a set of universal emotional expressions that humans would display and could recognize regardless of culture and language.

Emotion	Action Tendency	Function
Desire	Approach	Permits consummatory behavior
Fear	Avoidance	Protection
Enjoyment, Confidence	Being-with	Permits consummatory activity
Interest	Attending	Orientation to stimuli
Disgust	Rejecting	Protection
Indifference	Non-attending	Selection
Anger	Agonistic (attack/threat)	Regaining control
Shock, Surprise	Interrupting	Reorientation
Arrogance	Dominating	Generalized control
Humility, Resignation	Submitting	Secondary control

Table 2.1: Emotion and its associated action tendency and function

*Source: Adapted from Frijda (1987)*

Proponents of the Darwinian camp, including Ekman (1971), Izard (1971) and Plutchik (1984), have built on Darwin's theory of emotion by postulating a set of universal emotions, also known as basic or prototypical emotions, and defining the patterns associated with this set of emotions. Figure 2.1 shows the degree of overlap between what Ekman (Ekman et al., 1987), Izard (Izard, 1971, 1994) and Plutchik (Plutchik, 1962) consider to be "basic" emotions. Ekman's six basic emotions are *happiness*, *surprise*, *sadness*, *fear*, *disgust*, and *anger*. Plutchik's model is an expansion of Ekman's basic emotions through the addition of *trust* and *anticipation* in his eight basic emotions, while Izard's ten basic emotions also include *guilt* and *shame*. Only Ekman, Izard and Plutchik's emotion theories are discussed in this chapter because of their

dominant influence on sentiment analysis research (Mulcrone, 2012). It is important to note that Ekman, Plutchik and Izard derive their basic emotions based on the universality of facial expressions of emotion (Ekman & Rosenberg, 1997; Izard, 1971; Plutchik, 1962). They claimed that observers from different cultures achieved high agreement when asked to identify the facial expressions showing these basic emotions on posed photographs (Ekman & Friesen, 1971).

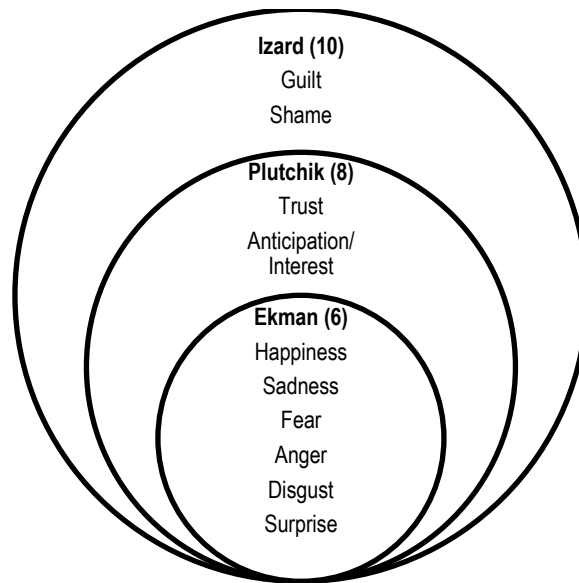


Figure 2.1: Ekman, Plutchik, and Izard's basic emotions

In line with the Darwinian perspective on a small set of emotional expressions being universal, sentiment analysis researchers seem to make the implicit assumption that the principle of “universality” also applies to emotions in text without considering the repertoire of linguistic devices writers use in expressing their emotions verbally in written form. Sentiment analysis researchers invoke the Darwinian framework merely to identify the emotion labels to use for the development of automatic emotion detectors without further application of the components specified in the theory. The definition of each emotion and the semantic difference between the emotion labels are rarely explained and are assumed to be universally understood. (Alm et al., 2005), Mohammad (2012a), Aman & Szpakowicz (2008), Chaumartin (2007), Liu, Lieberman, & Selker (2003), Strapparava & Mihalcea (2007, 2008), and Zhe & Boucouvalas

(2002) adopted Ekman's six basic emotions (*happiness, sadness, fear, anger, disgust, and surprise*) as the emotion labels of interest in their studies. Only Alm et al. (2005) made a slight modification to the list of six emotions by expanding the emotion "*surprise*" into two sub-emotion labels (*positively surprised* and *negatively surprised*) in her study to detect emotions in fairy tales.

Researchers who chose to use Plutchik's eight basic emotions (Kennedy et al., 2011; Mohammad, 2011, 2012b, 2012c; Mohammad & Yang, 2011) or a subset of Izard's ten basic emotions (Neviarouskaya, Prendinger, & Ishizuka, 2007b, 2007d) also demonstrated a similarly low-level of engagement with these emotion theories. One exception is Suttles & Ide (2013), who demonstrated greater application of Plutchik's wheel of emotion in framing the emotion classification problem. Following Plutchik's tenet that emotions represented by spatial opposition in the wheel of emotion were polar opposites and mutually exclusive (Plutchik, 1984), they examined the potential of building binary classifiers for each opposing emotion pair.

Although the Darwinian perspective puts a significant emphasis on discovering universal patterns of expression associated with an emotion, there has been little effort to date by researchers who chose to build automatic emotion detectors based on the basic emotions to map out the general linguistic patterns that can be used to reliably detect the expression of each basic emotion in text. The question of whether or not we can define a set of linguistic patterns the same way as Ekman has defined the facial muscle movement patterns for each basic emotion (Ekman & Rosenberg, 1997) remains a gap to be addressed.

The Darwinian perspective has contributed a set of basic emotion labels to inform research on emotion detection in text, thus making automatic detection of the basic emotions in text possible. However, these computational models are still limited to detect only a small set of emotions, and ignore other complex and nuanced emotions that are often expressed in writing. Sentiment analysis has merely touched on a small tip of the iceberg in adapting emotion

theories from this camp for the development of automatic emotion detectors. Darwin and his followers offer a rich discourse on emotion-related principles, hypotheses, constructs, definitions, and other research components that are yet to be properly tested and applied in text. For example, the Darwinian theories come with elaborate descriptions of physical and behavioral reactions associated with each basic emotion, but the extent to which they have been adopted as part of everyday (or online) language or to express or describe emotions is still unknown. If there is indeed a set of universal emotions in text, a greater understanding and level of engagement with the Darwinian theories may lead to the development of potentially automatic emotion detectors that are generalizable across domains. If linguistic expressions of emotion do evolve over time, more robust approaches will need to be developed for automatic emotion detectors employed in longitudinal studies.

### **2.2.2 The Jamesian Perspective: Emotion as Embodiment**

The Jamesian perspective, named after William James (the pioneer of this school of thought) defines emotion as being “embodiments”, meaning that emotions are embodied within the peripheral physiology (Calvo & D’Mello, 2010). James argued that the experience of emotion is primarily the experience of “bodily changes” that follow directly from the perception of an exciting fact (James, 1884). The James-Lange theory claimed that the physiological changes happen first in our nervous system, which then leads to us feeling a certain emotion (Lange & James, 1922). Note that the Jamesian perspective agrees with the Darwinian perspective in that emotion is the product of our survival-related responses to the environment (Cornelius, 1996). However, these two perspectives disagree on the operationalization of emotion. Darwin views emotion as patterns of expression while James emphasizes the patterns of physiological changes associated with an emotion (Calvo & D’Mello, 2010).

Due to its focus on physiology (e.g., changes in skin temperature, heart rate and respiration), the Jamesian perspective has been deemed to be less relevant for emotion

detection in text. The emotion theories in the Jamesian camp aim to explain and measure the implicit emotional state within an individual, and not the emotions that are made explicit in a written form. This explains why the Jamesian perspective is not explicitly discussed among sentiment analysis researchers in their research to develop computational models for emotion detection in text.

### **2.2.3 The Cognitive Perspective: Emotion as Appraisal**

The cognitive perspective focuses on the role of thought in how we appraise situations in the environment (Cornelius, 1996). This perspective posits that emotions are generated through people's appraisal of an object or event that directly affects them, based on their goals, experience, and opportunities for action (Arnold, 1960; Scherer, 1999). To appraise an object or event emotionally involves more than just knowing about it objectively. It depends on how we react to the object or event based on the way we construe the eliciting situation (Ortony, Clore, & Collins, 1988). *Construal* is the cognitive interpretation of external reality, rather than directly from the reality itself (Ortony et al., 1988). To illustrate this point, two individuals watching the same sports game but who are supporting opposite teams may construe the game differently. The one supporting the winning team will experience the emotion “joy” while the other supporting the losing team will experience the emotion “disappointment”.

The cognitive perspective posits that there is a distinct set of appraisal patterns associated with each emotion. While there may be differences in how an individual appraises a situation, there is a “core relational theme” that underlies each specific kind of emotion (Lazarus, 1991). Table 2.2 shows the core relational themes for eight emotions. While the Darwinian perspective falls short in providing a set of linguistic patterns to distinguish distinct emotions in text, Ortony et al. (1988) laid the groundwork by characterizing the cognitive structure of twenty-two emotion types as shown in Figure 2.2. However, rather than focusing on the language of emotion, they focus on eliciting conditions that could distinguish one emotion type from another.

Eliciting conditions refer to “situational descriptions of the condition under which an emotion can be triggered” (Ortony et al., 1988, p. 15). An emotion type is a distinct term used to represent a family of related emotions. Emotions in the same family share the same eliciting conditions, but differ in intensity. As shown in Figure 2.2, an emotion is a valenced (positive or negative) reaction to either consequences of events, actions of agents or aspects of objects. An event refers to “people’s construals about things that happen”, an object refers to any material thing, and an agent refers to “people, nonhuman animate beings and inanimate objects or abstractions” (Ortony et al., 1988, p. 18). The different focus on events, agents, and objects will lead to distinct classes of emotional reactions.

Emotion	Core Relational Theme
Anger	A demeaning offense against oneself
Anxiety	Facing uncertain threat
Fright/Fear	Facing an immediate, concrete, overwhelming physical danger
Guilt	Transgressing a moral imperative
Sadness	Experiencing an irrevocable loss
Happiness	Making progress toward the realization of a goal
Love	Desiring or participating in affection, usually but not necessarily reciprocated
Compassion	Being moved by another’s suffering and wanting to help

Table 2.2: Emotion and its core relational theme

*Source: Adapted from Lazarus (1991)*

Ortony et al.'s (1988) global structure of emotion types (OCC model – short for Ortony, Clore and Collins) was designed with the goal of creating computer models that can understand and predict people’s emotional reactions in various conditions, thus paving the way for sentiment analysis researchers to transform the goal into reality. One of the earliest proof-of-concept was implemented by Elliott (1992) in the Affective Reasoner, an agent-based platform exploring reasoning about emotions by simulation. He implemented 24 emotion types in the OCC model (including *liking* and *disliking*).



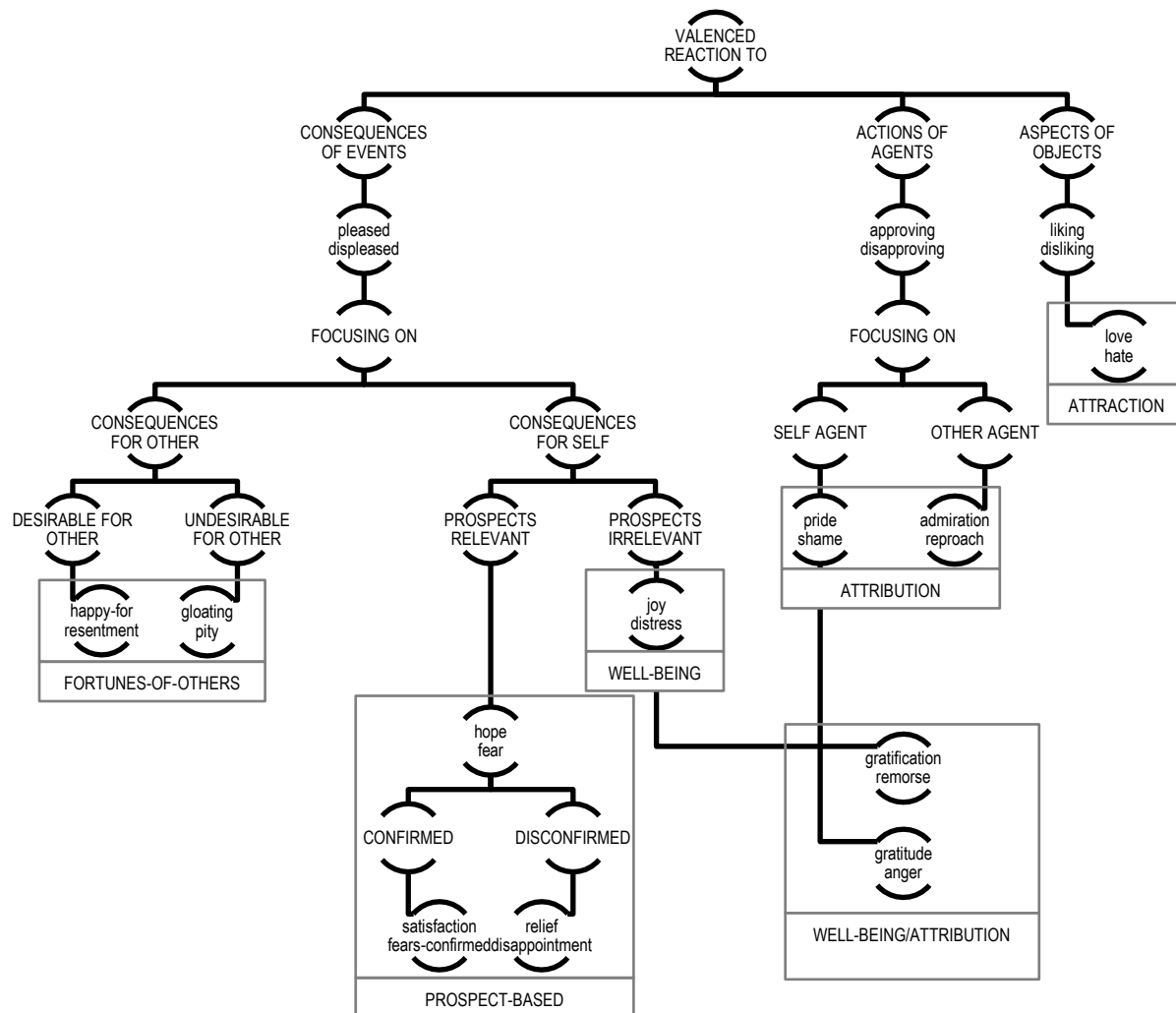


Figure 2.2: Global structure of emotion types

Source: Adapted from Ortony et al. (1988)

In the context of text analysis, Shaikh et al. proposed the implementation of a formal model using the 22 emotion types from the OCC model because they believed that Ekman's six basic emotions were insufficient for emotion detection in text (Shaikh, Helmut, & Ishizuka, 2006; Shaikh, Prendinger, & Mitsuru, 2007). They then transformed the appraisal criteria from the OCC model into rules for 8 out of 22 emotion types in the Emotion Sensitive News Agent (ESNA) to sense affective information from news text (Shaikh, Islam, et al., 2006; Shaikh,

Prendinger, & Ishizuka, 2007a). Detailed representations of all 22 emotion types were elaborated in Shaikh, Prendinger, & Ishizuka (2009).

Shaikh et al. (2009) has demonstrated successful application of the OCC's cognitive emotion theory in the implementation of automatic emotion detectors. At the other end of the continuum, sentiment analysis researchers have only referenced the OCC emotion types in their work without actual application of the theory (Tao, 2004; L. Zhang, 2013). For example, L. Zhang (2013) only referred to the OCC model to collapse a list of 15 emotions into three emotion labels: positive, negative, and neutral. Tao (2004) adopted the definition of emotion based on the OCC model but used the six basic emotions instead in the emotion estimation model.

The concept of emotion as appraisals also served as the structure of EmotiNet, a knowledge base containing common sense knowledge on concepts, their interactions, and affective consequences proposed by Balahur et al. (2011). Following the notion that emotion is an individual's reaction to an event of personal significance, the knowledge base consists of "action chains" representing actions that trigger an emotion on an actor (Balahur, Hermida, & Montoyo, 2012a, 2012b). The representation of these action chains are grounded in the cognitive perspective.

It is important to note that appraisal criteria in the cognitive perspective are concerned with the eliciting conditions that trigger an emotion, and do not take into account the language used to express emotion (Ortony et al., 1988). In fact, the theories in this camp are constructed to be language-independent. However, automatic emotion detectors process emotive information at the linguistic level. Therefore, sentiment analysis researchers using cognitive approaches still face a level of abstraction, and have to hash out the linguistic components to represent the appraisal criteria in the theories. For example, the emotion "joy" is elicited when the subject is pleased with an event. A computational model would have to understand the

repertoire of linguistic components that can be used to describe a “subject”, “pleased” and an “event” respectively.

The cognitive perspective offers automatic emotion detectors the opportunity to detect an extended set of emotions. However, it has been criticized for neglecting the social functions of emotion. Some have even claimed that it is too difficult, if not impossible, to construct a comprehensive knowledge base required for complex realistic situations (Calvo & D'Mello, 2010; Cornelius, 1996). Despite the complexity of the cognitive theories, this perspective provides researchers with another point of view on how emotions are expressed in text using an identifiable set of appraisal patterns.

#### **2.2.4 The Social Constructivist Perspective: Emotion as Social Constructs**

The social constructivist perspective asserts that emotions are social constructs (Calvo & D'Mello, 2010). Averill (1980a) claims that emotions are “cultural constructions”, and can only be understood from within the framework of a culture’s social practices. Unlike other perspectives, social constructivists are agnostic regarding the claim that emotion is a product of evolution. They postulate that emotions are rules learned by individuals to maintain the moral order of a culture. For example, a person from culture A may use a different set of rules to express anger than a person from culture B. It may be culturally acceptable for a person in culture A to engage in a public display of rage to express anger, whereas such a public display of emotion may be scorned in culture B. Language plays an important role in this perspective as emotion language influences the way we experience emotion. How we use words to express our emotions “embodies the meaning of emotion recognized by our culture” (Ochs & Schieffelin, 1989). Language can be used to form new concepts of emotion that are not tied to any biological or physiological functions (Russell, 2012), a main tenet that distinguishes the social constructivist perspective from the other three perspectives.

The social constructivists' emphasis on emotions being culturally-determined does not mean that the way people express emotions are completely arbitrary. On the contrary, as people from the same culture learn from a same set of rules, there are structured variations in how people in a culture use language to express their emotion. Rather than focusing on a set of universal emotions, this perspective shifts emphasis to the discovery of emotions with pertinent linguistic patterns within groups of people sharing the same culture or subculture, and how these patterns differ across various cultures.

Social constructivists maintain that there is no necessary core to emotion (Averill, 1980a) so it would not be appropriate to impose a predetermined set of emotions in a study. Instead, social constructivists attempt to discover these emotions inductively by observing the way speakers talk about and conceptualize emotions. This strategy is used by Brooks et al. (2013) and Scott et al. (2012) in their efforts to develop a taxonomy of affect in online distributed collaboration. They applied grounded theory (Corbin & Strauss, 2008) to identify a set of affect categories from data through content analysis. Subsequently, 40 affect categories emerged from the data, 13 of which were decided to be the most frequent. The social constructivist perspective provided an overarching framework to guide their discovery of emotion categories to be used in their computational model.

Although less explored by sentiment analysis researchers, the emphasis on language should make the social constructivist perspective a promising theoretical framework to discover the range of emotions that can be captured across different online communities and media. There have been few studies to date on how different cultural facets or social dimensions could be used in the construction of automatic emotion detectors, a gap that needs to be filled in order to create more robust automatic emotion detectors. More research is also needed to determine if culture-specific linguistic patterns associated with an emotion do exist, and the manner in which these linguistic patterns differ across cultures. After all, we may discover that underneath

the surface variation, emotions in all cultures may share a great deal of similarity (Cornelius, 1996).

## **2.3 Models of Emotion**

A starting point to build automatic emotion detectors is to determine how emotions can be classified. The categorization of emotion is largely based on two common models of emotion: 1) the categorical model, and 2) the dimensional model (Calvo & Mac Kim, 2012; Zachar & Ellis, 2012).

### **2.3.1 Categorical Model**

Emotions are classified into discrete categories, and each category represents a distinct emotion (Cowie, 2009). Each emotion category is characterized by a set of emotion patterns or structures that sets it apart from other categories. An emotion label is used to represent each category (e.g., happy, sad and angry) but there are various lexical realizations for each emotion label. For instance, the emotion label “*fear*” is associated with different words used to describe someone feeling threatened (e.g., “*terrified*”, “*afraid*”, and “*scared*”).

The basic emotion framework follows the categorical model, where emotion is organized and represented using a category system (Ekman, 1999; Izard, 1971; Plutchik, 1962). Each category represents a prototypical emotion that is defined by a set of features. Using a hierarchical classification approach, Shaver, Schwartz, Kirson, & O'Connor (2001) expanded the basic emotions into 25 finer categories through similarity sorting of 135 emotion words. These finer categories are more representative of the emotions that can be expressed using English words.

There are two advantages that come with using categorical labels: 1) using intuitive labels makes it easier to understand the emotion associated with the label, and 2) researchers have the flexibility to use different dimensions or criteria to define each emotion category. On

the other hand, it is crucial for researchers to draw clear distinctions between different emotion categories to avoid any confusion in the interpretation of the emotion labels.

In emotion detection in text, assigning only one emotion category to a text excerpt is the simplest type of implementation (Alm et al., 2005; Aman & Szpakowicz, 2007, 2008; Mohammad, 2012a). A more complex approach allows multiple emotion categories to be assigned to a text excerpt (Cherry, Mohammad, & de Bruijn, 2012). Another extension of categorical labels, known as “soft vector” (Calvo & D’Mello, 2010) uses a vector that consists of multiple emotion labels. Each emotion label in the vector is represented by a numerical estimate indicating the magnitude that a relevant emotion is present in a text excerpt (Liu et al., 2003; Neviarouskaya et al., 2007d; Strapparava & Mihalcea, 2007).

### **2.3.2 Dimensional Model**

Emotion is measured as a “coincidence of values on a number of strategic dimensions” (Bradley & Lang, 1999, p. 1). The dimensional model aims to account for all emotions in simpler and more general dimensions as opposed to discrete emotion categories. It holds that all emotional phenomena share the same fundamental structure, and can be identified from the composition of two or more independent dimensions (Zachar & Ellis, 2012). Russell & Mehrabian (1977) postulated three bipolar dimensions that are necessary and sufficient to adequately detect all emotions: valence, arousal, and dominance/submissiveness. Valence also referred to as “polarity” measures whether an emotion is pleasant or unpleasant. Arousal measures the degree of activation, which can range from calm to excited. Dominance/submissiveness, which is less commonly used in the sentiment analysis literature, measures the extent of control one has on events or surroundings, and can range from feeling a total lack of control to feeling extremely in control. Russell (1980) subsequently proposed the circumplex model of emotion mapping affect terms into a two dimensional space (valence and

arousal) but added four additional variables (excitement, contentment, distress and depression) to further define the quadrants of the space.

Valence is typically framed as a text classification problem: a text segment is either assigned a “positive”, “negative” or “neutral” label. Arousal can also be measured similarly using labels representing varying intensities (e.g., low, moderate or high) or a numerical scale. The dimensional measures allow researchers to capture more nuanced differences of emotions in text without the constraint of fitting all emotional phenomena into a limited set of categories. However, dimensional labels are less intuitive and more ambiguous to a lay person compared to categorical labels (Read, 2004). In addition, identifying the minimal number of dimensions to adequately define all emotions remains a difficult challenge to address.

## **2.4 Application of Emotion Theories in Automatic Emotion Detection in Text: A Summary**

Having described the perspectives and models of emotion, Table 2.3 summarizes their use in the research on automatic emotion detection in text. Research on emotion theories from psychology offers a wealth of scientific knowledge about human emotions that sentiment analysis researchers have not fully taken advantage of. In the early days, sentiment analysis researchers favored the basic or prototypical view of emotion because it offers a simple list of emotion categories for the classification of emotions in text. This has remained as the state-of-the-art as shown by the sheer frequency of appearance in Table 2.3. Prior work has also attempted to incorporate both the categorical and dimensional models in a hierarchical fashion to deal with the detection of emotion at different levels of granularity. Categorical emotion labels are considered to be more fine-grained, and are frequently grouped into the coarse-grained dimensional emotion labels.

Perspective	Theory	Scholars	Model	
			Categorical	Dimensional
Darwinian	Ekman (1973)	Alm, Roth, & Sproat (2005)	angry, disgusted, fearful, happy, sad, positively surprised, negatively surprised	valence: positive/negative/neutral
		Mohammad (2012a)	anger, disgust, fear, joy, sadness, surprise	
		Aman & Szpakowicz (2007, 2008)	happiness, sadness, anger, disgust, surprise, fear, mixed emotion	intensity: high/medium/low
		Ghazi, Inkpen, & Szpakowicz (2010)	happiness, sadness, anger, disgust, surprise, fear	
		Chaumartin (2007)	anger, disgust, fear, joy, sadness, surprise	degree emotional load (i.e., intensity): no emotion/maximum emotion load valence: positive/negative
		Strapparava & Mihalcea (2007, 2008)	anger, disgust, fear, joy, sadness, surprise	degree emotional load (i.e., intensity): no emotion/maximum emotion load valence: positive/negative
		Zhe & Boucouvalas (2002)	anger, disgust, fear, joy, sadness, surprise	intensity duration of expression
		Liu et al. (2003)	happy, sad, angry, fearful, disgusted, surprised	
	Izard (1971, 1977)	Neviarouskaya, Prendinger, & Ishizuka (2007a, 2007b, 2007c, 2007d)	anger, disgust, fear, guilt, interest, joy, sadness/distress, shame, surprise	intensity: very weak – very strong
	Plutchik (1962, 1980)	Mohammad & Turney (2010, 2012); Mohammad & Yang, (2011); Mohammad (2011, 2012b, 2012c)	joy, sadness, anger, fear, trust, disgust, surprise, anticipation	valence: positive/negative
		Kennedy et al. (2011)	joy, sadness, fear, surprise, disgust, anger, trust, anticipation	valence: positive/negative
Jamesian	Lange & James (1922)	None		



Perspective	Theory	Scholars	Model	
			Categorical	Dimensional
Cognitive	Scherer (1993)	Balahur et al. (2012a, 2012b, 2011); Balahur & Hermida (2012)	anger, fear, disgust, shame, sadness, joy, guilt	
	Ortony et al. (1988)	L. Zhang (2013)	approval, disapproval, angry, grateful, regretful, happy, sad, worried, stressful, sympathetic, embarrassed, praising, threatening, caring	valence: positive/negative/neutral
		Shaikh et al. (2009); Shaikh, Prendinger, & Ishizuka (2007a)	distress, sorry-for, resentment, gloating, hope, fear, satisfaction, fears-confirmed, relief, disappointment, pride, shame, admiration, reproach, love, hate, gratification (joy and pride), remorse (distress and shame), gratitude (joy and admiration), anger (distress and reproach).	valence: positive/negative intensity
Social constructivist	Averill (1980b)	Brooks et al. (2013); Scott et al. (2012); L. Zhang & Barnden (2012)	interest, amusement, considering, agreement, annoyance, confusion, acceptance, apprehension, frustration, supportive, surprise, anticipation, serenity	

Table 2.3: Summary of commonly-used emotion theories and models among sentiment analysis researchers

It is important to point out that there is no consensus thus far on a unifying theory or meaning of emotion especially one that can be used to explain emotion expressions in text. There is also no consensus on the number of emotion classes to use (Farzindar & Inkpen, 2015). We observe a strong reliance of prior research on the basic emotions grounded on visual displays of emotion, which may not be the best fit when directly adopted to represent emotion expressions in text. This is a weakness we aim to address in this research.

## 2.5 Distinguishing Emotion from Related Terms

Emotion<sup>4</sup> is a common concept but it is also an ill-defined one. This section attempts to better distinguish emotion from several terms often used interchangeably with emotion.

Term	Definitions	
	Dictionary: Merriam-Webster	Literature
<b>Sentiment</b>	"An attitude, opinion or judgment prompted by feeling" (Merriam-Webster, 2013).	[1] "Personal belief or judgment that is not founded on proof or certainty" (Balahur et al., 2012a, p. 89). [2] "Organized systems of emotional tendencies centered about some object" (Kövecses & Palmer, 1999, pp. 2–3). [3] "A settled opinion reflective of one's feelings" (Balahur et al., 2012b, p. 742).
<b>Affect</b>	"A set of observable manifestations of a subjectively experienced emotion" (Merriam-Webster, 2013).	[1] "An inclusive concept spanning emotions and feelings distinct from cognition" (Russ, 1993, p. 7), and more pervasive than the neurophysiological experiences of emotions (Moore & Isen, 1990; Scott et al., 2012). [2] "The subjective states that observers ascribe to a person on the basis of the person's conduct" (Besnier, 1990, p. 421)
<b>Emotion</b>	"A conscious mental reaction subjectively experienced as strong feeling usually directed towards a specific object and typically accompanied by physiological and behavioral changes in the body" (Merriam-Webster, 2013).	[1] "An episode of interrelated, synchronized changes in the states of all or most of the five organismic subsystems in response to the evaluation of an external or internal stimulus event as relevant to major concerns of the organism" (Balahur et al., 2012a, p. 89). [2] "Emotion refers to the process whereby an elicitor is appraised automatically or in an extended fashion, an affect program may or may not be set off, organized responses may occur, albeit more or less managed by attempts to control emotional behavior" (Ekman, 1977, p. 30)

Table 2.4: Comparison between dictionary and common literature definitions of sentiment, affect, and emotion

There is yet to be uniform terminology in this emerging research area. Emotion, affect and sentiment have often been used interchangeably by researchers to mean roughly the same thing. This is problematic for two reasons: 1) it may obscure differences that are important to an understanding of emotional phenomena, and 2) blurry concept definition boundaries introduce

<sup>4</sup> Emotion is a problematic term as its meaning is still being debated even among psychologists. The term is problematic not because it has no clear meaning. Rather, emotion has many meanings and consensus has yet to be achieved in the research community (Dixon, 2012; Izard, 2010).

unwanted noise into a research investigation which might lower the performance of automatic emotion detectors.

To illustrate the distinction between the terms related to emotion, the dictionary definitions and common literature definitions of sentiment, affect, and emotion are compared in Table 2.4. Only two definitions of emotion are provided in Table 2.4. A compilation of 92 definitions of emotion in the psychology literature can be found in Kleinginna Jr & Kleinginna (1981).

### **2.5.1 Emotion and Subjectivity**

Before discussing further the definition of the terms, it is crucial to introduce the notion of *subjectivity*. The broadest term related to emotion, subjectivity is defined as all things that are “based on feelings or opinions rather than facts” (Merriam-Webster, 2013). It was initially used to draw a distinction from the concept of objectivity (Wiebe et al., 2004). Wiebe, Wilson, & Cardie (2005) used subjectivity as an all-inclusive umbrella term to cover “opinions, emotions, sentiments, speculations, evaluations and other private states”. Thus, emotion is a subset of subjectivity, and subjectivity acts as a parent term when the distinction between sentiment, affect, and emotion is not a concern in a study.

### **2.5.2 Emotion and Sentiment**

Comparing the definitions of sentiment and emotion in Table 2.4, there are four elements that set the meaning of emotion apart from sentiment. First, emotions are states of an individual (see Table 2.4 emotion literature definition [1]), while sentiments are properties someone assigns to an object, an entity or a topic (Brave & Nass, 2009). Sentiments are often tied to people’s attitudes and opinions. Attitude has a broader meaning encompassing the way people think and feel about an object or person, while opinion focuses on the thinking dimension. Second, emotions are reactions to events concerning the individual (see Table 2.4 emotion

literature definition [1]), whereas sentiments are “emotional tendencies” or opinions on various events that may or may not be concerned with the person’s well-being (see Table 2.4 sentiment literature definition [2] and [3]).

a) *John is upset because the engine of his Toyota car would not start.*

b) *John does not own a Toyota but he thinks Toyota cars are great.*

To illustrate both points, suppose that John is expressing an emotion in statement (a) because he describes a change in his emotional state caused by an unmet expectation he sees as having a negative effect on his well-being. In statement (b), John is merely expressing his sentiments regarding Toyota cars. One can predict how he would feel about Toyota cars (i.e., joy) but there is no indicator of his emotional state in statement (b). Sentiment can affect a person’s emotion, and vice versa. Knowing that John has a positive sentiment on Toyota cars, he would most likely be happy (emotion) when he owns a new Toyota car, but his sentiment on Toyota cars may change from positive to negative after being upset (emotion) by the broken engine of his Toyota car.

The third element touches upon the persistence of the feeling. Emotions are fleeting, which means an emotional episode can last for a few seconds or a few hours, but sentiments persist indefinitely (see Table 2.4 sentiment literature definition [3]), and affects our everyday decisions to either seek out or avoid certain objects or situations (Brave & Nass, 2009). Fourth, emotion is caused by a stimulus (i.e., object, person or event) but sentiments can come from direct experience, subsequent generalization or external influences. Using the same example above, John’s agitation is caused by a specific object (i.e., his Toyota car), while John’s positive sentiment about Toyota cars may be based on his friends’ positive stories about driving Toyota cars.

Distinguishing the definitions between emotion and sentiment is important as emotion provides a finer-grained characterization of someone’s reaction to an identifiable object, person or event. Often times, researchers want to find out the direct impact of a particular stimulus to

an affected audience (i.e., actual audience reaction or outcome toward a specific stimulus). It may not be sufficient to know how people would react (i.e., prediction or emotional tendency) to a stimulus based on the audience's general sentiment. For example, does the release of a new product or marketing campaign cause frustration or anger? How does the public feel about a politician's introduction of a new bill? Another advantage relates to the handling of what seems to be contradictory subjective signals expressed within a statement. For example, a customer loyal to a brand may express positive sentiment towards the company but may be angered by a rude customer service representative. Being able to differentiate between sentiment and emotion is crucial for the company to identify the factor that causes the negative emotion, which in this case is the rude customer service representative, and not the company.

### **2.5.3 Emotion and Affect**

Affect is another umbrella term that encompasses emotions, moods and feelings (Russell, 2003). Unlike sentiment, affect focuses only on the "feeling" dimension, and not the "thinking" dimension (see Table 2.4 affect literature definition [1]). Care must be taken when interpreting the term "feeling". Not all feelings are emotions. Feeling is a generic term used to define "a broad category of person-centered psychophysiological sensations" (Besnier, 1990, p. 421). For example, hunger is a feeling but not an emotion. Excitement is considered to be an emotional feeling (Cowie, 2009). Affect can be considered a parent term to emotion, in the sense that all emotions are affective states, but not all affective states are emotions (Clore & Ortony, 1988).

Emotion is also often confused with another common affect term, mood. One core distinction between emotion and mood is the element of object-directedness (Brave & Nass, 2009; Nissenbaum, 1985). Emotion is usually directed at something or someone (see Table 2.4 emotion literature definition [2]). On the other hand, mood has a "global, free-floating quality" that tends to affect reactions to events people encounter while in that mood (Parrott, 2001). The

onset of a mood does not require any stimulus, whereas an emotion is triggered by a stimulus (P. Zhang, 2013). Therefore, identifying the presence of the emotion stimulus is the key criterion distinguishing these two terms. For coarse-grained level of analysis, it may be acceptable for researchers to use the term affect especially when the stimulus does not play any role in the research inquiry. This interchangeable use of affect and emotion is currently common practice among sentiment analysis researchers. However, when people's emotional reaction to a particular stimulus becomes the focal point of the study, the use of the term affect may be misleading.

## 2.5.4 Hierarchical Organization of Related Terms

Figure 2.3 shows how different commonly-used terms in sentiment analysis relate to one another and the position of emotion within this hierarchy of terms. This representation is not comprehensive, but helps to clarify the relationships among terms emerging in this research area.

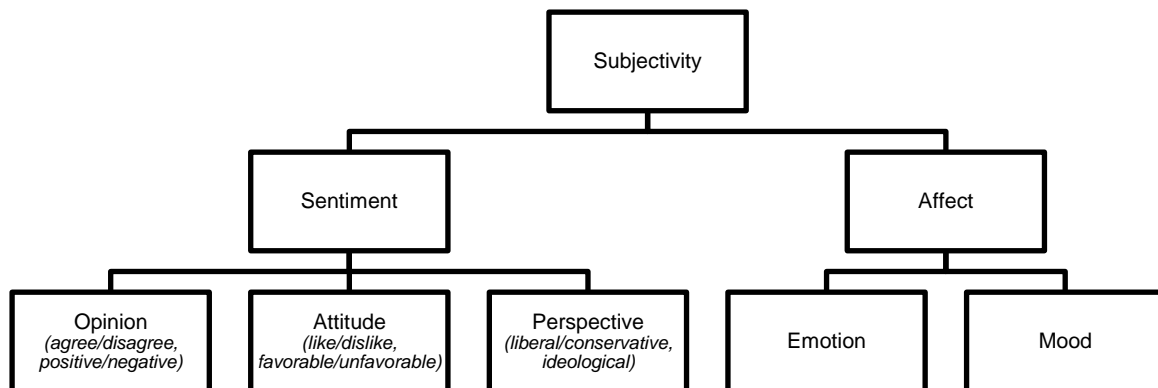


Figure 2.3: Commonly-used terms related to emotion in sentiment analysis

Subjectivity is used as the all-inclusive parent term to draw distinction between objective and subjective phenomena. The middle level in the hierarchy is conceived at the temporal

constraint dimension, which is based on the persistence of the subjective element across time (Clore et al., 2001; Clore & Schnall, 2005; P. Zhang, 2013). Affect can be observed in episodes, which are more ephemeral in nature. Sentiments are evaluative tendencies that are more stable, and remain within an individual for a longer period of time. Concepts that are grouped under “sentiment” include opinion, attitude, and perspective, while emotion and mood belong to the parent concept “affect”.

## **2.6 Conceptualizing Emotion in Text**

In this section, we provide a definition of emotion in text that integrates the different conceptualizations of emotions drawn from the four perspectives of emotion described in Section 2.2. The definition of concepts is an important first step in understanding how “emotion in text” is conceptualized, describing what is not part of the concept, and providing high level cues as to how to identify instances of emotion in text. In everyday language, people use the term emotion to refer to prototypes of common emotions such as happiness, sadness, and anger (Fehr & Russell, 1984). From a psychological perspective, emotion is generally defined as “ongoing states of mind that are marked by mental, bodily or behavioral symptoms” (Parrott, 2001, p. 3). The psychological definition focuses on the genesis of emotion, while emotion in text deals with how people talk about their emotions.

Kövecses & Palmer assert that “an emotion concept typically integrates content pertaining to all spheres of experience: social, cognitive, and physical” and “also invokes imagery pertaining to language and discourse” (Kövecses & Palmer, 1999, p. 253). The emotion language must reflect the blend of “universal experiences of physiological functions with culturally specific models and interpretations” (Kövecses & Palmer, 1999, p. 238). In line with this assertion, we adopt a more integrative view to define emotion in text. We define emotion in text as “a subset of particularly visible and identifiable feelings” (Besnier, 1990, p. 421; Kagan, 1978) that are expressed in written form through descriptions of expressive reactions (e.g.,

furrowed brow, smile), physiological reactions (e.g., increase in heart rate, teeth grinding), cognitions (e.g., thoughts of abandonment), behaviors (e.g., escape, attack, avoidance) as well as other socially prescribed set of responses (Averill, 1980a, 1980c; Cornelius, 1996).

To better distinguish the concept of emotion in text from the general concept of emotion, Figure 2.4 depicts the processes by which a writer generates text that expresses emotion, and by which a reader understands that expression. The target emotion to be detected is the one that is expressed by the writer. Emotional information gets encoded in text when writers are conscious of their emotional state, and verbalize their emotional experience in strings of characters or words. It is also possible for writers to subconsciously encode emotional information through their selection of words. Readers at the opposite end attempt to infer the writers' emotional state by decoding the emotional information in text. Each of these concepts is further described below.

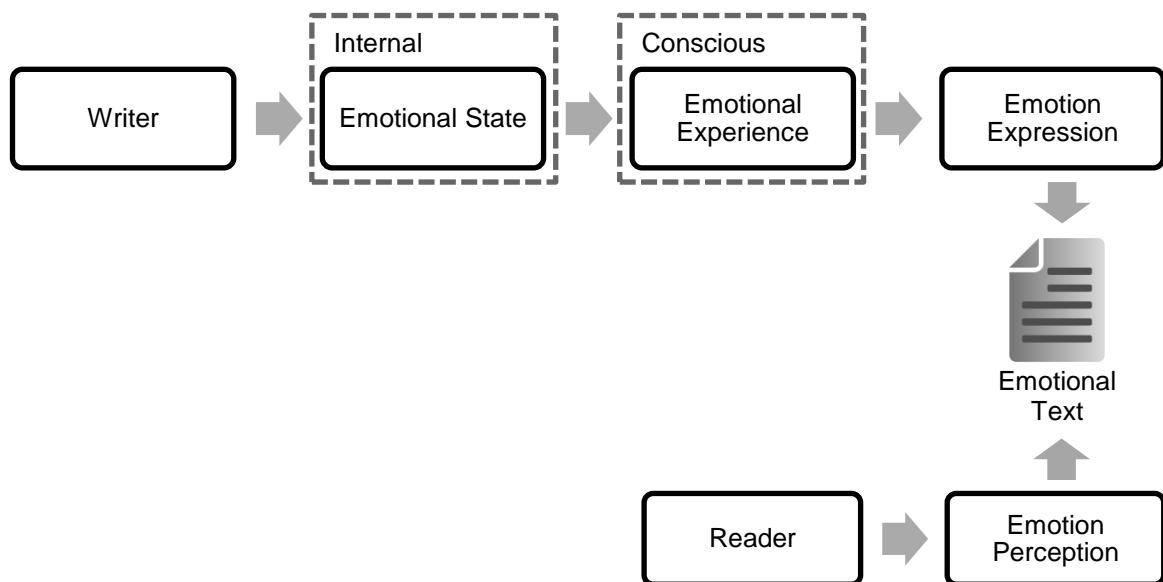


Figure 2.4: Concepts related to emotion in text

**Emotional State:** An emotional state is the smallest unit used to describe an emotion being felt by an individual. Picard (1998) defines emotional state as “internal dynamics” when a person



feels an emotion. Emotional state cannot be directly observed by another person, and can only be inferred through physical, physiological or behavioral observations. For instance, fear can be inferred through the physiological observation of a trembling body caused by terror, skin paling, sweat breaking out and hair bristling (Darwin, 1872).

***Emotional Experience:*** Emotional experience refers to “all that one consciously perceived as his or her emotional state” (Picard, 1998, p. 24). The individual is cognizant of the emotion being felt, and can recognize or articulate the emotional state. In addition, an emotional experience is often caused by some emotion stimuli that bring awareness to an individual’s emotional state. Emotional experiences are not limited to only the emotional state being activated at present. People also talk about their emotional experiences based on past events, although the intensity of emotion felt may not be the same (Fussell, 2002).

***Emotion Expression:*** Emotion expression consists of “signs that people give in various emotional states”, usually with the intention to be potentially perceived or understood by the others (Cowie, 2009). People express emotional states through different non-verbal (e.g., facial expression, vocal intonation, and gestures) and verbal (e.g., text, spoken words) manifestations. Specifically, emotion expression in text is the writers’ descriptions of the emotional experiences or feelings of their own or of others. It is important to note that emotion expression only provides a window to a person’s emotional experience depending on what the individual chooses to reveal to the others. It may not be depictions of a person’s actual emotional state, which is a limitation to the study of emotion in text (Calvo & D’Mello, 2010).

***Emotion Perception:*** Emotion perception is defined as “signs of emotion people can detect when they are alerted to them” (Cowie, 2009, p. 3515). Human perceivers can recognize another person’s emotions consciously when given appropriate guidance. Such recognition is improved in the presence of salient emotional signs or cues. For example, people perceive a person crying to mean that he or she is sad. Similarly, readers of a piece of text can detect

emotions by interpreting emotional cues given by a writer. If the writer states that “I am *happy* I won the lottery”, readers will most likely perceive the writer to be expressing happiness.

**Emotional Text:** Emotional text contains emotional cues that are expressed by the writer, and/or are perceived by the reader. Verbal cues in text are observed from the choices of words and types of sentences appearing in written text (Isbister & Nass, 2000). Traditional theories of communications such as the social presence theory (Short, Williams, & Christie, 1976) claimed that text lacks emotional expressions due to the reduction of non-verbal cues. The absence of vocal inflection, facial expression, and bodily movements causes text to be devoid of emotions. However, Wiener & Mehrabian (1968) argued that emotions can be expressed verbally through language variations. (Walther et al., 2005) supported this claim through the social information processing theory, which posited that verbal messages (i.e., text) contained emotions because writers would adapt to the medium by describing emotional cues in the form of words or other visual forms. For example, the use of adjectives and adverbs (descriptive words) can make the text more expressive (Benamara, Cesarano, Picariello, Recupero, & Subrahmanian, 2007), and can provide textual emotional cues.

## 2.7 Approaches to Identify Emotions in Text

Researchers have employed three different approaches to identify emotion in text: 1) emotion as declared by the writer, 2) emotion as perceived by the reader, and 3) triangulation between emotion declared by the writer and emotion perceived by the reader on the same text segment.

In the first approach, emotions are captured directly from the writer at the point when the text is written. Typically, writers are either asked to make explicit their emotions when they are posting a piece of text (e.g., tagging an text excerpt with an emotion label) (Keshtkar & Inkpen, 2012; Mishne, 2005) or describe their emotional experience given a set of emotions (HUMAINE, 2013). This seems to be the ideal way to capture the writers’ actual emotional state, but texts

with writers' own emotion annotations are hard to come by. It is useful only as an experimental protocol because it is not natural for writers to explicitly tag text segments with emotion labels unless they are specifically asked to do so. Also, existing systems need to be modified in order to capture, represent and store additional emotion metadata. Apart from that, different writers may not share the same conceptions and ways of using a particular emotion label, thus introducing a greater amount of noise in the data that can make scientific investigation of emotion in text more challenging.

The second approach captures the writer's emotion based on the perception of the reader. Readers or more commonly referred to as annotators are prompted to read a text segment, and then asked to identify or infer the emotion expressed by the writer in text. This approach provides a more practical approach for researchers to study emotion in text, and is dominantly used to mark up emotion for use in developing computational models (Alm et al., 2005; Aman & Szpakowicz, 2007; Scott et al., 2012). One drawback of the second approach is that emotion perceptions may be colored by the annotator's emotional state, background, and culture. This could potentially introduce greater noise in the study of emotion in text. One way to handle this issue is to provide annotators with more extensive training to reduce as much as possible the variance between annotator or the deviation from ground truth.

One criticism of both approaches is that each represents only a single angle to study emotion in text, and the reader's emotion perception may not be in-sync with the actual emotion expressed by the writer. The third approach takes on a more holistic view of emotion, and aims to integrate the first two approaches to ensure that the reader's emotion perception is reflective of the writer's emotion expression. This third approach is also less popular because such studies have to be conducted in a more controlled setting, and require more effort to recruit the writers and readers as participants.

## 2.8 Linguistic Representations of Emotion

A first step to building computational models that can detect emotions in text is to understand the aspects of language used to express emotion or what is hereafter referred to as “emotion language” (Kövecses, 2007). Emotion language is composed of two main elements: explicit emotion cues and implicit emotion cues.

Explicit emotion cues consist of words or typographical symbols that denote emotions. As shown in Table 2.5, emotion words can be further divided into two categories: expressive and descriptive (Kövecses, 2007). Expressive emotion words are words used to predicate the writer’s emotional experience (e.g., using the interjection “yuck!” when expressing disgust) while descriptive emotion words are nouns and adjectives used to describe emotions. Emoticons are typographical symbols invented to represent feelings or emotions in online interactions (Rezabeck & Cochenour, 1995).

Explicit Emotional Cue	Linguistic Unit	Examples
Descriptive emotion words	Word	happy, sad, surprise, hope, pride, love
Expressive emotion words	Word	yuck, wow, shit, haha
Emoticons	Typographical symbol	:-), :), :-(, :(, :O

Table 2.5: Examples of explicit emotion cues

On the other hand, implicit emotion cues are figurative descriptions of emotion, which are expressed implicitly through the use of a broad range of linguistic devices such as metaphors, metonymies, similes, idioms, etc. (see Table 2.6). In addition, implicit emotion cues can also be embedded within the linguistic structure. For example, emotions may be expressed in the form of actions associated with an actor towards an object (e.g., “*I cried because I failed my exam*” is an expression of sadness) (Balahur et al., 2012b) or the relationships between different entities (e.g., “*the mother scolded her daughter*” is an expression of anger) (Kövecses, 1990).

Many researchers assume that emotions are expressed explicitly using a handful of emotion words. This is the current dominant view in sentiment analysis as evidenced by the prior efforts to build automatic emotion detectors using only the emotion lexicons reviewed in Section 2.9.1. However, emotion words make up only a small fraction of emotion language. The richness of language allows people to express their emotions in many ways that do not use emotion words. According to Pennebaker, Mehl, & Niederhoffer (2003), emotion words account for only 4% of written words in text. Kövecses (2007) claims that implicit emotion cues make up the largest group of emotion expressions in language. Yet, it is the group that has received the least attention. The prevalence of implicit emotion cues in text, especially the manner in which these implicit emotion cues are structured in tweets, is unknown.

Implicit Emotional Cue	Description	General Examples
Metaphor	An object, activity, or idea that is used as a symbol of something else <sup>5</sup>	anger: blood is boiling [anger is heat] love: crazy about you [love is insanity]
Metonymy	Use of the name of one thing for that of another of which it is an attribute or with which it is associated <sup>5</sup>	anger: to see red fear: to have cold feet
Simile	Comparing two unlike things <sup>5</sup>	disgust: shoes smell like rotten egg love: love is like oxygen
Idiom	An expression with two or more words that has a meaning of its own and cannot be understood from the literal meanings of its separate words <sup>5</sup>	anger: chip on your shoulder jealousy: green-eyed monster happiness: in seventh heaven

Table 2.6: Descriptions and examples of implicit emotion cues

## 2.9 Automatic Emotion Detection in Text

Automatic emotion detection in text is framed as the problem of using a computational model to recognize segments of text expressing emotion. Growth in this area is fueled by the ready availability of subjective content on the Web, as well as the need to scale emotion

<sup>5</sup> Merriam-Webster: <http://www.merriam-webster.com/>

detection in text in a cost-effective manner since manual annotation is expensive. This section reviews the natural language processing techniques that have been employed in the development of automatic emotion detectors. Existing automatic methods can be classified into five main categories: 1) lexicon-based, 2) learning-based, 3) manually constructed rules, 4) knowledge-based, and 5) hybrid.

A key part to the discussion of automatic methods is the evaluation of automatic emotion detector performance. Generally, the performance of an automatic emotion detector is evaluated by comparing its predictions with gold standard data, corpora in which the “correct” answers have been marked. Performance is generally measured using four common metrics: 1) accuracy, 2) precision, 3) recall and 4) F-score.

- 1) Accuracy: Proportion of predictions that are correct
- 2) Precision: Proportion of the positive predictions that are correct
- 3) Recall: Proportion of positive cases identified correctly
- 4) F-score: Harmonic mean of precision and recall

While these measures are common, many others are used as well. See Sokolova & Lapalme, (2009) or Witten & Frank (2005) for a review. As sentiment analysis researchers have been using the term emotion and affect interchangeably in literature, the original terms from the source are retained in the following sections, and no distinction is drawn between affect and emotion.

### **2.9.1 Lexicon-based Approach**

Lexicon-based methods use a lexicon<sup>6</sup> to detect emotions in text and are considered to be easier to implement than other approaches. This approach is based on the assumption that individual words carry emotional coloring (Pajupuu, Kerge, & Altrov, 2012), and that emotions

---

<sup>6</sup> A lexicon is a dictionary of words for a particular language. An emotion lexicon is a specific type of lexicon that contains an inventory of words or lexemes related to emotion.

expressed in text can be adequately represented at the word level. Also known as keyword spotting, the lexicon-based approach is among the earliest approaches used for automatic emotion detection in text, having appeared in the early 2000s.

Lexicon Name	Scholar	Type	Context	Size	Original Word Source(s)
ANEW: Affective Norms for English Words	Bradley & Lang (1999)	Affect	General	1,034 words	<ul style="list-style-type: none"> <li>• 150 words: Mehrabian &amp; Russell (1974)</li> <li>• 450 words: (Bellezza, Greenwald, &amp; Banaji, 1986)</li> </ul>
LIWC: Linguistic Inquiry and Word Count Dictionary	Pennebaker et al. (2007)	Emotion, cognition, structural components	General	915 words (affective)	<ul style="list-style-type: none"> <li>• PANAS: (Watson, Clark, &amp; Tellegen, 1988)</li> <li>• Roget's Thesaurus</li> <li>• Standard English Dictionaries</li> </ul>
NRC Emotion Lexicon (EmoLex)	Mohammad & Turney (2008, 2010)	Emotion, sentiment polarity (positive, negative)	General	14,182 words	<ul style="list-style-type: none"> <li>• Macquarie Thesaurus</li> <li>• WordNet-Affect: Strapparava &amp; Valitutti (2004)</li> <li>• General Inquirer: (Stone, Dunphy, &amp; Smith, 1966)</li> </ul>
WordNet-Affect	Strapparava & Valitutti (2004)	Affect	General	4,787 words	<ul style="list-style-type: none"> <li>• AFFECT (1,903 words)</li> <li>• Dictionaries</li> <li>• WordNet: Miller (1995)</li> </ul>
AFINN	Nielsen (2011)	Affect	Twitter	2,477 words	<ul style="list-style-type: none"> <li>• Original Balanced Affective</li> <li>• Word List: Siegle (1994)</li> <li>• Urban Dictionary</li> <li>• The Compass DeRose Guide to Emotion Words: DeRose (2005)</li> <li>• Wikitionary</li> </ul>
Affect Database	Neviarouskaya et al. (2007a)	Affect	Instant Messaging	364 emoticons 337 acronyms & abbreviations 1,620 words	<ul style="list-style-type: none"> <li>• WordNet-Affect: Strapparava &amp; Valitutti (2004)</li> </ul>
Fuzzy Affect Lexicon	Subasic & Huettnner (2001)	Affect	General	3,876 words	<ul style="list-style-type: none"> <li>• Affect wordlist from newspaper articles by Mark Kantrowitz of Justsystem Pittsburgh Research Center</li> </ul>
Depheche Mood	Staiano & Guerini (2014)	Mood	News Articles	37,771 words	<ul style="list-style-type: none"> <li>• 13.5 million words from news articles on rappler.com</li> </ul>

Table 2.7: List of popular emotion/affect lexicons

Emotion lexicons play a central role in the lexicon-based approach. The performance of lexicon-based systems is dependent on the coverage and quality of the selected emotion lexicon (Neviarouskaya, Prendinger, & Ishizuka, 2011b). Commonly-used emotion and affect lexicons are summarized in Table 2.7.

Given an emotion lexicon, lexicon-based emotion detectors use a simple matching algorithm to extract emotion keywords from text based on the list of words found in the lexicon (Kao, Liu, Yang, Hsieh, & Soo, 2009). Text is first tokenized. Stemming or lemmatization can be used to reduce morphological variants of a word in the lexicon and text into its base form. Typically, the algorithm performs exact word matching (stems if stemming is applied or lemmas if lemmatization is applied), and tags the matching words in text with certain emotion-related attributes. Text is classified into different emotion categories based on emotion words. Various scoring methods have also been proposed to quantify the frequency of emotion words occurring in a text segment. For example, Grefenstette, Qu, Shanahan, & Evans (2004) determined the score for an entity (e.g, person) by dividing the number of positive affect words by the number of negative emotion words identified from a corpus of newspaper articles while Park et al. (2012) used the Linguistic Inquiry and Word Count (LIWC) program to generate a summary sentiment score that represented the percentage of total words that the positive and negative emotion categories accounted for in a sample of analyzed tweets.

In another more sophisticated lexicon-based method, Subasic & Huettnner (2001) employed what they called “*fuzzy semantic typing*” to first construct a fuzzy semantic lexicon (see Table 2.7) containing not only affect words, but also additional properties to represent the meaning of the affect words. Each affect word was associated with a part-of-speech (POS) tag, a centrality score (degree to relatedness to each affect class), and an intensity score (strength of word in an affect class). Each emotion word found in a document was tagged with its appropriate centrality and intensity scores. Combination of centrality scores, intensity scores,



and the document affect set was used to compute the emotional profiles of news and movie review documents.

Exact keyword matching has been criticized as being too simplistic, and results in poor performance when the meaning of the word is changed by the context in which it is used. For instance, the automatic emotion detector would not be able to handle negations (e.g., “not” and “never”). Keyword spotting only relies on obvious surface cues, and also cannot deal with text segments that convey emotion through the use of non-emotional words (Aman & Szpakowicz, 2008). To capture contextual information and deal with non-obvious emotion words, Tao (2004) added modifier words (e.g., very, too, and not) and metaphor words into the lexicon. He considered the syntactic structure of the sentences to identify the occurrences of modifiers and metaphors surrounding the emotion word. This approach yielded roughly 70% average precision based on text from a spontaneous speech corpus. Chuang & Wu (2004) also took into account positive and negative emotion modifiers in the detection of emotion in broadcast drama text but only reported an average recognition rate (i.e., number of true positives predicted out of the number of positive examples) across six basic emotion categories and a neutral category to be 65%.

Proponents of the lexicon-based approach have demonstrated its robustness by applying it on texts from different domains (Taboada, Brooke, Tofiloski, Voll, & Stede, 2011), as well as texts in a different language (Pajupuu et al., 2012). Unlike traditional exact keyword matching, the more advanced lexicon-based methods incorporated more sophisticated mechanisms to detect valence shifters including negations, intensifiers (e.g., very, more), and downtoners (e.g., slightly, less). Average accuracy reported using these augmented techniques is in the range of 70% to 80%.

The lexicon-based approach, being one of the earliest methods used for emotion detection in text has evolved significantly over the years. While emotion words are important explicit indicators of emotion, they are by no means the only type of emotion indicator in text.

Lexicon-based methods perform poorly in situations where emotion words are scarce and have been criticized for not being able to handle more complex linguistic structures of text.

## **2.9.2 Learning-based Approach**

Learning-based methods can be divided into two groups: supervised and unsupervised. Supervised learning techniques use marked up training data with pre-defined labels. Unsupervised learning searches for similarity between pieces of data to determine if they can be characterized as belonging to a cluster. The ability for learning-based methods to take into account contextual information and to capture emotional cues in segments longer than a word (i.e., sentence and message) makes it an appealing candidate to handle text with more nuanced emotional coloring.

### **2.9.2.1 Supervised Learning**

Supervised learning is concerned with the construction of computational models that can learn from training data. Supervised machine learning methods are more common than unsupervised for automatic emotion detection in text. A human-annotated corpus (i.e., gold standard data) is required to first train and evaluate a machine learning model. An emotion corpus contains text segments that are manually annotated with a pre-defined set of emotion categories. The machine learning algorithm then learns patterns associated with different emotion categories. Examples of corpora used for emotion analysis are shown in Table 2.8.

A set of features (e.g., individual words, also known as bag-of-words) are extracted from the training data set to be fed to a classifier. A classifier must be general enough to make accurate predictions not only on the training and testing data sets, but also subsequently on other unlabeled data sets. To evaluate the performance of a classifier, it is common to hold out a subset of the annotated data as the test set. This can be done via random splits or cross validation. Random split is a simple strategy that splits the data set in training and test subsets

based on pre-determined percentages. On the other hand, *k-fold* cross validation splits the data into *k* splits and, and runs the classifier *k* times, each time only training on *k-1* subsets and testing on the remaining single subset. This ensures each instance in the data set is used for training and testing equal number of times. In the training phase, the classifier learns patterns for each emotion category based on the set of identified features from the text. The classifier generated in the training phase will be used on the test data set, and its performance is evaluated through various metrics (including accuracy, precision, and recall).

Scholars	Categories	Context	Corpus Size	Unit of Analysis	# of Annotators
Wiebe et al. (2005)	Private state frames: polarity of attitude type (positive, negative, other, none)	News articles	535 documents (10,657 sentences)	Sentence	3
Alm et al. (2005)	7 emotion categories: anger, disgust, fear, happy, sad, positive surprise, negative surprise	Children stories	185 stories	Sentence	2
Aman & Szpakowicz (2007)	6 emotion categories: anger, disgust, fear, happy, sad, surprise	Blog posts	173 posts (5,205 sentences)	Sentence	4
Strapparava & Mihalcea (2007)	6 emotion categories: anger, disgust, fear, happy, sad, surprise	Newspaper headlines	1,250 headlines	Sentence	6
Brooks et al. (2013)	13 affect categories: interest, annoyance, amusement, surprise, anticipation, frustration, etc.	Chat logs	35,614 messages	Document	8
Gupta, Gilbert, & Di Fabbri (2010)	Emotional, non-emotional	Customer care emails	1,077 emails	Document	2
Rubin, Stanton, & Liddy (2004)	8 emotion octants with 38 sub-categories	Customer reviews	50 documents	Document	110
Pestian et al. (2012)	13 of 15 emotion categories: abuse, anger, fear, love, pride, etc.	Suicide notes	900 notes	Sentence	64
Mohammad et al., (2014)	8 emotion categories: anger, disgust, fear, happy, sad, surprise, anticipation, trust	Tweets	2,000 tweets	Document	Dozens

Table 2.8: Examples of emotion corpora

Scholars (Domain)	Labels	Classifier	ML Features	Results
Alm et al. (2005) Fairy tales	Case 1: Emotional (E), Non-Emotional (NE) Case 2: Neutral (N), Positive (PE), Negative (NE)	Winnow linear classifier – 10 fold cross- validation (90% train, 10% test)	<b>Content:</b> BoW by POS <b>Syntactic:</b> sentence length in words, verb count in sentence excluding participles, percent POS <b>Word list:</b> GI positive and negative word counts, WordNet emotion words, interjections, affective words <b>Orthographic:</b> special punctuations (! and ?), complete upper-case words, sentence quotes <b>Contextual:</b> ranges of story progress, first sentence in story, thematic story type (3 top and 15 sub-types) <b>Conjunction:</b> conjunctions of selected features	Case1: Avg accuracy (NE, E) = 0.63, 0.63 Avg error = 0.37, 0.37 Avg precision = 0.66, 0.56 Avg recall = 0.75, 0.42 Avg F-score = 0.7, 0.47 Case 2: Avg precision (N, NE, PE) = 0.64, 0.45, 0.13 Avg recall = 0.75, 0.27, 0.19 Avg F-score = 0.69, 0.32, 0.13
Holzman & Pottenger, (2003) Chat messages	Angry (An), Sad (Sa), Afraid (F), Disgusted (D), Happy (H), Surprise (Su)	K-nearest neighbor (KNN) – 10-fold cross validation	<b>Content:</b> average word length, maximum word length <b>Phonetic:</b> Phoneme counts <b>Orthographic:</b> special punctuations (., !, ?)	Case 1: Happy, Neutral Precision = 0.859, 0.839 Recall = 0.595, 0.987 F-beta = 0.703, 0.91 Case 2: Emotional, Neutral Precision = 0.804, 0.844 Recall = 0.345, 0.977 F-beta = 0.482, 0.905 Case 3: H, An, Su Precision = 0.528, 0.643, 0.321 Recall = 0.647, 0.850, 0.230 F-beta = 0.581, 0.732, 0.263
Brooks et al., (2013) Chat messages	Interest, Amusement, Considering, Agreement, Annoyance, Confusion, Acceptance, Apprehension, Frustration, Supportive, Surprise, Anticipation, Serenity	SVM – linear kernel (SMO) – 10-fold cross validation	<b>Content:</b> BoW (stemmed, lowercase) <b>Syntactic:</b> number of pronouns <b>Word list:</b> number of negation words, swear words, and known people names <b>Orthographic:</b> number and length of punctuations, number and length of capital letters, "hmmm"-variants, laughter phrases, and repeated letter sequences 3 or longer, number of emoticons <b>Document:</b> duration, length, characters/second, average rate of messages in the segment	Accuracy = 0.761 Precision = 0.766 Recall = 0.751 F-measure = 0.759

Scholars (Domain)	Labels	Classifier	ML Features	Results
Mishne (2005) Blogs	40 Mood Categories	SVM (400 test, 400 – 6400 train)	<b>Content:</b> unigram word count, POS tag count, word lemma frequency <b>Syntactic:</b> average sentence length in bytes, average word count per sentence, PMI-IR <b>Word list:</b> total semantic orientation of a post, and average word orientation in the blog based on a list verbs, nouns, and adjectives <b>Orthographic:</b> emphasized words (frequency of each emphasized word in a post + total number of stressed words per post), special symbols (frequencies of 15 punctuations and 9 emoticons) <b>Document:</b> length in bytes, number of words in a post	Accuracy (6400 train set) Confused = 0.66 Curious = 0.63 Happy = 0.61 Amused = 0.61 Sad = 0.6 Excited = 0.6 Annoyed = 0.59 Love = 0.58 Hopeful = 0.58 Accomplished = 0.56 Bored = 0.55 Anxious = 0.54 Exhausted = 0.53 Calm = 0.49
Aman & Szpakowicz (2007) Blogs	Happiness, Sadness, Anger, Disgust, Surprise, Fear, Mixed emotion, No emotion	Naïve Bayes, SVM – 10-fold cross validation	<b>Word list:</b> emotion words from GI (EMOT, Pos/Pstv, Neg, Ngvtv, Intrj, Pleasure, Pain), WordNet-Affect <b>Orthographic:</b> emoticons, !, ?	Accuracy Naïve Bayes = 0.72 SVM = 0.74
Aman & Szpakowicz (2008) Blogs	Happiness (H), Sadness (Sa), Anger (An), Disgust (Di), Surprise (Su), Fear (Fe), Mixed emotion, No emotion (NE)	SVM (SMO in Weka) – 10-fold cross validation	<b>Content:</b> unigrams (frequency > 3, stopwords) <b>Word list:</b> emotion words from Roget's Thesaurus and WordNet-Affect	Precision (H, Sa, An, Di, Su, Fe, NE) = 0.813, 0.605, 0.650, 0.672, 0.723, 0.868, 0.587 Recall = 0.698, 0.416, 0.436, 0.488, 0.409, 0.513, 0.625 F-measure = 0.751, 0.493, 0.522, 0.566, 0.522, 0.645, 0.605
Keshtkar & Inkpen (2012) Blogs	132 mood categories	SMO in Weka	<b>Content:</b> frequency of words, POS tag counts <b>Document:</b> document length, number of sentences, average number of word <b>Word list:</b> positive and negative words from GI, Kim & Hovy's list and Turney & Littman's list <b>Orthographic:</b> 9 emoticons	Global Accuracy Flat = 0.247 Hierarchical = 0.799
Gupta et al. (2010) Emails	Emotional, Non-emotional	Boostexter - cross validation (620 train + 457 test)	<b>Content:</b> n-grams (unigrams, bigrams, trigrams) <b>Word list:</b> presence of words/phrases from specific salient features dictionaries	Precision = 0.81 Recall = 0.65 F-measure = 0.72

Scholars (Domain)	Labels	Classifier	ML Features	Results
Mohammad (2012c) News headlines & Blogs	Anger (A), Disgust (D), Fear (F), Joy (J), Sadness (Sa), Surprise (Su)	Logistic Regression, SVM	<b>Content:</b> unigrams, bigrams <b>Word list:</b> emotion words from NRC-10, NRC-6 and WordNet-Affect	Precision = 0.506 Recall = 0.544 F-measure = 0.524 Precision (A, D, F, J, Sa, Su) = 0.42, 0.47, 0.59, 0.51, 0.66, 0.21) Recall = 0.35, 0.15, 0.8, 0.68, 0.68, 0.33 F-measure = 0.38, 0.23, 0.68, 0.58, 0.67, 0.25
Mohammad, (2012a) Tweets	Anger (A), Disgust (D), Fear (F), Joy (J), Sadness (Sa), Surprise (Su)	SVM (SMO) – 10-fold cross validation	<b>Content:</b> unigrams, bigrams (presence/absence) – frequency > 1	Precision = 0.551 Recall = 0.456 F-measure = 49.9 Precision (A, D, F, J, Sa, Su) = 0.37, 0.31, 0.6, 0.65, 0.42, 0.51) Recall = 0.22, 0.13, 0.44, 0.6, 0.36, 0.41) F-measure = 0.28, 0.19, 0.51, 0.62, 0.39, 0.45
Mohammad et al. (2014) Tweets	Anger, Anticipation, Disgust, Fear, Joy, Sadness, Surprise, Trust	SVM: LibSVM – simple linear kernel (10-fold cross validation)	<b>Content:</b> unigrams, bigrams (stemmed - Porter) <b>Word list:</b> NRC (frequency of emotion words), Osgood's semantic differential categories for Wordnet and GI (frequency of adjective or adverb sense) <b>Orthographic:</b> frequency contiguous sequences of !, ?, and combination of ! and ?, presence/absence of positive and negative emoticons, elongated words <b>Negation:</b> presence of negators, proximity of negator to emotion word <b>Contextual:</b> position of feature terms (appear at the beginning or end of tweet) <b>Conjunction:</b> presence of emotion words from multiple emotion categories (word list)	Accuracy = 0.568
Hasan, Rundensteiner, & Agu (2014) Tweets	Happy-Active, Happy-Inactive, Unhappy-Active, Unhappy-Inactive	SVM (SVM-light), Naïve Bayes, Decision Trees, K-Nearest Neighbors	<b>Word list:</b> LIWC dictionary (emotion-indicative categories and negation) <b>Orthographic:</b> emoticons, punctuations	F-measure = 0.901 (KNN) Precision = 0.902 (SVM) Recall = 0.901 (KNN)

Scholars (Domain)	Labels	Classifier	ML Features	Results
Roberts et al. (2012) Tweets	Anger (A), Disgust (D), Fear (F), Joy (J), Love (L), Sadness (Sa), Surprise (Su)	SVM	<b>Content:</b> unigrams, bigrams, trigrams, significant words <b>Word list:</b> WordNet synsets and hypernyms <b>Orthographic:</b> ! and ? presence <b>Contextual:</b> topic scores	Macro-average Precision = 0.721 Recall = 0.627 F-measure = 0.668 Precision (A, D, F, J, L, Sa, Su) = 0.672, 0.717, 0.897, 0.656, 0.725, 0.747, 0.631 Recall = 0.615, 0.622, 0.629, 0.697, 0.599, 0.637, 0.587 F-measure = 0.642, 0.666, 0.74, 0.676, 0.656, 0.688, 0.608
Wang et al., (2012) Tweets	Joy (J), Sadness (Sa), Anger (A), Love (L), Fear (F), Thankfulness (T), Surprise (Su)	LIBLINEAR, Multinomial Naïve Bayes	<b>Content:</b> unigrams, bigrams, trigrams and their combinations (f > 5, Boolean) <b>Syntactic:</b> percentage of words for each POS, adjectives <b>Word list:</b> percentage of positive and negative emotion words per tweet from LIWC and MPQA, count of WordNet-Affect emotion words <b>Orthographic:</b> punctuations, emoticons <b>Contextual:</b> n-gram position (1 <sup>st</sup> or 2 <sup>nd</sup> half of tweet)	Accuracy = 0.616 Precision (J, Sa, A, L, F, T, Su) = 0.676, 0.626, 0.698, 0.581, 0.597, 0.666, 0.447 Recall = 0.773, 0.668, 0.733, 0.462, 0.347, 0.5, 0.082 F-Measure = 0.721, 0.647, 0.715, 0.515, 0.439, 0.571, 0.139
Cherry et al. (2012) Suicide notes	Abuse, Anger, Blame, Fear, Forgiveness, Guilt, Happy/Peaceful, Hopeful, Hopeless, Information, Instructions, Love, Pride, Sorrow, Thankful	Linear SVM	<b>Content:</b> BoW (lowercase, unigrams, bigrams) <b>Syntactic:</b> sentence length in tokens, presence of manually-designed word classes, capitalized words, anonymized names, future tense verbs. <b>Word list:</b> count of words matching each category from Roget's Thesaurus <b>Orthographic:</b> set of cased character 4-grams, general upper/lowercase patterns <b>Document:</b> length	Precision = 0.674 Recall = 0.649 F-measure = 0.614

Table 2.9: Feature sets for supervised machine learning

Text is typically segmented into sentences, but the size of a text segment depends on the unit of analysis determined by the researchers. In binary classification, a text segment is classified as either being a positive or negative example of an emotion category. Identifying if a

text segment is emotional or non-emotional is an example of binary classification (Alm et al., 2005; Gupta et al., 2010). For sentences that contain more than one emotion, researchers have either included them in a separate category labeled as “mixed emotions” (Aman & Szpakowicz, 2008) or allowed multiple labels to be assigned to each sentence (i.e., multi-label classification problem) (Alm et al., 2005).

When training a classifier, selecting a set of representative features from which to construct the classifier is a central problem in machine learning. Researchers have used various feature sets as shown in Table 2.9. The performance numbers reported in Table 2.9 and discussed in this section are based on the exact numbers used by each author. Both bag-of-words (BoW) and word n-grams are popular features for emotion detection in text. BoW has been proven to be a successful feature set in sentiment analysis (Pang et al., 2002; Salvetti, Reichenbach, & Lewis, 2006). In terms of machine learning algorithms, support vector machines (SVMs) are popular for this problem space as they can scale to a large number of features and can outperform other classifiers for text classification (Yang & Liu, 1999). Chaffar & Inkpen (2011) showed that SVMs performed and generalized well on unseen data in emotion classification. They investigated the performance of classifiers using three different machine learning algorithms (i.e., Naïve Bayes, decision trees and SVM) on a heterogeneous corpus (i.e., news headlines, fairy tales and blogs) annotated with six basic emotions and reported that SVM yielded the greatest accuracy improvement compared to the baseline.

Alm et al. (2005) examined a stacked model to classify emotion in children’s fairy tales. The fine-grained set of emotion categories used suffered from a sparsity of data so the emotion categories were collapsed into coarser-grained categories: positive, negative and neutral. The machine classification of negative emotion category managed to achieve an average F1 of 0.32 while the positive emotion category only managed to achieve an average F1 of 0.13 due to limited amounts of training data. Using the same set of features and classifier at an even



coarser-grained level of classification (i.e., either a sentence is emotional or non-emotional) increased the average F1 to 0.47 (average accuracy = 0.63) for the combined emotion class.

Based on the claim that only obvious emotion words bear weight in characterizing emotion expressions in text, Aman & Szpakowicz (2007) used only emotion words from two lexicons (General Inquirer and WordNet-Affect), and punctuations as features for the machine learning algorithm used to classify emotions in blog posts. They achieved an average accuracy of 0.74. Further investigation on finer-grained emotion classification on blog posts using SVM showed that a combination of corpus-based and emotion lexicon features yielded slightly better results compared to using only corpus-based or emotion lexicon features alone (Aman & Szpakowicz, 2008).

Similarly, Mohammad (2012c) demonstrated that emotion lexicon features provided significant gains in classification accuracy when combined with corpus-based features (e.g., unigrams, bigrams) for classification of news headlines. Such gains were only observed when training and test sets were drawn from the same domain. When a model trained on news headlines was applied to blog posts, using only lexicon features produced the best results, thus supporting the contention that lexicon features are more portable for classification across domains. In contrast, Abbasi, Chen, Thoms, & Fu (2008) found that n-grams used in conjunction with automatically-generated lexicons for the prediction of emotion intensity in short stories, blogs and forums did not outperform the use of n-grams alone. More research is needed to examine the cause of this contradictory finding. It is possible that emotion lexicon features may only be advantageous in the classification of nominal emotion variables (i.e., categorical labels) and not continuous ones (i.e., numerical values such as emotion arousal or intensity measured on scale).

Gupta et al. (2010) reported that using a set of salient features extracted from customer care emails for emotion classification yielded better performance when compared to using unigram features, thus providing empirical evidence that emotion-specific features are more

effective than unigram features. A combination of salient and unigram features produced only a slight 2% increase in the F-measure. A general observation from the literature is that machine learning models seem to respond positively to the use of specific emotion-related features.

Tackling another set of challenges caused by the short message length in chat data, Brooks et al. (2013) segmented the corpus by combining chat messages within a particular time frame together to increase contextual cues for a unit of analysis. Features were derived from observation of language use and text characteristics in this domain. Using binary classifiers for multi-label classification of 13 affect classes (i.e., training one classifier for each emotion class), F-measure of the classifier ranged from 0.63 to 0.93. Binary classification was also shown to be effective in the multi-label classification of sentences from suicide notes for 15 emotion categories (Cherry et al., 2012). Using a combination of bag-of-words, thesaurus, character and document features, the classifier achieved an average F-score of 0.55. Brooks et al. and Cherry et al. both had to deal with the issue of class imbalance as the number of negative examples in the corpus far outweighed the number of positive examples. The former adopted a downsampling strategy (i.e., randomly removing negative examples) while the latter adjusted the class weights to reduce the impact of negative examples on learning.

To deal with misspellings and non-standard words in chat messages, Holzman & Pottenger (2003) used reproduction of speech phonemes from text combined with other statistics extracted from the training set as features to train a k-nearest neighbor (KNN) classifier to classify six basic emotions. They observed a relationship between phoneme counts and emotion class, and reported that including the phonetic feature proved to be useful in helping the classifier distinguish between different emotion classes given proper representation in the training set. The KNN classifier exceeded an accuracy of 0.9.

In another study related to the detection of a writer's mood from blog posts, Mishne (2005) experimented with a blog corpus that has already been tagged with the writer's mood from LiveJournal. More advanced features such as Pointwise Mutual Information and

Information Retrieval (PMI-IR) (Turney, 2001), and the semantic orientation of the blog were tested but yielded only accuracy ranging from 0.5 to 0.65. Also, empirical evidence showed that human performance on the classification of mood in blog post was not substantially better than machine performance. This study shows that using a corpus that is pre-tagged by the writer can be problematic as distinct writers may interpret and use the mood labels in different and inconsistent ways, thus introducing a greater amount of noise in the data. Mishne & De Rijke (2006) further showed that machine learning models could be trained to predict mood levels over time spans more accurately than predicting the mood of a single blog post.

Unlike Mishne (2005) who adopted a flat classification approach (i.e., treating each mood as a discrete category), Keshtkar & Inkpen (2012) experimented with mood classification in blogs using a hierarchical approach. In an attempt to perform classification on 132 mood categories, the intuition behind the hierarchical approach was to reduce the complexity of the problem by first allowing the classifier to learn coarse-grained distinctions, and then focus on the fine-grained distinctions at the lower levels of the hierarchy. They created a 5-level classifier, wherein the top most level was first trained to distinguish between 15 mood categories that were not closely related to one another. The hierarchical classification strategy yielded overall accuracy (i.e., taking into account of errors from all levels) of 0.799. Flat classification performed on the same data set resulted in only an accuracy of 0.247. The authors also replicated the classifier reported in Mishne (2005), and claimed that an accuracy improvement of 27% was obtained. However, the hierarchical structure must first be defined before training the classifiers.

To summarize, classifiers built using supervised learning techniques have yielded moderate to high accuracy (0.63 – 0.76), precision (0.56 – 0.77), recall (0.42 – 0.81) and F1 (0.47 – 0.76) (Alm et al., 2005; Aman & Szpakowicz, 2007; Brooks et al., 2013; Cherry et al., 2012; Gupta et al., 2010; Pestian et al., 2012; Yang, Willis, de Roeck, & Nuseibeh, 2012). There is a considerable range in the difficulty of detecting some emotion categories; some emotion categories have shown higher performance than others. For instance, “*interest*” yielded an F1 of

0.93, while the F1 for “*apprehension*” was only 0.64 in Brooks et al. (2013). Classifiers have also yielded poor performance for some emotion categories such as “*pride*” (F1 = 0.21) and “*abuse*” (F1 = 0.2) in Yang et al. (2012). The intention here is not to compare across the classification performance of different emotion categories since it is hard to compare results from different techniques. However, it is important to be aware that there are considerable variations observed in existing classification results, and it may be worthwhile to seek out the underlying cause to such discrepancies.

In order to produce a high-performing classifier, there must be a large enough sample of each label for training. Yang et al. (2012) claimed that the scarcity of training samples is the main reason behind very low classification results (e.g., emotion categories with less than 100 instances in gold standard data), while Mishne (2005), Mohammad (2012a) and Wang et al. (2012) demonstrated that the performance of classifiers can be improved by increasing the amount of training data. Emotion categories with a larger set of training samples have fared better than those that suffer from a dearth of samples. Therefore, sufficiently large human-annotated data for training are necessary in order to produce usable machine learning models. Having humans annotate hundreds of samples of each emotion category may not be feasible due to time and cost constraints. Performance of classifiers also depends on the number of emotion categories (learning to distinguish between 50 categories is harder than 5).

Also, classifiers generally perform well when tested on data from the domain that they are trained on. Their performance suffers when trained and tested using different domains (Pang & Lee, 2008). Feature sets developed for one domain often do not work well in other domains. The most predictive features may be strongly correlated to human cognitive processing of linguistic cues in text since humans do reasonably well in seeking out patterns, including the more complex ones. If so, researchers have to find out how to surface the most salient and useful features from the linguistic cues that humans generally use to express emotions in various text domains. This can potentially increase the methodological validity of a

classifier, as well as to improve interpretation of the classification results. There is no common agreement thus far as to which features are the most relevant and useful to define emotion in general, as well as each emotion category in specific. Extracting features that are generalizable across domains remains a challenge to the proponents of supervised learning methods.

To overcome the current limitations of supervised learning methods, it is important to find efficient ways to generate larger gold standard data and explore different strategies to increase the scalability of supervised learning methods across different domains.

### **2.9.2.2 Unsupervised Learning**

Unsupervised learning methods for emotion detection in text have only emerged fairly recently. Most of these methods are proposed to handle detection of emotions that are expressed implicitly in text (i.e., expressions that do not contain obvious emotion keywords). One popular unsupervised learning method in this problem space is latent semantic analysis (LSA). Strapparava & Mihalcea (2008) assessed the semantic similarity among the terms in a given text and emotion concepts using a variation of latent semantic analysis (LSA), an unsupervised learning method. LSA allows vectors containing emotion words, its synonyms or synsets and document vectors containing generic terms to be mapped into a concept space (i.e., a smaller and more compact space that is intended to preserve the ability to discriminate important concepts). Of the five methods tested by Strapparava & Mihalcea (2008), the LSA methods resulted in relatively higher recall and F-score than both lexicon-based and supervised learning-based methods but achieved the worst precision.

L. Zhang (2013) also applied LSA to inform affect processing of an intelligent agent in a role-playing virtual drama application. The intelligent agent was programmed to detect implicit emotion expressions of the human characters in a session, and produce appropriate responses based on the detected affect. LSA was used to identify discussion themes and target audiences in predefined scenarios. Terms in the documents with discussion inputs were mapped in a

concept space. Within this concept space, the similarity score between each document input and topic terms is generated. This technique used to deal with text without strong affect indicators achieved overall average precision and recall of over 80%. While LSA was not directly used to detect emotions in text, the researchers believed that the ability to identify the topic of discussion would increase accurate interpretation of the emotional context.

LSA has also been employed in Ahmad & Laroche (2015) to measure emotion in Amazon customer reviews. Specific words denoting four emotions of interest as well as their associated synonyms (i.e., happiness, hope, disgust and anxiety) and the consumer review were represented as vectors in concept space. The distance score between the emotion vector and a given customer review vector was then computed to determine the emotion class to be assigned to the customer review. Each emotion vector was constructed using words denoting the emotion as well as their associated synonyms.

### **2.9.3 Manually Constructed Rules**

The manually constructed rule-based approach uses rules to decide if a text segment contains emotion or not. First, rules are generally defined manually from an initial data set. Researchers have to analyze sample text to look for grammatical patterns associated with each emotion category or derive patterns based on a theoretical framework. These patterns are manually converted into a list of rules, which acts as the basis for a rule engine or inference engine. Rules need not be limited to lexical cues (e.g., keywords) in text, but can also deal with the more complex syntactic and semantic structures of a sentence. Syntactic (e.g., parts of speech) and semantic (e.g., semantic role labeling) information is obtained by running texts through a parser.

Automatic emotion detection in text using manually constructed rules is also one of the early approaches that emerged along with the lexicon-based approach. Many manually constructed rule-based methods develop complex rules based on emotion lexicons to deal with

the complexity of language. Zhe & Boucouvalas (2002) set up syntactic rules to include only emotion words expressed in first person form, took into account present continuous and perfect continuous tense as an indicator of emotion intensity, and excluded conditional sentences in an Internet chat environment. To detect anger in newsgroup, Donath, Karahalios, & Viégas (1999) set up rules to detect phrases in all capital letters, excessive punctuations, and profanities. In processing news titles, Chaumartin (2007) leveraged syntactic rules to determine the subject of the news title, as well as to detect contrasts and accentuations between good news and bad news.

Semantic rules often attempt to capture the core <subject><action><object> structure within a sentence. Liu et al. (2003) defined four rules to represent affective commonsense sentences from the Open Mind Commonsense Corpus. These four rules aimed to cover varying aspects of the core semantic structure in a sentence. Rules ranged from being very specific to preserve the accuracy of the affective knowledge to more general ones defining how to deal with different affective concepts. Shaikh et al. (2009) used a series of rules to implement a linguistic version of the OCC model<sup>7</sup> (Ortony et al., 1988) for emotion detection in text. The OCC model was originally conceived as a blueprint for rule-based emotion reasoning systems so the model can be converted into high-level rules. However, the OCC model did not include the linguistic details required to handle detection of emotion in text. Various linguistic resources including a semantic parser, scored POS lists, ConceptNet (Liu & Singh, 2004), and SenseNet (Shaikh, Prendinger, & Ishizuka, 2007b) were utilized to map the linguistic cues to the OCC concepts. Neviarouskaya, Prendinger, & Ishizuka (2011a) proposed a rule-based method that processed sentences in five stages according to the different unit of analysis. Symbols and abbreviations were processed first, and then followed by word, phrase, and sentence-level analyses.

---

<sup>7</sup> Ortony et al.'s (1988) global structure of emotion types (OCC model – short for Ortony, Clore and Collins) was designed with the goal to create computer models that can understand and predict people's emotional reactions in various conditions.

Syntactic rules are fairly straightforward to implement especially with the availability of state-of-the-art parsers. However, parsers are not perfect, and the results may be affected by errors introduced by the parsers. The linguistic implementation of semantic rules is more complex. Semantic rules are defined in terms of concepts (e.g., <subject><action><object>) so it is necessary for researchers to define the linguistic representations of these concepts. The more abstract a concept, the more complex it is to define all the linguistic representations of the concept.

The strength of the manually constructed rules lies in its more transparent representation of emotion patterns in text, at least for relatively small rule sets. Explanations can be generated for most instances captured by the rules because each rule pattern is clearly defined. This also applies to the interpretation of incorrectly identified instances. Researchers can refer to the pattern definition of a rule to find the cause of an error. However, it is impossible to capture instances of emotion not defined by any rules. Most often, only a limited number of rules are defined to capture the obvious and non-ambiguous patterns. The generalizability of rules is also a cause for concern. If rules are defined narrowly to work for a particular domain, they may perform poorly when applied in other domains. Since defining rules manually is a tedious task, it is difficult to define a comprehensive set of rules to cover all patterns of emotion expressions. For these reasons, the approach using manually constructed rules has not gained as much popularity compared to other approaches in the research community. Attempting to define all the linguistic rules of emotion is indeed a challenging endeavor as there are too many ways for emotions to be expressed in text.

#### **2.9.4 Ontology-based Approach**

The ontology-based approach focuses on the creation of a machine-readable formal representation of human emotions. Ontology is an “explicit specification of conceptualization” for a particular domain (Gruber, 1995). This structural representation includes a domain



vocabulary, descriptions of concepts and attributes, as well as the relations between concepts. Unlike lexicons, ontologies do not operate on a word-level (i.e., low-level linguistic cues). Rather, they are defined in terms of high-level concepts. Concepts are connected through taxonomic relations (e.g., subclass-superclass), and semantic relations (e.g., part-of, has-a). Motivation for researchers to adopt this approach mainly stemmed from the lack of agreement in how emotion is defined in the research community. Proponents of the ontology-based approach aim to define a standard set of descriptors that can help reduce the ambiguity in the interpretation of emotion expressed in text.

Ontology-based methods are concerned with the creation, modification, and testing of emotion ontologies. The adoption of ontology-based methods for emotion detection in text is still fairly new, and has started to appear in the literature only a few years ago. One of the earliest attempts to build an emotion ontology came from Grassi (2009). Grassi (2009) defined only high-level emotion concepts and properties in the Human Emotions Ontology (HEO). The concepts, properties, and relations were derived from multiple emotion theories well-known in psychology. Shivhare & Khethawat (2012) proposed a simple emotion ontology based on Parrott's emotion word hierarchy (Parrott, 2001).

Emotion ontologies can also be modeled based on common sense knowledge (Balahur et al., 2011). Grounded on appraisal theories (Scherer, 1999), Balahur et al. (2012a) modeled situations as "action chains" and their corresponding emotion using an ontology representation. The ontology, known as EmotiNet, was designed to address the problem of detecting implicit emotions. These action chains represent a sequence of actions that trigger an emotion. While the core of the ontology was designed manually, concepts within this core were populated semi-automatically using different existing knowledge bases. Compared to previous emotion ontologies, this type of ontology provides a greater amount of details through the definitions and interactions of low-level concepts and high-level concepts, but also introduces a greater level of complexity. For EmotiNet to be useful, it needs to be extended using existing knowledge bases

to cover as many emotion-triggering situations as possible. Therefore, it is limited to the knowledge that current knowledge bases provide. Performance of EmotiNet for emotion detection in text is reported to be at par with supervised learning methods (Balahur et al., 2012b).

As the ontology-based approach operates on the semantic level, it suffers from the same problems as semantic rules from manually constructed rules. Texts have to first go through a semantic parser in order to match linguistic elements in text to their corresponding concepts. This means that the performance of ontology-based approaches is affected by the accuracy of the semantic parsers. Errors introduced by the semantic parsers may result in overall poor performance of the emotion detector. Although ontologies serve to provide some form of standardization on the knowledge of emotion, extensive efforts are needed to build a consistent, if not a comprehensive one. Initial deep understanding and analysis of emotional text is required for the construction of emotion ontologies because researchers will need to model how different emotion concepts relate to one another. Furthermore, ontologies are domain-specific, and may not generalize well across domains different from the one it is built upon.

### **2.9.5 Hybrid Approach**

Hybrid approaches combine at least two of the four main approaches used for emotion detection in text: lexicon-based, learning-based, manually constructed rules, and ontology-based. A hybrid approach aims to strategically leverage the strengths of different selected approaches in an integrative framework. For example, Ma, Prendinger, & Ishizuka (2005) constructed a textual emotion estimation engine for a chat system with animated agent by combining keyword spotting for emotion estimation of words and a set of rules for emotion estimation of sentences.

The surge of hybrid approaches is apparent in more recent research. In the 2011 Medical Natural Language Processing Challenge organized by the i2b2/VA/Cincinnati to assign

emotions to suicide notes, many of the proposed automatic emotion detectors are implemented using the hybrid approach (Pestian et al., 2012). Yang et al. (2012) designed a voting-based system to pick emotions for each sentence based on outputs from a combination of keyword spotting, conditional random field (CRF), and supervised machine learning methods. To address the same problem, Nikfarjam, Emadzadeh, & Gonzalez (2012) first used rules to filter out sentences with obvious emotional cues, and passed the uncertain cases to a supervised machine learning model for a final decision. Sohn et al. (2012) also concluded that the union of manually constructed rules and supervised machine learning methods resulted in better performance compared to using rules or machine learning alone.

Narducci, de Gemmis, & Lops (2015) used a combination of classifiers and a thesaurus to create an emotion analyzer, a core component in their recommender system. The thesaurus was generated through a synonym enrichment procedure that extracted synonyms of a set of emotion seed words from WordNet. The synonyms were represented as dimensions in an emotion vector, and the emotion vectors were then mapped into a multidimensional vector space. The emotion label for a text segment was determined by computing the cosine similarity of the targeted text vector with respect to the emotion vectors. Three classifiers (i.e., Naïve Bayes, SVM and Random Forest) were also used to respectively assign an emotion label to each text segment. The final emotion label was determined using a voting algorithm. A comparison between the emotion assigned by the system and the emotion declared by the user revealed that the performance of the emotion analyzer varied across different emotion categories with “joy” yielding the best performance ( $F1 = 0.79$ ) and anger showing the worst performance ( $F1 = 0.33$ ).

Hybrid methods provide an alternative approach to combine strengths of different approaches together or use the strengths of one approach to overcome the weaknesses of another, thus creating more optimal and efficient automatic emotion detectors. For example, filtering out obvious emotional cues using rules or lexicons first can reduce the load for machine

learning, and increase the speed of classifiers. Determining which combination of approaches work optimally together remains a challenge for the research community.

## **2.10 Automatic Emotion Detection on Twitter**

Automatic emotion detection on Twitter is a fairly new area of research but is gaining traction in computational linguistics and social computing. Current research efforts are mainly aimed at: 1) analyzing emotional content on Twitter, or 2) creating language resources to support large-scale understanding of emotions expressed on Twitter.

Keyword spotting (i.e., lexicon-based approach) is the dominant method used for the analysis of emotional content on Twitter. In a study to examine the correlation between collective mood states on Twitter and stock market fluctuations over time, Bollen, Mao, et al. (2011) employed the OpinionFinder subjectivity lexicon and Google Profile of Mood States (GPOMS) emotion terms to detect six mood states (calm, alert, sure, vital, kind, and happy) in tweets. In another study to model collective emotion trends on Twitter, Bollen, Pepe, et al. (2011) used the Profile of Mood States (POMS) scoring function, which matched terms extracted from each tweet to an extended version of the POMS emotion terms for six mood states (tension, depression, anger, vigor, fatigue, and confusion).

Keyword spotting has also been applied in the analysis of more specific emotions on Twitter such as detecting anxiety during disasters, and sadness following the death of a famous celebrity. A list of English and Japanese keywords associated with anxiety events was used to extract tweets containing expressions of anxiety to investigate how patterns of public anxiety change throughout an earthquake (Doan, Vo, & Collier, 2012). Kim, Gilbert, Edwards, & Graeff (2009) performed an analysis of the average emotion valence, arousal, and dominance ratings generated based on the ANEW lexicon (Bradley & Lang, 1999) on a sample of tweets about Michael Jackson's death to examine if Twitter users were more likely to use more negative emotion words when tweeting about the death of a prominent public figure. Using a similar

approach to automatically detect negative emotion words, Park et al. (2012) scored tweets based on LIWC (Pennebaker et al., 2007) to examine if depressed users have a tendency to use more negative words on Twitter.

Although easy to implement, keyword spotting is a naïve approach that simply assumes that users express emotions using only emotion words. Consequently, only a small portion of emotional tweets are captured using keyword spotting. As each tweet is restricted to only 140 characters, many users tend to use irregular and shorter expressions to save space. Cui, Zhang, Liu, & Ma (2011) found that about one-third of the tweets in their sample contain at least one emotion token (i.e., emotion symbols, irregular forms of words and combined punctuations). To deal with these natural language forms that are not commonly included in domain-general emotion lexicons, Cui et al. (2011) constructed an emotion token lexicon (SentiLexicon) automatically from a sample of 5 million tweets. The polarity of each emotion token was determined using a graph propagation algorithm. The use of emotion tokens is not only prevalent in English tweets but similar usage trends are noted in non-English tweets. Therefore, the emotion token lexicon can be used for emotion detection in text regardless of language. However, the scoring function based on positive and negative scores of the emotion tokens in a tweet is only able to determine if a tweet is emotionally positive, negative or neutral.

Using Twitter, researchers have explored different strategies to automatically harness large volumes of data automatically for emotion classification. Using a method known as “distant supervision”, Pak & Paroubek (2010) applied a method similar to Read (2005) to extract tweets containing happy emoticons to represent positive sentiment, and sad emoticons to represent negative sentiment. Such a method allows for fast collection of a large self-labeled corpus without the need for manual annotation, but is limited in a sense that it enables the emotion classifier to detect only happiness and sadness. Furthermore, users may utilize emoticons in different and complex ways to express their emotions. Assuming that the emoticon represents the emotion in the overall tweet may be flawed without referring to the words in context as

emotions expressed in the text may not be in-sync with the emotion represented by the emoticon (e.g., sarcastic remarks).

Mohammad (2012) and Wang, Chen, Thirunarayan, & Sheth (2012) applied an improved method to create a large corpus of self-labeled tweets for emotion classification. Twitter allows the use of hashtags (words that begin with the # sign) as topic indicators. Extracting tweets that contain a predefined list of emotion words appearing in the form of hashtags was used to collect data for these studies. Mohammad (2012a) only extracted tweets with emotion hashtags corresponding to Ekman's six basic emotions (#anger, #disgust, #fear, #joy, #sadness, and #surprise) while Wang et al. (2012) expanded the predefined hashtag list to include emotion words associated with an emotion category, as well as the lexical variants of these emotion words. Wang et al. (2012) achieved slightly higher average performance (F1) across all emotion classes compared to Mohammad (2012a). F1 in both studies ranged from 0.1 – 0.7. Such corpus construction approach allows researchers to take advantage of the huge amount of data available on Twitter to train machine learning models. Statistical methods can be used to identify words that frequently co-occur with the emotion hashtags but little is known about the actual linguistic properties that are associated with these emotion categories. Also, this data collection method is biased towards users who choose to express their emotions explicitly in tweets.

As tweets in these corpora are extracted using common emotion hashtags, the data may not be representative of the range of emotions expressed on Twitter. To increase the emotion coverage of a tweet corpus, Hasan et al. (2014) extended the hashtags to include 28 affect words from the circumplex model of emotion (Russell, 1980) and their WordNet synsets. The circumplex model characterizes emotional states along two dimensions (i.e., valence and arousal). The affect words were mapped into four classes: Happy-Active, Happy-Inactive, Unhappy-Active, and Unhappy-Inactive, and a machine learning model was trained to detect these four emotion labels. The resulting model is limited in the sense that it can only detect

emotions at a coarse-grained level and it is trained based on straightforward positive examples of each emotion. The model is not trained to handle ambiguous tweets such as those containing multiple emotion hashtags from two different classes.

To address some of the criticism associated with the “distant supervision” method, Purver & Battersby (2012) investigated if the classifiers trained using automatically annotated data (i.e., noisy labels) are recognizing the actual underlying emotion class by cross validating models trained with different hashtag and emoticon labels or markers. A corpus of tweets was first collected using a predefined list of emotion markers, which only included emoticons and emotion word hashtags that were considered to be conventional markers for six emotion classes (i.e., happy, sad, anger, fear, surprise and disgust). The classifiers demonstrated reasonable performance when trained and tested on tweets containing the same label convention or emotion marker. Classifier performance was less reliable across label conventions (i.e., training on one emotion marker and testing on the others) and against a set of manually annotated examples. Such method was suitable for only some emotions like *happiness*, *sadness* and *anger* but did poorly in distinguishing other emotions.

Suttles & Ide (2013) experimented with a similar approach but included also emojis as emotion markers on top of the traditional emoticons and hashtags. Unlike prior research, they formulated the classification problem differently. Instead of training binary classifiers to identify whether an instance is an example of “Emotion-X” or “Not-Emotion-X”, they framed the task as a binary classification problem for four opposing emotion pairs (e.g., *joy* versus *sadness*). The mutually exclusive emotion pairs were determined using Plutchik’s wheel of emotion. The best performing classifiers yielded accuracies between 0.75 – 0.91, and the combination of hashtags, emoticons and emojis produced better results compared to previous distant supervision studies. Nonetheless, the classifiers were not trained to handle non-emotion content. Adding this layer of complexity may cause the classifiers to behave differently.

Manual efforts have also been used in the development of emotion tweet corpora. Roberts et al. (2012) annotated a tweet corpus sampled by topics expected to evoke emotions with seven emotion categories (anger, disgust, fear, joy, love, sadness, and surprise). With the exception of love, the other six emotion categories are adopted from Ekman's six basic emotions. While the data may not be representative of Twitter as a whole, manual annotation allows for tweets that are not explicitly tagged with an emotion word (#emotion) to be included in the training data. This way, the machine learning model can also learn from tweets containing implicit expressions of emotion.

Framing emotion detection as a semantic role labeling problem, Mohammad et al. (2014) applied a more complex structure in the emotion annotation task for a corpus containing tweets on the 2012 US Presidential Elections. On top of identifying an emotion response expressed in a tweet from a category of eight emotions (anger, anticipation, disgust, fear, joy, sadness, surprise, and trust), annotators in Amazon Mechanical Turk<sup>8</sup> were also asked to identify the individual experiencing the emotion and the stimulus causing the emotion to occur. About 88% of the tweets in the corpus were marked as containing only one emotion, thus indicating that topics related to politics contained high emotional content. The three roles explored in this study serve as a good starting point in providing semantic information useful for emotion analysis. However, machine learning models trained on such topic-specific data may not be generalizable to other topics.

Previous studies reviewed so far employed only a small set of emotion categories in automatic detection of emotion in tweets. Researchers have come to a realization that existing emotion categories are too limited to capture the richness of emotions expressed in tweets. Efforts to expand the emotion categories for fine-grained detection of emotion in text are shown in two recent studies by Sintsova, Musat, & Pu Faltings (2013) and Mohammad & Kiritchenko

---

<sup>8</sup> Amazon Mechanical Turk is an online crowdsourcing service that allows users to harness human intelligence from a large pool of workers to perform tasks that computers are unable to accomplish (<https://requester.mturk.com/>).



(2014). Sintsova et al. (2013) created a domain-specific emotion lexicon for sports events with 20 emotion categories of the Geneva Emotion Wheel (Scherer, 2005). Through crowdsourcing, annotators were given the task to select an emotion category and mark the emotion indicators based on the most prevailing emotion expressed in a tweet. The emotion indicators, which consisted of words or word sequences indicative of emotions, were then used to construct an emotion lexicon. This more fine-grained lexicon not only contained words, but also higher order n-grams (up to 5). An important finding highlighted in this study is that pride is expressed frequently when users are tweeting about a sports event, an emotion in which automatic emotion detectors from earlier work were not trained to recognize.

Mohammad & Kiritchenko (2014) found that fine-grained emotions are useful in personality detection. Using emotion hashtags, they collected a self-labeled tweet corpus using 585 fine-grained emotion hashtags, and subsequently constructed a word-emotion association lexicon (FineEmo) from the corpus. Experiments utilizing the FineEmo lexicon as features for a machine learning model to detect personality traits in essays and Facebook status updates yielded better performance compared to using coarser affect categories as features.

Unsupervised learning-based methods are less commonly used for automatic emotion detection in tweets. One such attempt is noted in a study to determine emotion shifts among participants in Twitter conversations. Kim, Bak, & Oh (2012) proposed a semi-supervised method using unannotated data for emotion classification. They first applied Latent Dirichlet Allocation (LDA) to discover 200 topics from a corpus of tweet conversations (i.e., a sequence of replies between users), and determined emotions from the discovered topics by calculating the pointwise mutual information (PMI) score for each emotion from a list of eight emotions (anticipation, joy, anger, surprise, fear, sadness, disgust, acceptance) given a topic. Evaluation of this method using a corpus of manually annotated tweets obtained through crowdsourcing revealed that this automatic emotion detector only managed to correctly classify 30% of tweets from the test dataset.

## 2.11 Conclusion

This chapter reviewed the extent to which emotion theories are used to inform research in automatic emotion detection in text, the conceptualizations of emotion in text, the conceptual differences between emotion and other related terms, the methods used to build automatic detectors in general, and specifically automatic emotion detection in tweets. The literature review drawn from computational linguistics, psychology, and linguistics not only reveals the diversity of research in this area but also exposes gaps that should be addressed moving forward.

In summary, three observations of research gaps are made from the literature review. First, researchers most often adopt a single emotion perspective (i.e., Darwinian, cognitive or social constructivist) in their conceptualization of emotion in text. The notion of an emotion being contained in a single lexical unit is still the dominant conception of emotion in text, and implicit expression of emotion has yet to be explored thoroughly. Second, machine learning methods are popularly used in automatic emotion classification but the focus has always been on building bigger corpora with the hope to improve classifier performance. Machine learning algorithms despite being effective often times produce no humanly understandable results. A promising research direction is to obtain a better understanding of what constitutes emotional cues in text, and how they can be used to inform features to improve classifier performance in this problem space. Third, in the context of emotion detection on an emotion-rich resource like Twitter, little has been done to build automatic emotion detectors that can recognize more fine-grained emotions expressed beyond the basic emotions, a problem that is addressed in this thesis.

# Chapter 3: Methodology

## 3.1 Introduction

The general premise of this thesis is that there is a richer set of emotions expressed on Twitter than current automatic emotion detectors are trained to identify. If the linguistic patterns associated with these more fine-grained emotion categories can be identified, it will be possible to develop automatic emotion detectors that capture a broader range of emotions expressed in microblogs. In this thesis, four specific research questions are addressed in three phases:

- *R1: What emotions can humans detect in microblog text?*
- *R2: What salient linguistic cues are associated with each emotion?*
- *R3: Do the salient cues humans associate with each emotion serve as better features for machine learning classification of emotion in text?*
- *R4: How do current machine learning techniques perform on more fine-grained categories of emotion?*

An overview of the three-phase study designed to address these questions is shown in Figure 3.1. The first two research questions (R1 and R2) were addressed in Phase 1 and Phase 2 in which we acquired a set of fine-grained emotion categories and annotations. To address R1, we characterized emotions using discrete categories based on layman's knowledge of emotion. R2 was investigated by analyzing the linguistic cues that humans relied on to recognize emotion expressions in text.

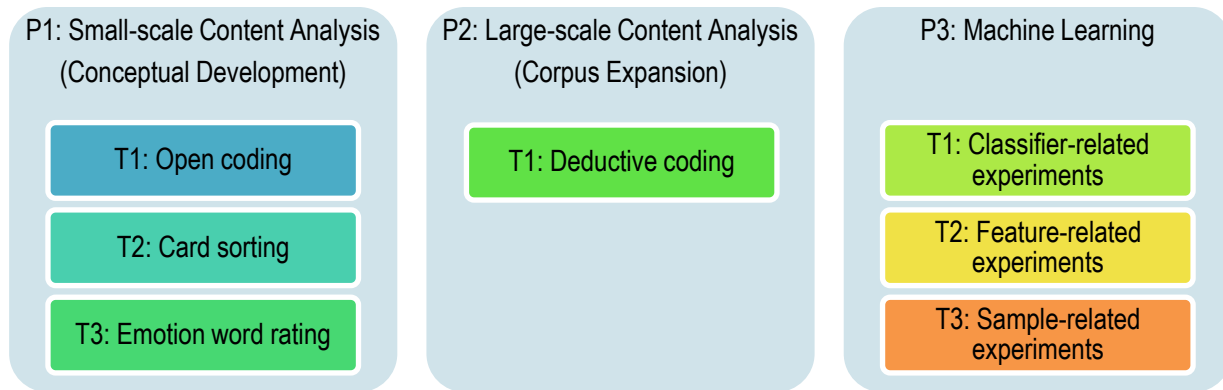


Figure 3.1: Overview of three-phase study

A small-scale content analysis was conducted in Phase 1 by training a group of annotators to annotate a sample of tweets for code book development and testing. The primary researcher worked closely with the annotators to develop and refine the annotation scheme based on the collective emotion knowledge of the group. The main goal of Phase 1 was to identify a stable set of emotion categories that is representative of the range of emotions expressed in tweets.

A larger corpus of 10,000 tweets was annotated using Amazon Mechanical Turk (AMT) in Phase 2 to develop a gold standard for machine learning experiments. The emotion categories that emerged in Phase 1 were further tested to determine how intuitive and representative they were of the range of emotions expressed on Twitter.

Questions R3 and R4 were addressed in Phase 3 through supervised machine learning experiments. Annotated data from Phase 1 and Phase 2 were used as gold standard data to train and evaluate supervised machine learning models on the emotion classification task. Different features were tested and their effects on the classifiers were examined.

### 3.2 Data Collection

The corpus consists of tweets (i.e., microblog posts) retrieved from Twitter. Data collected from microblogs is noisy and the frequency of emotional tweets in a sample may differ depending on the query terms used to retrieve the tweets. Mohammad, Zhu, & Martin (2014)

reported that 90% of their Twitter sample retrieved using 2012 US presidential elections query terms contained emotion, while Qu, Huang, Zhang, & Zhang (2011) reported that only 16% of their Sina-Weibo<sup>9</sup> sample about earthquakes in China were emotion-related. Also, the distribution of emotions is highly skewed (Kim, Bak, & Oh, 2012).

On one hand, it is important to ensure that a sample is representative of the population on Twitter but, on the other hand, it is also important to include as many tweets that contain emotion expressions as possible as the goal of this study is to discover the variability of emotions expressed on Twitter. To balance both factors, four different sampling strategies were used to retrieve the tweets to be included in the corpus: random sampling (RANDOM), sampling by topic (TOPIC), and two variations of sampling by user type (SEN-USER and AVG-USER). Topic sampling was done by retrieving tweets that contain selected topical hashtags or keywords. Sampling by user type retrieved tweets using selected user names (@usernames). One user sample contained tweets retrieved from US Senators (SEN-USER). Tweets from the second user sample were retrieved using randomly selected user names (AVG-USER). Tweets were either retrieved using the Twitter API or acquired from publicly available data sets.

Tweets were pre-processed to remove spam, duplicates, repeated retweets, and non-English tweets. A total of 15,553 tweets were included in the corpus, where 5,553 tweets were annotated in Phase 1 and 10,000 tweets were annotated in Phase 2. The distribution of tweets for each sample is shown in Table 3.1. The sample in Phase 1 consists of tweets annotated through open coding in Task 1 (P1-T1) and after the card sorting activity in Task 2 (P1-T2).

Sample	Sample Size				
	P1-T1	P1-T2	P1-ALL	P2	Total
RANDOM	1000	450	1450	2500	3950
TOPIC	1010	300	1310	2500	3810
SEN-USER	1000	493	1493	2500	3993
AVG-USER	1000	300	1300	2500	3800
<b>Total</b>	<b>4010</b>	<b>1543</b>	<b>5553</b>	<b>10000</b>	<b>15553</b>

Table 3.1: Distribution of tweets for 4 samples

<sup>9</sup> Sina-Weibo: A popular microblogging site like Twitter in China.

### 3.2.1 Random Sampling [RANDOM]

The first sampling strategy was intended to collect a random sample of tweets that is representative of the overall population on Twitter. The sample produced using this strategy might not be as rich with emotional content as the other samples. Since the Twitter API required query terms to retrieve tweets, nine stopwords (the, be, to, of, and, a, in, that, have) reported to be words most frequently used on Twitter were used to retrieve tweets for the random sample. An initial sample of 48,577 tweets was collected. Then, a random number generator was used to select tweets to be included in the corpus. The tweets were created between May – July 2014.

### 3.2.2 Sampling by Topic [TOPIC]

The second sampling strategy was based on topics or events. Tweets were sampled based on hashtags of events expected to contain emotional content. A wide range of topics were included to reduce the effect of emotional biases associated with certain topics (e.g., disaster-related topics are more likely to contain more negative emotions).

Data Source	Topic Description	Available	Sample Size	
			P1	P2
SemEval 2014	Topics related to famous characters (e.g., Gadafi, Steve Jobs), products (e.g. Kindle, Android phone), and events (e.g., Japan earthquake, NHL playoffs)	9520	910	400
2012 US presidential elections	#4moreyears, #Barack #election2012, #ObamaBiden2012, #mitt2012, #dems2012, #gop2012, etc.	168975	200	1100
Twitter API	#Sochi2014, #Oscar2014, #PrayForMH370, #MH17, #ValentinesDay, #anniversary, #graduation, #americanairlines, #jetblue, #unitedairlines, #usairways, #BlackFriday, #Thanksgiving, #vacation, #Gaza, #Israel, #Taliban, #PeshawarAttack, #RobinWilliams	6621	200	1000

Table 3.2: Description of topics included in TOPIC

The tweets for this sample were sampled from three sources: 1) the SemEval 2014<sup>10</sup> tweet data set (Nakov et al., 2013; Rosenthal, Nakov, Ritter, & Stoyanov, 2014), 2) the 2012 US presidential elections data set<sup>11</sup> (Mohammad et al., 2014), and 3) tweets retrieved using the Twitter API from February – December 2014 using query terms shown in Table 3.2.

### 3.2.3 Sampling by User

The final two sampling strategies were based on usernames. These two sampling strategies were aimed at striking a balance between including users who were representative of “average” Twitter users and active users who generated a relatively large number of tweets for analysis. One sample was collected from “average” Twitter users [AVG-USER] and another sample was collected from US political leaders (active Twitter users) [SEN-USER]. While sampling tweets from selected individuals limits the generality of findings, it allows exploration of the emotion variation and distribution in individual streams of tweets and examination of any differences when compared to the TOPIC and RANDOM samples.

**[SEN-USER]:** We first collected the @usernames of 89 US Senators, who were active users with a large number of followers from [www.tweetcongress.org](http://www.tweetcongress.org). The tweet streams were then collected from the Twitter API using the @usernames as the query terms. The number of tweets retrieved for each @username ranged between 43 and 386. We drew a sample from a total of 16,393 tweets created between March 2008 and April 2013.

**[AVG-USER]:** Another random sample of 10,000 tweets was collected using the same technique described in RANDOM. From this sample, we randomly selected 82 @usernames belonging to individuals and not organizations or news agencies. We then collected tweet streams using these 82 @usernames as the query terms from the Twitter API. The number of tweets for each @username ranged between 2 and 248. Similar to SEN-USER, we drew a

---

<sup>10</sup> SemEval 2014 corpus: <http://alt.qcri.org/semeval2014/task9/index.php?id=data-and-tools>

<sup>11</sup> 2012 US presidential elections corpus: <http://www.purl.org/net/PoliticalTweets2012>

sample of tweets to be annotated from 31,556 tweets created between July – August 2014. We included only users with at least 100 retrieved tweets in the sample.

### **3.3 Corpus Development**

To develop the corpus annotators were instructed to examine text in a systematic fashion, and assign the appropriate “labels” or “codes” to relevant linguistic expressions in a process known as “coding” (Krippendorff, 2004). Phase 1 combined both inductive and deductive methods to develop an annotation scheme comprising different facets of emotion. Inductive content analysis, wherein annotators were not given a pre-defined classification scheme, was used in Phase 1 to uncover a set of fine-grained emotion categories from data. As the category set was refined during Phase 1, deductive coding (fixed category set) was used to ensure that all tweets used the most current set of categories. Content analysis in Phase 2 was done in full deductive fashion with the classification scheme obtained from Phase 1. Although annotators in Phase 2 were asked to annotate using the Phase 1 categories, we did allow them the option to propose new categories. Tweets were segmented at the message-level. This level of analysis provided annotators with sufficient context to identify emotion expressed each tweet.

#### **3.3.1 Phase 1: Small-scale Content Analysis**

If we are to build machine learning classifiers that can recognize the emotions represented in tweets we need a set of suitable emotion categories. Prior work has mainly focused on classifying the emotion expressed in unstructured data into a small set of six to eight basic emotions adopted from existing theories from psychology. These theories were developed mainly to characterize physical expressions of emotion and may be a poor fit to represent the expression of emotions in text. The content analysis in Phase 1 was designed to develop an annotation scheme suited to emotions expressed in text. Three tasks were completed to



uncover this set of emotion categories: 1) inductive coding, 2) card sorting, and 3) emotion word rating.

### 3.3.1.1 Task 1: Inductive Coding

The annotation scheme comprised the four facets of emotion listed in Table 3.3. Two emotion dimensions, valence and arousal, commonly found in the literature (Russell, 1980) were included in the annotation scheme. Emotion valence can be positive, negative or neutral. Positive emotions are evoked by events causing one to express pleasure (e.g., happy, relaxed, fascination, love) while negative emotions are evoked by events causing one to express displeasure (e.g., anger, fear, sad). Emotions that were neither positive nor negative were considered to be neutral (e.g. surprise). Valence was useful to help annotators distinguish between tweets that contained emotion and those that did not.

Dimension	Description	Codes
Valence	Expressing pleasure or displeasure towards events, objects or situations	Positive: Expressing pleasure Negative: Expressing displeasure Neutral: Emotion expressed is neither positive nor negative No Emotion
Arousal	Level of arousal/activation to the stimuli	1: Calm (Very low intensity) 2: Low intensity 3: Moderate intensity 4: High intensity 5: Very high intensity
Emotion Tag	Emotion category or word that best describes the emotion expressed in a tweet	Open coding
Emotion Cues	Words/phrases that influence annotators to annotate the tweet with a particular emotion tag	Open coding

Table 3.3: Classification schemes for four facets of emotion

Arousal represents the degree of activation towards an emotion-causing stimulus. It is measured using a five point scale that ranges from calm to excited. These two facets of emotion can be used to characterize the properties of finer-grained emotion categories and facilitate comparison with earlier work.

Of more direct interest for the current research are the facets identified using inductive coding: annotators were asked to suggest the best-fitting emotion tags to describe the emotion expressed in each tweet, and to highlight portions of the tweets that serve as cues for recognizing the emotion. The inductive approach derives the annotation scheme through observation of content (Potter & Levine-Donnerstein, 1999). Construction of the classification scheme did not start from a theoretical framework. Instead, annotators began by looking for themes in the data and then moved to empirical generalization. The classification scheme was refined through an iterative process until a stable set of categories were finalized.

In essence, we used an adapted grounded theory approach developed by Glaser & Strauss (1967) for the purpose of building theory that emerges from the data. Although grounded theory was originally developed to generate a theory or model of the data, Scott et al. (2012) demonstrated how this approach can be adapted to derive a taxonomy of affect from collaborative online chat. Leveraging grounded theory's closeness to the data and its ability to capture nuanced concepts in a structured manner, they applied this method as an intermediate step to derive 40 codes or categories to characterize affect<sup>12</sup> expressions in text. Five annotators first annotated the data openly for anything of interest. In subsequent iterations of coding, they began to focus on grouping and refining a set of core categories that emerged from data until the core "affect" category was formed. When a stable version of the affect taxonomy was obtained, they finalized the codes by cross checking with Plutchik's emotion classification (Plutchik, 1962).

Our adaptation of grounded theory to expose a set of fine-grained emotion categories from tweets followed procedures similar to Scott et al. (2012) although we started with a clearer focus on the core theme of emotion. Annotators engaged in three coding activities central to this method: open coding, axial coding, and selective coding (Corbin & Strauss, 2008). In open

---

<sup>12</sup> As discussed in Section 2.5.3, affect is an umbrella term that encompasses emotions, moods and feelings (Russell, 2003). Affect can be considered a parent term to emotion, in a sense that all emotions are affective states, but not all affective states are emotions (Clore & Ortony, 1988).

coding, annotators read the content of each tweet to capture all possible meanings, and took a first pass at assigning concepts to describe the interpretation of the data. No restriction was posed on analysis in this phase, and minimal instructions were provided to avoid predisposing annotators. Axial coding then involved the process of drawing the relationships between concepts and categories. Based on their knowledge of emotion, annotators started with a small set of self-defined emotion tags. They then met in groups with the primary researcher to start drawing relationships between different emotion tags suggested by individuals in the group. Emotion tags were examined, accepted, modified, and discarded. Discrete emotion categories started to form in this phase, and were systematically applied to more data. Annotators switched back and forth between axial coding and open coding until a stable set of categories was identified. Finally, selective coding represented an integration phase where the identified discrete categories were further developed, defined, refined, and brought together in a unifying theme of emotion. Annotators then continued to validate the classification scheme by applying and refining it on more data until a point of saturation was reached (Corbin & Strauss, 2008).

<b>Demographic Aspect</b>	<b># of Annotators</b>
Gender	
Female	11
Male	7
Geographic region of origin	
USA	3
China	9
India	3
Southeast Asia	2
Middle East	1

Table 3.4: Demographic information of annotators in Phase 1

Graduate students who were interested in undertaking the task as part of a class project (e.g., Natural Language Processing course) or to gain research experience in content analysis (e.g., independent study) were recruited as annotators in Phase 1. Annotators were not expected to possess special skills except for the required abilities to read and interpret English text. A total of eighteen annotators worked on the annotation task over a period of ten months.

Annotators' demographic information is summarized in Table 3.4. To derive an emotion framework based on collective knowledge, each tweet was annotated by at least three annotators. Thus, annotators were divided into groups of at least three. Each group was assigned to work on one of the four samples.

All the annotators went through the same training procedures to reduce as much as possible the variation among different individuals. Each annotator first attended a one hour training session to discuss the concept of emotion with the researcher and to receive instructions on how to perform annotations of the tweets. In this exploratory stage, all annotations were collected on Excel spreadsheets.

Annotators were first instructed to annotate the valence of a tweet. If “No Emotion” was selected, annotators would not be prompted to provide annotations for arousal, emotion tag, and emotion cues. Annotators were required to provide an arousal rating, and identify emotion tag and emotion cues when valence for a tweet was labeled as either “Positive”, “Negative” or “Neutral”. For emotion tag, annotators were instructed to assign an emotion label that best described the overall emotion expressed by the tweeter (see Example 3.1). In cases where a tweet contained multiple emotions, annotators were asked to first identify the primary emotion expressed in the tweet, and then also include the other emotions observed (see Example 3.2). For such cases, the emotion cues for each emotion tag were specified.

**Example 3.1:** Alaska is so proud of our Spartans! The 4-25 executed every mission in Afghanistan with honor & now, they're home <http://t.co/r8pLpnud>

Valence: Positive

Arousal: 3

Emotion Tag [Emotion Cues]: Pride [so proud of, with honor]

**Example 3.2:** Saw Argo yesterday, a movie about the 1979 Iranian Revolution. Chilling, sobering, and inspirational at the same time.

Valence: Positive, Negative

Arousal: 4

Emotion Tag [Emotion Cues]: Inspiration [inspirational], Fear [Chilling, sobering]

The annotation scheme captured all expressions and descriptions of emotion in the broadest sense. A tweet contained emotion if it matched any one of the criteria below:

- Tweeter expressed his or her own emotion.
  - *Example: I'm happy to call you a friend.*
  - *Example: Hurrah! Nice catch Austin!*
- Tweeter described another person's emotion
  - *Example: He got scared of the penny.*
- Tweeter described an emotion-related phenomenon.
  - *Example: i can irritate the hell out of Saifullah tmr.*

Tweets that contained the use of sarcasm were challenging to annotate. To handle such tweets, annotators were instructed to assign emotion tags to only tweets with clear emotion cues. In Example 3.3, the cue “GEE THANKS” was not used to express gratitude. Rather, it was a sarcastic remark. The presence of the cue “JERK” at the end of the tweet further confirmed that the tweeter was expressing anger. Sarcastic tweets with high ambiguity were labeled as “No Emotion” (see Example 3.4).

**Example 3.3:** @SunnyKoda GEE THANKS FOR THE HEADS UP JERK [**Anger**]

**Example 3.4:** Ronaldo is 100% Flirtin' with the referee! He loves to chase him [**No Emotion**]

In the first iteration, also referred to as the training round, all annotators annotated the same sample of 300 tweets from the SEN-USER sample. Annotators were expected to achieve at least 70% pairwise agreement for valence with the primary researcher in order to move forward. The annotators achieved a mean pairwise agreement of 82% with the researcher.

Upon passing the training round, annotators were assigned to annotate at least 1,000 tweets from one of the four samples (RANDOM, TOPIC, AVG-USER or SEN-USER) in subsequent iterations. Every week, annotators worked independently on annotating a subset of 150 – 200 tweets but met with the researcher in groups to discuss disagreements, and 100% agreement for valence and emotion tag was achieved after discussion. In these weekly meetings, the researcher also facilitated the discussions among annotators working on the same sample to merge, remove, and refine suggested emotion tags. The output of Task 1 included 4,010 annotated tweets in the gold standard corpus and 246 distinct emotion tags.

### 3.3.1.2 Task 2: Card Sorting

Some of the 246 emotion tags were simply morphological variations and many were semantically similar. Task 2 served as an intermediate step to refine the emotion tags emerging from data into a more manageable set of higher level emotion categories. Annotators were asked to perform a card sorting exercise in different teams to group emotion tags that are variants of the same root word or semantically similar into the same category. Annotators were divided into 5 teams, and each team received a pack of 1' x 5' cards containing only the emotion tags used by the all members in their respective teams.

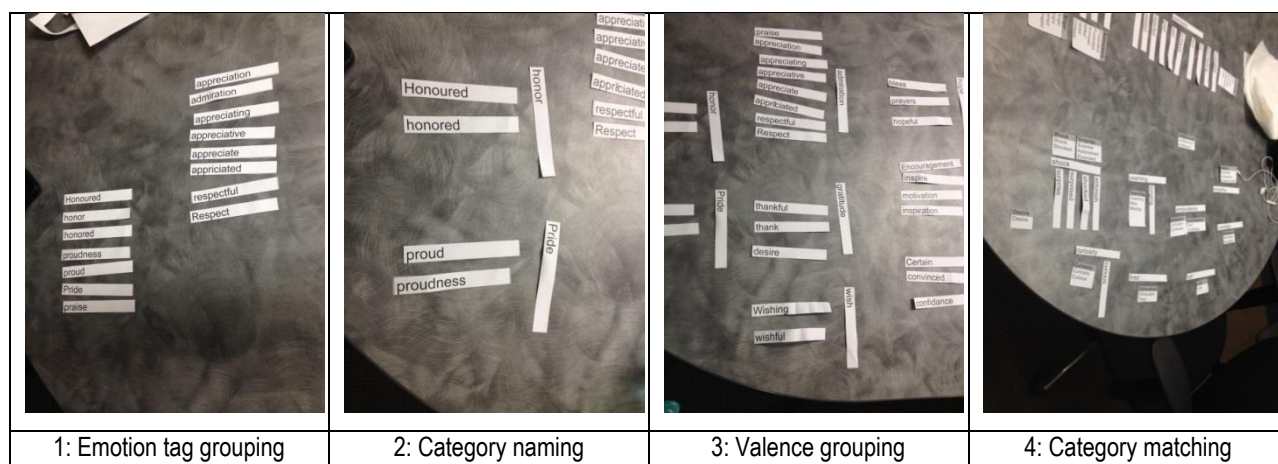


Figure 3.2: Four-step procedure in the card sorting activity

Each team consisted of 2 - 3 members who worked on the same sample. Teams were instructed to follow the four-step procedures described below (illustrated in Figure 3.2):

- 1) Group all the emotion tags into categories. Members were allowed to create a “Not Emotion” category if needed.
- 2) Decide a name for the emotion category. Collectively pick the most descriptive emotion tag or suggest a new name to represent each category.
- 3) Group all the emotion categories based on valence: positive, negative and neutral.
- 4) Match emotion categories generated from other team’s card sorting activity to the emotion categories proposed by your team.

Members in the same team were allowed to discuss their decisions with each other during the card sorting exercise with minimal intervention from the researcher. The session concluded when all members completed the four-step procedure and reached a consensus on final groupings of the emotion tags. No limit was placed on the number of categories or the number of emotion tags within each category so the number of categories proposed varied across the five teams as shown in Table 3.5. Some teams decided to put the emotion tags into fewer higher-level categories, while others who chose to capture more subtle emotions generated more emotion categories. Finally, the researcher merged, divided, and verified the final emotion categories to be included in the classification scheme.

Team	Sample	Number of Emotion Categories			
		Positive	Negative	Neutral	Total
G1	SEN-USER	8	13	2	23
G2	TOPIC	16	14	5	35
G3	TOPIC	16	18	8	42
G4	AVG-USER	14	18	15	47
G5	RANDOM	14	16	9	39

Table 3.5: Number of categories proposed by each card sorting team

Once the final 48 emotion categories shown in Figure 3.3 were identified, the original emotion tag labels generated from the open coding exercise were systematically replaced by the appropriate emotion category labels. After all emotion tags were resolved into one of the 48

emotion categories, we generated three inter-annotator reliability scores: percent agreement, Fleiss' kappa (Fleiss, 1971) and Krippendorff's alpha (Krippendorff, 2004). Annotators then incrementally annotated more tweets (150 - 200 tweets per round) to ensure that a point of saturation was reached. No new emotion category emerged from data in this coding phase. Another 1,543 annotated tweets with gold labels were added to the corpus.

Positive = 16	Negative = 21	Neutral = 11
<ul style="list-style-type: none"> <li>•Admiration</li> <li>•Love</li> <li>•Like</li> <li>•Fascination</li> <li>•Gratitude</li> <li>•Pride</li> <li>•Pleased</li> <li>•Happiness</li> <li>•Amusement</li> <li>•Relaxed</li> <li>•Relief</li> <li>•Excitement</li> <li>•Anticipation</li> <li>•Hope</li> <li>•Confidence</li> <li>•Inspiration</li> </ul>	<ul style="list-style-type: none"> <li>•Sadness</li> <li>•Sympathy</li> <li>•Yearning</li> <li>•Disappointment</li> <li>•Displeased</li> <li>•Annoyance</li> <li>•Anger</li> <li>•Boredom</li> <li>•Exhaustion</li> <li>•Guilt</li> <li>•Doubt</li> <li>•Shame</li> <li>•Regret</li> <li>•Desperation</li> <li>•Dread</li> <li>•Awkward</li> <li>•Fear</li> <li>•Worry</li> <li>•Hate</li> <li>•Disgust</li> <li>•Jealousy</li> </ul>	<ul style="list-style-type: none"> <li>•Surprise</li> <li>•Shock</li> <li>•Amazement</li> <li>•Empathy</li> <li>•Curiosity</li> <li>•Confusion</li> <li>•Indifference</li> <li>•Nostalgia</li> <li>•Ambivalence</li> <li>•Desire</li> <li>•Lust</li> </ul>

Figure 3.3: List of 48 emotion categories

### 3.3.1.3 Task 3: Emotion Word Rating

Task 3 was designed to further refine and reduce the number of discrete emotion categories to be used in subsequent phases of the research. While it was plausible to train a small group of regular annotators to be experts in applying a classification scheme with 48 emotion categories, growing the size of the corpus with 48 emotion categories on a larger scale



posed several methodological challenges. First, it would be challenging and time consuming to provide rigorous training to a large number of annotators. Second, high cognitive load would be imposed on a person attempting to annotate a tweet with 48 available options especially for non-experts, thus reducing their effectiveness and efficiency in performing the task. Furthermore, preliminary analysis on the 4,010 tweets from Task 1 revealed that certain pairs of emotion categories had relatively higher similarity scores based on the Jaccard similarity coefficient computed based on co-occurrences of emotion categories assigned by pairs of annotators, suggesting that some categories could be further merged.

A word rating study was conducted as a systematic method to merge and distill the number of categories into a more manageable set. The motivation behind the word rating study came from prior studies showing that emotion words with greater similarity tends to be in close proximity to one another on a two-dimensional pleasure and degree of arousal space (Russell, 1980; Russell & Pratt, 1980). Figure 3.4 shows the mapping of 28 emotion terms in a two-dimensional space with the x-axis representing the pleasure-displeasure dimension and the y-axis representing the degree of arousal from a study conducted by Russell (1980). The study first asked participants to group together emotion terms that are similar so that the similarity between each pair of terms could be assessed. A unidimensional scaling procedure was then used to map the 28 terms into a two-dimensional space. The closer the proximity of the terms in the space, the more similar they are. Following this line of reasoning and using the two-dimensional plot as a means to visualize the emotion categories, we examined if we could halve the number of categories by combining categories into families of related emotions based on their proximity on the two-dimensional space as well as their semantic similarity.

The emotion categories derived from data in this research do not map directly onto the 28 emotion terms in Figure 3.4. Only a quarter of the terms in Figure 3.4 could be mapped into our set of 48 emotion categories. In order to plot our emotion categories in this two-dimensional space, we opt to collect the pleasure and arousal ratings for each emotion category. A set of 50

emotion words were selected for the emotion rating task. We included the 48 emotion category names and added 2 emotion words from the original list of emotion tags that were believed to be more appropriate category names than the ones determined by the annotators in Task 2. These two words were “*longing*” found in the category “*yearning*” and “*torn*” found in the category “*ambivalence*”. Pleasure and arousal ratings were collected from two sources: 1) Affective Norms for English Words (ANEW) lexicon (Bradley & Lang, 1999), and 2) a word rating study conducted using Amazon Mechanical Turk (AMT).

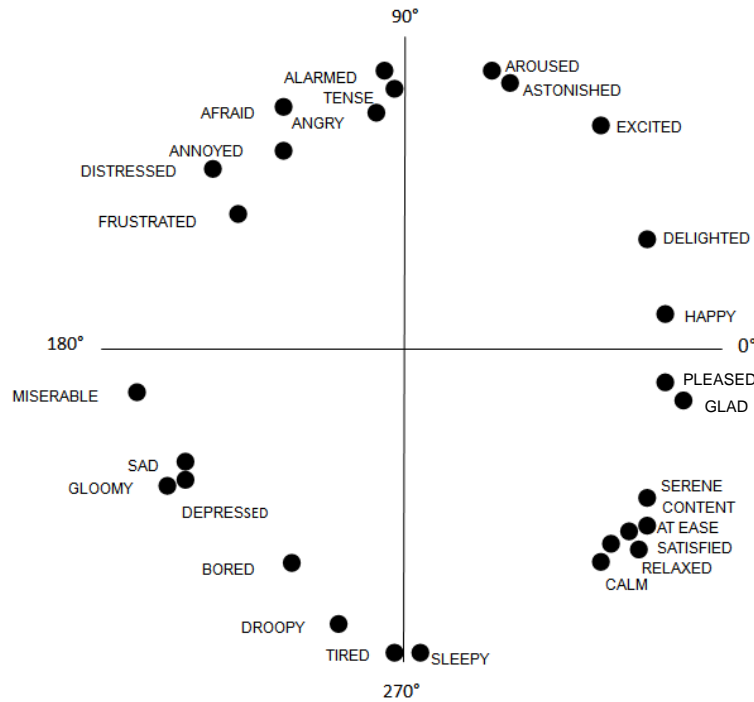


Figure 3.4: Scaling coordinates for 28 affect words in two-dimensional space (x-axis represents pleasure and y-axis represents arousal) (Russell, 1980, p. 1167)

Before plotting the mean pleasure and mean arousal on a two-dimensional scatterplot, we used feature scaling ( $x'$ ) to normalize the data so both 9-point scales ranged from 0 – 1. The spread of data became clearer after normalization.

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)}$$

## A) ANEW Ratings

We first attempted to collect pleasure and arousal ratings for each emotion word from the Affective Norms for English Words (ANEW) lexicon. The ANEW lexicon contains mean rating and standard deviation for three affective dimensions: pleasure, arousal, and dominance for 1,034 English words.

Emotion word	Present	ANEW word(s)	Emotion word	Present	ANEW word(s)
Admiration	Yes	Admire	Guilt	Yes	Guilty
Amazement	No		Happiness	Yes	Happy
Ambivalence	No		Hate	Yes	Hate
Amusement	Yes	Humored	Hope	Yes	Hope
Anger	Yes	Anger	Indifference	Yes	Indifferent
Annoyance	Yes	Annoy	Inspiration	Yes	Inspired
Anticipation	No		Jealousy	Yes	Jealousy
Awkward	Yes	Uncomfortable	Like	No	
Boredom	Yes	Bored	Love	Yes	Love
Confidence	Yes	Confident	Lust	Yes	Lust
Confusion	Yes	Confused	Nostalgia	No	
Curiosity	Yes	Curious	Pleased	Yes	Pleasure
Desire	Yes	Desire	Pride	Yes	Pride
Desperation	No		Regret	Yes	Regretful
Disappointment	Yes	Disappointment	Relaxed	Yes	Relaxed
Disgust	Yes	Disgusted	Relief	No	
Displeased	Yes	Displeased	Sadness	Yes	Sad
Doubt	No		Shame	Yes	Shame, Embarrassed
Dread	No		Shock	No	
Empathy	No		Surprise	Yes	Surprised
Excitement	Yes	Excitement	Sympathy	Yes	Pity
Exhaustion	No		Worry	Yes	Anxious
Fascination	Yes	Fascination, Interest, Impressed	Yearning	No	
Fear	Yes	Fear	Longing	No	
Gratitude	Yes	Grateful	Torn	No	

Table 3.6: Mapping between 50 emotion words and ANEW words

Each word in ANEW was rated on a graphical 9-point scale for each dimension using an affective rating instrument called the Self-Assessment Manikin (SAM) (Bradley & Lang, 1994).

Out of the 50 emotion words, we were able to obtain pleasure and arousal ratings for only 35 words from ANEW.

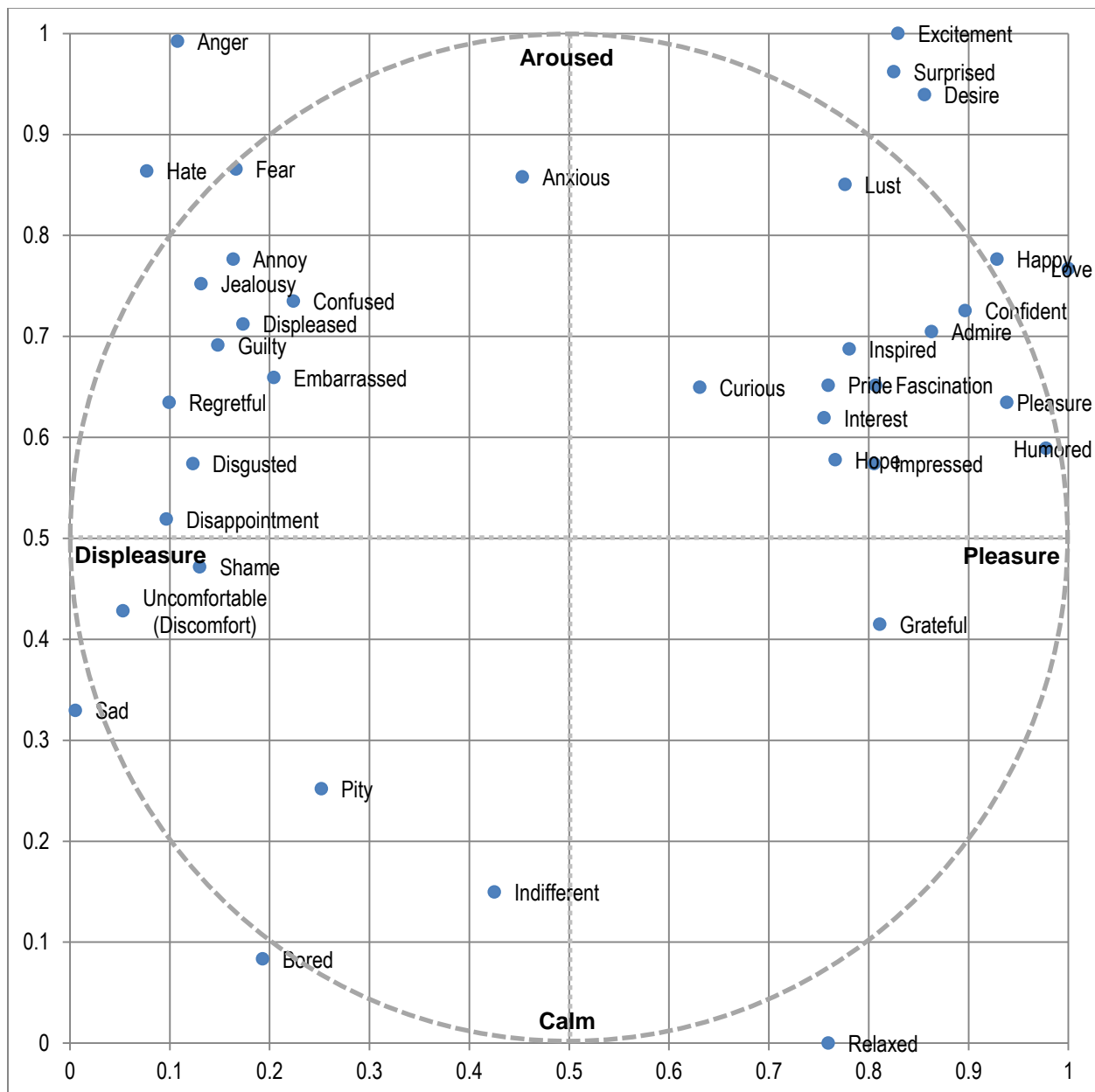


Figure 3.5: Two-dimensional pleasure and arousal plot for 38 ANEW words representing 35 emotion categories based on ANEW ratings (x-axis represents pleasure, y-axis represents arousal)

Figure 3.5 shows the two-dimensional plot based on ANEW ratings. The emotion labels on the plot were selected from words in the lexicon that were most similar to the emotion category names. The mapping between 50 emotion words in the set and the ANEW words selected to represent each emotion word is shown in Table 3.6. First round of selections were based on exact matching between the emotion words in the set with the words in the lexicon. If exact matches between the emotion words and ANEW words were not found, we selected from ANEW either a morphological variation or synonym<sup>13</sup> of the emotion word. Our selection criteria yielded ANEW ratings for only 35 emotion words.

## **B) AMT Ratings**

To obtain a complete set of pleasure and arousal ratings for our set of 50 emotion words, we conducted an emotion word rating study on AMT. We adapted the instrument that was used in Bradley & Lang (1999) to collect the ANEW ratings in our word rating study. We implemented the study using exactly the same 9-point scale for the pleasure and arousal ratings (see Appendix A). The only difference is that we used radio buttons instead of the graphic SAM figures to depict the values along the pleasure and arousal dimensions. The validity of the scales are described in Bradley & Lang (1994). The same set of instructions was reused but modified accordingly to fit the crowdsourcing context.

Human raters were recruited from the pool of workers available on AMT. The rating instrument was offered to the workers via a Human Intelligence Task (HIT), and workers would receive payment of US\$ 0.20 upon completion and approval of the HIT. HITs were restricted to workers in the US to increase the likelihood that ratings came from native English speakers. Each respondent first read the instructions on how to use the pleasure and arousal scales. Respondents were then instructed to make a pleasure rating and an arousal rating for each of the 50 emotion words on a 9-point scale. To ensure respondents read and understood the

---

<sup>13</sup> Synonyms were referenced from Roget's Thesaurus: <http://www.thesaurus.com>

instructions, they were explicitly asked to rate “*boredom*” and “*relaxed*” at the calm end, and “*anger*” and “*excitement*” at the aroused end of the arousal scale. Responses that failed this check were rejected. After removing incomplete and rejected responses, mean rating and standard deviation were computed from 76 usable responses.

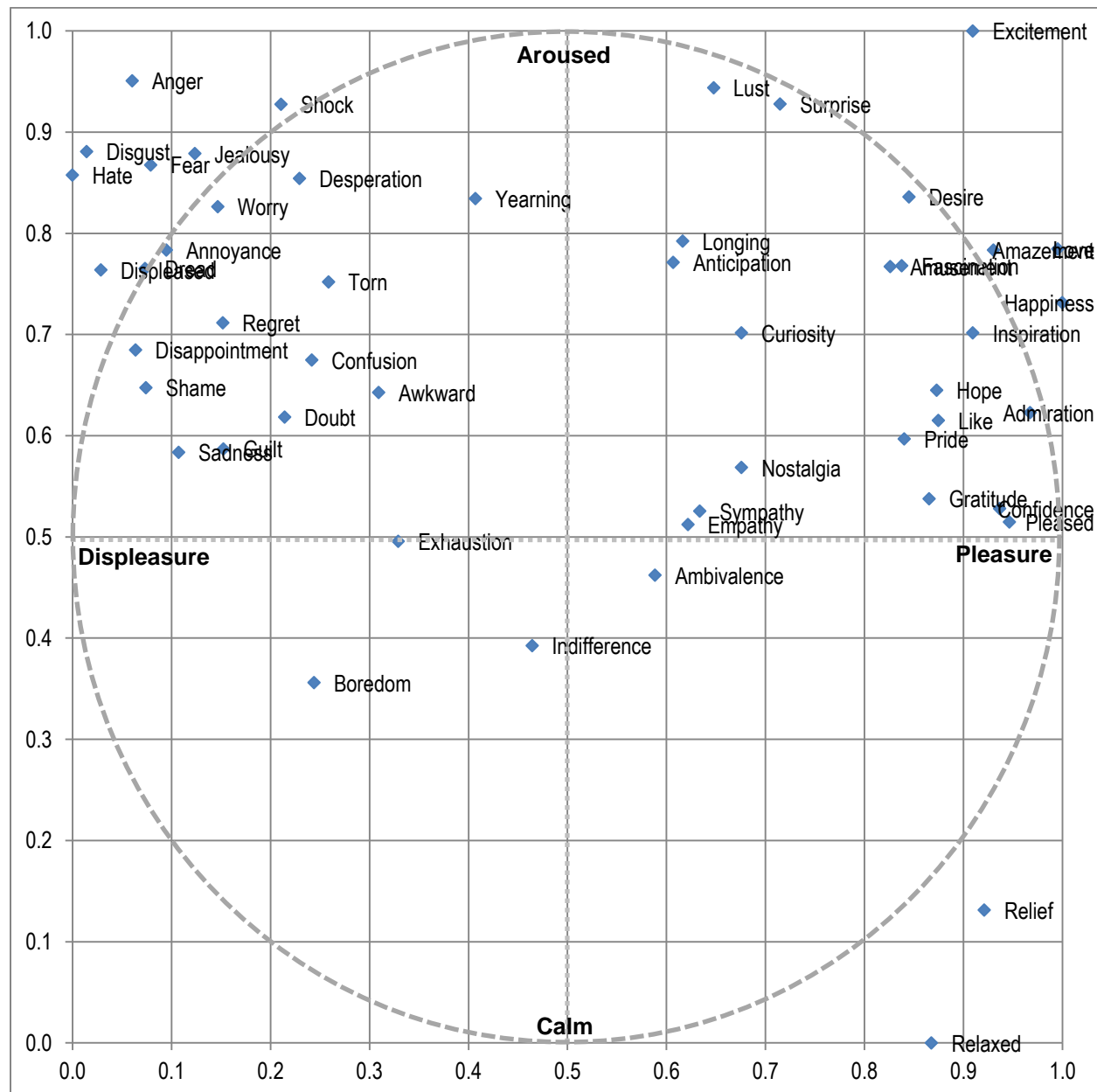


Figure 3.6 shows the plot for all 50 emotion words based on the AMT ratings. The relative position of each emotion word in the plots shown in Figure 3.5 and 3.6 is quite similar even though the coordinates scattered more densely around the upper quadrants in the AMT plot (Figure 3.6). It is interesting to note that the coordinates fall around the space in a circular fashion and do not just cluster around the axes which is consistent with the circumplex model of affect (Russell, 1980). Four quadrants are defined in the two-dimensional space. The upper right quadrant characterized pleasurable and aroused emotions, while the upper left quadrant characterized unpleasurable and aroused emotions. The lower right quadrant described pleasurable and calm emotions, while the lower left quadrant described unpleasurable and calm emotions.

### **C) Category Merging**

We decided to merge the categories corresponding to emotion words using three criteria: 1) the emotion words must fall in the same quadrant on both plots, 2) the emotion words should be relatively close in proximity to one another on both plots, and 3) the emotion words must belong to the same emotion cluster identified by Shaver et al. (2001, pp. 34–35). The 135 emotion words grouped into 25 clusters using hierarchical clustering in Shaver et al. (2001) overlapped with many of the emotion words in our study.

In the circumplex model, the two dimensions used to characterize emotions were bipolar, meaning that the two ends of each dimension were supposed to be polar opposites so emotion words located in different quadrants should be distinct. The second criterion worked on the basis that words closer together on the circle described emotions that were more similar. Some emotion words might appear close to one another in the same quadrant but were dissimilar in meaning. For example, “amusement” and “fascination” appeared very close together in Figure 3.6 but their meanings were distinct enough to keep them separate as each a

discrete category. The third criterion ensured that only emotion words that were semantically similar were combined (e.g., “hate” and “disgust”).

Emotion-Category-28	Emotion-Category-48	Emotion-Category-28	Emotion-Category-48
Admiration	Admiration	Hate	Hate, Disgust
Amusement	Amusement	Hope	Hope
Anger	Anger, Annoyance, Displeased, Disappointment	Indifference	Indifference
Boredom	Boredom	Inspiration	Inspiration
Confidence	Confidence	Jealousy	Jealousy
Curiosity	Curiosity	Longing	*Longing, Nostalgia
Desperation	Desperation	Love	Love, Like
Doubt	Doubt, Confusion, *Torn	Pride	Pride
Excitement	Excitement, Anticipation	Regret	Regret, Guilt
Exhaustion	Exhaustion	Relaxed	Relaxed, Relief
Fascination	Fascination, Amazement	Sadness	Sadness
Fear	Fear, Dread, Worry	Shame	Shame, Awkward
Gratitude	Gratitude	Surprise	Surprise, Shock
Happiness	Happiness, Pleased	Sympathy	Sympathy, Empathy

Table 3.7: Mapping between the final set of 28 emotion categories to the original set of 48  
(category names preceded by \* were modified)

Using these three criteria, we reduced the emotion categories from 48 to a final set of 28 shown in Table 3.7, which were employed in Phase 2 to annotate the gold standard corpus. The names of two emotion categories from the original set of 48 were changed to allow merging to happen without violating criterion 1. Category name “ambivalence” was substituted by its more descriptive member term, “torn” and “yearning” was substituted by “longing”. Also, two emotion categories from the original 48, “desire” and “lust” were dropped altogether from the final set of 28 because whether or not they should be considered as emotional states were still debatable (Ortony & Turner, 1990). Based on how these two categories were conceptualized in our annotation scheme, they were considered to be more general feelings of wanting rather than emotional states. Finally, the 48 emotion category labels in the corpus were systematically replaced by the appropriate 28 emotion category labels.



### 3.3.2 Phase 2: Large-scale Content Analysis

Using the annotation scheme described in the last section, a larger set of manual annotations was obtained using Amazon Mechanical Turk (AMT) in Phase 2. The goal was to add 10,000 annotated tweets to the gold standard corpus in the shortest time possible. AMT enabled us to collect manual annotations of emotions on a large-scale, thus allowing for the creation of a corpus that was large enough to support machine learning experiments, to improve estimates of the distribution of emotion categories using larger and more representative samples, and to further evaluate the effect of individual differences on annotator performance.

Careful considerations were given to the design of the annotation task in Phase 2 since the amount of training undertaken by the AMT workers was more limited in a crowdsourcing environment. The main goal was to scale up emotion annotations in tweets while minimizing threats to reliability. To streamline the annotation process across a large pool of annotators, we developed a Web annotation application shown in Figure 3.7, which was tailored to our annotation scheme. The facets of emotion to be annotated were presented as a series of questions. Apart from the four questions meant to elicit the valence (the term “polarity” was used as a more descriptive term to a lay person), arousal, emotion tag and emotion cues for each tweet, workers were also asked to indicate the source of the emotion (author of tweet or someone else) and if the tweet contained more than one emotion. Workers were able to make quick references to the definition of each code for the different classification schemes through tool tips that would appear on screen by hovering on the option labels.

For emotion tag, workers were given a set of 28 emotion categories to choose from plus an “other” option with a text box so they were allowed to suggest a new emotion tag for any tweets where none of the listed emotion category was applicable. The order in which the emotion categories were presented to the workers was randomized across the four samples in order to control for order effect. If a tweet was flagged as containing multiple emotions,

annotators were asked to provide all relevant emotion tags, and highlight the emotion cues for each selected emotion.

The screenshot shows a web application for emotion annotation. At the top, a tweet is displayed: "Joyce shares one of the worst things you can do in your relationships. Check it out <http://t.co/5jnSSzljvt>". Below the tweet, there are several sections for annotation:

- Polarity**: What is the polarity of emotion expressed or described? Options: ☒ Positive, ☐ Negative, ☐ Neutral, ☐ No Emotion.
- Arousal**: What is the degree of emotion arousal? (1: Very Calm to 5: Very Activated) Options: ☐ 1, ☐ 2, ☐ 3, ☐ 4, ☐ 5.
- Emotion Tag**: What emotion is expressed or described? Choose one of the options below that best represents the emotion. If none of the options apply, choose Other and suggest your own emotion tag. Options: ☐ Anger, ☐ Sadness, ☐ Surprise, ☐ Curiosity, ☐ Fascination, ☐ Happiness, ☐ Admiration, ☐ Sympathy, ☐ Regret, ☐ Hate, ☐ Longing, ☐ Love, ☐ Amusement, ☐ Inspiration, ☐ Doubt, ☐ Desperation, ☐ Fear, ☐ Indifference, ☐ Gratitude, ☐ Excitement, ☐ Hope, ☐ Shame, ☐ Jealousy, ☐ Boredom, ☐ Exhaustion, ☐ Pride, ☐ Relaxed, ☐ Confidence, ☐ Other .
- Emotion Cues**: Highlight all the cues (characters, words, phrases, etc.) that are indicators of the emotion from the tweet above and they will appear below. You can also type the cues into the box. You can remove cues by deleting them below. A large text area for input is provided.
- Emotion Source**: Whose emotion is expressed or described? Options: ☐ Tweeter, ☐ Other Person, ☐ No One.
- Multiple Emotions**: Is there another emotion expressed or described? Options: ☐ Yes, ☐ No.

At the bottom, there is a "submit" button and a status bar that says "You have coded 0 tweets in this session. [ Quit this task ]".

Figure 3.7: Web annotation application for data collection in Phase 2

The annotation for emotion source emerged in Phase 1. Expert annotators found that pinpointing the source of emotion was helpful to identify emotions expressed in the tweet. The task to identify the source of emotion was reasonably simple so we also included this task in Phase 2. It was more challenging to identify the target or stimulus causing the emotion to be expressed given the short length of the tweets so the task was excluded from the annotation scheme.

Recruitment of workers was done through Human Intelligence Tasks (HITs) on the online AMT platform. AMT workers must fulfill at least the basic requirement of being able to read and understand English text. Therefore, the HITs were only made available to workers who

were geographically located in the United States (US), Australia (AU), Canada (CA), Ireland (IE), New Zealand (NZ) and Great Britain (GB) to increase the likelihood of recruiting native English speakers. This was intended to reduce possible variance caused by cultural differences. Also, we set the HIT approval rate<sup>14</sup> for all requesters' HITs to greater than or equal to 95% and the number of HITs approved to greater than or equal to 1000 to increase the probability of recruiting first-rate workers.

The screenshot shows a web interface for an emotion annotation task. It features a blue header with the title "Instructions". Below this, there is a text block explaining the goal of the task: to detect emotions in tweets. It then provides instructions on how to proceed, including selecting a link, reading task descriptions, and entering a batch ID. A warning is given not to use the browser's back button. A compensation of US\$ 0.50 is mentioned for completing 30 tweets. An email address (jliewsue@syr.edu) is provided for reporting problems. A final instruction states to keep the window open and paste a validation code upon completion. Below the instructions, there is a form with three fields: "Annotation site link:" with a URL, "Batch ID:" with the value 13475, and "Provide the validation code here:" with a text input field containing the placeholder "e.g. 123456". A blue "Submit" button is located at the bottom right of the form.

**Instructions**

The goal of the emotion annotation task is to detect emotions expressed or described in tweets collected from Twitter. This hand-annotated data can be used to identify linguistic patterns associated with each emotion type, and to train computer systems to automatically detect emotion in text.

Select the link below to go to the emotion annotation site. First, read the task and code descriptions for polarity, arousal, emotion tag, emotion cues, emotion source, and multiple emotions. You will then be prompted to enter a batch ID. Copy and paste the batch ID provided below into the annotation site.

Read and annotate all 30 tweets in one single session. Annotate each tweet carefully as you will not be able to correct submitted tweets (please do not use your Web browser's "Back" button). At the end of the task, you will receive a validation code to paste into the box below to receive US\$ 0.50 as compensation for annotating 30 tweets. Please do not hesitate to send me an email ([jliewsue@syr.edu](mailto:jliewsue@syr.edu)) if you encounter any problems or have any suggestions to improve the task.

**Make sure to leave this window open as you complete the annotation task.** When you are finished, you will return to this page to paste the code into the box.

**Annotation site link:** <http://yorick.syr.edu/emotiontext/ect-example/ect-random/codebookex.html>

**Batch ID:** 13475

**Provide the validation code here:**

**Submit**

Figure 3.8: Design of the HIT for the emotion annotation task on AMT

In the design of the HIT, workers were provided clear and simple instructions describing the task, the annotation site link, as well as a batch id required to retrieve a subset of 30 tweets to work on (see Figure 3.8). The annotation task was not embedded in the HIT so the annotation link was provided to direct workers to our external Web annotation application. We had more flexibility to experiment with different quality monitoring mechanisms and controls using our own Web annotation application.

<sup>14</sup> HIT approval rate shows the number of HITs approved out of all the HITs submitted by an AMT worker: <http://docs.aws.amazon.com/AWSMechTurk/latest/RequesterUI/mechanical-turk-concepts.html>. The HIT approval rate allows requesters to recruit serious workers who are more likely to produce high quality work.

Of the 30 tweets in one HIT, 25 were new tweets and 5 were gold standard tweets intended to be used for quality control. Upon clicking the annotation site link and entering a valid batch id, annotators were first presented with a code book providing the definitions and examples for all classification schemes relevant to the task. The definitions and coding procedures were explained in layman's terms to reduce potential confusion surrounding the task since interaction with the researcher was limited. Each HIT was assigned to three different annotators. Each HIT bundled a different subset of 30 tweets so a worker could attempt more than one HIT.

Each batch of 30 tweets included five gold standard tweets from Phase 1 to compare the consistency of annotations between Phase 1 and Phase 2 and to ensure the validity of the annotations collected from AMT. A set of 707 tweets that obtained full agreement on emotion tag among 3 annotators in Phase 1 were selected to be included in the validation sample for Phase 2. Tweets with multiple emotions were excluded from the validation sample in order to avoid introducing greater complexity to the validation procedure. We made sure that each batch of tweets contained a diverse set of five gold standard tweets (i.e., tweets in each batch have a variety of valence and emotion tag gold labels).

We performed a few pilot experiments to identify how to best incorporate the gold standard tweets to produce high or at least reasonable quality annotations from AMT. In one case (C1), the gold standard tweets were included as the last 5 tweets in a batch of 30, and were treated as a form of validation test. The gold standard tweets were presented exactly the same way as unannotated tweets to avoid any bias in the treatment of tweets within the batch. In line with the expected percent agreement (70%) between annotators expected in Phase 1, workers would have to correctly annotate at least 4 of the 5 tweets for valence and emotion tag respectively in order to pass the validation test and have their HIT submission approved. Out of 20 batches of tweets across all four samples, applying this validation rule would yield an approval rate of only 15%. In C1, workers were able to correctly identify 3 of the 5 valence gold

labels and 2 of the 5 emotion tag gold labels on average so setting such a high bar proved to be unrealistic for this task. Furthermore, such a high rejection rate would make it more difficult to recruit workers and drastically increase the time required for the study.

gotta appreciate leorio omg

**Polarity** What is the polarity of emotion expressed or described?  
*\* This is an example of positive*  
☒ Positive   ☐ Negative   ☐ Neutral   ☐ No Emotion

**Arousal** What is the degree of emotion arousal? (1: Very Calm to 5: Very Activated)  
☐ 1   ☐ 2   ☐ 3   ☐ 4   ☐ 5

**Emotion Tag** What emotion is expressed or described? Choose one of the options below that best represents the emotion. If none of the options apply, choose Other and suggest your own emotion tag.  
*\* This is an example of gratitude*

☐ Anger

☐ Sadness

☐ Surprise

☐ Curiosity

☐ Fascination

☐ Happiness

☐ Admiration

☐ Sympathy

☐ Regret

☐ Hate

☐ Longing

☐ Love

☐ Amusement

☐ Inspiration

☐ Doubt

☐ Desperation

☐ Fear

☐ Indifference

☐ Gratitude

☐ Excitement

☐ Hope

☐ Shame

☐ Jealousy

☐ Boredom

☐ Exhaustion

☐ Pride

☐ Relaxed

☐ Confidence

☐ Other

**Emotion Cues** Highlight all the cues (characters, words, phrases, etc.) that are indicators of the emotion from the tweet above and they will appear below. You can also type the cues into the box. You can remove cues by deleting them below.

**Emotion Source** Whose emotion is expressed or described?  
☐ Tweeter   ☐ Other Person   ☐ No One

**Multiple Emotions** Is there another emotion expressed or described?  
☐ Yes   ☐ No

  You have coded 0 tweets in this session. [ Quit this task ]

Figure 3.9: Gold labels for valence and emotion tags obtained from Phase 1's ground truth displayed in the first five tweets in a batch

In another case (C2), the gold standard tweets were included as the first 5 tweets in a batch of 30, and served as training examples as well as a mechanism to filter out workers who failed to read and follow instructions. The gold labels for valence and emotion tag were displayed right below their respective questions (see Figure 3.9). It was assumed that workers who actually paid attention to the instructions would correctly pick options matching the gold labels provided as part of the instructions. More importantly, the gold standard tweets provided

workers the opportunity to learn how to apply the annotation scheme using real examples. Compared to C1, workers were able to correctly identify 4 of the 5 valence gold labels and 3 of the 5 emotion tag gold labels on average based on another set of 20 batches of tweets across all four samples. Given the highly subjective nature of the task, we acknowledge the possibility that novice annotators might not agree with the gold standard labels as they might not interpret the emotion category labels exactly the same way as expert annotators. Also, we presented the gold labels in a subtle way so it is also possible for workers to pay less attention to them.

We proceeded with C2 to collect annotations for the rest of the tweets in Phase 2 as this approach was more promising in reducing inconsistencies between Phase 1 and Phase 2. C2 allowed us to analyze the degree of deviation from ground truth as well as the degree of inconsistencies between expert and novice annotators. For consistency checking, some duplicate tweets were included intentionally (e.g., some appeared in the form of a retweet) in the actual annotation task, so as to examine if the duplicates were assigned exactly the same labels. We exercised greater leniency in approving HIT submissions as we were also interested to investigate how good or bad novice annotators were at performing emotion classification on tweets using a set of fine-grained emotions. HIT submissions were only rejected when one of the three conditions below was true:

- All tweets in a batch were assigned the same valence and/or emotion tag especially if tweets in a batch were annotated as not containing emotion (i.e., “none” was selected for valence) when a variety of labels were found in annotations provided by others for the same batch of tweets.
- None of the gold labels for both valence and emotion tags were correctly identified.
- Unusually high disagreement was observed between one particular worker (i.e., rogue annotator) compared with the others for the same batch of tweets (e.g., zero or very few matches between the rogue annotator with the others).

Workers were paid US\$ 0.50 for every completed and approved HIT containing 30 tweets. In a similar annotation task, Sintsova et al. (2013) offered a payment of \$0.04 for each HIT with 25 tweets, and obtained inter-annotator agreement that was only slightly worse than trained annotators. The payment offered in this study was considerably higher than what was offered in the earlier study by Sintsova et al. (2013). The total cost of annotations was about US\$ 800.00.

To speed up annotation in Phase 2, three graduate students also volunteered to work on annotating a subset of the corpus. Volunteers followed the same instructions and procedures as workers on AMT with the exception of the monetary reward. Tweets in this subset were singly annotated. The primary researcher annotated at least one batch of tweets with each of the volunteers at the beginning to ensure volunteers were fit for the task and reasonable inter-annotator reliability was achieved. Volunteers contributed annotations for 15% of the corpus in Phase 2, while the remaining 85% was annotated by AMT workers (see Table 3.8). A total of 206 workers completed annotation HITs on AMT. Workers submitted 1 HIT at minimum and 302 HITs at maximum. It took roughly 2 months for AMT workers and volunteers to complete annotations for all 10,000 tweets in Phase 2.

<b>Sample</b>	<b># of Batches (AMT)</b>	<b># of Batches (Volunteers)</b>	<b>Total</b>
TOPIC	80	20	100
RANDOM	90	10	100
SEUSER	84	16	100
AVGUSER	85	15	100
<b>Total Batches</b>	<b>339</b>	<b>61</b>	<b>400</b>
<b>Total Tweets</b>	<b>8475</b>	<b>1525</b>	<b>10000</b>

Table 3.8: Batches of annotation contributed by AMT workers and volunteers

About one third of the tweets had full agreement for emotion tag among all annotators (32%). To avoid throwing away any data and to make sure ground truth was obtained for machine learning experiments in Phase 3, the primary researcher manually reviewed all annotations and resolved the disagreements. Such effort was deemed necessary to reduce as

much noise as possible in the corpus, and to ensure that the classification schemes were applied consistently across the two phases of data collection. Similar to the Phase 1, each tweet in Phase 2 was assigned final labels for valence, arousal (mean arousal across all annotators) and emotion tag. Emotion cues provided by the annotators were also reviewed and finalized.

We mainly used the Natural Language Toolkit<sup>15</sup> (NLTK) to study the linguistic properties of the emotion cues (see Section 4.7). We used NLTK's Twitter-aware tokenizer<sup>16</sup> on the emotion cues to generate the unigram tokens. We also used the collocation package for the generation of bigrams and trigrams.

### 3.4 Phase 3: Machine Learning Experiments

The purpose of Phase 3 was to compare the performance of supervised machine learning with that of human annotators in detecting these fine-grained emotion categories in text. Using supervised learning, machine learning algorithms were trained to classify 28 emotion categories using the corpus containing 15,553 annotated examples from Phase 1 and Phase 2. We experimented with two ways to frame the emotion classification task:

- **Multi-class:** A tweet  $x$  is assigned with a single label from 29 classes (28 emotion category labels and no emotion). We used this framing in our preliminary experiments as a means to simplify the classification problem. Since each tweet can only be assigned a single label, we only kept the first (primary) label assigned to tweets with multiple emotions and dropped all other labels. Using this framing, we evaluated the overall performance of classifiers on a per instance basis, i.e., for tweet  $x$ , what is the classifier performance on predicting the correct label for the tweet.
- **Multi-label:** In the corpus, a tweet might be assigned multiple emotion categories. Hence, a more appropriate approach to handle tweets assigned with multiple labels was to frame the classification problems as a multi-label classification task, where each

---

<sup>15</sup> Natural Language Toolkit (NLTK): <http://www.nltk.org/>

<sup>16</sup> NLTK Twitter-aware tokenizer: <http://www.nltk.org/api/nltk.tokenize.html>



instance could be assigned to more than one emotion category label. To handle multi-label classification, a separate binary classifier was built for each emotion category to detect if an emotion category were present or absent in a tweet (emotion X or not emotion X) (Cherry et al., 2012; Joachims, 1998). A tweet x therefore has 28 labels, one from each binary classifier indicating whether each emotion category is present or absent. Tweet x has no emotion if it is not assigned any emotion label from the 28 binary classifiers. Using this framing, we first evaluated the binary classifier performance on each emotion category label. We evaluated the overall performance of the classifiers by averaging over the measures generated all 28 binary classifiers.

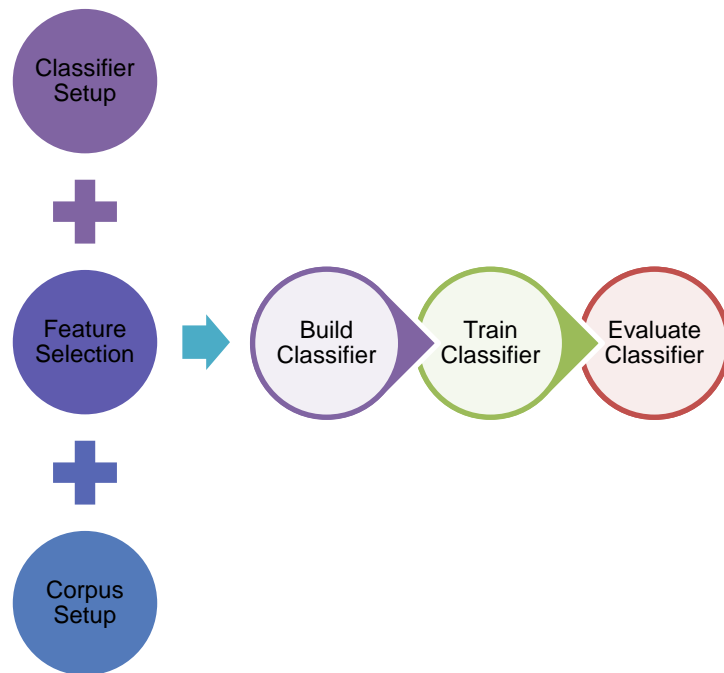


Figure 3.10: Processes for running machine learning experiments

Phase 3 followed the processes shown in Figure 3.10. Building an effective machine learning model for emotion classification required experimentation on the choice of machine learning algorithm (i.e., classifier setup), the selection of features, as well as the training sample. Each tweet was represented by a vector of feature values and a class label. We use the term “class label” to refer to the 28 emotion category labels and none (no emotion). Typical

evaluation metrics (i.e., accuracy, precision, recall, and F-measure) were used to assess the performance of the classifiers.

We used the data mining software, Weka (Hall et al., 2009) for our experiments. Weka came with a collection of machine learning algorithms and functionalities that were sufficient to support the machine learning experiments in this study. To build an effective machine learning model for emotion classification, we conducted machine learning experiments in three parts: 1) classifier-related experiments, 2) feature-related experiments, and 3) data-related experiments.

### **3.4.1 Task 1: Classifier-related Experiments**

The first set of experiments aimed to identify a set of machine learning algorithms that generally perform well for this task. These classifiers would then be used in subsequent tests. Four machine learning algorithms were found to perform well in this problem space: support vector machines (SVM) (Alm et al., 2005; Aman & Szpakowicz, 2007; Brooks et al., 2013; Cherry et al., 2012), Bayesian networks (Sohn et al., 2012; Strapparava & Mihalcea, 2008), decision trees (Hasan, Rundensteiner, et al., 2014), and k-nearest neighbor (KNN) (Hasan, Rundensteiner, et al., 2014; Holzman & Pottenger, 2003). Some amount of parameter tuning for the classifiers was performed in order to produce reasonable performance for the feature selection experiments in Task 2. The features used in Task 1 were held constant across different classifiers in the candidate set. As a starting point, a unigram (i.e., bag-of-words) model, which has been shown to work reasonably well for text classification in sentiment analysis (Pang et al., 2002; Salvetti et al., 2006), was chosen.

We tokenized the text in the corpus and extracted all unique terms as features. We created a custom tokenizer to better handle elements that are common in tweets. In particular, the tokenizer recognizes emoticons, emojis, URLs and HTML encoding. The tokenizer also handles common abbreviations and contractions. Text was encoded in UTF-8 in order to preserve the emojis.

We then evaluated the effect of case normalization (i.e., lowercasing), stemming, and a minimum word frequency threshold ( $f = 1, 3, 5$  and  $10$ ) as a means to reduce the number of features. Classifiers were evaluated using 10-fold cross validation and compared with three baselines (i.e., majority-class, random and OneR). The majority-class baseline, implemented using the ZeroR classifier in Weka, merely assigns all instances in the test set with the majority class label, which is the “none” (no emotion) label. The random baseline picks a class at random. OneR uses a single feature with minimum error for classification.

To make the experiments more manageable in Task 1, we utilized only the corpus of 5,553 tweets developed in Phase 1 (P1). Since the distribution of the emotion categories were similar between Phase 1 and Phase 2 (P1 + P2), we expected our P1 classifiers to exhibit comparable performance with classifiers trained with the full dataset. We experimented with two main experimental setups:

- Multiple-class classification (*multi-class*): Each tweet was assigned to only one emotion label. For tweets with multiple labels, only the primary label (i.e., first label) was assigned to the tweet, and the other labels were ignored. We carried out two sets of experiments. First, we created one single classifier (***multi-class-single***) to distinguish between 29 classes (i.e., 28 emotion categories and *no emotion*). Second, we ran experiments using Weka’s MultiClassClassifier, a meta-classifier that mapped a multi-class dataset into multiple two-class classifiers (***multi-class-binary***), one for each emotion and one for *no emotion*, thus resulting in a setup with 29 binary classifiers in total. Unfortunately, this setup was not designed to handle instances with multiple labels but it offered a straightforward implementation of multiple binary classifications for preliminary analysis. About 92% of the corpus contained instances with only a single label so overall classification performance is close to that of a multi-label classifier.
- Binary classification for multi-label classification (***multi-label***): A binary classifier is built for each emotion category. This setup was able to handle tweets with multiple labels. A

tweet assigned with multiple emotion labels will occur as a positive instance for more than one classifier. For example, a tweet tagged with *happiness* and *love* would appear as a positive instance in both the happiness and love binary classifiers. Each emotion classifier was configured to recognize one emotion against all others. In other words, the instances tagged as Emotion X were positive examples for classifier X, and all instances tagged with labels other than emotion X including *no emotion* were considered as negative examples. Unlike *multi-class*, the *multi-label* setup consists of 28 binary classifiers (i.e., one for each emotion category) and excludes a binary classifier to distinguish between tweets with emotion and *no emotion*.

Initial experiments in Task 1 used *multi-class* classifiers to identify a candidate list of machine learning algorithms. The effectiveness of these algorithms was then evaluated with *multi-label* classifiers. We switched to fully using the *multi-label* setup in Task 2 and Task 3.

Classifier	<i>multi-class-single</i>	<i>multi-class-binary</i>	<i>multi-label</i>
BayesNet	0.533	<b>0.574</b>	<b>0.611</b>
SMO	<b>0.571</b>	0.529	0.593
J48	0.567	0.520	0.561
KNN ( $k = 1$ )	0.391	0.391	0.438

Table 3.9: Micro-averaged F1 of BayesNet, SMO, J48 and KNN for the three experimental setups in Task 1

We found that the use of stemming and case normalization and applying a word frequency threshold of 3 produced consistently good results. Based on our experiments, the two machine learning algorithms that yielded the best performance in this problem space were Sequential Minimal Optimization (SMO), an algorithm for training SVM (Platt, 1998) and Bayesian Networks (BayesNet) (Bouckaert, 1967). Based on the micro-averaged F1 shown in Table 3.9, the performance ranking was similar between the four different classifiers across the three experimental setups except that SMO was the top performing classifier while BayesNet was the second best in the *multi-class-single* setup. For the sake of comparison with *multi-*

*class-single* and *multi-class-binary*, the micro-averaged F1 for multi-label reported in Table 3.9 is computed based on 29 classes including no emotion,

Classifier Type	Classifier	multi-class-single	multi-two-class	multi-label-binary
Bayesian	BayesNet	7	13	11
	NaiveBayes	1	2	0
Support Vector Machines (SVM)	SMO	13	3	7
	LibSVM	3	5	4
Decision Trees	J48	1	0	2
k-Nearest Neighbors	KNN	2	2	0
One Rule	OneR	1	3	4
<b>Total Emotion Categories</b>		<b>28</b>	<b>28</b>	<b>28</b>

Table 3.10: Frequency counts based on best performing classifier (F1) for each emotion category

A more in-depth analysis of the best performing classifier for each emotion category also showed that BayesNet and SMO were well-suited for a majority of the emotion categories as shown in Table 3.10. Based on these initial evaluations, we focused on the SMO and BayesNet classifiers for further experimentation. We also explored two other techniques that might be effective for this task:

- **Ensemble methods:** Bagging, boosting, stacking and voting schemes were explored to combine the decisions from different models into a single classification result. Bagging, short for “bootstrap aggregating” was employed to create separate samples of the training set and to generate a classifier for each sample. Results of multiple classifiers were then aggregated (Breiman, 1996). Bagging assumed that each sample of the training data was different, thus allowing each classifier to learn subtly different focus and perspective of the problem. On the other hand, boosting first created a base classifier from the training data, and subsequent classifiers were created to focus on the instances that were misclassified by the previous classifier (Freund & Schapire, 1996). Since we observed that different emotion categories seemed to respond favorably to different classifiers (see Table 3.10), stacking and voting were also examined to obtain

the best prediction label for each tweet using a combination of different machine learning algorithms found to perform well in this problem space.

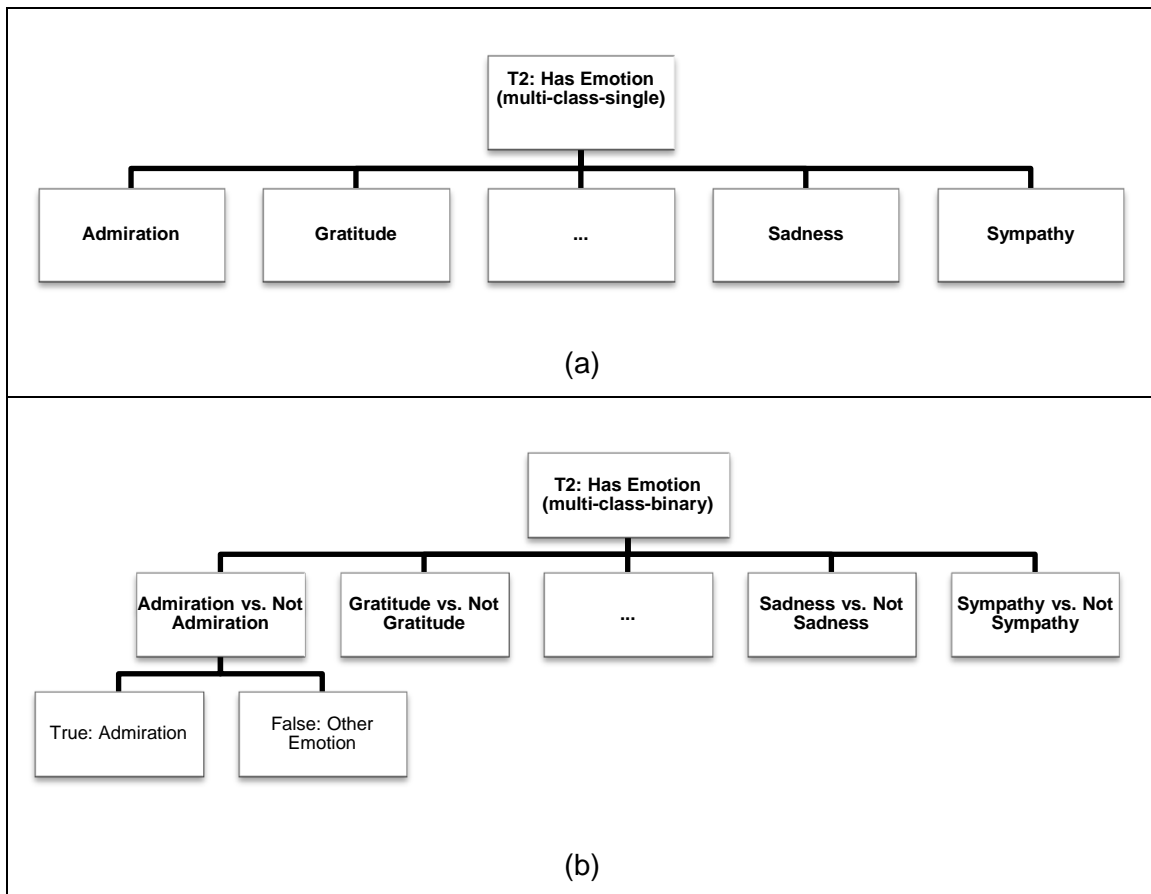


Figure 3.11: Two Tier 2 approaches to train classifiers that distinguish only instances containing emotion: a) *multi-class-single* classification and b) *multi-class-binary* classification

- **Tiered model:** As there were significantly more tweets containing no emotion (i.e., almost half of the corpus), we experimented with the notion of building a stacked or tiered classification model. The classifier in the first tier (Tier 1) was first trained to discriminate between tweets that contained emotion and those that did not. The second tier (Tier 2) then focused solely on distinguishing the 28 emotion classes. For Tier 2, we examined two different approaches: 1) creating a single classifier to discriminate all 28 emotion classes (*multi-class-single*, Figure 3.11a), and 2) creating 28 binary classifiers, each configured to distinguish between emotion X versus all other emotions that are not

X (*multi-class-binary*, Figure 3.11b). We removed all the instances labeled with no emotion for Tier 2 experiments to examine if training more focused classifiers to recognize emotion X versus other emotion would yield better performance compared to classifiers fed with negative examples comprising other and no emotion.

### 3.4.2 Task 2: Feature-related Experiments

A second important factor that affects the performance of machine learning algorithms is the selection of features considered to be good predictor of emotion in text.

Feature Group	Description of Feature Sets
Corpus-based (based on terms in the corpus)	C1: All unigrams (bag-of-words) C2: Bigrams C3: Unigrams with part-of-speech (POS) tags C4: Presence of URL (hasURL)
Lexicon-based (based on terms in the NRC EmoLex <sup>17</sup> )	L1: Words associated with emotion and non-emotion (full lexicon) L2: Words associated with emotion (partial lexicon) L3: Count of words associated with semantic orientation (i.e., positive and negative), anger, anticipation, disgust, fear, joy, sadness, surprise and trust (10 features)
Cue-based (based on terms in emotion cues)	E1: Unigrams from all 28 emotion categories (joined) E2: Unigrams and phrases (bigram and trigrams) from all 28 emotion categories (joined) E3: Unigrams from each emotion category (custom) E4: Unigrams and phrases (bigrams) from each emotion category (custom)

Table 3.11: Feature groups and the description of features

Our goal for Task 2 was to find a representative set of features to produce consistently good classification performance. One key question explored in this study relates to the predictive power of the emotion cues identified by annotators in earlier content analysis phases. We experimented with three feature groups described in Table 3.11.

<sup>17</sup> NRC Word-Emotion Association Lexicon (EmoLex): Mohammad, S. M., & Turney, P. D. (2013). Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3), 436–465.

### 3.4.2.1 Corpus-based Features

Corpus-based features consist of statistical features generated from a large portion of, if not the entire vocabulary of the training corpus. We tested four different features generated from the corpus: C1: unigrams, C2) bigrams, C3) unigrams associated with their POS tags and C4) a binary feature indicating the presence of absence of URLs in each tweet (hasURL). For both the unigram and bigram features, we captured the frequency of occurrences in each tweet and substituted all URLs with a generic tag. The POS tags were generated using GATE Twitter POS tagger<sup>18</sup> (Derczynski, Ritter, Clark, & Bontcheva, 2013). The features were stemmed and lowercased. We only included terms that occur more than 3 times in the corpus as features. We evaluated the classifiers using corpus-based features using 10-fold cross validation.

### 3.4.2.2 Lexicon-based Features

Lexicon-based features are generated based on the terms in the NRC-Word-Emotion Association Lexicon (see Table 2.7 for a more detailed description of the lexicon). The lexicon contained 14,182 terms and a binary marker indicating each term's association with eight emotions (i.e., anger, anticipation, disgust, fear, joy, sadness, surprise and trust) and two sentiments (i.e., positive and negative). For L1, we used all terms in the lexicon as features including both words associated with emotion and words that were not associated with emotion. We also examined using only the subset of words associated with emotion as features for L2. These two sets of features served as more fine-grained representation of emotion. We performed tokenization and case normalization on the tweets. The frequency of occurrence of each term from the lexicon in a given tweet was captured in L1 and L2. On the other hand, L3 reflected more coarse-grained representation of emotion. L3 consisted of ten features with each feature generated by computing the count of words associated with each emotion (eight

---

<sup>18</sup> GATE Twitter part-of-speech tagger: <https://gate.ac.uk/wiki/twitter-postagger.html>



features) as well as sentiment (two features). Classifiers using lexicon-based features were also evaluated using 10-fold cross validation.

### **3.4.2.3 Cue-based Features**

Cue-based features are extracted from the emotion cues identified by annotators. This feature group falls somewhere in the middle of the corpus-based and lexicon-based features, wherein features are carved out from a small subset of the corpus and comprise more than just emotion words. We only utilized the emotion cues reviewed and finalized by the primary researcher (i.e., gold cues). The gold cues are considered to be the ground truth obtained after a systematic review of all the annotators' emotion cues.

For cue-based features, we split the corpus into three parts for feature development (dev set = 30% of the corpus), training (train set = 40% of the corpus) and testing (test set = 30% of the corpus). We used stratified sampling to select the instances to be included in each set to ensure the classes in the dev, train and test sets are roughly balanced. To avoid optimistically biased performance estimates, cue-based features were only selected from the emotion cues of the tweets in the dev set. The classifiers were then trained on the train set and evaluated on the test set.

We tested two different implementations of the cue-based features. In the first implementation (joined), we experimented with using the same set of features across all the 28 emotion categories. Features were extracted from a combined set of emotion cues from all the emotion categories. The second implementation (custom) focused on creating a custom set of features for each emotion category. For both implementations, we generated unigrams and phrases from the emotion cues as features. All unigrams and phrases are lowercased. We experimented with both stemmed and non-stemmed versions of features to examine if preserving the morphological variations of the stems has any effect on classification performance.

In the joined feature implementation, we extracted all unigrams from the emotion cues in the dev set (E1). To generate the phrases from the joined emotion cues in the dev set for E2, we first handpicked 177 phrases from the top 50 most frequent bigrams, top 50 most frequent trigrams, and collocations that commonly co-occur but not necessarily related to the frequency of words. We then expanded our list of phrases to include negations. An automatic approach is used to extract all bigrams and trigrams starting with a common negation word (refer to full list of negation words in Appendix B) from the full set of emotion cues (i.e., not restricted to only the dev set). A total of 814 negated phrases are extracted from this procedure. The full list of phrases contains 991 multi-word items.

**\*\* Extract bigrams from emotion cues and negative examples of an emotion category**

For each emotion category (C):

    cue\_token = Tokenize dev emotion cues (p)

    neg\_token = Tokenize dev negative examples (n)

    corpus\_token = Tokenize corpus (x)

    cue\_bigrams = Extract top 1000 bigrams sorted by frequency in p

    neg\_bigrams = Extract top 2000 bigrams sorted by frequency in n

    For each cue\_token:

        cue\_term\_weight =  $\log(\text{token frequency in p} / (\text{token frequency in x} * \% C \text{ in x}))$

        if (cue\_term\_weight > 0) and (cue\_term\_weight < max(cue\_term\_weight))

            pos\_phrase = Extract all bigrams in p containing token

            neg\_phrase = Extract all bigrams in n containing token

Figure 3.12: Pseudocode on bigram extraction for custom phrase cues

The custom feature implementation follows a different procedure. For each emotion category, E3 is made up of only the cue unigrams associated with the category. The number of cue unigrams used as features is different for each emotion category. We extracted phrases consisting of collocations from the emotion cues (i.e., represents positive examples) as well as the negative examples in the dev set. The intuition is that capturing the context surrounding a term that appears in both the positive and negative examples for an emotion category can help

the binary classifier learn to better distinguish between the positive class and the negative class. The automatic phrase extraction procedure follows the steps in the pseudocode shown in Figure 3.12.

The rationale of using cue term weight is further described in Chapter 4. We set the cutoff point of cue\_bigrams to 1000 and the neg\_bigrams to 2000 as there were more bigrams generated from the negative examples. We only included bigrams composed of words and removed the ones containing period, comma, emoticons, emojis and hashtags.

#### **3.4.2.4 Cross-Group Combinations**

We also experimented with cross-group combinations of feature sets. To avoid redundant features, we attempted combinations consisting of only vocabulary features (i.e., actual terms that occur in the tweets) with non-vocabulary features (i.e., C4 and L3).

### **3.4.3 Task 3: Sample-related Experiments**

In Task 3, we explored two types of experiments related to the samples used as training data. The first type of experiment focuses on class imbalance strategies. Due to the imbalance distribution of class labels observed in the corpus (i.e., there were significantly more negative examples than positive examples in the corpus), we experimented with downsampling to create more balanced datasets. The size of negative examples in the training data is reduced to match the positive examples. We adjust the ratio of the positive and negative classes using Weka's *SpreadSubsample* filter. Holding the size of positive examples as constant in the training data, we systematically increase the ratio of negative examples to positive examples to examine how class imbalance affects the behavior of classifiers.

The second type of experiment examined if the different sampling strategies utilized in this study had any effect on the performance of the classifier. Our intuition is that a classifier trained with a greater diversity of examples from four different sampling strategies would be

superior compared to any individual classifier trained on a specific sampling strategy. Five classifiers trained using different training samples are listed below:

- Classifier trained solely on RANDOM training data
- Classifier trained solely on TOPIC training data
- Classifier trained solely on AVG-USER training data
- Classifier trained solely on SEN-USER training data
- Classifier trained on combined RANDOM, TOPIC, AVG-USER and SEN-USER training data

Classifiers in Task 3 were evaluated using the train/test splits. We used the same training and test sets from our experiments using cue-based features. We varied the sample in the training data for the experiments in Task 3 but all classifiers were evaluated on the same test set.

### **3.5 Conclusion**

Chapter 3 describes the details of the three-phase methodology followed in this research study to uncover a set of fine-grained emotions from tweets, to reveal the linguistic cues associated with each emotion and to explore how well classifiers can be trained to recognize this set of emotions. Phase 1 applied grounded theory in a small-scale content analysis to expose a set of fine-grained emotion categories inductively from data and to develop a more elaborate annotation scheme for the study of emotion in text. Phase 2 employed large-scale content analysis to further test the annotation scheme developed in Phase 1 and to expand the size of the corpus for classification experiments. Phase 3 utilized various supervised learning methods in search of the machine learning model that would yield the best performance for fine-grained emotion classification using this carefully curated corpus.

# Chapter 4: Conceptual Analysis

This chapter presents the results of the work undertaken to answer the first two research questions:

- *R1: What emotions can humans detect in microblog text?*
- *R2: What salient linguistic cues are associated with each emotion?*

To answer R1, we first characterize the emotion categories included in our annotation scheme. Three measures are used to determine the emotion categories that humans can detect in microblog text: class distinctiveness, level of agreement and class intuitiveness. Section 4.1 shows the distinctiveness of the emotion categories that emerged from the tweet data. Having data obtained from both expert and novice annotators allowed us to compare the level of agreement achieved by the two groups in the emotion classification task, thus providing insights to the performance we can realistically expect from humans on the subjective task of detecting fine-grained emotions in text. Section 4.2 presents a detailed analysis of the overall inter-annotator agreement and the level of full agreement annotators achieved in identifying the emotion categories. Section 4.3 compares initial human judgments to the ground truth as a means to gauge how intuitive are the emotion categories.

To address R2, we describe the salient linguistic cues associated with each emotion category from our analysis of the emotion cues. We begin the second part of the chapter by presenting the characteristics of the corpus (Section 4.5) and emotion cues (Section 4.6). Salient linguistic cues refer to linguistic features in text that serve as important or notable indicators of an emotion. We perform linguistic analysis at the lexical-level to identify salient linguistic cues that are key to distinguishing the emotion categories. Findings from the linguistic analysis are presented in Section 4.7.

## 4.1 Class Distinctiveness

Class distinctiveness measures how well humans are able to distinguish each emotion category as evidenced by how reliable annotators are at recognizing the categories. Distinctiveness is measured using inter-annotator reliability measures at the category level. We compare the distinctiveness of a set of 48 emotion categories obtained from inductive coding (Task 1) and card sorting (Task 2) with a refined set of 28 emotion categories obtained from the word rating study (Task 3) in Phase 1.

### 4.1.1 Emotion Categories: Set-48

Annotators suggested a total of 246 emotion tags from the inductive coding task (Phase 1: Task 1) across four samples (4010 tweets). The emotion tags are emotion labels suggested by annotators. These 246 emotion tags were then grouped into 48 emotion categories through the card sorting exercise (Phase 1: Task 2). Another 23 emotion tags were added from Phase 2 (tags indicated by \* in Table 4.1, Table 4.2 and Table 4.3). We were able to match these 23 emotion tags from Phase 2 to equivalent existing emotion tags found in Phase 1. All 269 emotion tags were sorted into 48 emotion categories. Table 4.1 shows the positive emotion categories (17) and the list of emotion tags associated with each emotion category, Table 4.2 shows the neutral emotion categories (10) and Table 4.3 shows the negative emotion categories (21).

Annotators used a variety of nouns, adjectives and verbs as emotion tags. Most of the emotion tags in an emotion category are morphological variations of the category name and its synonyms. A few emotion categories contain emotion tags that are action words used to describe the emotion (e.g., the emotion tag *celebrate* in *happiness* describes the act of acknowledging a happy event with an enjoyable activity). The number of emotion tags in the emotion categories ranges from a minimum of 1 (*relief*, *dread* and *lust*) to a maximum of 21

(*sadness*). This suggests that some emotions (e.g., *sadness* and *happiness*) are described with a larger vocabulary of words than others.

Category Name	Emotion Tag(s)	Count
<b>Happiness</b>	Happiness, Happy, Cheerful, Cheering, Joy, Joyful, Delight, Delighted, Elated, Blessed, Enjoyment, Beatific, Congratulation, Congratulations, Congrats, Celebrate, <i>Celebratory*</i> , <i>Gratification*</i>	18
<b>Inspiration</b>	Inspiration, Inspired, Inspiring, Moved, Motivation, <i>Motivated*</i> , Encouragement, Encouraged, <i>Supportive*</i> , Touched	10
<b>Admiration</b>	Impressed, Veneration, Admiration, Appreciation, Honorable, Honored, Admire, Respect, Respectful	9
<b>Gratitude</b>	Gratitude, Grateful, Thank, Thankfulness, Thankful, Appreciate, Appreciative, Blessed	8
<b>Pleased</b>	Pleased, Pleasure, Glad, Satisfied, Satisfaction, Content, Contented	7
<b>Amusement</b>	Amused, Amusement, Fun, Funny, Humorous, Humored, <i>Humor*</i>	7
<b>Excitement</b>	Excitement, Excited, Exciting, Enthusiastic, Energetic, Enthused, Aroused	7
<b>Anticipation</b>	Anticipation, Anticipated, Expected, Expect, Expecting, Eager, Keen	7
<b>Love</b>	Love, Loved, Loving, Obsessed, Affection, Bonding	6
<b>Relaxed</b>	Relaxed, Relax, Comfortable, Calm, Serene, At ease	6
<b>Pride</b>	Pride, Proud, Proudness, Accomplished, Praiseful	5
<b>Like</b>	Like, Liking, Fond, Affinity	4
<b>Hope</b>	Hope, Hopeful, Optimistic, Optimism	4
<b>Amazement</b>	Amazed, Amazing, Amazement, Awed	4
<b>Fascination</b>	Fascination, Interest, Interested	3
<b>Confidence</b>	Confidence, Confident	2
<b>Relief</b>	Relief	1

Table 4.1: Emotion tags associated with each of the 17 positive emotion categories

Category Name	Emotion Tag(s)	Count
<b>Surprise</b>	Surprise, Astonish, Astonished, Surprised, Unexpected, Unbelievable, <i>Disbelief*</i>	7
<b>Desire</b>	Desire, Ambitious, Wish, Wishes, Wishful, Wishing	6
<b>Nostalgia</b>	Nostalgia, Nostalgic, <i>Reminiscent*</i> , <i>Reminiscing*</i>	4
<b>Empathy</b>	Empathetic, Compassion, Compassionate	3
<b>Confusion</b>	Confusion, Confused, Confuse	3
<b>Curiosity</b>	Curiosity, Curious	2
<b>Exhaustion</b>	Exhausted, Tired	2
<b>Indifference</b>	Indifference, Indifferent	2
<b>Ambivalence</b>	Torn, Conflicted	2
<b>Lust</b>	Lust	1

Table 4.2: Emotion tags associated with each of the 10 neutral emotion categories

Category Name	Emotion Tag(s)	Count
<b>Sadness</b>	Sad, Saddened, Sadness, Sorrow, Distress, Distressed, Depressed, <i>Depression*</i> , Grief, Dejected, Miserable, Pain, Gloomy, Cry, Hurt, Somber, <i>Lonely*</i> , <i>Loneliness*</i> , <i>Sick*</i> , <i>Resigned*</i> , <i>Commemoration*</i>	21
<b>Worry</b>	Worried, Worry, Anxious, Anxiety, Concerned, Concern, Concerns, Urgent, Tense, Mad, Restless, <i>Agitated*</i> , <i>Stress*</i> , <i>Worrisome*</i>	14
<b>Annoyance</b>	Frustration, Frustrated, Annoy, Annoyed, Annoying, Annoyance, Irritated, Irritation, Aggravated, Miffed, Upset, Unhappy, Offended	13
<b>Hate</b>	Dislike, Hate, Hatred, Aversion, Disdain, Averse, Self-loathing, Revenge, Vengeance, <i>Contempt*</i> , <i>Condescension*</i> , <i>Scorn*</i>	12
<b>Fear</b>	Fear, Scary, Frightened, Afraid, Scare, Scared, Crazy, Craze, Alarm, Caution, Danger	11
<b>Anger</b>	Anger, Angry, Blame, Blamed, Outraged, Aggressive, Pissed, <i>Indignation*</i> , <i>Resentment*</i>	9
<b>Displeased</b>	Displeased, Dissatisfaction, Dissatisfied, Unsatisfied, Disapproval, Discontent	6
<b>Regret</b>	Regret, Regretful, Sorry, Remorse, Remorseful	5
<b>Yearning</b>	Yearning, Longing, Miss, Missing	4
<b>Shame</b>	Shame, Shameful, Embarrassed, Embarrassment	4
<b>Sympathy</b>	Sympathy, Sorry, Sympathetic, Pity	4
<b>Awkward</b>	Awkward, Uncomfortable, Weird, Strange	4
<b>Disappointment</b>	Disappointment, Disappointing, Disappointed	3
<b>Doubt</b>	Doubtful, Doubt, Pessimism	3
<b>Desperation</b>	Desperation, Desperate, <i>Hopeless*</i>	3
<b>Disgust</b>	Disgust, Disgusted, Degradation	3
<b>Shock</b>	Shock, Shocked, Dismayed	3
<b>Guilt</b>	Guilt, Guilty	2
<b>Boredom</b>	Boring, Bored	2
<b>Jealousy</b>	Jealousy, Jealous	2
<b>Dread</b>	Dread	1

Table 4.3: Emotion tags associated with each of the 21 negative emotion categories

Using Phase 1 (P1) data (5553 tweets), we compute both Fleiss' kappa ( $\kappa$ ) and Krippendorff's alpha ( $\alpha$ ) for each emotion category among three annotators to determine how well-defined each emotion category is and how reliable the annotators are at recognizing the emotion categories. Similar to the approach used in Teufel, Siddharthan, & Tidhar (2006) to determine the distinctiveness of a category in content analysis, we first create artificial splits of the data into a binary representation for each emotion category (i.e., the emotion category is represented by 1 while all other labels were collapsed and represented by 0). The  $\kappa$  and  $\alpha$



values are exactly the same so we reported only  $\kappa$  in Table 4.4. Higher  $\kappa$  indicates higher reliability among annotators in distinguishing an emotion category.

Category	n	$\kappa$	PI	BI	Category	n	$\kappa$	PI	BI
Mean-48	3476	0.315			Mean-48	3476	0.315		
Gratitude	221	0.791	0.930	0.002	Admiration	158	0.315	0.971	0.004
Jealousy	4	0.667	0.999	0.000	Anger	91	0.3	0.976	0.005
Lust	12	0.555	0.998	0.000	Awkward	10	0.291	0.997	0.000
Amusement	237	0.543	0.944	0.002	Shock	16	0.279	0.997	0.000
Exhaustion	10	0.541	0.997	0.001	Inspiration	21	0.273	0.995	0.001
Pride	85	0.511	0.973	0.002	Pleased	323	0.268	0.907	0.020
Regret	42	0.509	0.990	0.001	Disgust	19	0.253	0.995	0.001
Yearning	31	0.47	0.994	0.001	Annoyance	155	0.241	0.956	0.005
Sadness	158	0.457	0.952	0.002	Worry	34	0.236	0.990	0.000
Hope	185	0.447	0.950	0.006	Confusion	25	0.225	0.991	0.001
Love	113	0.447	0.967	0.001	Displeased	149	0.216	0.956	0.012
Curiosity	30	0.442	0.991	0.001	Amazement	42	0.206	0.995	0.002
Surprise	77	0.434	0.979	0.003	Sympathy	30	0.201	0.991	0.001
Like	120	0.425	0.980	0.004	Relaxed	13	0.194	0.995	0.000
Indifference	28	0.424	0.994	0.002	Nostalgia	6	0.181	0.999	0.001
Fear	36	0.411	0.991	0.003	Confidence	19	0.16	0.997	0.001
Excitement	175	0.394	0.947	0.000	Disappointment	49	0.115	0.987	0.003
Anticipation	89	0.392	0.975	0.005	Guilt	7	0.071	0.999	0.001
Boredom	12	0.391	0.998	0.001	Relief	13	0.071	0.999	0.000
Desire	60	0.386	0.987	0.000	Fascination	12	0.066	0.998	0.001
Happiness	455	0.351	0.856	0.010	Dread	7	0	1.000	0.000
Desperation	8	0.333	0.999	0.000	Ambivalence	12	-0.001	0.999	0.001
Shame	16	0.321	0.997	0.000	Doubt	13	-0.001	0.998	0.001
Hate	44	0.32	0.989	0.000	Empathy	4	-0.001	0.999	0.001

Table 4.4: Distinctiveness of 48 emotion categories based on kappa ( $\kappa$ )

Mean  $\kappa$  achieved across all 48 emotion categories is 0.315. Of the 48 emotion categories, annotators are most reliable in recognizing *gratitude* ( $\kappa = 0.79$ ) and *jealousy* ( $\kappa = 0.67$ ), and achieve moderate reliability (0.41 – 0.6) for 14 emotion categories. The  $\kappa$  for a majority of the categories ranges from 0.2 – 0.4. On the other hand, annotators are the least reliable in recognizing *dread*, *ambivalence*, *doubt* and *empathy* ( $\kappa \leq 0$ ). The  $\kappa$  for the 4 least reliable emotion categories falls below zero (i.e., an indicator that agreement is no better than expected by chance) because no full agreement is observed among three annotators for even

one of the few instances identified in the categories respectively.  $\kappa$  plunges below zero when no full agreement exists in the data especially for rare categories.

There are several reasons why fair to moderate agreement is observed for most of the emotion categories in this set of 48. First, some emotion categories still share close semantic ties, thus making it difficult for annotators to set them apart. For example, *annoyance* is often confused with *anger* as *annoyance* is a manifestation of *anger* expressed with lower intensity (see Example 4.1 and Example 4.2).

**Example 4.1:** I saw 3 things yesterday that made me want 2 slit my wrists: the Jets game; Looper & #FoxNews Sunday w/ #ChrisWallace #MorningJoe #GOP #tcot  
**[Annoyance]**

**Example 4.2:** This is America speak English or GTFO!!! @Samaanthalove **[Anger]**

To visualize the degree of confusion between each pair of emotion categories as a heatmap, we first build a co-occurrence matrix based on frequencies of paired annotator (A) assignment of labels (i.e., <A1, A2>, <A2, A3> and <A1, A3>) for all 4010 tweets annotated with a single emotion category in the inductive coding task (see Figure 4.1). Tweets annotated with multiple emotions are excluded to avoid adding another layer of complexity to the analysis. Using the same approach, we also construct another co-occurrence matrix shown in Figure 4.2 based on frequencies of paired gold and annotator labels (i.e., <Gold, A1>, <Gold, A2> and <Gold, A3>) to examine how often annotator labels match the gold labels. The co-occurrence frequencies in the heatmaps are represented in different shades of gray, the darker the color, the higher the co-occurrence between an emotion category pair.

If annotators are truly able to distinguish an emotion category, the color of the main diagonal cells in both Figure 4.1 and Figure 4.2 would be the darkest as in the case of *gratitude* and *amusement*. With the exception of *desperation*, the presence of shades of very light gray on the main diagonal in Figure 4.1 indicates the absence of agreement between any annotator pairs for 7 emotion categories (i.e., *ambivalence*, *doubt*, *dread*, *empathy*, *fascination*, *guilt* and

*relief*), which coincides with the 7 least distinctive emotion categories in Table 4.4. As shown in Figure 4.2, at least one annotator correctly identified the gold labels for 6 of the 7 least distinctive emotion categories. None of the annotators are able to correctly recognize *dread*. The moderate shade of gray between *annoyance* and *anger* in both heatmaps further verify that annotators had difficulty telling them apart. While the set of 48 emotion categories capture emotions expressed at a very fine-grained level, the annotation data reveals that some emotion categories are problematic and need to be merged or removed.

Care must be taken when interpreting the magnitude of  $\kappa$ . Another reason to explain the fair to moderate agreement observed is that the  $\kappa$  coefficient is influenced by the effects of prevalence and bias (Byrt, Bishop, & Carlin, 1993; Sim & Wright, 2005). The prevalence effect, measured using the prevalence index<sup>19</sup>, exists when the proportion of agreements for the positive class (i.e., the emotion category represented by 1) is unevenly distributed compared with that of the negative class (i.e., not the emotion category represented by 0). Such cases would produce a relatively high value of expected agreement ( $P_e$ ) and the magnitude of  $\kappa$  is reduced accordingly, thus yielding a high prevalence index. In our case, the number of agreements for the negative class is significantly greater than the number of agreements for the positive class. Therefore, low magnitudes of  $\kappa$  can be partially explained by the large prevalence index (PI) values (i.e., shown in Table 4.4).

The prevalence effect is especially apparent for the sparse categories. For sparse categories,  $\kappa$  tends to underestimate the agreement (Viera & Garrett, 2005). We divide the emotion categories into four groups based on class frequency ( $n$ ) and  $\kappa$  as shown in Figure 4.3. Indeed, a majority of the low frequency emotion categories fall into the *low n, low  $\kappa$*  group (i.e., bottom left quadrant C) while a majority of the high frequency ones fall into in to *high n, high  $\kappa$*  group (i.e., top right quadrant B).

---

<sup>19</sup> *Prevalence Index (PI)* =  $\frac{|\text{Frequency of positive class agreement} - \text{Frequency of negative class agreement}|}{n}$

\* Value of PI ranges from 0 – 1, where 0 reflects low prevalence and 1 reflects high prevalence

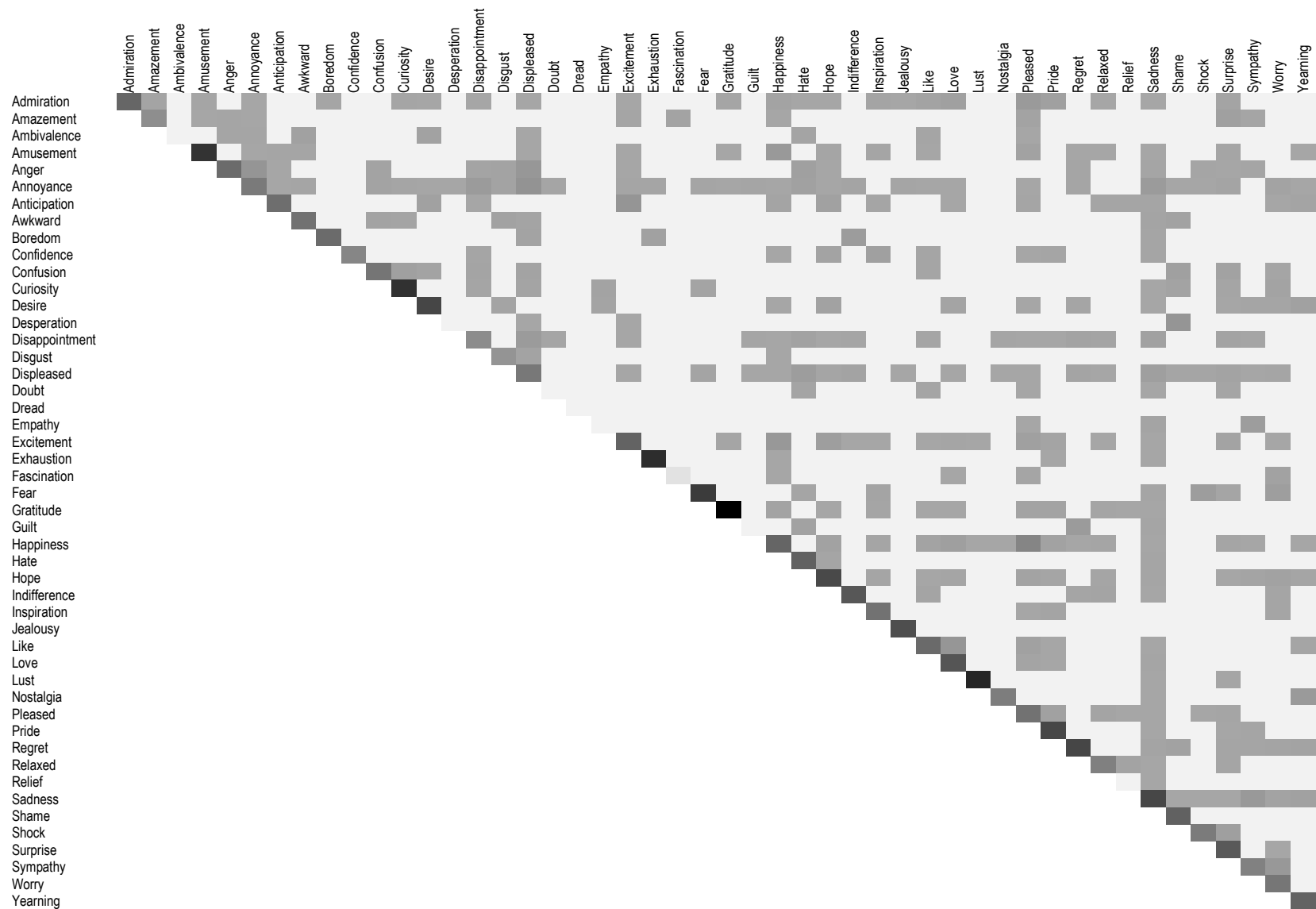


Figure 4.1: Heatmap showing co-occurrence frequencies of paired annotator labels

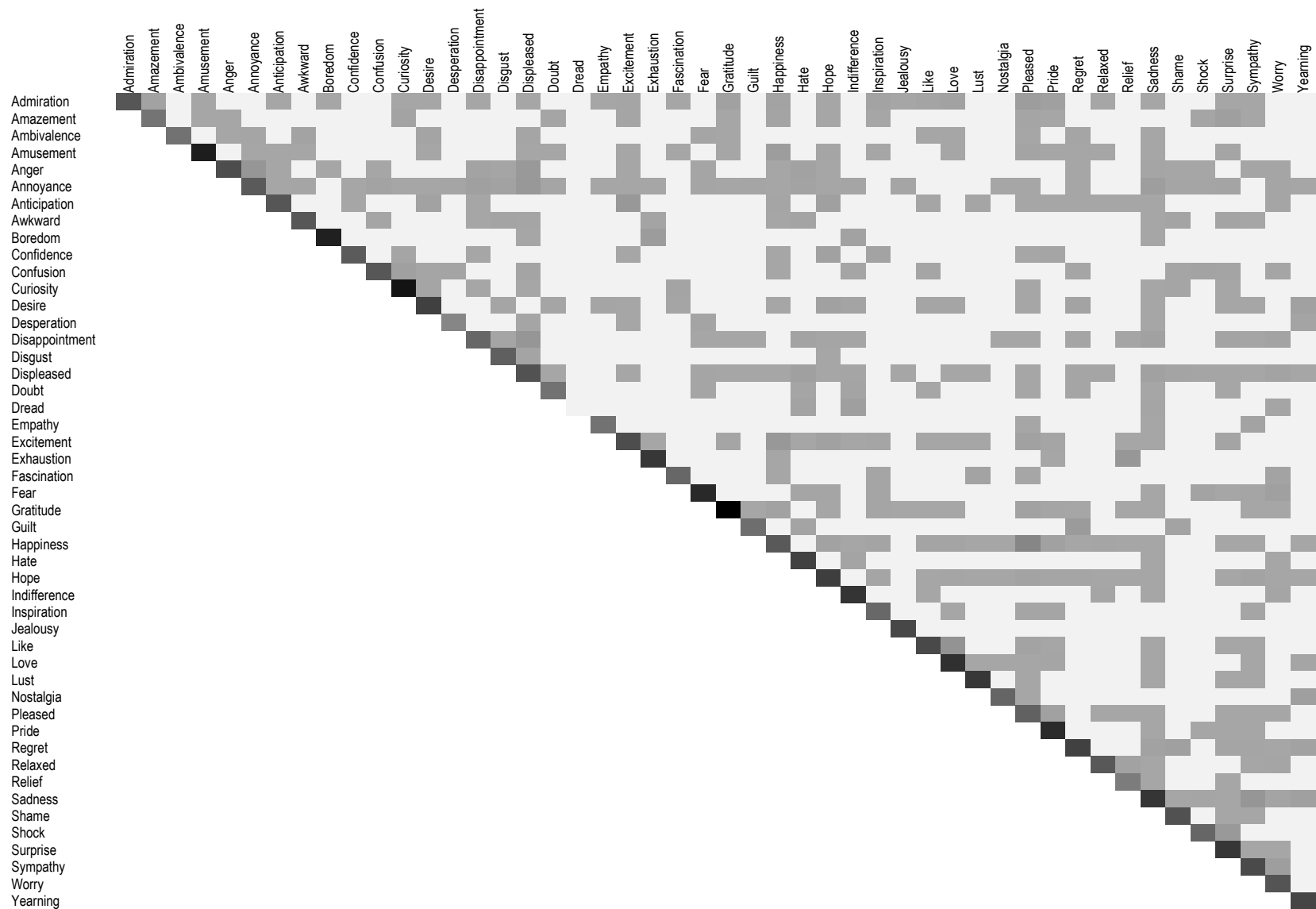


Figure 4.2: Heatmap showing co-occurrence frequencies of paired gold and annotator labels

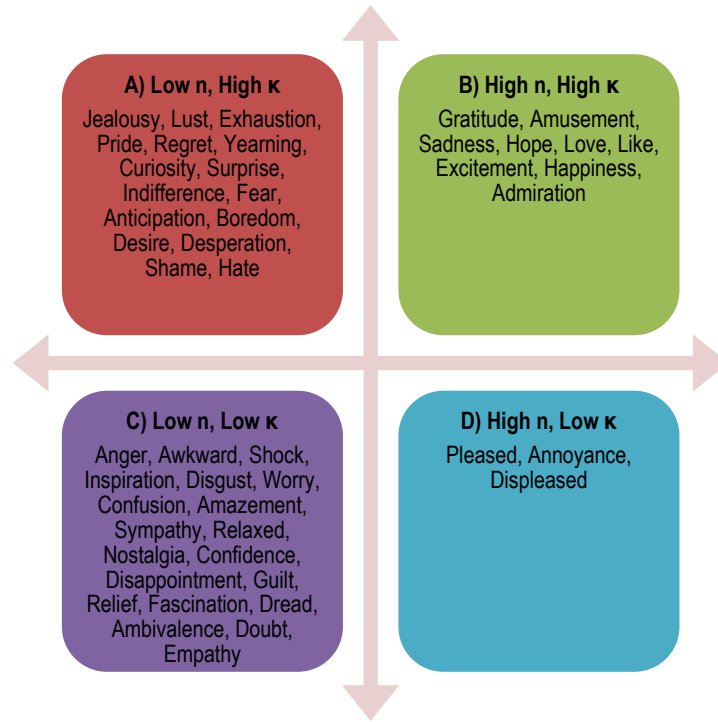


Figure 4.3: Four groupings of emotion categories based on class frequency and  $\kappa$

Interestingly, there are also emotion categories with low frequency that managed to achieve relatively high  $\kappa$  (see emotion categories listed on the top left quadrant A). For instance, annotators are able to reliably recognize expressions of *jealousy* although the emotion category occurred only four times in the data. Each emotion in quadrant A has a distinctive set of patterns that set it apart from other categories; even if the category occurs rarely, it can be recognized without difficulty. Three emotion categories (i.e., *pleased*, *annoyance* and *displeased*) exhibit low  $\kappa$  values despite occurring with higher frequency (bottom left quadrant D). One potential interpretation of this observation is that these emotion categories are not as distinct or as well defined as other high frequency categories. Therefore, the emotion categories in quadrant D join the emotion categories in quadrant C as candidates to be merged or removed.

The bias effect, measured by the bias index<sup>20</sup>, is caused by the uneven proportion of disagreements between the positive cases and the negative cases (Sim & Wright, 2005). Low bias yields lower  $\kappa$ . The low bias index (BI) values shown in Table 4.4 can partially explain the reason for low magnitudes of  $\kappa$  for each emotion category.

As  $\kappa$  is susceptible to the effects of prevalence and bias, low  $\kappa$  may not necessarily reflect low agreement following common  $\kappa$  magnitude guidelines (Landis & Koch, 1977) especially given the nature of the data. Nonetheless, the ranking of  $\kappa$  across the emotion categories still provides meaningful information to help us identify the emotion categories that are relatively easier for annotators to recognize compared with the others (e.g., annotators are better at recognizing *gratitude* as opposed to *doubt*).

#### 4.1.2 Emotion Categories: Set-28

Task 3 in Phase 1 (i.e., the word rating task) further refined the 48 emotion categories to a set of 28. Of the 48 emotion categories, 18 are merged with another category and 2 are dropped. Table 4.5 shows the general description of the final set of 28 emotion categories. To determine how reliable annotators are in recognizing each of the 28 emotion categories, we compute the  $\kappa$  and  $\alpha$  per category among three annotators using P1 and P2 data respectively. The  $\kappa$  values and frequencies for the set of 28 emotions are displayed in Table 4.6. By merging semantically related emotion categories from the set of 48 together, increases in  $\kappa$  are observed for the majority of merged categories in P1 except for *fear*, *regret* and *relaxed*. This serves as an indicator that the merging process helps to increase the distinctiveness of the emotion categories. The  $\kappa$  scores for *fear*, *regret* and *relaxed* are pulled down slightly by one of its low-scoring merged member.

---

<sup>20</sup>  $Bias\ Index\ (BI) = \frac{|Frequency\ of\ positive\ class\ disagreement - Frequency\ of\ negative\ class\ disagreement|}{n}$

\* Value of BI ranges from 0 – 1, where 0 reflects low bias and 1 reflects high bias

The annotations collected from AMT (P2) show that novice annotators are less able to distinguish the 28 emotion categories. Lower  $\kappa$  scores for every category are observed in P2 compared with P1 as shown in Table 4.6. Mean  $\kappa$  for P1 across the 28 categories was 0.328 while mean  $\kappa$  for P2 was 0.165. As observed in P1, *gratitude* was also the easiest emotion to recognize with fairly high reliability among novice annotators, followed by *love*, *sadness*, *excitement*, *amusement*, *happiness* and *jealousy*. Novice annotators have more difficulty recognizing *fascination* and *indifference*. Novice annotators most often mix up *indifference* with *no emotion*. None of the emotion categories fall below chance agreement.

Recognizing 28 emotion categories from tweets is a highly subjective and challenging task for novice annotators (Antoine, Villaneau, & Lefevre, 2014). Although category definitions are provided, it is impossible to ensure that all novice annotators share the exact same interpretation and understanding of each of the 28 emotion categories as the expert annotators. Distinguishing between 28 emotion categories also imposes high cognitive load and there is relatively little that can be done to mitigate any disruptions or confusions faced by the annotators in a crowdsourcing environment. We did find that novice annotators who submitted multiple HITs tended to perform better than those who attempted the task only one time.

As shown in Figure 4.4, agreement for the emotion categories ranges from slight to substantial. When P1 and P2 data are combined, one emotion category (*gratitude*) achieves substantial agreement (0.61 – 0.8) while seven emotion categories (*jealousy*, *love*, *amusement*, *sadness*, *happiness*, *excitement* and *pride*) attain moderate agreement (0.41 – 0.6). However, given sufficient training (P1), we observe that slightly over half of the categories (16) can be identified with at least moderate agreement among three annotators. With limited training, annotators are, at best, able to recognize two emotion categories with moderate agreement. Four emotion categories achieve only slight agreement (0 – 0.2) in both P1 and P2: *relaxed*, *confidence*, *doubt* and *fascination*. These four emotion categories are among the hardest for both expert and novice annotators to recognize.



Code	Description
Admiration	Someone or something regarded as impressive or worthy of respect. Honoring or looking up to someone.
Amusement	State of finding something funny or entertaining.
Anger	Feeling of disappointment, displeasure, dissatisfaction, annoyance, frustration, hostility or rage caused by the non-fulfillment of one's hopes/expectations or about an undesirable event.
Boredom	Feeling dull, uninterested or left without anything in particular to do.
Confidence	Feeling of self-assurance arising from one's appreciation of one's own abilities or qualities. Feeling one can trust or rely on someone or something.
Curiosity	Strong desire to know or learn something.
Desperation	Feeling complete loss of hope or despair, typically one that results in rash or extreme behavior. Suffering or driven by great need or distress.
Doubt	State of being bewildered, confused, uncertain or unclear about something. Having mixed feelings about someone or something. Feeling of distrust, suspicion or one cannot rely on someone or something.
Excitement	Feeling great enthusiasm and anticipation in considering some expected or longed-for good event.
Exhaustion	State of physical or mental fatigue or feeling tired.
Fascination	State of being fascinated, amazed or interested in something. Feeling of great wonder or awe.
Fear	Feeling caused by the belief that someone or something is dangerous, likely to cause pain, or a threat. Feeling dread or anticipate with great apprehension or fear. Feeling anxious or worried over actual or potential problems.
Gratitude	State of being thankful or readiness to show appreciation for and to return kindness.
Happiness	Feeling pleased, satisfied, happy or delighted about a desirable event.
Hate	Feeling of dislike, distaste or aversion towards a person, event or thing. Feeling of disgust or profound disapproval aroused by something unpleasant or offensive.
Hope	Feeling of expectation and desire for a certain event to happen or grounds for believing something good will happen.
Indifference	Lack of interest, concern, or sympathy.
Inspiration	Feeling that makes someone want to do something or that gives someone an idea about what to do or create.
Jealousy	Feeling or showing envy of someone or their achievements and advantages. Feeling or showing suspicion of someone's unfaithfulness in a relationship.
Longing	Yearning for or missing someone or something that one cannot have or cannot get easily. Feeling nostalgic, sentimental longing or wistful affection for the past, typically for a period or place with personal associations.
Love	Feeling of affection or natural liking towards another person, event or thing.
Pride	Deep pleasure derived from one's own achievements, the achievements of those with whom one is closely associated, or from qualities or possessions that are widely admired.
Regret	Feeling remorse or repentance over something that has happened or has been done. Feeling guilty of having done wrong or failed in an obligation.
Relaxed	Feeling calm, at ease. Relief following release from anxiety or distress.
Sadness	Feeling of loss, helplessness or sorrow for own misfortune.
Shame	Humiliation or embarrassment caused by the consciousness of wrong or foolish behavior. Feeling uncomfortable or awkward in a situation.
Surprise	Unexpected or astonishing event, fact or thing. Sudden shocking event or experience.
Sympathy	Feeling of pity and sorrow for someone else's misfortune. Feeling empathy and expressing the ability to understand and share the feelings of another.

Table 4.5: General description of 28 emotion categories

	P1		P2	
Category	n	$\kappa$	n	$\kappa$
Mean-28	3414	0.328	5916	0.165
Gratitude	221	0.791	300	0.572
Jealousy	5	0.667	29	0.321
Amusement	237	0.543	423	0.337
Exhaustion	10	0.541	39	0.174
Love	234	0.522	447	0.408
Pride	85	0.511	128	0.291
Regret	49	0.497	104	0.084
Happiness	778	0.49	1009	0.326
Excitement	265	0.468	421	0.345
Longing	41	0.467	80	0.223
Anger	444	0.458	757	0.308
Sadness	158	0.457	363	0.38
Surprise	93	0.451	173	0.217
Hope	187	0.447	335	0.249
Curiosity	30	0.442	63	0.16
Indifference	28	0.424	40	0.04
Boredom	12	0.391	36	0.262
Hate	63	0.384	129	0.234
Fear	77	0.341	162	0.242
Desperation	8	0.333	50	0.047
Shame	26	0.325	64	0.095
Admiration	158	0.315	245	0.183
Inspiration	21	0.273	54	0.09
Sympathy	35	0.222	66	0.166
Fascination	54	0.188	150	0.018
Relaxed	26	0.182	51	0.075
Doubt	50	0.167	108	0.071
Confidence	19	0.16	91	0.088

Table 4.6: Distinctiveness of 28 emotion categories based on  $\kappa$

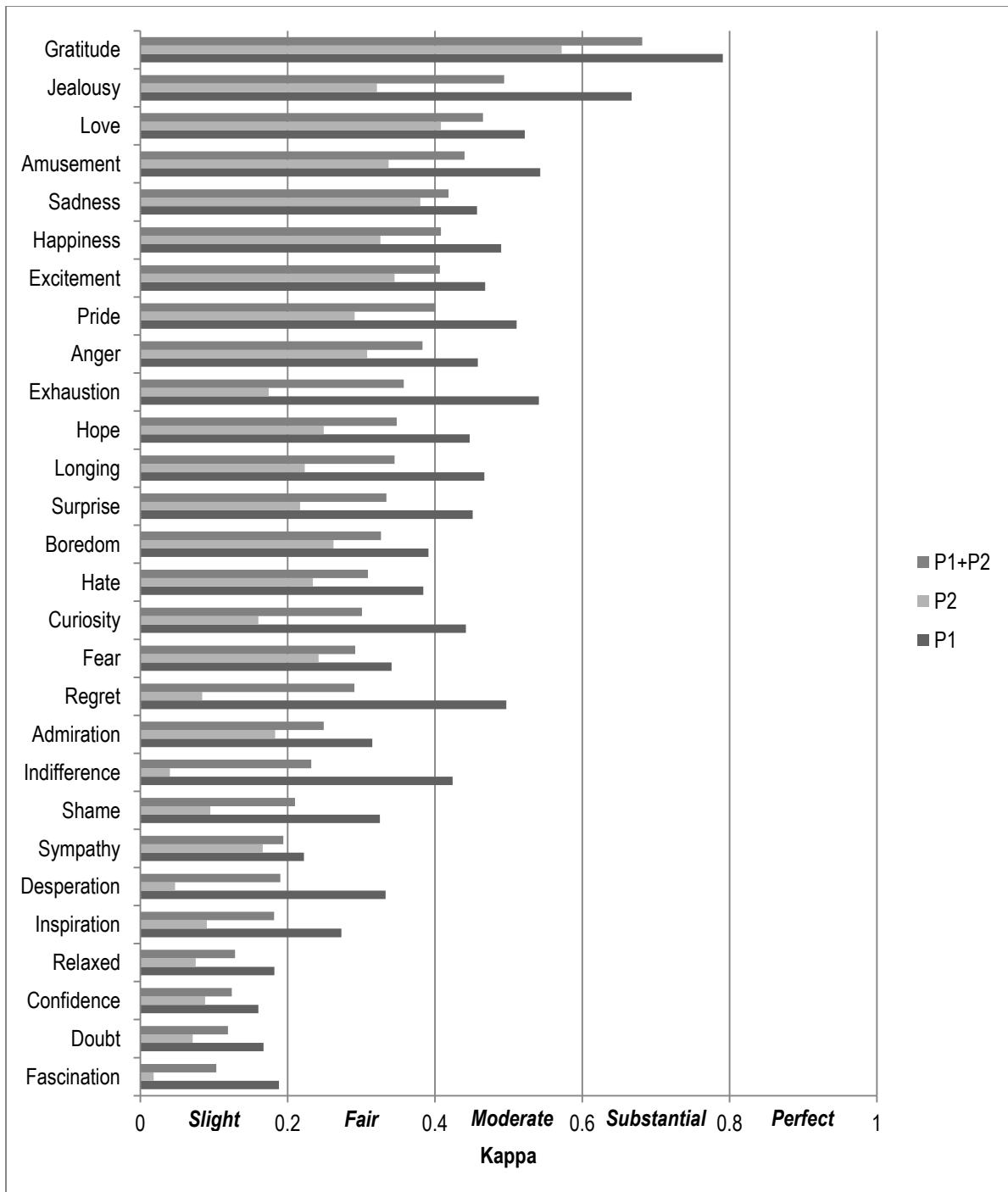


Figure 4.4: Comparing  $\kappa$  for 28 emotion categories across P1, P2 and P1+P2

Generally, achieving an acceptable level of inter-annotator agreement is deemed important to establish the basic validity of the annotation scheme and to ensure that multiple annotators are calibrated to interpret the annotation scheme similarly, thus reducing individual

subjectivity to a minimum. High agreement among annotators is desirable to ensure the reliability of the data (Krippendorff, 2004; Neuendorf, 2002). For example, a general rule of thumb for what is considered to be acceptable inter-annotator reliability in the social sciences is  $\kappa$  of 0.75 or greater (Fleiss, Levin, & Paik, 2013) or  $\alpha$  of 0.8 or greater (Krippendorff, 2004) to indicate substantial agreement. Conclusions that are drawn from variables or classes with  $\kappa$  and  $\alpha$  below these cutoff points are highly tentative and are to be treated with great caution. However, given the highly subjective nature of the task of identifying emotions in text and also the exploratory nature of the study, it is not realistic to apply the strict cutoffs to the emotion categories in this study.

We acknowledge that not all the emotion categories achieved substantial agreement among annotators. Yet, we report the inter-annotator scores for all 28 emotion categories to highlight the range human performance that can be expected given this set of classes. Furthermore, our goal is not to draw direct conclusions from the data. The data is still useful as ground truth to train machine learning models to perform the emotion classification task. Regardless of the amount of noise or level of disagreement among annotators, we can still turn to machine learning models to automatically discover or induce patterns from data that may not be obvious to the human eye.

<b>Class Distinctiveness</b>	<b>Emotion Categories</b>	<b>Count</b>
High	Gratitude, Jealousy, Love, Amusement, Sadness, Happiness, Excitement, Pride	8
Medium	Anger, Exhaustion, Hope, Longing, Surprise, Boredom, Hate, Curiosity, Fear, Regret, Admiration, Indifference, Shame	13
Low	Sympathy, Desperation, Inspiration, Relaxed, Confidence, Doubt, Fascination	7

Table 4.7: Emotion categories with high, medium and low levels of class distinctiveness

To summarize, based on the measure of class distinctiveness, 8 of the 28 emotion categories can be characterized as moderately to highly distinctive, 13 categories are fairly distinctive, and 7 categories are only slightly distinctive as shown in Table 4.7. This grouping is

somewhat arbitrary but represents a reasonable organization of the emotion categories into three levels of class distinctiveness.

## 4.2 Level of Agreement

We use level of agreement, based on a variety of agreement measures, to give more insight on the reliability we can expect from humans on the emotion classification task. We first present the overall inter-annotator agreement for the four facets of emotion included in the annotation scheme. We then discuss the level of agreement achieved at the category level and specifically focus on the proportion of full agreement obtained for each emotion category.

### 4.2.1 Overall Inter-annotator Agreement

On a broad level, we compare the overall inter-annotator reliability statistics across the four facets of emotion (i.e., *valence*, *arousal*, *emotion category* and *emotion cue*). Krippendorff's alpha ( $\alpha$ ) is used as the primary measure of agreement for valence, arousal and emotion category as  $\alpha$  can be applied for any number of annotators as well as for both nominal and ordinal variables. Percent agreement (%) and Fleiss' Kappa ( $\kappa$ ) are also presented alongside  $\alpha$  as a means to compare our results with those reported in related work.

Emotion cue captures the segment of text marked by annotators as the indicator of an emotion category. Unlike valence, arousal and emotion categories, emotion cue does not have a pre-defined set of categories and the boundary of the marked up text is not fixed. The size of an emotion cue varies from a single word to long strings of words within a tweet. As such,  $\alpha$  cannot be computed directly for emotion cue. Krippendorff's  $\alpha$  can only be computed directly for binary, nominal, ordinal, interval and ratio variables (Krippendorff, 2004). We adopt the measure of agreement on set-valued items (MASI) to determine the agreement between sets of text spans among multiple annotators for each tweet (Aman & Szpakowicz, 2007). MASI has been

applied previously to quantify the reliability in co-reference annotation (Passonneau, 2004) and automatic summarization (Passonneau, 2006).

MASI measures the distance between two sets, where a value of 1 indicates the two sets being identical while 0 indicates the two sets being disjoint. For set A and set B, MASI is defined as:

$$MASI = J * M$$

where J stands for the Jaccard metric and M stands for monotonicity.

$$J = \frac{|A \cap B|}{|A \cup B|}$$

$$M = \begin{cases} 1, & \text{if } A = B \\ 2/3, & \text{if } A \subset B \text{ or } B \subset A \\ 1/3, & \text{if } A \cap B \neq \emptyset, A - B \neq \emptyset, \text{ and } B - A \neq \emptyset \\ 0, & \text{if } A \cap B = \emptyset \end{cases}$$

Each emotion cue is split into a string of words and every word is represented as an element in a set. The Jaccard metric is used to weigh the difference between the elements in Set A and Set B. The similarity weight is then adjusted based on the monotonicity of Set A and Set B. Set A shares a monotonic relationship with Set B when Set B contains all of the elements of Set A and vice versa. For example, if Set A contains {of, enthusiasm} and set B contains {lots, of, enthusiasm}, Set A is monotonic with respect to Set B since all the words in A are also in Set B. If Set A contains {good, work} and Set B contains {good, day}, Set A and Set B are non-monotonic sets and would be penalized more than the former. MASI is computed for each annotation pair per tweet.

Table 4.8 presents the inter-annotator agreement statistics for presence or absence of emotion, valence, arousal, 48 emotion categories, 28 emotion categories and emotion cues for all tweets with three annotations. P1-Train shows the agreement scores computed from the training set annotated by all 18 expert annotators. For P1, we report the agreement scores from

before (P1-IND: inductive coding) and after (P1-DED: deductive coding) the emotion categories are defined.

Facet	Emo/Non-Emo			Valence			Arousal	EmoCat-48			EmoCat-28			EmoCues
Sample	%	$\kappa$	$\alpha$	%	$\kappa$	$\alpha$	$\alpha$	%	$\kappa$	$\alpha$	%	$\kappa$	$\alpha$	MASI
P1-Train	81	0.63	0.63	78	0.63	0.63	0.68	62	0.42	0.42	66	0.46	0.46	0.55
<b>P1</b>	<b>81</b>	<b>0.62</b>	<b>0.62</b>	<b>77</b>	<b>0.61</b>	<b>0.61</b>	<b>0.59</b>	<b>62</b>	<b>0.45</b>	<b>0.45</b>	<b>66</b>	<b>0.50</b>	<b>0.50</b>	<b>0.55</b>
<b>P1-IND</b>	<b>80</b>	<b>0.59</b>	<b>0.59</b>	<b>74</b>	<b>0.59</b>	<b>0.59</b>	<b>0.56</b>	<b>58</b>	<b>0.41</b>	<b>0.41</b>	<b>63</b>	<b>0.47</b>	<b>0.47</b>	<b>0.51</b>
AVGUSER	78	0.56	0.56	70	0.56	0.56	0.51	54	0.40	0.39	59	0.45	0.44	0.44
RANDOM	79	0.57	0.57	75	0.57	0.58	0.54	61	0.42	0.42	65	0.46	0.46	0.41
SEUSER	80	0.61	0.61	78	0.61	0.61	0.65	64	0.46	0.46	69	0.52	0.52	0.63
TOPIC	81	0.62	0.62	74	0.61	0.61	0.55	52	0.38	0.38	59	0.46	0.46	0.55
<b>P1-DED</b>	<b>83</b>	<b>0.65</b>	<b>0.65</b>	<b>79</b>	<b>0.64</b>	<b>0.64</b>	<b>0.63</b>	<b>66</b>	<b>0.48</b>	<b>0.48</b>	<b>70</b>	<b>0.53</b>	<b>0.53</b>	<b>0.59</b>
AVGUSER	84	0.67	0.68	78	0.66	0.66	0.67	63	0.48	0.48	68	0.54	0.54	0.48
RANDOM	82	0.63	0.63	79	0.64	0.64	0.63	63	0.46	0.45	67	0.51	0.51	0.63
SEUSER	83	0.59	0.59	82	0.59	0.59	0.56	74	0.46	0.46	76	0.51	0.51	0.55
TOPIC	85	0.69	0.69	79	0.67	0.67	0.65	64	0.52	0.52	68	0.56	0.56	0.69
<b>P2</b>	<b>66</b>	<b>0.29</b>	<b>0.29</b>	<b>60</b>	<b>0.34</b>	<b>0.34</b>	<b>0.32</b>				<b>51</b>	<b>0.28</b>	<b>0.28</b>	<b>0.48</b>
AVGUSER	63	0.26	0.26	57	0.34	0.34	0.30				46	0.27	0.27	0.49
RANDOM	64	0.26	0.26	58	0.29	0.30	0.29				50	0.24	0.24	0.46
SEUSER	70	0.35	0.35	67	0.37	0.37	0.39				59	0.30	0.30	0.50
TOPIC	65	0.31	0.31	59	0.36	0.36	0.32				49	0.31	0.31	0.47
<b>P1+P2</b>	<b>76</b>	<b>0.51</b>	<b>0.51</b>	<b>71</b>	<b>0.52</b>	<b>0.52</b>	<b>0.50</b>				<b>61</b>	<b>0.43</b>	<b>0.43</b>	<b>0.52</b>

Table 4.8: Inter-annotator agreement statistics for emotion/non-emotion, valence, arousal, emotion category, and emotion cue

Generally, we observe higher  $\alpha$  for all facets of emotion in P1-DED compared with P1-IND especially for the two sets of emotions categories. The increase in  $\alpha$  from P1-IND to P1-DED is expected when the emotion becomes more well-defined and as annotators receive more training.

It is interesting to note that agreements are not appallingly low when expert annotators are asked to suggest their own emotion tags to describe the emotion in the tweets. In fact, expert annotators scored higher  $\alpha$  for the first round of inductive coding ( $\alpha = 0.39$ ) (see Round 1 in Table 4.9) compared with novice annotators who were asked to make a selection from a pre-defined set of emotion categories ( $\alpha = 0.28$ ).

Mean  $\alpha$  for 28 emotion categories across all rounds in P1-DED is 0.53. Table 4.9 presents the percent agreement and  $\alpha$  for the first three rounds in P1-IND (i.e., 1, 2 and 3) and the last two rounds in P1-DED (i.e., 4 and 5). With continuous training provided to annotators on a weekly basis, agreement continues to climb until  $\alpha$  reaches 0.61 and percent agreement exceeds 70%. Although emotion annotation is a subjective and difficult task, it is possible to generate reliable data when annotators are given sufficient training. With limited training,  $\alpha$  scores in P2 decrease almost by half for all facets of emotion as shown in Table 4.8.

Measure	% Agreement					$\alpha$				
Round	1	2	3	4	5	1	2	3	4	5
AVGUSER	45%	56%	59%	69%	72%	0.31	0.41	0.45	0.56	0.58
SEUSER	65%	71%	78%	81%	77%	0.47	0.41	0.56	0.58	0.63
RANDOM	68%	66%	65%	69%	72%	0.38	0.46	0.49	0.52	0.60
TOPIC	54%	64%	61%	67%	74%	0.39	0.50	0.51	0.55	0.64
Mean	58%	64%	66%	72%	74%	0.39	0.45	0.50	0.55	0.61

Table 4.9: Agreement per round for 28 emotion categories among expert annotators

Since  $\alpha$  and  $\kappa$  are affected by dissimilar scales and the number of categories, care must be taken when making comparisons across different facets of emotion. Typically, a larger number of categories would lead to more disagreements, and thus lower  $\alpha$  (Sim & Wright, 2005). Our P1 results are consistent with this general observation. However, we observe an anomaly in that valence annotation (4 classes: “positive”, “negative”, “neutral” and “no emotion”) in P2 obtains slightly higher  $\alpha$  compared with the binary emotion versus non-emotion annotation (2 classes: “has emotion” and “no emotion”). Essentially, valence is a more fine-grained scale that breaks down the “has emotion” class from the emotion/non-emotion scale into three sub-categories: positive, negative and neutral. This led us to conclude there are high enough agreements among annotators in making the distinction between positive, negative and neutral instances to offset some of the disagreements in the binary emotion versus non-emotion annotation.



It is important to note that agreement based on 28 emotion categories is not a great deal lower than that observed for other more coarse-grained facets of emotion. Annotators across P1 and P2 achieve overall 61% agreement when asked to identify 28 emotion categories, which is not a drastic drop compared with 71% agreement obtained from the four-class valence annotation. MASI scores for the emotion cues are more stable across P1 (MASI = 0.55) and P2 (MASI = 0.48), thus showing that there is less discord among expert and novice annotators when asked to identify written linguistic cues associated with emotion.

Table 4.10 summarizes the agreement statistics reported in related work on emotion annotation. The purpose is not to draw direct comparisons with our results given that the annotation context in previous studies may differ from ours. No benchmark on inter-annotator reliability has been agreed upon given the difficulty of the task. Therefore, Table 4.10 serves as a point of reference to help us interpret our results in the context of what is considered to be the state-of-the-art. Note that the inter-annotator agreement for emotion category ranges from low to moderate. Generally, the larger the number of classes, the lower the inter-annotator reliability scores.

Our results for different facets of emotion are comparable to the agreement statistics reported in existing literature even though we use a larger number of emotion categories (28). Half of the 20 emotion categories used in Sintsova, Musat, & Pu Faltings (2013) coincide with our 28 emotion categories but their study obtained slightly lower  $\kappa$  among annotators recruited from AMT. Our study employs the greatest number of emotion categories but we managed to achieve inter-annotator reliability scores (% Agreement = 61%,  $\kappa$  = 0.43 and  $\alpha$  = 0.43) that are comparable to other studies with far fewer emotion categories. With sufficient training, it is possible for fine-grained emotion annotation with 28 emotion categories to achieve inter-annotator agreement on par with emotion annotation using only a small set of basic emotion categories.

Related Work	Concept	Context	Average Agreement Score(s)
Mohammad, Zhu, & Martin (2014)	Emotion	Tweet	% Agreement (category-8) = 59.59%
Sintsova et al., (2013)	Emotion	Tweet	% Agreement (polarity) = 75.7 – 78.5% % Agreement (category-20) = 29.3 – 38.5 % Kappa (category-20) = 0.24
Roberts, Roach, Johnson, Guthrie, & Harabagiu (2012)	Emotion	Tweet	Kappa (category-7) = 0.67
Alm, Roth, & Sproat (2005)	Emotion	Fairy tale	% Agreement (category-8) = 45 – 64% Kappa (category-8) = 0.24 – 0.51
Aman & Szpakowicz (2007)	Emotion	Blog	Kappa (category-7) = 0.43 – 0.79 Kappa (intensity) = 0.37 – 0.72 Kappa (indicator) = 0.66 MASI (indicator) = 0.61
Strapparava & Mihalcea (2007)	Emotion	News headline	Pearson correlation (category-6) = 36.07 – 68.19 Pearson correlation (valence) = 78.01
Gupta, Gilbert, & Di Fabbri (2010)	Emotion	Email	Kappa (salient features) = 0.814
Rubin, Stanton, & Liddy (2004)	Emotion	Customer review	Agreement (octant-8) = 70.7% SD = 21.5%
Pestian et al. (2012)	Emotion	Suicide note	Krippendorff's alpha (category-16) = 0.546
Brooks et al. (2013)	Affect	Chat	Modified Kappa (category-13) = 0.49 – 0.81
Neviarouskaya, Prendinger, & Ishizuka (2007)	Affect	Text message	Kappa (emoticons) = 0.94 Kappa (abbreviations) = 0.93
Wiebe, Wilson, & Cardie (2005)	Private states	News	agr (indicator) = 0.72 – 0.82

Table 4.10: Inter-annotator reliability statistics from related work on emotion annotation

We acknowledge that overall inter-annotator agreement in detecting the 28 emotion categories is at best fair to good (kappa between 0.40 – 0.75) according to the guidelines described in Fleiss et al. (2013). The overall inter-annotator agreement scores could be increased by removing some of the emotion categories with poor agreement or retraining annotators until a kappa of above 0.75 is achieved for all facets of emotion. However, the use of inter-annotator agreement here is intended to develop a realistic assessment of human performance in annotating the emotion categories that emerged from the inductive coding task.

### 4.2.2 Category Level Agreement

Next, we examine how often three annotators agree with one another on the set of 28 emotion categories. Table 4.11 shows the proportion of tweets with full, partial and no agreement on emotion category among 3 annotators in the corpus. Slightly over half of the tweets annotated by expert annotators (P1) have full agreement while only one third from novice annotators (P2) shows full agreement. On the other hand, a greater portion of tweets with partial agreement (i.e., two out of three annotator labels are the same) is observed in P2 (45%) compared with P1 (36%). Tweets with full disagreement (i.e., all three annotator labels are different) make up less than 20% of the corpus, also with a higher portion coming from P2. Indeed, judgements from annotators with limited training come with a higher level of full disagreement.

Agreement Level	P1: n	P1: %	P2: n	P2: %	P1+P2: n	P1+P2: %
Full Agreement	2886	52%	2704	33%	5590	41%
- <i>Emotion</i>	797	28%	604	22%	1401	25%
- <i>Non-emotion</i>	2089	72%	2100	78%	4189	75%
Partial Agreement	2011	36%	3709	45%	5720	42%
Full Disagreement	656	12%	1770	22%	2426	18%
<b>Total</b>	<b>5553</b>		<b>8183</b>		<b>13736</b>	

Table 4.11: Proportion of full, partial and no agreement for 28 emotion categories among three annotators

Of particular interest are the emotion categories for which full agreement is often observed. Tweets containing emotion only make up a quarter of tweets with full agreement, but a high frequency of full agreement for an emotion category suggests that the category is more reliably recognized by humans. The proportion of annotator labels with full agreement for each emotion category is presented in Table 4.12. Again, we see the usual suspects at the top of the list: gratitude, pride, excitement, jealousy and happiness although their ranks slightly differ from the top five emotion categories based on class distinctiveness. The bottom of the list reflecting

emotion categories with the lowest proportion of full agreement is also consistent with the emotion categories with lowest  $\kappa$  scores in Table 4.6.

Category	n: Cat	n: FA	% FA/Cat	Category	n: Cat	n: FA	% FA/Cat
Gratitude	521	247	47%	Boredom	48	6	13%
Pride	213	65	31%	Fear	239	25	10%
Excitement	686	167	24%	Exhaustion	49	5	10%
Happiness	1787	387	22%	Indifference	68	6	9%
Jealousy	34	7	21%	Admiration	403	35	9%
Sadness	521	107	21%	Regret	153	13	8%
Amusement	660	116	18%	Inspiration	75	5	7%
Love	681	109	16%	Shame	90	6	7%
Hope	522	79	15%	Sympathy	101	6	6%
Curiosity	93	14	15%	Relaxed	77	3	4%
Anger	1201	174	14%	Confidence	110	2	2%
Hate	192	26	14%	Desperation	58	1	2%
Longing	121	16	13%	Doubt	158	2	1%
Surprise	266	35	13%	Fascination	204	2	1%

Table 4.12: Proportion of full agreement (FA) for 28 emotion categories (Cat)

Based on a manual review of a sample of tweets, the tweets that contain obvious or explicit emotion interjections (Example 4.3), words (Example 4.4.) or phrases (Example 4.5) are more likely to obtain full agreement among multiple annotators.

**Example 4.3:** @CorrinCampbell haha!!!! Yes let's get together and play some intense dress up! **[Amusement]**

**Example 4.4:** I seriously love my coworkers #evenontaxfreeweekend **[Love]**

**Example 4.5:** I'm going to punch someone in the face **[Anger]**

Table 4.13 shows the 28 emotion categories grouped into three levels based on the proportion of full agreement per category. We conclude that 8 emotion categories have high level of full agreement, 12 have moderate level of full agreement and the remaining 8 are considered to have low level of full agreement.

Level of Full Agreement	Emotion Categories	Count
High	Gratitude, Pride, Excitement, Happiness, Jealousy, Sadness, Amusement, Love	8
Medium	Hope, Curiosity, Anger, Hate, Longing, Surprise, Boredom, Fear, Exhaustion, Indifference, Admiration, Regret	12
Low	Inspiration, Shame, Sympathy, Relaxed, Confidence, Desperation, Doubt, Fascination	8

Table 4.13: Emotion categories with high, medium and low levels of full agreement

### 4.3 Class Intuitiveness

Annotator labels capture spontaneous judgments while the gold labels represent standards that have been established by experts. All tweets in the corpus are assigned gold labels, which act as ground truth for the machine learning experiments. Comparing spontaneous human judgments to ground truth can be used to gauge how intuitive the emotion categories are or their face validity (i.e., the extent to which a category seems to capture the desired emotion). Spontaneous human judgments are annotations collected in the first pass from annotators. Class intuitiveness assesses how likely each emotion category is applied correctly by the annotators.

Annotator Label	Gold Label					
	P1		P2		P1 + P2	
	Match	Deviation	Match	Deviation	Match	Deviation
Full Agreement	3058 (96%)	127 (4%)	2688 (83%)	556 (17%)	5746 (89%)	693 (11%)
Partial Agreement	2359 (94%)	158 (6%)	3840 (87%)	554 (13%)	6199 (90%)	712 (10%)
- Majority	1703 (68%)		2979 (67%)		4682 (68%)	
- Minority	656 (26%)		861 (20%)		1517 (22%)	
Full Disagreement	453 (91%)	47 (9%)	1290 (79%)	337 (21%)	1743 (82%)	384 (18%)
All	5870 (95%)	332 (5%)	7818 (84%)	1475 (16%)	13688 (88%)	1789 (12%)

Table 4.14: Percent matches and deviation between annotator labels and gold labels

Table 4.14 presents the matches and deviations between annotator labels and gold labels based on tweets with full, partial and no agreement among three annotators. A match means that the gold label is the same as at least one of the annotator labels for a particular

tweet. On the other hand, a deviation means that the gold label matches none of the annotator labels.

For P1, all disagreements were first resolved through discussion with expert annotators. Essentially, expert annotators achieved 100% agreement in P1. We then reviewed all the tweets in P1 to ensure consistency of the annotations between different groups of expert annotators assigned to work on distinct samples. The deviation between the annotator label and gold label for P1 shown in Table 4.14 is obtained by comparing the initial judgments provided by annotators before discussion to the gold labels all expert annotators agreed upon after discussion. The 4% deviation between expert annotator labels with full agreement and the gold labels in P1 is caused by systematic adjustments to the data. First, this systematic adjustment can be triggered by heuristics resulting from the disagreement discussion. Second, recall that the emotion categories might be formed at different annotation rounds in the open coding task. To ensure that emotion categories formed at later annotation rounds were also reflected on the data from earlier rounds, we had to perform a systematic review on all annotated data every time a new category came to light. As a result, some tweets that annotators all agreed contained no emotion in earlier annotation rounds were subsequently changed.

For P2, we assigned the gold labels after manually reviewing all 10,000 annotations provided by AMT workers. The manual review procedure was necessary to reduce as much as possible the noise from a large group of novice annotators. The deviation between annotator labels with full agreement and gold labels is higher (17%) in P2 than P1. The main reason for this is that AMT workers have a tendency to miss multiple emotions being expressed in a tweet. In such cases, the primary annotator label with full agreement matches one of gold labels for a tweet. The other gold labels that all the annotators failed to recognize (i.e., full agreement) are counted as non-matches.

In P1, 95% of the gold labels match at least one of the three labels provided by the expert annotators. Only 5% of the gold labels did not originate from any one the annotator

labels. The percentage of deviation (16%) increases by three folds in P2, reflecting that there is a higher likelihood for novice annotators to select a label other than the gold label. Nonetheless, over 80% of the gold labels match at least one of the annotator labels. This shows that novice annotators can do a fairly decent job in identifying the gold labels with 28 emotion categories.

For tweets with partial agreement, not all the gold labels originate from the majority annotator labels. Only 68% of the annotator and gold labels come from the majority (i.e., two of the three matching annotator labels), 22% come from the minority (i.e., the single non-matching annotator label), and 10% come from none of the annotators. We also observe that a high percentage (82%) of gold labels from the set of tweets with no agreement at all among three annotators matches at least one of the annotator labels. With 88% of the overall gold labels matching at least one of the annotator labels, we can conclude that at least one out of three annotators is able to accurately recognize the emotion expressed in the tweet in most cases.

Category	% Agreement (Annotator Label x Gold Label)	Category	% Agreement (Annotator Label x Gold Label)
Gratitude	76%	Jealousy	42%
Pride	62%	Surprise	42%
Excitement	61%	Sympathy	41%
Exhaustion	55%	Inspiration	39%
Happiness	54%	Regret	37%
Amusement	54%	Admiration	36%
Curiosity	53%	Indifference	36%
Sadness	53%	Fear	35%
Anger	49%	Shame	32%
Boredom	48%	Doubt	28%
Hate	48%	Confidence	26%
Love	47%	Relaxed	26%
Hope	45%	Desperation	25%
Longing	44%	Fascination	15%

Table 4.15: Mean pairwise agreement between annotator and gold labels per emotion category

To determine how intuitive each emotion category is, we present in Table 4.15 the mean percent agreement between each stream of annotator labels and the gold labels per emotion category (i.e., <A1, Gold>, <A2, Gold> and <A3, Gold>). The higher the mean pairwise percent

agreement between the annotator labels and gold labels, the more intuitive the category and thus, the more likely the category can be recognized with greater spontaneity. Once again, *gratitude*, *pride*, *excitement*, *happiness*, *amusement* and *sadness* appear at the top of the list but are joined by *exhaustion* and *curiosity*. The least intuitive emotion categories are *confidence*, *relaxed*, *desperation* and *fascination*. Based on the mean pairwise percent agreement scores between the annotator-gold pairs, we group the emotion categories into three levels of intuitiveness as shown in Table 4.16.

Class Intuitiveness	Emotion Categories	Count
High	Gratitude, Pride, Excitement, Exhaustion, Happiness, Amusement, Curiosity, Sadness	8
Medium	Anger, Boredom, Hate, Love, Hope, Longing, Jealousy, Surprise, Sympathy, Inspiration, Regret, Admiration, Indifference, Fear, Shame	15
Low	Doubt, Confidence, Relaxed, Desperation, Fascination	5

Table 4.16: Emotion categories with high, medium and low levels of intuitiveness

## 4.4 Summary: Human Recognition of Emotion Categories in Tweets

We derive from tweets a set of 28 fine-grained emotion categories that humans can detect in microblog text. Three measures (i.e., class distinctiveness, level of full agreement and class intuitiveness) are used to characterize human performance in recognizing the 28 emotion categories. Results from all three measures, which offer different perspective on the emotion categories, are summarized in Table 4.17.

We can conclude that annotators perform the best at detecting 6 emotion categories in microblog text. The 6 emotion categories in no particular order are *amusement*, *excitement*, *gratitude*, *happiness*, *pride* and *sadness*. These emotion categories have very distinctive linguistic patterns that make them easy to recognize in text. With the exception of happiness and sadness, our top 6 emotion categories differ from what Ekman considers to be basic emotions (i.e., *happiness*, *sadness*, *anger*, *disgust*, *fear* and *surprise*), which suggests that easily recognized emotions in tweets are not necessarily the ones that are associated with



fundamental life tasks (e.g., facing an immediate danger). Many of these emotions are expressed for social purposes (e.g., expressing gratitude towards a good deed from another person).

Category	Distinctiveness	Full Agreement	Intuitiveness
Amusement	H	H	H
Excitement	H	H	H
Gratitude	H	H	H
Happiness	H	H	H
Pride	H	H	H
Sadness	H	H	H
Jealousy	H	H	M
Love	H	H	M
Anger	M	M	M
Curiosity	M	M	H
Exhaustion	M	M	H
Admiration	M	M	M
Boredom	M	M	M
Fear	M	M	M
Hate	M	M	M
Hope	M	M	M
Indifference	M	M	M
Longing	M	M	M
Surprise	M	M	M
Regret	M	M	M
Shame	M	L	M
Inspiration	L	L	M
Sympathy	L	L	M
Confidence	L	L	L
Desperation	L	L	L
Doubt	L	L	L
Fascination	L	L	L
Relaxed	L	L	L

Table 4.17: Triangulation of measures to determine the emotion categories humans can detect in microblog text

Overall, annotators perform moderately well in recognizing a majority of the emotion categories. The only exception is the 5 emotion categories with very low scores on all three

measures: *confidence*, *desperation*, *doubt*, *fascination* and *relaxed*. These low performing emotion categories may not be as well-defined and intuitive as the others. Due to their infrequent occurrences in the corpus, it is also possible that annotators pay less attention to these emotion categories or find it difficult to learn to better recognize them even over time. Nonetheless, we kept all 28 emotion categories for the machine learning experiments so we could examine the performance of automatic classification on emotion categories with varying degrees of agreement.

## 4.5 EmoTweet-28: Corpus Characteristics

The final corpus, EmoTweet-28 contains 15,553 tweets from P1 and P2. Overall, the corpus is composed of 247,872 words, in which 42,620 are unique terms. Message length is short with 16 words on average per tweet. The shortest tweet contains only one word while the longest tweet contains 40 words. The word composition of the four samples (i.e., RANDOM, TOPIC, SEN-USER and AVG-USER) in the corpus is shown in Table 4.18.

Sample	Word Count	Average Word Count/Tweet
RANDOM	64793	16
TOPIC	67717	18
SEN-USER	68255	17
AVG-USER	47107	12
<b>Total</b>	<b>247872</b>	<b>16</b>

Table 4.18: Word composition of the four samples in the corpus

### 4.5.1 Emotion Distributions

This section describes the distribution of gold labels among the facets of emotion. As shown in Table 4.19, the overall distribution between tweets containing emotion and those that do not is roughly balanced; slightly over half of the tweets (51%) contain emotion. The ratios between emotional and non-emotional tweets respectively for RANDOM, TOPIC, SEN-USER and AVG-USER are similar. The biggest contribution of emotional tweets comes from TOPIC,

and the lowest from SENUSER. The number of emotional tweets exceeds the number of non-emotional tweets in TOPIC and AVG-USER but the reverse is observed for RANDOM and SEN-USER.

Class	P1	P2	P1+P2	RANDOM	TOPIC	SEN-USER	AVG-USER
Emotion	2916 (53%)	4953 (50%)	7869 (51%)	1775 (45%)	2281 (60%)	1615 (40%)	2198 (58%)
Non-Emotion	2637 (47%)	5047 (50%)	7684 (49%)	2175 (55%)	1529 (40%)	2378 (60%)	1602 (42%)
<b>Total</b>	<b>5553</b>	<b>10000</b>	<b>15553</b>	<b>3950</b>	<b>3810</b>	<b>3993</b>	<b>3800</b>

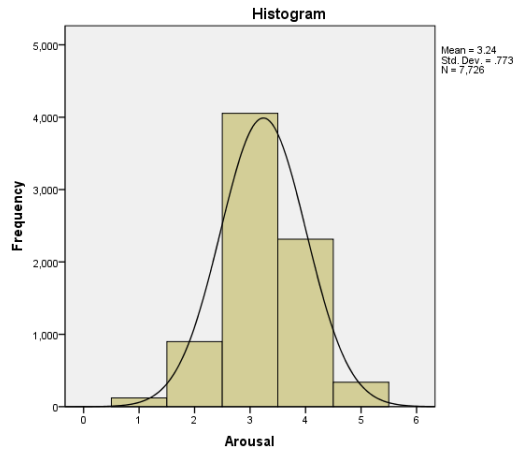
Table 4.19: Distribution of emotional and non-emotional tweets

Table 4.20 summarizes results for emotion valence. The overall corpus contains more than twice as many positive tweets than negative. This skew is especially apparent for SEN-USER with three quarters of the tweets annotated as positive and barely any as neutral. RANDOM, TOPIC and AVG-USER samples are similar in the proportion of positive, negative, and neutral tweets and are likely to be more representative samples of the true distribution on Twitter. About 7% of the corpus consists of tweets assigned with multiple valence labels (e.g., presence of positive and negative emotions in the same tweet).

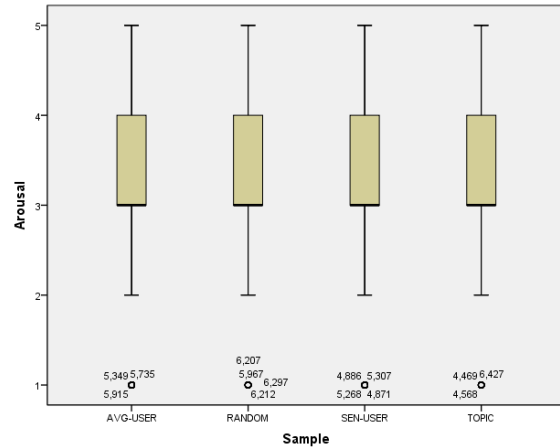
Class	P1	P2	P1+P2	RANDOM	TOPIC	SEN-USER	AVG-USER
Positive	1840 (63%)	2846 (57%)	4686 (60%)	1022 (58%)	1306 (57%)	1259 (78%)	1099 (50%)
Negative	744 (26%)	1493 (30%)	2237 (28%)	538 (30%)	689 (30%)	276 (17%)	734 (33%)
Neutral	155 (5%)	222 (4%)	377 (5%)	87 (5%)	107 (5%)	24 (1%)	159 (7%)
Multiple Valence	177 (6%)	392 (8%)	569 (7%)	128 (7%)	179 (8%)	56 (3%)	206 (9%)
<b>Total</b>	<b>2916</b>	<b>4953</b>	<b>7869</b>	<b>1775</b>	<b>2281</b>	<b>1615</b>	<b>2198</b>

Table 4.20: Distribution of tweets based on emotion valence

Each tweet containing emotion is assigned a final arousal score, which is computed based on the mean arousal ratings provided by all the annotators. The data follows a roughly normal distribution with a slight skew to the right. Mean arousal is 3.24 (see Figure 4.5a). The spread of data across the four samples is similar as shown in Figure 4.5b).



(a)



(b)

Figure 4.5: Distribution of tweets based on mean arousal ratings

Table 4.21 summarizes the frequency distribution of emotion categories. Tweets that are assigned with multiple emotion categories are counted more than one time. As expected, the frequency of emotion classes becomes even more unbalanced and sparse with a greater number of classes compared to valence. Of the 28 emotion categories, the full corpus contains the highest instances of *happiness* (12%) and the lowest instances of *jealousy* (0.2%). All four samples share one similarity: *happiness* occurs the most frequently in each sample. Other than that, the proportion of emotion categories differs across the four samples. For example, political leaders (SEN-USER) express more gratitude and much less anger on Twitter than a typical user (AVG-USER) indicating that leaders take a more controlled and strategic approach when expressing their emotions on Twitter. RANDOM, TOPIC and AVG-USER contribute at least a few positive instances of each emotion category. Three emotion categories are notably absent from SEN-USER: *boredom*, *indifference* and *jealousy* (see Appendix C).

Category	P1		P2		P1 + P2	
	Frequency	% Occurrence	Frequency	% Occurrence	Frequency	% Occurrence
Admiration	158	2.8	245	2.5	403	2.6
Amusement	237	4.3	423	4.2	660	4.2
Anger	444	8.0	757	7.6	1201	7.7
Boredom	12	0.2	36	0.4	48	0.3
Confidence	19	0.3	91	0.9	110	0.7
Curiosity	30	0.5	63	0.6	93	0.6
Desperation	8	0.1	50	0.5	58	0.4
Doubt	50	0.9	108	1.1	158	1.0
Excitement	265	4.8	421	4.2	686	4.4
Exhaustion	10	0.2	39	0.4	49	0.3
Fascination	54	1.0	150	1.5	204	1.3
Fear	77	1.4	162	1.6	239	1.5
Gratitude	221	4.0	300	3.0	521	3.3
Happiness	778	14.0	1009	10.1	1787	11.5
Hate	63	1.1	129	1.3	192	1.2
Hope	187	3.4	335	3.4	522	3.4
Indifference	28	0.5	40	0.4	68	0.4
Inspiration	21	0.4	54	0.5	75	0.5
Jealousy	5	0.1	29	0.3	34	0.2
Longing	41	0.7	80	0.8	121	0.8
Love	234	4.2	447	4.5	681	4.4
Pride	85	1.5	128	1.3	213	1.4
Regret	49	0.9	104	1.0	153	1.0
Relaxed	26	0.5	51	0.5	77	0.5
Sadness	158	2.8	363	3.6	521	3.3
Shame	26	0.5	64	0.6	90	0.6
Surprise	93	1.7	173	1.7	266	1.7
Sympathy	35	0.6	66	0.7	101	0.6

Table 4.21: Frequency distribution of all emotion categories in the corpus

#### 4.5.2 Multiple Emotions in a Tweet

Although tweets are short and contain only 140 characters at maximum, we also captured tweets tagged with multiple emotion categories during the annotation process. People can be very expressive in conveying their emotions on Twitter even in such a short span of text. Such tweets have usually been excluded from existing gold standard corpora (Hasan, Rundensteiner, et al., 2014; Mohammad et al., 2014) to reduce complexity. In fine-grained emotion analysis, multiple emotions occur naturally so a corpus should represent the occurrences of such cases and not ignore them because it is easier. If the portion of tweets

containing multiple emotion categories is high, including them in the corpus would help increase the number of positive examples for each emotion category.

Tweets that contain multiple emotion can be characterized in two ways: 1) expression of multiple emotions with the same valence being labeled as *multiple emotions* (see Example 4.6), and 2) expressing multiple emotions with distinct valence being labeled as *multiple valence* (see Example 4.7). For instance, the tweeter in Example 4.6 expressed three positive emotions in a single tweet: *gratitude* (*thank you so much*), *love* (*As a fan of the series*), and *excitement* (*i'm really looking forward to*). In Example 4.7, the tweeter expressed both a positive emotion, *happiness* (*Yay freedom!*) and a negative emotion, *anger* (*Fffffff*) in the same tweet.

**Example 4.6:** @yenpress thank you so much for licensing kagerou project!!! As a fan of the series i'm really really looking forward to the release!!!!

**[Multiple Emotions Same Valence: Gratitude, Love, Excitement]**

**Example 4.7:** Yay freedom! \*looks at traffic map\* Fffffff-

**[Multiple Emotions Different Valence: Happiness, Anger]**

Category Count/Tweet	P1	P2	P1+P2	RANDOM	TOPIC	SEN-USER	AVG-USER
Single	5102 (92%)	9135 (91%)	14237 (92%)	3652 (92%)	3398 (89%)	3736 (94%)	3451 (91%)
Multiple	451 (8%)	865 (9%)	1316 (8%)	298 (8%)	412 (11%)	257 (6%)	349 (9%)
- Multiple: Same Valence	274 (5%)	467 (5%)	741 (5%)	165 (4%)	232 (6%)	201 (5%)	143 (4%)
- Multiple: Different Valence	177 (3%)	398 (4%)	575 (3%)	133 (4%)	180 (5%)	56 (1%)	206 (5%)
Total	5553	10000	15553	3950	3810	3993	3800

Table 4.22: Distribution of tweets containing single and multiple emotion categories

As shown in Table 4.22, the corpus contains a significant portion of tweets tagged with a single emotion category (92%) and only 8% of tweets tagged with more than one emotion category. Mohammad et al. (2014) reported 2% of their 2012 US presidential elections corpus comprises of tweets with two or more contrasting emotions. Our findings are consistent with previous observation although the proportion of tweets with multiple emotion categories is

higher in our corpus (8%). The emotion categories in our annotation scheme are more fine-grained which naturally lead to more tweets being tagged with multiple emotion categories.

Although tweets containing multiple emotions represent only 8% of the corpus, including such tweets in the corpus leads to over 40% overall increase in the number of positive examples (i.e., instances of an emotion category). Tweets annotated with only a single emotion produces only 6553 positive examples. The inclusion of tweets annotated with multiple emotions increases the number of positive examples to 9331. This is especially beneficial for categories that suffer from sparseness of positive examples such as *jealousy*, *boredom* and *exhaustion*. Overall, including tweets containing multiple emotions gives each emotion category a boost in frequency, notably for *happiness* and *love*.

### 4.5.3 Emotion Expressions and Descriptions in Tweets

In this study we cast a broad net to capture as many textual emotion signals as possible. While a majority of the tweets contain expressions of the tweeter's own emotional experience (self-reference) as seen in Example 4.8, there are two other notable forms in which emotions are expressed in tweets: 1) description of emotion attributed to other individuals or entities, and 2) description of an emotion-related phenomenon. Since the goal of the research is to study the full range of emotion expressions in tweets, we did not limit annotators to identify only a single form of emotion expression.

It is common for tweeters to talk about the emotional experiences of other individuals. For self-expression of emotions tweets are written in first person. Tweets containing emotion description of others, however, are usually written in third person as illustrated in Examples 4.9 and 4.10. In Example 4.9, the tweeter is not expressing his or her emotion but is describing the *fear* experienced by the nephew. Similarly, the tweeter is describing Mark's emotion of *gratitude* in Example 4.10.

**Example 4.8:** Excited for @LSUfbal v #Alabama. Saturday night in Death Valley will be the loudest place in the country. #LSURoar #Beatbama #LSU [**Self: Excitement**]

**Example 4.9:** Awake at 5am because my nephew isn't used to sleeping by himself so he runs crying into my room as if someone died. [**Other: Fear**]

**Example 4.10:** It was moments like this that Mark appreciated his kids. Grateful they were there to untie him after burglars ransacked the house. [**Other: Gratitude**]

In the second form of emotion description, tweeters use strong emotion words or indicators to describe an emotion-related phenomenon. The presence of such emotion indicators neither describes the tweeter's own emotion nor someone else's emotion. Some of the patterns we have observed from the tweets in the corpus include stating an attempt to make someone feel certain emotion (Example 4.11), a general description of how someone feels in the onset of an emotion (Example 4.12), and describing how one is expected to feel in a particular situation (Example 4.13).

**Example 4.11:** tonight, i should learn how to rap like nicki minaj so that i can irritate the hell out of Saifullah tmr. [**Description: Anger**]

**Example 4.12:** That awkward moment when your stalking someones instagram and you like a pic from 8238.3 weeks ago [**Description: Shame**]

**Example 4.13:** @DanielleCasting That's a story for another day. Today if I were you I'd be celebrating what another amazing Workshop I had. => [**Description: Happiness**]

These three examples show that the presence of emotion words in a tweet is not always indicative of the tweeter's own emotion. Emotions words can be used in many ways to describe emotional phenomena in everyday communication. In applications where the goal is to detect how a person is reacting emotionally to a particular stimulus, automatic emotion detectors that recognize only emotion words will also capture descriptions of emotion-related phenomena as positive examples of an emotion category, and thus will yield more false positives (i.e., instances incorrectly identified as being an example of a category).



## 4.6 Emotion Cues

Emotion cues consist of all text in a tweet identified as indicators of an emotion. In Example 4.14, the emotion cues are “*thank you*” and “*i really appreciate it*”. We use the term *cue segment* to refer to each marked word sequence in a tweet that is associated with an emotion. Using the same example, the tweet contains two cue segments. The first cue segment is “*thank you*” and the second one is “*i really appreciate it*”.

**Example 4.14:** Thank you to @Mark\_Sanchez and @nickmangold for signing my jets football today i really appreciate it you guys are class acts! **[Gratitude]**

### 4.6.1 Emotion Cue Characteristics

Table 4.23 an overview of the characteristics of emotion cues identified by annotators (annotator cues) as well as the emotion cues finalized as ground truth (gold cues). The mean length of both annotator and gold cues is 3 words although maximum length for annotator cues is longer. The longest annotator cue almost encompasses the entire tweet. Cue segments are shorter with a mean length of 2 words. Cue segments consisting of between one and three words make up a large portion of annotator cues and gold cues as shown in Figure 4.6. We can thus infer that the unit of meaning for emotions in text is not limited to only single words and can be substantially captured within a window size of three-word sequence. Cues with more than 10 words form a long tail of rare occurrences in both annotator cues and gold cues.

Word Count	Cue: Annotator	Cue: Gold	Segment: Gold
Mean	3	3	2
Minimum	1	1	1
Maximum	37	21	17
Total	33486	5917	14059

Table 4.23: Emotion cue and segment statistics

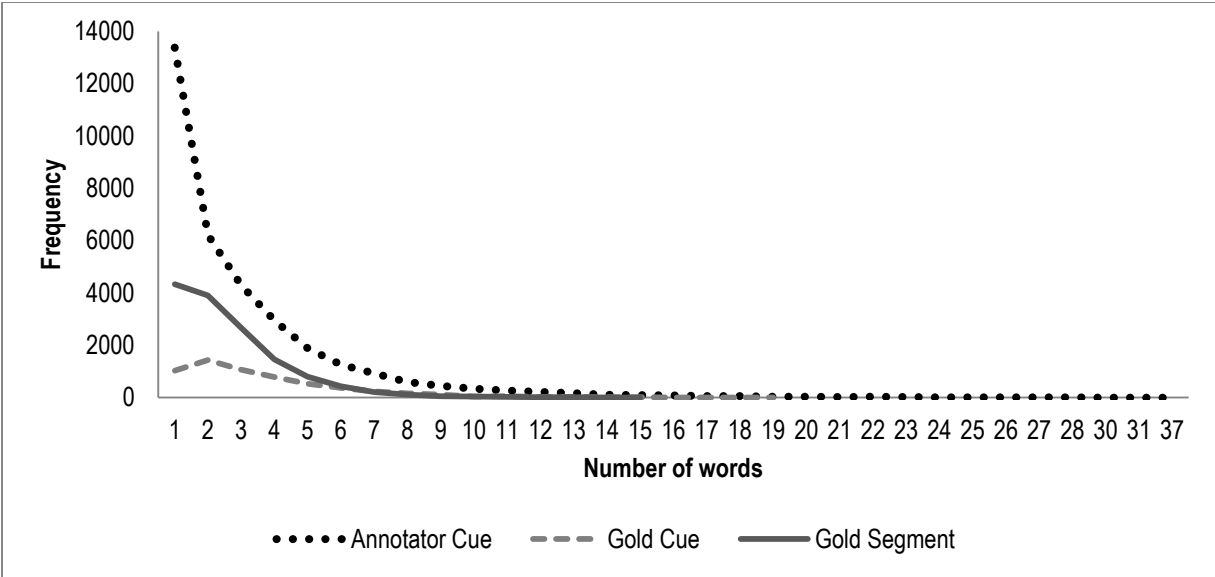


Figure 4.6: Frequency of emotion cue and segment length based on word count

Token Type	Count	%
Alphanumeric	32545	90
Hashtag	436	1
Punctuation	1771	5
Emoticon	483	1
Emoji	1062	3
Total	36297	

Table 4.24: Composition of token types

Table 4.24 shows the composition of five token types in the gold cues: alphanumeric word, hashtag (#keyword) commonly used as a topic indicator in tweets, punctuation mark, emoticon and emoji. A large portion of the textual emotion signals (91%) consist of words (i.e., alphanumeric and hashtag tokens) while 9% of the emotion cues consist of non-alphabetical symbols, which include punctuations marks, emoticons and emojis. Emoticons (e.g., “:”) and “:”) are combinations of punctuation marks commonly used to represent facial expressions in online communication whereas emojis (e.g., 😊 and 😊) are the more expressive successors of emoticons expressed in the form of picture characters that come with standard Unicode

encoding. Both are commonly used as pictorial or ideographic representations of emotion in text. Emojis used to express emotions in the emotion cues are not only limited to smiley faces.

POS Tags	Description	Examples	Count	%
NOUN	Noun	<i>heart, life, champion</i>	10459	27%
VERB	Verb	<i>look, love, missed</i>	7049	18%
ADJ	Adjective	<i>happy, sad, beautiful</i>	4007	10%
PRON	Pronoun	<i>I, you, me, we</i>	3344	9%
ADV	Adverb	<i>really, very, just</i>	3015	8%
ADP	Adposition/Preposition	<i>in, on, up, under</i>	2534	7%
DET	Determiner	<i>the, a, an</i>	1685	4%
CONJ	Conjunction	<i>and, or, but</i>	1180	3%
PRT	Particle	<i>at, on, out, with</i>	1339	3%
.	Punctuation Mark	<i>!, ?, ...</i>	3067	8%
NUM	Numeral	<i>one, 2</i>	195	1%
X	Other		659	2%

Table 4.25: Composition of POS tags

As shown in Table 4.25, nouns, verbs, and adjectives account for slightly more than half of all words contained in gold cues. Emotions are not expressed solely using common emotion adjectives occurring in the predicative position of a sentence (e.g., “I’M SO HAPPY”). The gold cues also capture many other forms of expression using action words (e.g., “CRYING HIS EYES OUT”), modifiers (e.g., “GREAT DAY” and “WORST DAY”), intensifiers (“VERY GRATEFUL”) and negations (“NOT SATISFIED”).

## 4.7 Linguistic Analysis

In this section, we focus on eliciting linguistic patterns associated with each emotion category from the gold cues. Since the cue segments consist of words or word sequences no longer than a sentence, we perform analysis both at the lexical level and phrase level on the emotion cues. Analyzing the underlying linguistic properties of each emotion category serves two purposes: 1) to uncover lexical items associated with each emotion category and 2) to identify useful features for automatic emotion classification. In computational linguistics, *lexical*

*items or units of meaning* is used as a broad term that refers to “single words, compounds, multiword units, phrases and even idioms” (Halliday, Cermáková, Teubert, & Yallop, 2004, pp. 2–3).

#### 4.7.1 Lexical Diversity and Density

Words are building blocks of language. We first examine the repertoire of terms (i.e., words and symbols) used to describe each emotion category. We use the surface forms of the terms and not the stems or the base forms. Lexical diversity measures how varied and broad the vocabulary is for each emotion category. We use the type-token ratio (TTR) as the most basic measure of lexical diversity. TTR is the ratio of the number of distinct terms to the total number of tokens. The higher the TTR score, the more diverse the vocabulary for an emotion category, which simply means that more distinct terms are used to express the particular emotion.

Lexical diversity takes into account both content and function words. Since function words play a peripheral role in the understanding of content, lexical density is also reported alongside lexical diversity in Table 4.26. Lexical density measures the proportion of content words in the emotion cues for each emotion category. The content words are the words that contribute meaning to the concept. Function words (e.g., *am*, *to*, *so*, etc.) alone have little meaning but these words play a role in indicating how words relate to one another. A stop word list is used to remove function words from the emotion cues when computing lexical density.

The terms used to describe each emotion category are not equally diverse. Based on the lexical diversity scores in Table 4.26, some emotion categories exhibit a richer and more varied set of distinct terms, notably for *exhaustion*, *desperation* and *inspiration*. The same terms are seldom repeated across the positive instances for these emotion categories. For instance, the root word “exhaust” occurs only 3 times in the set of emotion cues for *exhaustion*. A variety

of other terms used express *exhaustion* include “#yawn”, “#yawnagain”, “completely drained”, “tired”, etc. Many of these terms also occur sparsely in the emotion cues.

Category	# Segment	# Token	# Distinct Term	Lexical Diversity	Distinct Content Words	Distinct Stop Words	Lexical Density
All	14059	36297	4895	0.13	4779	116	0.68
Desperation	89	274	172	0.63	131	41	0.61
Exhaustion	73	207	130	0.63	97	33	0.67
Inspiration	97	277	168	0.61	135	33	0.68
Boredom	64	185	110	0.59	77	33	0.66
Shame	119	334	196	0.59	153	43	0.67
Relaxed	112	319	185	0.58	141	44	0.64
Indifference	79	272	146	0.54	109	37	0.63
Jealousy	59	224	122	0.54	84	38	0.56
Fear	350	992	479	0.48	406	73	0.67
Fascination	284	713	322	0.45	266	56	0.71
Hate	268	648	284	0.44	227	57	0.67
Confidence	160	541	230	0.43	173	57	0.56
Regret	227	690	298	0.43	232	66	0.61
Doubt	223	754	311	0.41	249	62	0.6
Longing	176	574	222	0.39	164	58	0.57
Surprise	335	747	284	0.38	235	49	0.76
Curiosity	123	372	137	0.37	89	48	0.5
Admiration	660	1807	659	0.36	580	79	0.66
Sadness	826	2085	706	0.34	619	87	0.7
Sympathy	182	555	189	0.34	147	42	0.57
Anger	2049	5706	1740	0.3	1633	107	0.65
Pride	271	674	184	0.27	139	45	0.58
Amusement	833	1460	376	0.26	322	54	0.88
Excitement	1211	3050	731	0.24	656	75	0.72
Hope	770	2190	514	0.23	442	72	0.66
Love	949	2581	555	0.22	473	82	0.61
Happiness	2806	6608	1327	0.2	1231	96	0.74
Gratitude	660	1446	217	0.15	165	52	0.61

Table 4.26: Lexical composition for each emotion category

On the other hand, the terms used to describe *gratitude* and *happiness* are far less varied. Tweeters usually stick to a relatively small set of conventional terms when expressing such emotions. For *gratitude*, the root word “thank” occur 457 times in the emotion cues. It is

interesting to note that the emotion categories with lower lexical diversity are also the categories that human annotators can recognize with greater reliability. Naturally, it is easier for annotators to remember a smaller set of terms associated with a category in their working memory while performing the annotation task. Furthermore, the repetition of same terms can help improve the recognition process over time.

With the exception of *curiosity*, each emotion category has a greater share of content words as opposed to function words. The function words in the emotion cues play a more significant role than merely acting as glue to string words together in a grammatically correct manner. Some functions words surrounding the content words can subtly change the emotion being communicated in the tweets. In Example 4.15, the position of “to” appearing after “honored” (*honored to*) shows that the tweeter is expressing *pride* whereas the position of “to” the other way around (*to honor*) is used rather as an expression of admiration. Based on the lexical density scores presented in Table 4.26, *amusement* has the least number of function words present in the emotion cues.

**Example 4.15:** I'm deeply honored to serve another term. Now, it's time to get back to work ensuring NJ remains a great place to live, work & raise a fam. **[Pride]**

**Example 4.16:** Prince George's Co Commission women today honored me as their champion, but I want to honor women of PG county. <http://t.co/o5fLcq mL> **[Admiration]**

### 4.7.2 Lexical Uniqueness

We next examine if the terms occurring in the cues for an emotion category are salient indicators of the category and how much overlap exists among terms from different emotion categories. If a term is used for an emotion category more often than expected, the term is likely to be a salient indicator for that category. A high degree of salience can be established if a substantial proportion of all term used in the corpus are in the cues for a given category. The

term can then serve as a basic lexical pattern that humans and computers can leverage to recognize the emotion category of interest.

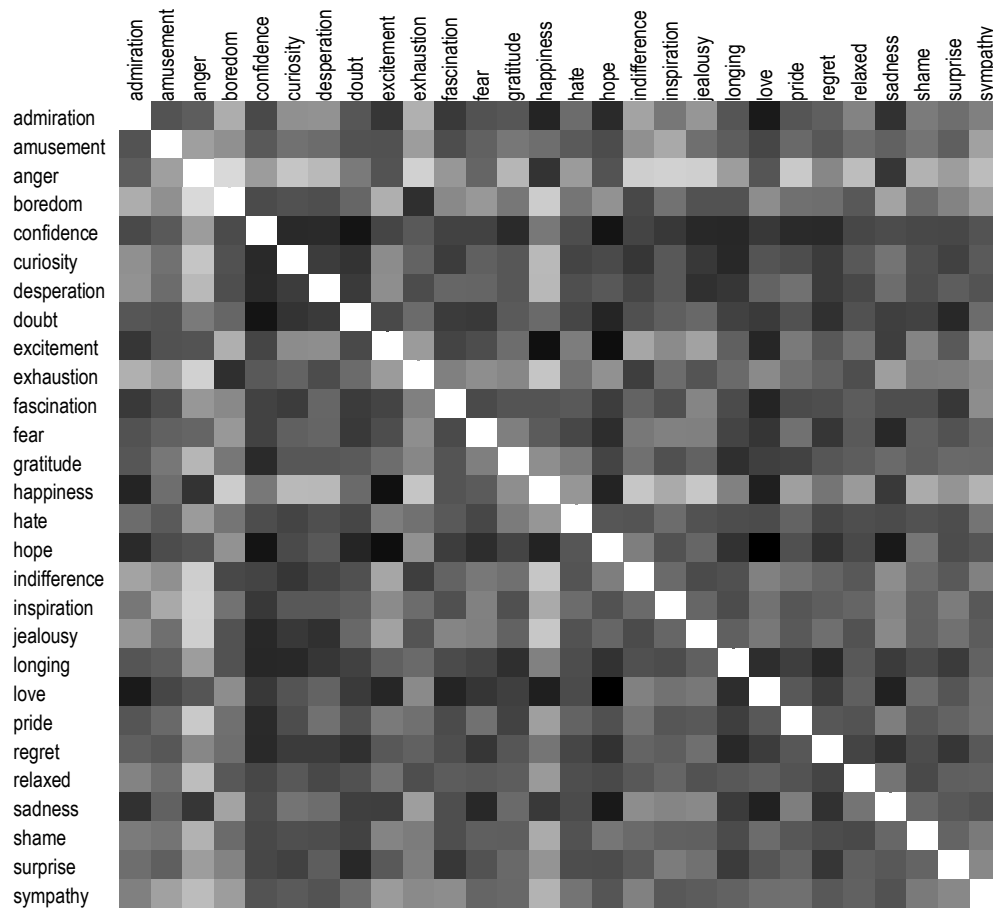


Figure 4.7: Heatmap based on term similarity between emotion category pairs

The proportion of shared terms between an emotion category and another may reveal interesting insights on the relationships among the 28 emotion categories. The overlap or similarity of terms between each emotion category pair is measured using the Jaccard measure. We compare a set of distinct cue terms for an emotion category to the set of cue terms from every other category. The similarity scores between each category pair are presented in a matrix visualized as a heatmap shown in Figure 4.7. The matrix is symmetric, meaning that the values above and below the main diagonal are exactly the same. We exclude the similarity scores on the main diagonal since all cells have similarity scores of 1. The maximum similarity

score is 0.2, which shows that the proportion of similar terms between the emotion-category pairs is still low.

Each shaded cell represents a pair of distinct emotion categories. The cell shading reflects the amount of overlapping terms found in an emotion category pair, the darker the shading, the greater the amount of overlapping terms. The frequencies used to generate the heatmap include function words so a small amount of term overlap between categories is expected. A majority of the emotion category pairs share only a small amount of similar terms as indicated by the lightly shaded cells. Of greater interest are the darkly shaded cells in the heatmap where there is a relatively high overlap between two emotion categories. The seven category pairs with high proportions of shared terms (i.e., similarity score = 0.2) are *confidence-doubt*, *excitement-happiness*, *excitement-hope*, *confidence-hope*, *love-hope*, *sadness-hope* and *admiration-love*.

A closer analysis of the shared terms in the seven category pairs reveals several reasons for the higher similarity scores. First, two emotion categories can be polar opposites. The *confidence-doubt* pair is one example. Both category pair has many content words in common but these words are negated in one category and not in the other. In Example 4.17, the cue segment “*I’m sure*” indicates confidence but the negated form “*Im not sure*” serve as a cue for doubt. Other shared content words for confidence and doubt include “*trust*”, “*doubt*”, “*confidence*” and “*believe*”.

**Example 4.17:** I've watched 3 seasons in like 2 weeks so I'm sure I can do it lol  
[Confidence]

**Example 4.18:** Im not sure of what's real and what's not. [Doubt]

Second, two emotion categories sharing many similar terms may belong to a superclass or a larger family of emotions that have semantic ties. For instance, *excitement* comprises terms describing feelings of expectation and pleasure. *Happiness* is also described with terms describing pleasure. *Excitement* and *happiness* share similar adjectives that are used to



express pleasure (e.g., *great*, *fantastic*, *excellent*, etc.). Therefore, it is likely that the two categories belong to a pleasure superclass. *Excitement* may share a similar relation to hope as terms used to express expectations (e.g., *waiting*, *expecting*, etc.) are found in the emotion cues for both *excitement* and *hope*. Such relations are also observed in the *confidence-hope* and *admiration-love* pairs.

Third, common words in the emotion category pairs can have different senses based on the context in which the words are used. The hope-love and hope-sadness pairs have many overlapping words that are polysemes. To illustrate this point, the word “love” occurs in the emotion cues for both love and hope but the word means affection for a person in love (Example 4.19) while the same word when used in hope means to desire something very much (Example 4.20).

**Example 4.19:** going to sleep with a smile on my face. much love to you all. sleep tight

**[Love]**

**Example 4.20:** GO FOR IT SERGIO - I would love to see you win!

<http://t.co/iuNpmspOn2> **[Hope]**

Low term similarity scores between all emotion category pairs suggest that each category is characterized by a set of salient lexical items that can be extracted from the cues. However, not all terms contribute equally as salient indicators of an emotion category. The saliency of a term diminishes when it occurs in more than one emotion category (multi-category term). Therefore, the next logical step is to identify the terms (i.e., single words and symbols) that can serve as units of meaning for each emotion category.

### 4.7.3 Lexical Indicators

In this section, we explore the meaning of terms used in each emotion category to identify salient indicators for each category. We first extracted all symbols (punctuation marks, emoticons and emojis) (Table 4.27) and hashtags (Table 4.30) from the gold cues of each

emotion category. For words, we extracted the top 50 most frequent terms in the gold cues excluding function words. A subset of frequent content words for each emotion category is shown in Table 4.28.

#### 4.7.3.1 Symbols

Four particular punctuation marks are prevalent in the emotion cues: exclamation mark (!), question mark (?), ellipsis (...) and combination of exclamation and question marks (!?). These punctuation marks occur in almost all the emotion categories; none of them is specifically fixed to a particular emotion category. The presence of exclamation marks is more notable in emotions with higher intensity such as *excitement*, *happiness* and *anger*. On the other hand, question marks occur more frequently in emotion categories with high degree of unexpectedness: *curiosity*, *doubt* and *surprise*.

Emoticons exhibit similar characteristics in that they are not unique to a particular emotion category. Although emoticons may not, by themselves, be reliable indicators of a fine-grained emotion category, they can serve as indicators of emotion valence (i.e., whether an emotion category is positive, negative or neutral). Note that different variations of the *happy face* emoticons (e.g., “:)” and “:D”) appear in only a subset of emotion categories that are used to express pleasure (e.g., *amusement*, *happiness*, *excitement* and *hope*) while the *sad face* emoticons (e.g., “:(” and “:/”) do not occur at all in these emotion categories. The reverse is also true for the subset of emotion categories used to express displeasure (e.g., *anger* and *sadness*).

Emotion-related emojis, the more expressive successors of emoticons, are more informative in identifying expressions of emotion. These pictograms come with a richer set of facial expressions and descriptions of their intended use<sup>21</sup>. Some smiley emojis can, in principle, be mapped to a number of the emotion categories in this study based on the guidelines for their

---

<sup>21</sup> Emoji description and mapping to standardize Unicode characters: <http://apps.timwhitlock.info/emoji/tables/unicode>

intended use. For example, 😱 is intended to be used for the expression of *fear*. However, the actual use of emojis by tweeters does not always follow the prescribed guidelines. It is evident from Table 4.27 that 😱 does not exclusively appear in the emotion cues for fear but is also used to express surprise and curiosity. Based on the observed use of emojis in the corpus, it is not possible to establish a one-to-one mapping between an emoji and an emotion category. Unlike emoticons, emojis can be treated as indicators for a smaller subset of emotion categories as they occur in far fewer categories. To illustrate this point, when 😍 occurs in a tweet it is most likely to be expressing *love*, *admiration* or *happiness*.

The frequency of occurrences of an emoji in a certain emotion category does increase its saliency as a class indicator. Although 😄 occurs in *amusement* and *happiness*, its frequency in *amusement* (145) far exceeds that of *happiness* (4), thus making it a more salient indicator of *amusement*. Co-occurrence of certain emojis especially in a sequential fashion in a tweet can hold slightly different meaning from each individual emoji's intended meaning. For instance, 😄😄😄😄😄😄 is not used to express both *sadness* and *happiness* at the same time but rather it is a unit of meaning for *amusement*. In addition to smiley emojis, tweeters also use various heart and hand gesture emojis to express emotions in tweets.

The symbols listed in Table 4.27 make up part of the vocabulary of each emotion category and are useful in identifying emotion at a coarse-grained level. However, the symbols serve more of a complementary role in the detection of fine-grained emotions since they cannot be used to definitively determine a single category. In other words, the symbols, by themselves, are not sufficient to discern any of the 28 emotion categories.

Category	Punctuation <sup>22</sup>	Emoticon	Emoji: Unicode (Symbol)
Admiration	!*	:P, <3, =), :------)	f09f9898 (😄), f09f988d (😄), e29da4 (❤️), f09f92a9 (👑)
Amusement	!, ?	:), =), :'), :P, =P, :-p, :D, xD, X'D, X"D, :D, :3, (=, (X, (^_^), ^_^	f09f9882 (😄), f09f98ad (😄), f09f98a9 (😄), f09f98b4 (😄), f09f98b3 (😄), f09f9889 (😄), f09f988f (😄), f09f988a (😄), f09f9884 (😄)
Anger	!*, ?	:/, :-/, =/, :(, :-(, >:(, :S, :L, /:, -_-, \_/_., 0,0, ;_;	f09f9892 (😡)
Boredom	..., ?, !		f09f98b4 (😞)
Confidence	!	:), :)	f09f918d (👊), f09f988a (😄)
Curiosity	?*		f09f9987 (👁️), f09f98b1 (👁️), f09f98b3 (😄)
Desperation		-_-, u_u, :(	f09f98ad (😄), f09f9894 (😄)
Doubt	?, !	-_-	f09f9895 (😄)
Excitement	!*	:), :-), :3, ^_^	f09f98b3 (😄), f09f988d (😄), f09f9881 (😄), f09f988a (😄)
Exhaustion	!		f09f92a4 (zzz), f09f9890 (😄), f09f98ab (🙄)
Fascination	!, ..., ?	:D, xD, :), =)	
Fear	!	:(, :S, --, :/, :-(	f09f98b1 (👁️), f09f98a8 (😱), f09f98a2 (😄)
Gratitude	!*, ?	:D, :'), :)	f09f9297 (❤️)
Happiness	!*	:), :D, =p, =), :-), :'), (:, :p, :), :D, =], :3, ^_^, ^.^	f09f998c (👩), f09f988d (😄), f09f918d (👊), e298ba (😄), f09f9881 (😄), f09f9882 (😄), f09f9880 (😄), f09f918f (👏), f09f989c (👩)
Hate	!*, ?	:/, :(	
Hope	!	:), :-), =), ='), <3	f09f998f (👩)
Indifference			
Inspiration	!*, ...	=)	
Jealousy			
Longing	!*	:(, :), :), :/, D;	f09f9894 (😄), f09f9294 (❤️), f09f98ad (😄), f09f98a3 (😄)
Love	!*	:), :D, =), =D, :0), <3, &lt;3, >o<	f09f988d (😄), f09f9295 (❤️), e29da4 (❤️), e299a5 (❤️), f09f9299 (❤️), f09f998c (👩), f09f929c (👏), f09f9297 (❤️), f09f9296 (❤️), f09f9898 (😄)
Pride	!*	:), :')	
Regret	!	:/, :(, :-(	
Relaxed	!	:D, :')	f09f988a (😄), f09f988c (😄), f09f9883 (😄)
Sadness	!, ?	:(, :-(, :'), :'), :), =(, :/, :), ;_;	f09f98ad (😄), f09f9894 (😄), f09f98a9 (😄), f09f9294 (❤️), f09f98a2 (😄)
Shame	!	-_-	f09f98a9 (😄)
Surprise	!*, ?, ?!*	:o, :0, 0,0, O_o, o_O	f09f9180 (👁️), f09f98b3 (😄), f09f98b1 (👁️)
Sympathy	!	:(, <3	

Table 4.27: Symbols associated with each emotion category

<sup>22</sup> The \* symbol indicates single or more occurrences of the punctuation mark.

#### 4.7.3.2 Lemmas

Table 4.28 lists the top 50 most frequent interjections, abbreviations and lemmas (i.e., the canonical form of words) associated with each emotion category from the gold cues. We next examine the interjections, abbreviated forms and lemmas frequently used to describe each emotion category. The semantic properties of the words within an emotion category contribute to the overall abstract meaning of the category (Beck & Kumar, 1998).

Due to the 140 character limit imposed on a tweet, interjections and abbreviations are widely used as compact representations of emotions. For example, common sounds of laughter used to express *amusement* include interjections such as “*haha*”, “*hehe*” and “*hoho*” as well as abbreviations like “*lol*” (*laughing out loud*) and *lmao* (*laughing my ass off*). Only the shortest canonical representations are presented in Table 4.28 and the \* symbol indicates that various elongated forms of the interjection are found in the emotion cues. It is common for tweeters to elongate the interjections as well as the abbreviations (e.g., “*hahahaha*” and “*looooo*”) to emphasize the expression.

At the core of each emotion category are the emotion words, i.e., words that denote or describe emotion (e.g., *fear*, *love*, *anger*, *amusement* and so forth). Emotion words can be nouns, adjectives, or verbs (e.g., “*sadness*”, “*sad*” or “*sadden*”). Many emotion words within the same category are synonyms or near synonyms (e.g., “*shame*”, “*embarrass*”, “*humiliate*”, etc.). The words within each emotion category also share two common semantic properties. First, each emotion category also contains words describing actions and behaviors associated with emotions. Unlike emotion words, the meaning of the action words is connotative rather than denotative. For example, “*crying*” often connotes *sadness* or *desperation* while “*cheering*” connotes *happiness* or *excitement*. Second, content words in the emotion cues carry strong positive or negative connotative meaning that can influence the overall semantic orientation of

the tweet. For instance, content words in *anger* carry a negative connotation. The use of the term “*bitch*” to refer to a woman implies that the tweeter is displeased with the woman.

A word may belong to a single emotion category or multiple categories. Words that belong to a single category offer greater contribution as a salient indicator of that category. We will refer to these words as primary indicators of an emotion category. Their occurrence in a tweet almost always establishes the presence of a particular emotion category. Without knowledge of the context of use, multi-category words by themselves are ambiguous and cannot be used as a sole indicator of a particular emotion category. The emotive meaning of multi-category words depends on the contextual cues surrounding the words. We refer to these words as secondary indicators of an emotion category.

To distinguish between the primary and secondary indicators of each emotion category, we compute a cue term weight for each term in an emotion category. Cue term weight measures the importance of a term for an emotion category. It is a logarithmically scaled fraction of the observed frequency of a cue term in a category divided by its expected frequency in the category. If a term only occurs in a single emotion category and nowhere else, the term is considered to be a primary indicator for the particular emotion category.

The set of terms within each emotion category that fall above a certain weight threshold are the primary indicators. Otherwise, terms in the corpus occurring across multiple emotion categories would produce low cue term weights. For example, function words that occur very frequently in the corpus but are uniformly dispersed across multiple emotion categories would be expected to have weights near zero.

Category	Interjection/Abbrev.	Lemma
Admiration		honor, best, beautiful, love, respect, look, perfect, talent, good, tribute, cute, nice, incredible, hero, great, brilliant, adore, admire, well, talent
Amusement	haha*, hehe*, hoho*, lol*, lmao, lmfao	laugh, funny, fun, hilarious, crack, cool, cry, funniest, humor, amuse, joke, prank, entertaining, comical, best, pretty, cute
Anger	ugh, argh, wtf, smh, gtfo, stfu	fuck, shit, stop, hell, damn, suck, bitch, lie, upset, worst, angry, delay, piss, mad, stupid, ass, horrible, fail, weak, annoy, upset, disappoint, outrage
Boredom	ugh	bore, boredom, hour, tire, slow, tedious, unproductive, dull, moody, drag
Confidence		confident, faith, believe, better, sure, best, let's, stand, win, victory, brave, trust, queen, boss
Curiosity		wonder, curious, curiosity, happen, know, who
Desperation		desperate, need, stop, hopeless, help, protest, kill, hell, suicide, please, deprive, beg, cry
Doubt	idk	confuse, understand, believe, trust, want, sure, torn, doubt, maybe, may, baffle, know, decide, fuzzy
Excitement	omg, oh, woo, woop, yeah, yea, ya	wait, excite, go, look, forward, cheer, let's, pump, great, ready, tonight, blow, fire, new, win, enthusiasm, best, fun, thrill, touchdown, anticipate, awesome
Exhaustion	zz	tire, exhaust, sleep, asleep, sleepy, aching, energy, mile, run
Fascination	wow, waww, omg, omfg	amaze, amazing, interest, fascinate, beautiful, cool, stuff, look, story, good, awe, impress, awesome, strange, incredible, epic
Fear	eek, #yikes	concern, worry, fear, scare, anxiety, horrific, hope, terrify, creepy, screw, look, afraid, stress, anxiety, danger, risk, death, panic, threat, nightmare
Gratitude	thnx, thx, tysm, ty	thank, grateful, gratitude, mahalo, appreciate, bless
Happiness	yay, yeh, yiips, woop, wohooo, gr8	great, good, happy, happiness, congrats, best, nice, enjoy, glad, news, fun, birthday, love, beautiful, cheer, cute, win, smile, visit, awesome, celebrate
Hate	ew, ugh, wtf, h8	hate, disgust, gross, sick, suck, lie, despise, dislike, hatred, distaste, traitor, detest, fuck, ugly, shit
Hope		hope, god, good, bless, luck, great, best, wish, may, pray, day, fun, let's, come, keep, better, want, prayer, miracle, dream, safe, enjoy, love
Indifference	meh, cba, idc	don't, care, give, fuck, lazy, doesn't, bother, motivate
Inspiration		inspire, motivate, move, uplift, touch, heart, story, energy, best, beautiful
Jealousy		jealous, jealousy, boyfriend, bitch, girl
Longing		miss, long, yearn, old, memory, wish, remember, back, bring, good, time
Love	fav, ily, luv, ilysm	love, like, favorite, favourite, smile, fall, crush
Pride		proud, honored, honor, home, first, accomplish, pride, best
Regret		sorry, wish, bad, back, apology, regret, shame, forgive, miss, fault
Relaxed	whew, #whew	finally, good, relax, back, chilling, chillin, lay, done, sleep, lazy, comfortable, home, peace, relief
Sadness	rip	sad, sadness, sadden, cry, heart, miss, lost, tear, loss, depress, remember, sigh, tragedy, news, heartbreak, tragic, death, terrible, end, pain, hurt
Shame	oops, #oops	shame, embarrass, awkward, weird, humiliate, naked, dirty, disgrace
Surprise	wow, oh, omg, wtf, woah	believe, god, unbelievable, shock, expect, surprise, unreal, thought, astonish, blow, speechless, traumatize
Sympathy		prayer, thought, heart, condolence, lost, human, victim, need, bad, family, tragic, deepest, offer, tragedy, sympathy

Table 4.28: Frequent canonical form of words and abbreviations for each emotion category

Category	Primary (cue term weight)	Secondary (cue term weight)
Admiration	admire, impressed (3.6)	honoring (3.5), honour (3.4), finest (3.4), precious (3.4), honor (3.2), honored (2.0)
Amusement	lmfao, lmao, haha, hilarious, lol, amused (3.2)	funny (2.9), jk (2.9), entertaining (2.8), farts (2.8), laughing (2.7)
Anger	smh, disappointed, outrage, asshole, ignorant (2.56)	annoying (2.5), upset (2.5), bullshit (2.5), shitty (2.4), angry (2.3)
Boredom	bore, unfunny, boredom, tedious, unproductive (5.8)	boring (5.7), bored (5.7), drag (5.1), moody (4.7), dull (4.7)
Confidence	confident, determined (5.0)	rely (4.3), assure (4.3), certainty (4.0), faith (3.8), confidence (3.6)
Curiosity	curious, curiously, wondered (5.12)	wonder (5.0), wondering (4.7), curiosity (4.4), hm (4.0), strange (3.5)
Desperation	desperately, hopeless, pleading, doomed (5.5)	desperate (5.3), desperation (5.1), sos (4.82), stranded (4.8), begging (4.4)
Doubt	baffled, conflicting, confuse, uncertain (4.6)	confused (4.5), torn (4.1), traitors (3.9), snakes (3.9), fuzzy (3.9)
Excitement	thrilled, pumped, geaux, enthusiastic, rooting (3.1)	excited (3.1), exciting (3.0), excitement (3.0), hurry (2.9), touchdown (2.8)
Exhaustion	sleepy, exhausted, tiring, drained (5.8)	stressful (5.1), sore (5.1), asleep (5.0), tired (4.9), drove (4.4)
Fascination	amaze, awe, intrigued, interestingly, enthusiast (4.3)	interesting (4.2), amazing (3.9), thoughtful (3.6), fascinating (3.6), phenomenal (3.6)
Fear	anxiety, creeps, troubled, concern, eek, horrifying, haunt (4.2)	worried (4.1), nervous (4.0), fear (4.0), terrifying (4.0), panic (3.9), scared (3.9)
Gratitude	grateful, thnx, mahalo, thanked, thankful (3.4)	thank (3.4), thanks (3.4), thx (3.3), appreciate (3.0), ty (3.0)
Happiness	congrats, happiness, applauds, shoutout, happier (2.2)	glad (2.1), pleased (2.1), enjoyed (2.1), congratulations (2.1), happy (2.0), joy (1.9)
Hate	disgusting, ew, hatred, dislike, despise, gross, detest, h8 (4.4)	hate (4.3), hated (4.2), hates (4.2), messed (3.7), traitors (3.7), ughhh (3.7)
Hope	hopefully, hopeful, miracles, godspeed (3.4)	hope (3.3), hoping (3.3), luck (3.3), miracle (3.1), bless (2.9), pray (2.7)
Indifference	cba, unmotivated, pfft, meh, dgaf (5.5)	idc (5.2), fucks (4.8), faze (4.8), bothered (4.4), lazy (4.3), motivated (4.13)
Inspiration	inspired, inspiration, inspires, motivational, heartwarming (5.3)	inspiring (5.2), inspirational (5.1), inspire (5.0), motivation (5.0), uplifting (4.9), moved (4.0)
Jealousy	jealousy, envy, possessiveness (6.2)	jealous (6.1), chicks (5.5), sidelines (5.5), allowed (5.0), boyfriend (4.7)
Longing	yearning, crave, longs, sentimental (4.8)	unforgettable (4.4), miss (4.1), yearns (4.1), memories (3.8), wish (3.3)
Love	ilysm, ily (3.1)	favourite (3.0), luv (3.0), love (2.8), adore (2.7), lovers (2.7), liking (2.7)
Pride	proudly (4.3)	proud (4.2), honored (4.0), humbled (3.6), pride (3.0), honor (2.7)
Regret	apologies, sry, unhealthy (4.6)	regret (4.5), sorry (4.3), regrets (4.2), wished (3.9), guilt (3.9)
Relaxed	whew, relaxation, thankfully (5.3)	chillin (5.1), relaxing (4.9), mellow (4.6), calmer (4.6), relax (4.6), comfortably (4.2), chilling (4.2)
Sadness	saddened, sadly, heartbreaking, sadness, painful, depressing, saddest, cries (3.4)	sad (3.4), rip (3.2), poured (3.1), crying (3.1), mourns (3.1), cry (2.9), sigh (2.9)
Shame	embarrassed, shameful, ashamed, humiliates (5.1)	awkward (4.9), oops (4.8), shame (4.8), disgraceful (4.4), ruins (4.4), cringe (4.4)
Surprise	shocked, unbelievable, disbelief, stunned, yikes, astonishing, astounding (4.1)	shocking (3.8), whoa (3.8), shock (3.7), woah (3.7), wow (3.6), surprised (3.6)
Sympathy	sympathise, sympathies (5.1)	condolences (5.0), prayers (4.9), thoughts (4.5), tragic (4.1), praying (4.0), sympathy (3.7), sorry (2.7)

Table 4.29: Primary and secondary indicators of each emotion category



For each term ( $t$ ) in an emotion category ( $E$ ),

$$\text{Cue term weight } (t, E) = \log \frac{f_{t,cue}}{f_{t,corpus} * P_{E,corpus}}$$

where

$f_{t,cue}$  = Frequency of term in emotion cues for  $E$

$f_{t,corpus}$  = Frequency of term in the corpus

$$P_{E,corpus} = \frac{\text{Number of instances of } E \text{ in the corpus}}{\text{Number of instances in the corpus}}$$

Highly ranked primary and secondary indicators for each emotion category as well as their cue term weights are presented in Table 4.29. Each emotion category possesses only a small set of terms that are fixed to an emotion category (i.e., primary terms). The primary terms serve as salient indicators of an emotion category regardless of the context of use. We present only the terms that scored the maximum weight for each emotion category in Table 4.29. Many words that belong to this category are emotion words.

All other terms with weights that fall below a particular threshold are considered to be secondary terms. Table 4.29 shows the top secondary terms ranked below the maximum weight for each emotion category. We found secondary terms with weights that fall within the range of zero and the threshold to be more informative than the terms with negative weights. Based on the cue term weights, a significant portion of the terms can be characterized as secondary indicators as they occur in more than one emotion category. Secondary terms rely on other surrounding terms to form emotive meaning. Such terms can still serve as lexical clues or weak identifiers of an emotion category especially if the terms occur frequently in the category. Given the prevalence of secondary terms, it is evident that many emotion-related words have multiple senses. These words add a layer of ambiguity to the expression of emotion in text (e.g., “sorry” in *regret* refers to feeling regretful for an action while “sorry” in *sympathy* means feeling distressed by someone’s loss). Secondary terms can also express different emotions when

combined with other terms (e.g., “I am tired” is a cue for *exhaustion* and “got tired of my pet” is a cue for *boredom*).

Using the cue term weights, we can compare the importance of a term occurring in multiple emotion categories. For example, the term “honor” is weighted higher in *admiration* (3.2) as opposed to *pride* (2.7), which suggests that “honor” is a more important indicator of *admiration* than *pride*. On the other hand, the term “honored” has a higher weight in *pride* (4.0) than in *admiration* (2.0), making “honored” more important for *pride*. Although “honor” and “honored” are forms of the same lexeme with the dictionary meaning “regard with respect” knowing who is being regarded with respect makes a difference in distinguishing *admiration* and *pride*. If the tweeter is the one who feels that he or she is being regarded with respect, then the tweeter is expressing *pride* but if the tweeter is regarding someone else with respect, *admiration* is being expressed instead.

The primary and secondary terms form the foundation of our emotion lexicon. All terms are converted into features for the machine learning classifiers. In addition, the primary and secondary terms can function as seed words to enrich the vocabulary of each emotion category. The primary terms can be used to retrieve synonyms or other semantically related words associated with each emotion category while secondary terms can be used to capture the contextual cues surrounding the secondary terms.

#### **4.7.3.3 Hashtags**

Hashtags on Twitter serve as topical markers to enable convenient identification of tweets based on topic. Generally, hashtags are considered part of the content of a tweet and most often appear at the end of a tweet (Example 4.21) but can also appear at the beginning or anywhere in the middle of the tweet (Example 4.22).

Category	Hashtag
Admiration	#respect, #mcm, #mancrush*, #perfection, #handsome, #gergous, #fuckingbeautiful, #sexy, #legends, #ifancyourface, #legendary, #standup, #laugh, #gorgeous
Amusement	**funny*, #somerecordsmustbeproken, #hadtodoit, #lmao, #lmfao, #lol, #haha, #dying, #great, #oldjoke, #smiles
Anger	#growup, #notreally, #stop*, #*wrong, #*stupid*, #*fuck*, #*fail*, #ooops, #fraud, #cheating, #*sucks, #worst*ever, #*problems, #scandal, #areyoufreakingkiddingme, #wastedchance, #stinky*, #lame, #badsportsmanship, #dontevenbotheregethefacts, #didnthearhimcomplaining, #grr, #murderer, #youjustruinedmylife, #warcrimes, #getyourshittogether, #ridiculous, #killingit, #*bullshit, #horrible*, #nowwearelate, #never*again, #disappointed, #moron, #racism, #gettfoverit, #doublefacepeople, #pathetic, #nocustomerservice, #forpetesake, #*lies, #lyin*, #*hypocrisy, #secretive, #deceptive
Boredom	#bored, #suckyweather
Confidence	#yeswecan, #notconfusedanymore
Curiosity	#isitok, #justwondering, #magic
Desperation	#desperation, #desperate*, #pleasehelp, #dying
Doubt	#whytho, #thestruggle
Excitement	#excited, #*excited*, #geaux*, #go*, #touchdowns, #ponderthat, #omg, #takingittothehouse, #jacked, #ticktock, #getnthefuckoutofhere, #getpumped, #olympicspirit, #finally, #2moreweeks, #stoked, #oolegooo, #teambbringit, #shouldbegood, #walkoff, #whodat, #ohyeah, #feelthat, #keepchoppin, #cantwait, #lookingfoward, #longoverdue, #nervousmuch, #woohoo
Exhaustion	#tired, #yawn, #yawnagain
Fascination	#incredible, #amazing, #canadayum, #mythroathurts, #fromscreaming
Fear	#yikes, #nervousmuch, #terror, #scary, #dontdissapear
Gratitude	#thankyou, #*thank*, #gratitude, #yearoflivinggratefully, #tybg, #foreverthankful
Happiness	#happy, #*happy*, #happiness, #*best*, #welldeserved, #*cool, #fun*, #*good*, #*bless*, #longlive*, #worthit, #whatabadass, #*awesome*, #congrat*, #brofist, #biglove, #heaven, #hokie, #tgif, #dope, #gotime, #greatmemories, #handsdown, #bringiton, #bravo, #bangtidy, #excellent, #smile, #cutie, #*spirit, #bam, #win*, #moments, #together, #celebratorymoment, #finally, #thedayishere, #vacation, #spoiled
Hate	#*hate*, #bugseverywhere, #nothing, #blind, #uglysoul, #intolerance, #scumbag, #revenge, #warcrimes, #ihl
Hope	#hopeso, #morewins, #forward, #faith, #id, #lets, #go, #good, #myoctoberwish, #deserves, #great, #again, #wishfulthinking, #goals, #nevergiveup, #greatness, #prayfor*
Indifference	#justdontgiveafuck, #wedontcare, #zero, #didnteventry, #shit
Inspiration	#inspiring, #inspiration, #everyonehasahero
Jealousy	
Longing	#miss*, #lonely, #sigh, #takemeback
Love	#love, #*love*, #alittleobsessed, #heart, #mcm, #*crush, #whatababe, #ily, #favorite*, #stolen, #romance, #romantic
Pride	#proud, #*proud*, #honors, #madeinamerica
Regret	#help, #sorrycamloveyou, #mybad, #baby
Relaxed	#relax, #relaxing, #relaxation, #relax*, #whew, #lazynight, #restday
Sadness	#sadness, #neverforget, #destroyed, #pornharms, #memorial, #inmemory, #noonecares, #thestruggle, #rip*, #fangirlsproblem, #depressing, #depression, #nasty, #alone, #suckstosuck
Shame	#shameful, #noshame, #kingoftheflipflop, #oops, #awkward, #thisisweird
Surprise	#unbelievable, #wow, #what, #wasntexpectingthat, #shocker, #whatthe, #believe, #havesomerespect
Sympathy	#god, #bless, #pray, #prayfor*

Table 4.30: Hashtags associated with each emotion category

A hashtag can contain a single word (e.g., #amazing) or multiple words merged together as unit without spaces in between the individual words (e.g., #takingittothehouse). Hashtags are not purely informational and can be used as an emotion marker as evidenced by the presence of hashtags in the emotion cues. Table 4.30 shows all the hashtags associated with each emotion category extracted from the gold cues.

**Example 4.21:** She prob entered a train that is empty and lined with mirrors #uglysoul #intolerance #pseudoclass <http://t.co/ltZvGyQbH0> [**Hate**]

**Example 4.22:** #Obama the #MURDERER & #UN condemn #Hamas 4 violating the #ceasefire by capturing a soldier! What about #Israel killing more than 100 today! [**Anger**]

Emotion can be expressed through hashtags in many creative forms. The most straightforward forms of expression are the use of emotion words (e.g., #bored, #desperation, #happiness, #love) and interjections (e.g., #haha, #yikes, #sigh, #grr) as hashtags. Abbreviated forms are also used (e.g., #ily, #mcm, #omg, #lol), possibly to save space. In addition to emotion words, hashtags are also formed with the emotion word as part of a larger string of characters. In other words, the emotion word (i.e., primary terms) can be preceded by or followed by any number of words. Multiple words are merged into a hashtag in a meaningful way.

Common patterns observed are presented below. Single terms (e.g., emotion words and other linguistic elements) are encompassed within pointy brackets (<...>). The emotion words in the examples below are underlined. The asterisk symbol (\*) is used a wildcard to represent any character. The asterisk can appear before and/or after the term. The asterisk preceding a term shows any characters (usually a meaningful word) can be attached in front of the term while the asterisk following a term indicates any characters can be attached to the end of the term. We use the “+” symbol to indicate a join between two terms. The order of the terms is not important in the patterns.

### **Hashtag Pattern #1: #\*<emotion word>\***

#### **Pattern #1.1: #<intensifier>+<emotion word>**

- Examples: #nervousmuch, #soexcited

#### **Pattern #1.2: #<pronoun>+<emotion word>**

- Examples: #lovehim, #loveher, #lovethem

#### **Pattern #1.3: <content word>+<emotion word>**

- Examples: #yearoflivinggratefully, #ihatepeoplewho, #canadaproud

With the exception of *jealousy*, each emotion category is associated with a variety of hashtags not limited to only emotion words. Many of the hashtags also incorporate secondary terms (e.g., #justwondering, #missyou and #corporatebullshit) listed in Table 4.30. *Anger* and *happiness* contain the greatest variety of hashtags. In the case of anger, the use of swear words or curse words is apparent in hashtags. It is interesting to note that tweeters prefer to use hashtag words that connote anger, rather than words that denote anger as evidenced by the absence of hashtags such as #anger and #angry in Table 4.30. Hashtags in anger also tend to contain adjectives expressing negative quality whereas adjectives expressing positive quality are commonly included in the happiness hashtags.

### **Hashtag Pattern #2: #\*<curse word>\***

#### **Pattern #2.1: #<curse word>+<noun>**

- Examples: #fuckbankfees, #fuckyou, #teamfuckbarackobama

#### **Pattern 2.2: #get+\*<curse word>\***

- Examples: #getyourshittogether, #gettfoverit, #getfucked

### **Hashtag Pattern #3: #<quality adj>+<noun>**

#### **Pattern #3.1: #<positive quality adj>+<noun>**

- Examples: #funtimes, #greatmemories, #goodfeeling

#### **Pattern #3.2: #<negative quality adj>+<noun>**

- Examples: #badsportsmanship, #horribleservice

**Pattern #3.3:** #best/worst+<noun>+ever

- Examples: #worstpotusever, #worstexperienceever, #besttripever

Essentially, hashtags are emotion-related terms preceded by the # sign and function the same way as words to express emotion. The only difference between words and hashtags is that the unit of meaning of the latter may extend beyond the boundary of a single word. For a multi-word hashtag, it is difficult to decompose the hashtag into individual words due to the absence of any obvious separators between the words. We treat hashtags as a single lexical item regardless of the number of words they contain.

#### **4.7.3.4 Collocations**

The notion of single words as units of meaning is a good starting point but, as earlier examples have shown, the expression of emotion often depends in the context in which words occur. The importance of context is evidenced by the limited set of primary terms for each emotion category. Teubert (2004, p. 171) argues that phrasal units (i.e., compounds, multiword units and set phrases) are more common units of meaning as opposed to single words.

In this section, we extend our investigation to collocations (i.e., two or more words that habitually co-occur) that can serve as salient indicators for each emotion category. Our focus is on the units of meaning associated with emotion category that extend beyond a single word and not on the grammatical structure of phrasal units. Therefore, we use collocations rather than phrases. On one hand, this allows us to expand the lexical items associated with each emotion category as well as abstract helpful linguistic patterns to aid machine learning classification. On the other, it would be interesting to examine if the selection of words joined together as a unit of meaning in tweets is any different at all from conventional set phrases or compound words.

Collocations have to fulfill three criteria: 1) the collocated words must occur together significantly more often than would be expected based on individual word frequencies, 2) a

collocation has to be semantically relevant, and 3) a collocation has its own meaning and is not merely the sum of meaning from its parts. Analysis on the emotion cues is divided into two parts. The first part focuses on relevant bigrams and trigrams extracted from the emotion cues. We then switch our attention to negations in the second part of the section.

### **A) Bigrams and Trigrams**

We extracted the top 20 bigrams and trigrams based on frequency and pointwise mutual information (PMI) (Church & Hanks, 1990) from the gold cues of each emotion category. A sample of the highest ranked bigrams and trigrams are shown in Table 4.31. The bigrams and trigrams that occur in both the frequency and PMI lists are displayed in regular black font, items from only the frequency list are shown in gray font, and items from only the PMI list are italicized.

There are several interesting findings from Table 4.31. First, various types of collocations common in standard English are found although they occur more rarely in the emotion cues from our tweet corpus. Collocations include compound words (e.g. “role model”), phrasal verbs (e.g., “blowing up”, “fired up” and “piss [me] off”), as well as idioms and metaphors (e.g., “heart of gold”, “chokes me up” and “see fire”). Due to their idiosyncratic nature, collocations are directly included as lexical items in the emotion lexicon and are represented as patterns.

Second, relevant phrases used for informal expressions and exclamations are captured in the bigrams and trigrams. For example, exclamations like “well done” is used to express *happiness* while “way to go” is used to express either *happiness* or *excitement*. These lexical items are often followed by exclamation mark and can be treated as multiword interjections.

Category	Bigram	Trigram
Admiration	honor of, to honor, respect for, so beautiful, the best, <i>role model</i>	heart of gold, in honor of
Amusement	always fun, the best, <i>best thing, pretty funny, so funny</i>	
Anger	fuck off, going to, the fuck, the hell, tired of, what the, <i>another hour, deal with, fuck yaw, never fly, very disappointed, wake up</i>	don't have time, get out of, in the face, mad at you, piss me off, what the hell
Boredom	i'm bored, tired of	
Confidence	believe in, bring it, can do, do it, do this, have faith, i can, you can, the best	be the best, can do it
Curiosity	i wonder, are they, is it, to know, wonder how, wonder if, wonder what, wonder why	
Desperation	i need, to go	
Doubt	don't understand, i doubt, know what, not sure	not sure what
Excitement	can't wait, let's go, ready for, ready to, see you, so excited, to go, to see, wait for, wait to, <i>big day, big win, blowing up, fired up, i'm pumped, let's geaux, oh yeah, see fire</i>	go go go, here i come, look forward to, <i>all the way, to be crazy, way to go</i>
Exhaustion	fall asleep	can't keep up
Fascination	interested in, it's amazing, is amazing, so amazing, so beautiful, so cool, the amazing, was amazing	
Fear	at risk, i don't, make me	
Gratitude	be grateful, be thankful, big thanks, thank u, thank you, thankful for, thanks for, thanks so, thanks to, thx to, to thank, <i>many thanks, thnx 2</i>	
Happiness	a good, a great, be happy, congrats to, congratulations to, good news, good to, great to, happy birthday, so happy, the best, <i>an incredible, better than, brilliant film, ever seen, fun times, gold medal, keep up, long time, on point, shout out, well done</i>	of the best, <i>chokes me up, makes you happy, shout out to, way to go</i>
Hate	don't like, fucking hate, hate it, i hate, to hate	
Hope	a good, a great, best of, come back, god bless, good luck, have a, have fun, hope you, hoping for, i hope, i wish, please come, still hoping, the best, want to, <i>bless ya, could be, do happen, i'm hoping, keep on, keep praying, we can, wish them, would be</i>	all the best, best of luck, closed my eyes, for the best, god bless you, have a good, have a great, please come back, <i>want to be</i>
Indifference	don't care	
Inspiration		
Jealousy	isn't allowed, my boyfriend	
Longing	i could, i miss, i had, i want, i wish, miss him, miss me, miss my, miss you, missed you, so much, to miss, wish u, wish you, <i>wish i, want to</i>	i miss you, i wish you, want to be, wish i could, wish i was, wish you were
Love	i like, i love, in love, love it, love with, love you, my favorite, so much, we love, <i>a crush, a good, all time, crush on, fall in, my fav, my fave, my heart, new favorite, really like</i>	i love you, fall in love, we love you,
Pride	an honor, be proud, honor to, honored to, i'm honored, i'm proud, proud of, proud to, so proud, very proud	makes me proud, proud of you
Regret	feel bad, i wish, i'm sorry, my bad, my fault, sorry to, wish u, wish it	wish i could, wish i had, wish i was, wish it was
Relaxed		
Sadness	be missed, im crying, in peace, lost their, my heart, my life, so sad, will miss, <i>come back, deeply saddened, end of, is dead, saddest episode, your heart</i>	lost their lives, rest in peace, will be missed
Shame	awkward moment, so embarrassed	shame on you
Surprise	a surprise, can't believe, holy shit, never thought, omg, to believe, what ?, omg !, what the, wow !, wtf ?	oh my god
Sympathy	condolences to, deepest condolences, in need, in our, keep praying, my prayers, my thoughts, our thought, pray for, prayers to, praying for, sorry for	keep praying for, prayers are with, prayers go out, thoughts and prayers

Table 4.31: Top ranking bigrams and trigrams based on frequency and PMI



More informal expressions such as “oh my god” and “holy shit” are also commonly found in tweets. These two lexical items are closely associated with *surprise*. Using the two phrases as seeds, we can retrieve other variations used to convey the same emotional meaning (e.g., “oh my gosh”, “holy cow” and “oh my goodness”).

Third, based on repeated lexical patterns noted in the bigrams and trigrams across different emotion categories, general patterns of phrase structure can be constructed. The three emotion patterns below reference only the use of personal pronouns but the pattern set can be extended to include other forms of reference to a person.

#### **Emotion Pattern #1**

<personal pronoun: I, you, he, she, we, they> + <verb: to be> + <adjective: emotion word>

- Examples: I’m bored, we are happy, he is desperate

#### **Emotion Pattern #2**

<personal pronoun: I, you, he, she, we, they> + <verb: feel> + <adjective: emotion word, good, bad>

- Examples: I feel bad, she feels sad

#### **Emotion Pattern #3**

<personal pronoun: I, you, he, she, we, they> + <verb: love, hate, miss> + <noun: person/object>

- Examples: I love you, I hate it, we miss him

Also, some emotion categories such as *gratitude* and *sympathy* display obvious lexical patterns based on the high ranking bigrams and trigrams in Table 4.31. For example, the occurrence of “to” or “for” attached to a secondary term in these two emotion categories suggests that these emotions are express towards another person.

In the case of sympathy, the two phrase patterns below are abstracted based on the common positive examples found in the category.

### Sympathy Pattern #1

<noun: prayers, thoughts, prayers and thoughts> +  $\left\{ \begin{array}{l} \text{are with} \\ \text{go out to} \\ \text{to} \\ \text{for} \\ \text{with} \end{array} \right. + \text{<noun: person>}$

### Sympathy Pattern #2

<preposition: in> + <first person possessive pronoun: my, our> + <noun: prayers, thoughts, prayers and thoughts>

Fourth, the prevalence of pronouns, especially first person pronouns, occurring in close proximity to the emotion-related words implies a high degree of self-reference which suggests that the tweeters are most often expressing their own emotion. Who is being referenced together with an emotion-related term can play a significant role in making a distinction between the emotion categories. To illustrate this point, the use of “they” (i.e., third person pronoun) being “the best” in Example 4.23 means that the tweeter is expressing *admiration* towards the two people mentioned in the tweet. On the other hand, the use of “I” (i.e., first person pronoun) with the phrase “the best” in Example 4.24 shows that the tweeter is expressing *confidence* towards his ability to be a father.

**Example 4.23:** the only thing that was good for me was malik and hubert omfg they were the best [**Admiration**]

**Example 4.24:** I'm gonna be the best Dad [**Confidence**]

After filtering out bigrams and trigrams that occur in multiple categories, Table 4.32 lists all the multiword lexical items that can be used as primary indicators of each emotion category. Only the prototypical form of the lexical item is presented. Other morphological variations of the lexical items may be present in the corpus (e.g., “look forward to” is the prototypical form for “looking forward to”).

Category	Phrase
Admiration	the shit, heart of gold, role model
Amusement	crack up
Anger	tired of, shut up, fucked up, fuck off, fuck you, fuck w, what the heck, suck ass, not good, no cure, no way, shut the fuck up
Boredom	doing nothing, real drag
Confidence	don't give up, have faith, leap of faith, bring it, like a boss, hang in there, can do it, will work out
Curiosity	wonder if, wonder where, wonder why, wonder who, wonder what, wonder when, wonder how, wonder whether
Desperation	gave up, shoot me, serious need of, given up, would do anything, badly needed
Doubt	head scratcher, wtf is, what if, not sure, not quite sure, don't understand, maybe not, never trust
Excitement	look forward to, can't wait, let's go, lets go, ready for, ready to, blowing up, cheering for, cheer on, nail biter, counting down, fist pump, keen for, fired up, dying to
Exhaustion	killing me, blowed as fuck, aching body, on fire
Fascination	blows my mind, never get bored, so into
Fear	don't stigmatize, mental break down, heart attack, at risk, i'm screwed, not looking forward to, can't be dealing, avoid eye contact,
Gratitude	thank you, count your blessings
Happiness	good news, good work, good to, good day, good seeing, happy birthday, great to, gr8 to, great news, great pick, great crowd, great day, great time, great meeting, great pic, keep up, nice to, nice job, nice work, home run, well done, way to go
Hate	don't like, sick of
Hope	good luck, have fun, god bless, have a good, dont lose hope, fingers crossed, be great to, keep praying, don't stop praying, all the best, think positive, may god, bless ya
Indifference	don't care, doesn't care, giveth no fucks, don't give a fuck, not that bothered, does not care
Inspiration	words of encouragement, keep going
Jealousy	my boyfriend, isn't allowed
Longing	wish i could, brings back memories, old times, wish you were, wish i was, remember when, brings me back
Love	attracted to
Pride	honored to, honored 2, honor the past, an honor, you go girl, that's my boy
Regret	wish it was, feel so bad, my bad, missed a lot
Relaxed	not bad news, thank goodness, nothing to do, laxing back, laying in
Sadness	rest in peace, will be missed, feeling down, hearts out, not forgotten, teared up
Shame	shame on, feel weird
Surprise	oh my god, holy shit, can't believe, blows my mind, so much wtf, holy toledo, goodness gracious, hard 2 believe, hard to believe, refuse to believe, will not believe, no believe, never thought, sink in
Sympathy	thoughts and prayers, are with, go out to, heart is with, prayers are with, feel so bad for, feel bad for, pray for, prayer for, prayers to, paying tribute to

Table 4.32: Common collocated words or phrases associated with each emotion category

## **B) Negation**

Negation is a grammatical construct that reverses the truth value of a proposition (Miestamo, 2007). In text analysis, negation words (e.g., “no” and “not”) can modify the sentiment being expressed in a unit of meaning (Zhu, Guo, Mohammad, & Kiritchenko, 2014). Kennedy & Inkpen (2006) call negation words valence shifters that are used to change the semantic orientation of a neighboring lexical item. Previous studies have provided evidence emphasizing the importance of negation words in discriminating between positive and negative sentiment (Jia, Yu, & Meng, 2009; Alistair Kennedy & Inkpen, 2006; Zhu et al., 2014). We examine the qualitative behavior of negation words in the emotion cues to determine what role these words play.

Borrowing the terms from Zhu et al. (2014), we will use the term “negator” to refer to negative words (e.g., not), “argument” to mean the text span or lexical item being affected by the negator (e.g., happy), and “negated phrase” to refer to the phrase containing both the negator and argument (e.g., not happy). We first create a list of common negation words, adverbs and verbs (see Appendix B). Negators are then used as seed words to retrieve all relevant negated phrases from the gold cues. We extracted the bigrams, trigrams and 4-grams starting with the negators.

Table 4.33 shows a sample of negated phrases for four different negators (“no”, “not”, “doesn’t” and can’t) selected from all the gold cues. With a binary classification scheme (e.g., sentiment polarity detection), the negation of one class affirms the opposite class. This line of reasoning is not applicable when a more fine-grained classification scheme containing as many as 28 different classes is employed.

To a large extent, a negator neutralizes or alters the emotive meaning of an argument when the negator is attached to a primary or secondary indicator of an emotion category. For example, the occurrence of the primary indicator “happy” serves as a cue to affirm the

expression of *happiness* in a tweet whereas “not happy” nullifies the expression of *happiness*. It is clear that any instances of “not happy” can be interpreted as the tweeter not feeling the emotion *happiness*. Depending on how the negated phrase “not happy” is used in context, it can be used to affirm the expression of another emotion category as demonstrated by Example 4.25 (e.g., sadness or anger) or to diminish the emotive meaning in a tweet (Example 4.26).

**Example 4.25:** Here is the nip slip that happened yesterday on Good Morning America....and Nicki Minaj is not happy but ABC has... <http://fb.me/EVXID9e9> [**Anger**]

**Example 4.26:** "If We Date don't worry about me cheating. I'm with you for a reason. I want you and only you. And when I'm not happy, I'll let you know"-CO [**None**]

The negated phrase “not happy” in Example 4.26 merely invalidates the person feeling *happiness* and the tweet reveals no other clues as to whether the tweeter is describing *anger*, *sadness* or other emotions. Similarly, negators function the same way as emotion modifiers or neutralizers with arguments containing secondary indicators (e.g., “no good” and “doesn’t appeal”).

A more interesting finding is that negated phrases can also serve as units of meaning to affirm the presence of a particular emotion category. In such cases, the negator becomes a part of the lexical item associated to the emotion category and plays an important role to retain the meaning of the lexical item. One such example is the negated phrase “can’t wait”. As separate words, “can’t” and “wait” are not strong indicators of any emotion in particular. When joined together as a lexical item, “can’t wait” becomes a strong indicator of *excitement*. The negated phrase “doesn’t care” is another example demonstrating similar characteristic. This negated phrase is an important indicator of *indifference*. Important negated phrases that can be treated as lexical items to affirm an emotion category are listed in Table 4.32.

In the context of fine-grained emotion analysis, negators can affect the emotive meaning being expressed in tweets. Not all negators function to negate an emotion category. Negators can also be used to affirm an emotion category. Unlike sentiment polarity detection, it is not a

matter as simple as flipping the sentiment orientation of an argument attached to the negator. A negated phrase has to be interpreted according to the meaning of the lexical item it contains.

Negator	Affirm Emotion	Negate Emotion	
		[negator] + [primary term]	[negator] + [secondary term]
no	no desire no mercy no cure no way no better feeling than no greater feeling than no fucking way	no sympathy no doubt no love no regrets no thanks	no good no believe
not	not care not giving a fuck will not believe	not happy not be happy not satisfied not too pleased not liking not jealous not funny not concerned not looking forward to not particularly enjoyable not be trusted not trust	not good not so good not a good not as good not bad not so bad not ready not perfect not ok not looking great not worked not cool not sure not quite sure not know
doesn't	doesn't care doesn't make sense	doesn't like doesn't love	doesn't appeal
can't	can't believe can't wait can't stand can't put up can't deal can't keep up can't get enough can't tell if	can't trust	

Table 4.33: Negated phrases used to affirm and negate emotive meaning

#### 4.7.3.5 Part-of-speech (POS) Tags

Similar to words, each emotion category may demonstrate unique characteristics based on its POS tag composition. In this section, we examine if the POS tag composition in the emotion cues differ significantly across the 28 emotion categories and if any unique

characteristics can be associated with each emotion category. The POS tags are acquired using Stanford POS tagger in NLTK.

Category	NOUN	VERB	ADJ	ADV	PRON
All	27%	18%	10%	8%	9%
Admiration	29%	17%	15%	8%	7%
Amusement	33%	7%	7%	3%	3%
Anger	29%	19%	9%	9%	8%
Boredom	21%	26%	12%	13%	3%
Confidence	21%	26%	8%	8%	12%
Curiosity	12%	24%	9%	11%	18%
Desperation	24%	23%	9%	10%	9%
Doubt	17%	23%	8%	15%	11%
Excitement	21%	19%	8%	8%	4%
Exhaustion	28%	19%	10%	11%	6%
Fascination	28%	21%	13%	10%	5%
Fear	27%	22%	10%	9%	8%
Gratitude	27%	17%	6%	6%	16%
Happiness	32%	15%	15%	6%	5%
Hate	24%	24%	9%	7%	16%
Hope	26%	25%	11%	6%	10%
Indifference	27%	17%	9%	14%	8%
Inspiration	30%	23%	7%	6%	9%
Jealousy	23%	23%	11%	8%	14%
Longing	17%	24%	7%	10%	22%
Love	21%	17%	8%	6%	20%
Pride	21%	24%	10%	7%	5%
Regret	25%	23%	12%	9%	14%
Relaxed	30%	16%	8%	11%	6%
Sadness	28%	19%	11%	8%	8%
Shame	29%	18%	12%	10%	8%
Surprise	27%	17%	7%	12%	8%
Sympathy	32%	21%	6%	4%	10%

Table 4.34: POS tag composition based on content words in each emotion category

Most early work on the development of sentiment resources focus on adjectives as the primary indicator of emotion expressed in text (Hatzivassiloglou & McKeown, 1997; Taboada, Anthony, & Voll, 2006). Table 4.34 shows the percent occurrence of five POS tags based on the gold cues. The distribution of the POS tags in Table 4.34 indicates that emotion cues are not

only adjectives but also nouns, verbs, adverbs as well as pronouns associated with each emotion category. There is a higher portion of nouns and verbs than adjectives in the gold cues, suggesting that tweeters use more than just adjectives to express their emotions. Nouns and verbs occurring in the emotion cues show that they are equally as important as adjectives in determining the emotion category.

Several interesting observations can be made based on the composition of POS tags in the Table 4.34. For a majority of the emotion categories, the noun-verb-adjective ratio across the 28 emotion categories is similar in that nouns are the most frequent, followed by verbs and adjectives. Six emotion categories contain higher portion of verbs compared to nouns. These six emotion categories are *boredom*, *confidence*, *curiosity*, *doubt*, *longing* and *pride*. This suggests that these six categories are more likely to be expressed using action words. Two emotion categories, *longing* and *love*, display a higher percentage of pronouns whereas tweeters express *gratitude* and *sympathy* more commonly using nouns and verbs but with infrequent use of modifiers (i.e., adjectives and adverbs).

Overall, the POS tag composition varies across the 18 emotion categories but no stark differences are observed. The variation may be too small to help distinguish an emotion category from the others.

## **4.8 Summary: Salient Linguistic Cues for the Emotion Categories**

From the emotion cues marked by the annotators, we have identified a set of lexical terms that can serve as salient indicators for each of the 28 emotion categories. The lexical items include punctuation marks, emoticons, emojis, interjections, words, hashtags as well as collocations. Not all terms contribute equally as indicators of a particular emotion category. Naturally, terms that belong exclusively to a single emotion category have higher saliency than terms that occur in multiple categories. The importance of a term to a particular emotion



category is determined using a measure referred to as cue term weights. This weight is used to rank the importance of the terms in each emotion category.

The linguistic analysis presented in this section focuses on identifying patterns at the lexical level. It is possible to extract syntactic and semantic patterns associated with emotion category but that is beyond the scope of this thesis. All the lexical items are inserted into an emotion lexicon, which is used to inform features for our machine learning experiments.

## **4.9 Conclusion**

Part of the goal of this research is to better define the linguistic characteristics of a set of emotion categories representative of the range of emotions expressed in tweets so that computational models can take advantage of the information to improve classification performance. We have achieved the goal by first uncovering a set of 28 emotion categories from data and then define the linguistic characteristics of these categories from the emotion cues marked by annotators.

Humans can detect a wide range of emotions in tweets, all of which can be categorized into 28 emotion categories. This chapter addressed R1 by characterizing the 28 emotion categories that humans can detect in microblog text. We conclude that this set of 28 emotions categories offers the best compromise between informativeness and reliability. As summarized in Section 4.4 and Table 4.17, humans can recognize 6 emotion categories with high reliability, 17 categories with moderate reliability and 5 categories with low reliability.

We addressed R2 by presenting the salient linguistic cues associated with each emotion. We have shown that there are significant differences in the language used to express the different emotion categories (Section 4.7). Specifically, we have identified the lexical items that can be used as features for a machine learning classifier. The lexical items associated with each emotion category include punctuation marks, emojis and emoticons (Section 4.7.3.1), lemmas (Section 4.7.3.2) and hashtags (Section 4.7.3.3). We also show that collocations (i.e.,

bigrams, trigrams and negations) are important in making a distinction between the emotion categories (Section 4.7.3.4). Finally, we did not observe any stark differences in the POS composition of the emotion cues between the emotion categories (Section 4.7.3.5).

## Chapter 5: Machine Learning Results

The main goal of this chapter is to explore the performance of machine learning techniques for automatically identifying the expression of emotion in tweets. These automatic classification experiments use the EmoTweet-28 corpus to address the third and fourth research questions:

- *R3: Do the salient cues humans associate with each emotion serve as better features for machine learning classification of emotion in text?*
- *R4: How do current machine learning techniques perform on more fine-grained categories of emotion?*

We present the results of the machine learning experiments conducted using our carefully hand-crafted corpus, EmoTweet-28, in three parts. In the first part (Task 1), we conduct experiments to identify classifiers and parameter settings that perform consistently well for this problem space. One purpose is to determine if the machine learning techniques currently used for sentiment classification can be applied to such fine-grained set of emotion classes. Another purpose of Task 1 is to identify reasonable base classifiers to be used for more advanced experimentation on the features. The second part (Task 2) compares classifier performance on three different feature groups: corpus-based features, lexicon-based features and cue-based features. We empirically test if the salient cues humans associate with each emotion category (i.e., emotion cues) serve as better features for fine-grained emotion classification than corpus or lexicon based features. The third task (Task 3) investigates the effect of varying training samples on classifier performance.

## 5.1 Definition of Terms and Evaluation Metrics

We use the term “classifier” or “model” interchangeably to refer to a machine learning classifier. Ground truth is defined by the EmoTweet-28 corpus described in Chapter 4. The corpus was developed based on discussion and review by the expert annotators. Each tweet is assigned one or more emotion classes as gold labels. A match between the label predicted by an automatic classifier and the gold label is considered a successful prediction. We use the term “category labels” to refer to the 28 emotion category labels (excluding no emotion). The term “class labels” encompasses the 28 emotion categories plus an additional category for no emotion.

We use F1, the harmonic mean of precision and recall, as the primary measure to assess the performance of the classifiers. We also present recall and precision results when needed. For each binary classifier, precision, recall and F1 are defined as:

$$Precision = \frac{\# \text{ of correct positive predictions}}{\# \text{ of positive examples}}$$

$$Recall = \frac{\# \text{ of correct positive predictions}}{\# \text{ of positive predictions}}$$

$$F1 = 2 * \frac{(Precision * Recall)}{(Precision + Recall)}$$

We compute both the macro and micro averages to evaluate the overall performance of a classifier across all classes. Macro average gives equal weight to each class and is computed by simply averaging over all the classes. Micro average, which gives equal weight to each instance, is an average over all instances.

We frame the classification problem in two ways. To simplify the classification problem, we first frame it as a multi-class classification task. Using this framing, a tweet  $x$  is assigned with only one of the 29 labels (28 emotion categories and no emotion). The multi-class classifier does not handle tweets with multiple emotions so we kept the primary label assigned to the

tweet with multiple emotions and ignored all other labels. Tweets with more than one emotion category label make up only a small portion of the corpus (8%) so we expect the performance of the multi-class classifier to be a close approximation to a real classifier expected to predict the set of emotions for a tweet. To evaluate a multi-class classifier, accuracy, precision, recall and F1 are computed based on classifier performance on all class labels including none (29 classes). We discuss two sets of preliminary experiments using the multi-class setup in Section 5.2.1 and Section 5.2.2.

All other experiments in this chapter are based on multi-label classification. In a multi-label classification task, the classifier assigns one or more emotion categories to a tweet. To handle tweets with multiple emotions appropriately, we create a binary classifier for each emotion category. In this setup, a classification model consists of 28 binary classifiers. Given a tweet  $x$ , we measure the performance of the machine learning model in predicting whether or not the tweet expresses emotion  $y$ . For each binary classifier, we measure the precision, recall and F1 of the positive class. We do not take into account the performance measures associated with the negative class since the binary classifiers most often yield very high scores for the negative class due to high class imbalance. To compute a single aggregate measure that combines the individual measures of each binary classifier, we compute the macro average by taking an average of the measures over 28 binary classifiers and micro average based on the true positives, false positives and false negatives from each of the classifiers.

## 5.2 Task 1: Classifier-related Experiments

Task 1 uses only the 5,553 tweet dataset generated in Phase 1 (P1). Results reported in this section are based on the *multi-label* experiment setup (i.e., a binary classifier is trained for each emotion category). We only discuss results from our preliminary experiments using the *multi-class* setup in Section 5.1.1 (baseline comparison) and Section 5.2.2 (tiered model). The tiered model is specifically materialized through the *multi-class* setup.

Based on the results of preliminary machine learning experiments reported in Chapter 3, we settled on two classifiers that work consistently well for this problem space: Bayesian networks (BayesNet in Weka) and support vector machines using sequential minimal optimization (SMO in Weka). Both classifiers use the same basic feature set: unigrams that are stemmed and lowercased and that occur three or more times in the corpus. The default SMO and BayesNet parameters in Weka yield good performance. We did not attempt to tune classifier parameter settings for individual experiments but, preferred, instead, to use consistent parameter settings for all of the experiments. These two classifiers serve as the basic models for the experiments in Task 1. All the classifiers in Task 1 are evaluated using 10-fold cross validation.

### 5.2.1 Comparison with Baseline

Three baselines are first established as the basis of comparison for all other future classifiers. The three baselines are: majority-class, random and OneR (one-rule classifier).

- Majority-class baseline: The majority-class baseline simply assigns the majority class to each tweet. Weka's ZeroR implements this classifier.
- Random baseline: The random baseline guesses whether or not a tweet  $x$  contains emotion  $y$ . We adapted the approach used by Mohammad (2012) to compute the accuracy, precision, recall and F1 for the random baseline.
- OneR: OneR is a simple classifier that uses single feature with minimum error for classification.

We first compare our basic models (SMO and BayesNet) with three baselines in the *multi-label* setup. Recall that the overall performance of the machine learning model in the *multi-label* setup is reflected by average measures across all 28 binary classifiers.

In the *multi-label* setup, the majority class is always the negative class (i.e., not *emotion X*) for each of the binary classifiers. The negative examples make up over 90% in each binary

classifier. If a binary classifier predicts all instances as negative, it would correctly classify over 90% of the instances. Based on the average measures from the 28 binary classifiers, the model achieves high average accuracy but the average precision, recall and F-measure for the positive class are always zero. While average accuracy represents an important lower bound that “good” classifiers should surpass, it is not a particularly good measure of classifier performance in our case. For example, since the emotion category *fear* occurs in 1.5% of tweets, the majority class classifier achieves an accuracy of 98.5% while not classifying a single positive example correctly.

Classifier	Accuracy	Precision	Recall	F-Measure
<b><i>multi-label</i></b>				
Majority baseline: ZeroR	97.6	0	0	0
Random baseline	3.8	0.04	0.50	0.07
OneR	97.9	0.75	0.23	0.35
BayesNet	98.1	0.75	0.37	0.50
SMO	97.8	0.58	0.41	0.48
<b><i>multi-class-single</i></b>				
Majority baseline: ZeroR	47.4	0.23	0.47	0.31
OneR	49.8	0.26	0.50	0.34
BayesNet	60.1	0.54	0.60	0.51
SMO	58.9	0.57	0.59	0.57
<b><i>multi-class-binary</i></b>				
Majority baseline: ZeroR	47.4	0.23	0.47	0.31
OneR	51.7	0.56	0.52	0.46
BayesNet	63.0	0.60	0.63	0.57
SMO	48.9	0.61	0.49	0.53

Table 5.1: Comparison between basic models and baselines for emotion classification

The random baseline achieves an accuracy of 3.8% as shown in Table 5.1. The performance of the random baseline is far worse than the base models. In terms of accuracy, the OneR classifier performed slightly worse than BayesNet but slightly better than SMO. OneR sets a high bar on accuracy, making it difficult for “good classifiers” to surpass but, accuracy is not sensitive to what we want to measure, which is the ability for the classifier to correctly identify positive examples. The extremely high proportion of negative examples means that

accuracy is high, making it difficult to delineate any performance improvement that is related to the positive examples in an emotion category. Precision, recall and F1 for the positive class provide a more meaningful picture of classifier performance on the positive examples. In terms of F1, both BayesNet and SMO outperform OneR.

In the *multi-label* setup, Table 5.1 shows the average accuracy across 28 binary classifiers. Note that average accuracy across the 28 binary classifiers is a good characterization of the accuracy of the classifier set, but it does not reflect the expected performance of a combined model used to predict the set of emotions that are expressed in a given tweet. The accuracy of classifiers in the *multi-class* setup is a closer approximation of the performance we can expect from the combined model used to predict a set of emotions that are expressed in a given tweet (i.e., 28 emotion categories and no emotion).

In terms of accuracy, SMO and BayesNet outperform all three baselines in both multi-class-single and multi-class-binary. BayesNet correctly predicts roughly 60% of the instances while SMO correctly predicts roughly 50%. In terms of F1, SMO and BayesNet also exceed the performance of all three baselines.

## 5.2.2 Tiered Model Results

The purpose of the tiered model experiments is two-fold: 1) to examine classifier performance on three different levels of granularity, and 2) to investigate if flat classification (single decision to assign final label) or hierarchical classification (assign labels in stages) is more appropriate for this problem space. The *multi-class-single* and *multi-class-binary* setups are used to make it easier to compare across different levels of granularity and stages.

### 5.2.2.1 Levels of Granularity

The first set of experiments compares performance of classifiers with fine-grained versus coarser-grained class structures. The three levels of granularity are: 1) emotion



presence/absence (2 classes), 2) emotion valence (5 classes) and, 3) emotion category (28 classes).

From Table 5.2, it is evident that SMO and BayesNet perform significantly better than the majority baseline (ZeroR) across all three levels of granularity in a flat classification implementation. Comparing across the three levels of granularity, better performance is observed when there are fewer classes (coarser-grained). For example, a classifier trained to distinguish between emotion presence or absence (2 class labels) yields higher performance than a classifier trained to distinguish between the emotion categories (29 class labels). There are two main reasons why the classifiers perform better on coarser-grained classification schemes using this data set. One reason is that coarser-grained classification has fewer classes, which makes the classification task easier. Second, the distribution of classes is more balanced for emotion presence and emotion valence when compared to emotion category. Therefore, overall classifier performance is less affected by sparse classes with few training examples.

Level	SMO				BayesNet				ZeroR			
	A	P	R	F1	A	P	R	F1	A	P	R	F1
Emotion Presence: hasEmotion, None	72.7	0.73	0.73	0.73	72.2	0.73	0.72	0.72	52.6	0.28	0.53	0.36
Emotion Valence: Pos, Neg, Neu, Multiple, None	65.5	0.63	0.64	0.63	67.0	0.65	0.67	0.65	47.4	0.23	0.47	0.31
<b>Emotion Category</b>												
multi-class-single: 28-Emo-Cat, None	58.9	0.57	0.59	0.57	60.1	0.54	0.60	0.51	47.4	0.23	0.47	0.31
multi-class-binary: 28-Emo-Cat, None	48.9	0.61	0.49	0.53	63.0	0.60	0.63	0.57	47.4	0.23	0.47	0.31

Table 5.2: Accuracy (A), precision (P), recall (R) and F1 across classification schemes with different levels of granularity

The drop in classifier performance from coarser to finer levels of granularity is gradual. Note that the performance of a classifier trained to classify 29 classes (28 emotion categories and no emotion) is not a great deal worse than a classifier dealing with fewer classes (2 or 5). A

closer analysis of the F1 per class shows that the classifiers are able to correctly predict some classes better than the others. For example, SMO and BayesNet achieve F1 greater than 0.7 for the *gratitude* class. The performance measures in Table 5.2 are micro averages across all 28 emotion categories.

### 5.2.2.2 Flat versus Hierarchical Classification

The performance of SMO and BayesNet are comparable in both *multi-single-class* and *multi-class-binary* as shown in Table 5.2. However, the *multi-class-binary* setup is a closer reflection of our intent to handle multi-label tweets. We selected BayesNet to further examine if it is worth pursuing a two-tiered approach. In the two-tiered model, the emotion presence classifier in tier 1 (T1) is employed as a first pass to filter out instances that contain emotion from those that do not. The emotion category classifier in tier 2 (T2) is then used to determine the specific label among 28 emotion categories that is associated with an emotion instance (see Figure 5.1b).

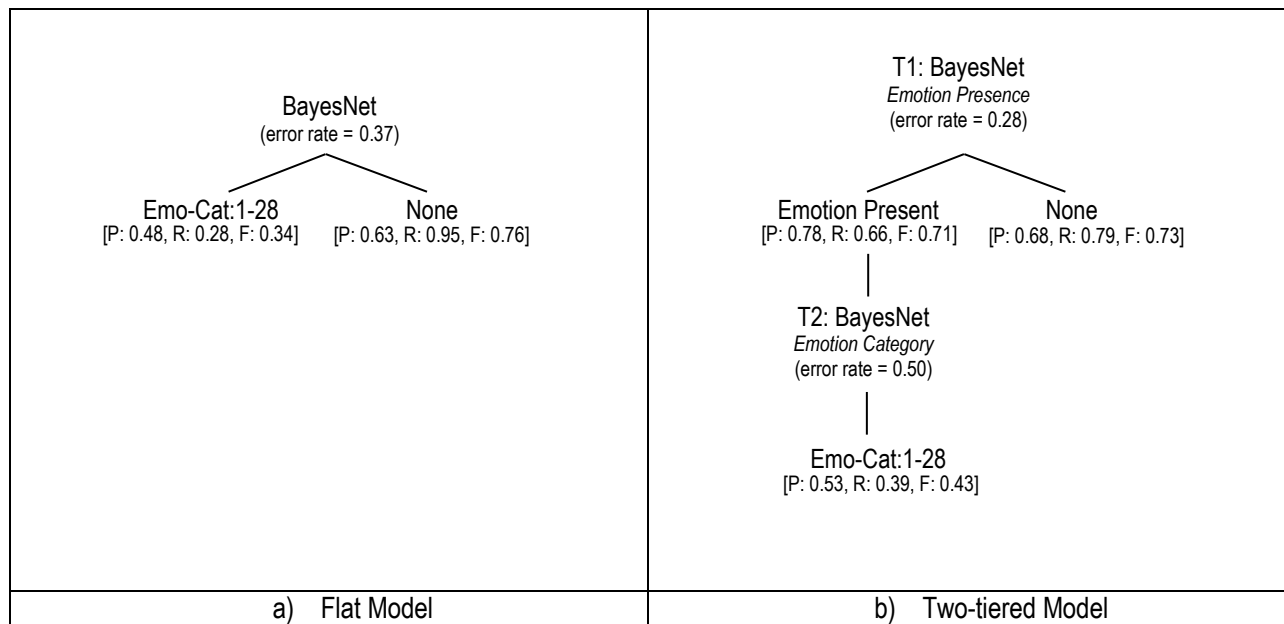


Figure 5.1: Comparing BayesNet performance in flat and two-tiered classification

Two measures are used as the basis of comparison between the flat and two-tiered classification models. First, we inspect the error rates of the classifiers in the overall model. The flat classification model (*multi-class-binary* BayesNet: 28 emotion categories and no emotion) has an overall error rate of 37%. As for the two-tiered model, overall performance relies heavily on the initial classification at the higher levels. Errors in the higher level classifiers are propagated to the lower level classifiers so any positive example that was incorrectly classified in the first tier is not available for the second tier classifier. The T1 classifier has an error rate of 28%.

To establish an upper bound on the level of performance that could be expected from T2 we tested a multiclass classifier using only the instances that express emotion. Essentially, this models the case where the T1 classifier is perfect – it is able to pass exactly the correct set to T2. Even assuming a perfect T1 classifier (error rate = 0%), the T2 classifier has an error rate of 50% (though this is measured only on the positive examples). Given that the flat classification model has an error rate of 37%, it is unlikely that the two tier model could perform better than the flat model.

Second, we examine the performance of the two approaches in predicting the no emotion class. Note that the recall for the “none” (i.e., no emotion) label is significantly higher in the flat classification model (recall = 0.95) compared with the T1 classifier in the two-tiered model (recall = 0.79). The flat classification model produces a greater number of true positives (TP = 2486) for the “none” class in contrast to the T1 classifier (TP = 2085) in the two-tiered model. Therefore, we continue other experiments using the flat classification structure.

### 5.2.3 Ensemble Methods

We utilized Weka’s *meta* classifiers to explore the effect of bagging (*meta: Bagging*), boosting (*meta: AdaBoost*), voting (*meta: Vote*) and stacking (*meta: Stacking*) (Section 3.4) on the basic models. Based on the results in Table 5.3, bagging can produce higher precision at

the expense of recall. Boosting on the other hand offers very little overall performance improvement to the basic models but does help to slightly increase the performance of a few emotion categories where the basic models are not able to identify any true positives ( $F1 = 0$ ).

An ensemble of SMO and BayesNet classifiers, where the individual decisions are combined based on majority voting, generated the highest F1 among the classifiers presented in Table 5.3. Basically, the two classifiers complement each other, in which the strengths of one classifier can be used to compensate for the weaknesses of another. The ensemble model performs at par with the basic SMO model in terms of recall. The precision is higher than that of the basic SMO classifier but not as high as the basic BayesNet model. Using logistic regression to classify decisions from both the SMO and BayesNet classifiers in the stacked model also yielded higher F1 compared to the basic models alone. The precision of the stacked model is better than each basic model in the classifier mix but the opposite is observed for recall.

Classifier	Precision	Recall	F1
<b>SMO</b>			
SMO	0.578	0.406	0.477
AdaBoost: SMO	0.443	<b>0.417</b>	0.430
Bagging: SMO	0.705	0.345	0.464
<b>BayesNet</b>			
BayesNet	0.754	0.368	0.495
AdaBoost: BayesNet	0.533	0.313	0.394
Bagging: BayesNet	<b>0.792</b>	0.329	0.464
<b>Multi-Classifier</b>			
Voting: SMO + BayesNet	0.669	0.406	<b>0.505</b>
Stacking: SMO + BayesNet	0.782	0.366	0.498

Table 5.3: Overall precision, recall and F1 using boosting, bagging, voting and stacking

Overall, ensemble methods can be used to improve precision of the basic models but do not contribute significantly to the improvement of recall. A closer analysis of F1 for each emotion category reveal that the basic models, especially BayesNet, still perform better in terms of both precision and recall for over half the emotion categories in the set. As shown in Table 5.4, boosting only significantly enhances the performance of *exhaustion* in SMO and *admiration*,

*jealousy* and *longing* in BayesNet. Other than that, boosting and bagging decrease the performance of a majority of the emotion categories.

Category	SMO	AdaBoost:SMO		Bagging:SMO		BN	AdaBoost:BN		Bagging:BN	
	S:F1	F1	Diff(S)	F1	Diff(S)	B:F1	F1	Diff(B)	F1	Diff(B)
Admiration	<b>0.289</b>	0.280	-0.01	0.228	-0.06	0.115	<b>0.253</b>	0.14	0.081	-0.03
Amusement	<b>0.628</b>	0.555	-0.07	0.604	-0.02	<b>0.761</b>	0.620	-0.14	0.751	-0.01
Anger	0.231	<b>0.248</b>	0.02	0.214	-0.02	0.063	<b>0.066</b>	0.00	0.063	0.00
Boredom	0.133	0.133	0.00	<b>0.154</b>	0.02	<b>0.667</b>	0.000	-0.67	0.353	-0.31
Confidence	<b>0.000</b>	0.000	0.00	0.000	0.00	<b>0.000</b>	0.000	0.00	0.000	0.00
Curiosity	0.586	0.586	0.00	<b>0.600</b>	0.01	<b>0.615</b>	0.190	-0.42	0.065	-0.55
Desperation	<b>0.222</b>	0.222	0.00	0.000	-0.22	<b>0.800</b>	0.200	-0.60	0.200	-0.60
Doubt	0.033	<b>0.104</b>	0.07	0.000	-0.03	0.000	<b>0.054</b>	0.05	0.000	0.00
Excitement	0.431	0.419	-0.01	<b>0.441</b>	0.01	<b>0.444</b>	0.392	-0.05	0.427	-0.02
Exhaustion	0.000	<b>0.200</b>	0.20	0.000	0.00	<b>0.000</b>	0.000	0.00	0.000	0.00
Fascination	<b>0.354</b>	0.327	-0.03	0.257	-0.10	<b>0.465</b>	0.184	-0.28	0.070	-0.39
Fear	0.113	<b>0.137</b>	0.02	0.070	-0.04	<b>0.302</b>	0.033	-0.27	0.000	-0.30
Gratitude	<b>0.917</b>	0.878	-0.04	0.900	-0.02	<b>0.899</b>	0.772	-0.13	0.895	0.00
Happiness	0.542	0.500	-0.04	<b>0.550</b>	0.01	<b>0.524</b>	0.519	-0.01	0.514	-0.01
Hate	<b>0.549</b>	0.537	-0.01	0.435	-0.11	<b>0.602</b>	0.255	-0.35	0.566	-0.04
Hope	0.560	0.511	-0.05	<b>0.584</b>	0.02	<b>0.622</b>	0.415	-0.21	0.613	-0.01
Indifference	0.176	<b>0.182</b>	0.01	0.067	-0.11	<b>0.293</b>	0.098	-0.20	0.000	-0.29
Inspiration	<b>0.722</b>	0.722	0.00	0.500	-0.22	0.722	0.067	-0.66	<b>0.743</b>	0.02
Jealousy	<b>0.571</b>	0.571	0.00	0.333	-0.24	0.000	<b>0.571</b>	0.57	0.000	0.00
Longing	0.316	<b>0.347</b>	0.03	0.085	-0.23	0.000	<b>0.203</b>	0.20	0.000	0.00
Love	<b>0.522</b>	0.408	-0.11	0.492	-0.03	0.529	0.428	-0.10	<b>0.546</b>	0.02
Pride	0.725	0.678	-0.05	<b>0.752</b>	0.03	0.760	0.471	-0.29	<b>0.770</b>	0.01
Regret	0.138	0.139	0.00	<b>0.148</b>	0.01	<b>0.550</b>	0.208	-0.34	0.382	-0.17
Relaxed	0.059	0.043	-0.02	<b>0.067</b>	0.01	0.000	<b>0.057</b>	0.06	0.000	0.00
Sadness	<b>0.412</b>	0.317	-0.09	0.325	-0.09	<b>0.439</b>	0.364	-0.07	0.360	-0.08
Shame	<b>0.378</b>	0.368	-0.01	0.069	-0.31	<b>0.368</b>	0.167	-0.20	0.000	-0.37
Surprise	<b>0.284</b>	0.277	-0.01	0.150	-0.13	0.167	<b>0.222</b>	0.06	0.119	-0.05
Sympathy	<b>0.519</b>	0.519	0.00	0.391	-0.13	<b>0.604</b>	0.436	-0.17	0.560	-0.04
Macro-F1	<b>0.372</b>	0.365	-0.01	0.300	-0.07	<b>0.404</b>	0.259	-0.15	0.289	-0.12
Micro-F1	<b>0.477</b>	0.430	-0.05	0.464	-0.01	<b>0.495</b>	0.394	-0.10	0.464	-0.03

Table 5.4: F1 of each emotion category based on boosting and bagging

Three outcomes are observed when an ensemble method is used to combine SMO and BayesNet. First, it is possible for the ensemble model to show an increase in performance compared to each of its member classifiers respectively such as in the case of *surprise*.

Second, the ensemble model may show an increase in performance from one of its members but also a decrease in performance from the other member (e.g., *admiration*). Third, it is also possible for the ensemble model to perform worse than both its members (e.g., *love* in voting as well as *indifference* and *inspiration* in stacking). The second outcome is more common while the third one is rare as shown in Table 5.5.

Category	SMO(S)	BN(B)	Voting:S+B	Diff(S)	Diff(B)	Stacking:S+B	Diff(S)	Diff(B)
Admiration	<b>0.289</b>	0.115	0.234	-0.06	0.12	0.200	-0.09	0.08
Amusement	0.628	<b>0.761</b>	0.710	0.08	-0.05	0.750	0.12	-0.01
Anger	<b>0.231</b>	0.063	0.170	-0.06	0.11	0.063	-0.17	0.00
Boredom	0.133	<b>0.667</b>	0.600	0.47	-0.07	0.609	0.48	-0.06
Confidence	0.000	0.000	0.000	0.00	0.00	0.000	0.00	0.00
Curiosity	0.586	0.615	0.635	0.05	0.02	<b>0.646</b>	0.06	0.03
Desperation	0.222	0.800	<b>0.857</b>	0.63	0.06	0.800	0.58	0.00
Doubt	0.033	0.000	<b>0.036</b>	0.00	0.04	0.000	-0.03	0.00
Excitement	0.431	0.444	<b>0.461</b>	0.03	0.02	0.437	0.01	-0.01
Exhaustion	<b>0.000</b>	<b>0.000</b>	0.000	0.00	0.00	0.000	0.00	0.00
Fascination	0.354	0.465	<b>0.530</b>	0.18	0.07	0.465	0.11	0.00
Fear	0.113	<b>0.302</b>	0.270	0.16	-0.03	0.024	-0.09	-0.28
Gratitude	<b>0.917</b>	0.899	0.915	0.00	0.02	0.907	-0.01	0.01
Happiness	0.542	0.524	0.538	0.00	0.01	<b>0.543</b>	0.00	0.02
Hate	0.549	<b>0.602</b>	0.588	0.04	-0.01	0.586	0.04	-0.02
Hope	0.560	0.622	0.610	0.05	-0.01	<b>0.625</b>	0.06	0.00
Indifference	0.176	0.293	<b>0.350</b>	0.17	0.06	0.114	-0.06	-0.18
Inspiration	0.722	0.722	<b>0.743</b>	0.02	0.02	0.706	-0.02	-0.02
Jealousy	<b>0.571</b>	0.000	0.571	0.00	0.57	0.000	-0.57	0.00
Longing	<b>0.316</b>	0.000	0.275	-0.04	0.27	0.204	-0.11	0.20
Love	0.522	0.529	0.508	-0.01	-0.02	<b>0.541</b>	0.02	0.01
Pride	0.725	<b>0.760</b>	0.731	0.01	-0.03	0.752	0.03	-0.01
Regret	0.138	<b>0.550</b>	0.507	0.37	-0.04	0.494	0.36	-0.06
Relaxed	<b>0.059</b>	0.000	0.000	-0.06	0.00	0.000	-0.06	0.00
Sadness	0.412	0.439	<b>0.479</b>	0.07	0.04	0.405	-0.01	-0.03
Shame	0.378	0.368	0.389	0.01	0.02	<b>0.450</b>	0.07	0.08
Surprise	0.284	0.167	<b>0.346</b>	0.06	0.18	0.309	0.03	0.14
Sympathy	0.519	0.604	<b>0.615</b>	0.10	0.01	0.549	0.03	-0.05
Macro-F1	0.372	0.404	<b>0.452</b>	0.08	0.05	0.399	0.03	0.00
Micro-F1	0.477	0.495	<b>0.505</b>	0.03	0.01	0.498	0.02	0.00

Table 5.5: F1 of each emotion category based on voting and stacking

The ensemble methods we examined show mixed results. Voting and stacking improve classifier performance slightly but the gain is not significant enough to outweigh the significantly higher computational costs for training.

#### **5.2.4 Summary: Classifier-related Experiments**

Our experiments in Task 1 conclude that using the basic SMO and BayesNet classifiers in a flat classification structure produces consistently good precision and recall. The overall performance of these two classifiers is better than the three baselines. Therefore, we continued our experiments in Task 2 and Task 3 using both SMO and BayesNet.

We have also justified the reason for choosing a flat, rather than tiered, classification structure. Finally, we found that the ensemble methods help boost the precision but make very little to no contribution to the recall observed with the basic SMO and BayesNet models. Two of the ensemble methods, voting and stacking, yielded only slight performance improvements compared to the basic models.

### **5.3 Task 2: Feature-related Experiments**

Our experiments for Task 1 used only data from P1 (P1: 5,553 tweets). The remainder of the experiments described here use the full EmoTweet-28 corpus combining gold standard data annotated in P1 and P2 (P1 + P2: 15,553 tweets).

Classifiers using corpus-based and lexicon-based features are evaluated using 10-fold cross validation. These features are selected statistically (corpus-based) or using resources that are not derived from the corpus (lexicon-based). The cue-based features, however, were selected by human annotators and represent judgments about what is significant in each tweet. To ensure that these human judgments did not bias the machine learning experiments, we split the corpus into three parts (development, training, and test splits). Cue-based features were selected only from the development set (30% of the corpus, 4670 tweets). These features were

then used to train classifiers on a training set (40% of the corpus, 6222 tweets) which were then evaluated using a test set (30% of the corpus, 4661 tweets). No part of the training or test data involves judgments based on that data.

### 5.3.1 Comparison of P1 and P2 Data

We first compare the performance of the basic models across P1, P2 and P1 + P2. The basic models in this section are slightly enhanced in that we normalized the hyperlinks (URLs)<sup>23</sup> in the tweets and include a feature to indicate the presence or absence of URL in a tweet. This is based on the intuition that many tweets that contain URLs tend to not contain any emotional content.

With the possible exception of sparse emotion categories, the behavior of the basic models across P1 and P2 was expected to be similar since the gold standard data is developed using the same annotation scheme. For the sparse emotion categories, it may be possible that there are simply too few positive examples in P1 so the classifiers lack representation of certain important aspect of the positive class. The greater amount of training data in P2 or P1 + P2 should reduce this problem.

The precision, recall and F1 for both SMO and BayesNet across P1, P2 and P1 + P2 are shown in Table 5.6 (precision), Table 5.7 (recall) and Table 5.8 (F1). A general upward trend in micro-precision, micro-recall, and micro-F1 are observed across the three data sets. The macro averages also follow this upward trend. The only exception is average precision for BayesNet (see Table 5.6).

A larger data set provides classifiers with more examples to learn from, and thus helps improve the performance of the classifiers in detecting the emotion category. For instance, classifiers in P1 fail to classify any instances correctly for the low frequency emotion categories such as *confidence*, *exhaustion*, *jealousy* and *relaxed*. However, we start to see some

---

<sup>23</sup> URLs in the tweets are normalized to <http://URL>.



improvement in the classifier performance as more data becomes available for training in P2 and P1 + P2.

<b>Precision</b>	<b>SMO</b>			<b>BayesNet</b>		
<b>Category</b>	<b>P1</b>	<b>P2</b>	<b>P1+P2</b>	<b>P1</b>	<b>P2</b>	<b>P1+P2</b>
Admiration	0.417	0.328	0.370	0.179	0.684	0.543
Amusement	0.744	0.888	0.869	0.849	0.940	0.899
Anger	0.288	0.495	0.478	0.336	0.603	0.548
Boredom	0.400	0.714	0.818	0.778	0.882	0.880
Confidence	0.000	0.286	0.303	0.000	0.750	0.533
Curiosity	0.586	0.591	0.638	0.593	0.500	0.859
Desperation	1.000	0.417	0.500	0.833	0.778	0.813
Doubt	0.125	0.256	0.269	0.000	0.486	0.291
Excitement	0.457	0.675	0.655	0.630	0.707	0.687
Exhaustion	0.000	0.706	0.611	0.000	0.500	0.875
Fascination	0.417	0.587	0.553	0.600	0.688	0.667
Fear	0.240	0.556	0.491	0.444	0.854	0.709
Gratitude	0.943	0.913	0.928	0.917	0.942	0.923
Happiness	0.589	0.596	0.622	0.705	0.725	0.764
Hate	0.778	0.812	0.788	0.769	0.897	0.861
Hope	0.660	0.781	0.781	0.848	0.820	0.816
Indifference	0.500	0.308	0.235	0.455	0.667	0.375
Inspiration	0.923	0.731	0.816	0.867	0.905	0.917
Jealousy	1.000	0.846	0.765	0.000	0.929	0.938
Longing	0.545	0.487	0.529	0.100	0.538	0.462
Love	0.608	0.645	0.659	0.750	0.704	0.698
Pride	0.817	0.907	0.862	0.864	0.943	0.923
Regret	0.500	0.571	0.514	0.655	0.631	0.595
Relaxed	0.200	0.550	0.737	0.000	0.692	0.789
Sadness	0.609	0.612	0.650	0.726	0.729	0.724
Shame	0.600	0.545	0.622	0.643	0.900	0.804
Surprise	0.342	0.627	0.556	0.591	0.632	0.590
Sympathy	0.813	0.625	0.705	0.889	0.673	0.595
<b>Macro-avg</b>	<b>0.539</b>	<b>0.609</b>	<b>0.619</b>	<b>0.536</b>	<b>0.739</b>	<b>0.717</b>
<b>Micro-avg</b>	<b>0.580</b>	<b>0.647</b>	<b>0.656</b>	<b>0.718</b>	<b>0.761</b>	<b>0.741</b>

Table 5.6: Precision of basic SMO and BayesNet classifiers across P1, P2 and P1+P2

SMO generally produces higher recall while BayesNet yields higher precision in overall performance. Based on F1 of each emotion category, the positive correlation between corpus

size and the classifier performance is more apparent for SMO. Of the 28 emotion categories, the best performing category in both SMO and BayesNet is *gratitude* while the worst performing category is *indifference* in SMO and *admiration* in BayesNet.

Recall	SMO			BayesNet		
Category	P1	P2	P1+P2	P1	P2	P1 + P2
Admiration	0.190	0.155	0.201	0.032	0.053	0.062
Amusement	0.515	0.617	0.645	0.688	0.740	0.759
Anger	0.203	0.346	0.321	0.081	0.291	0.261
Boredom	0.167	0.417	0.375	0.583	0.417	0.458
Confidence	0.000	0.088	0.091	0.000	0.099	0.073
Curiosity	0.567	0.413	0.548	0.533	0.222	0.591
Desperation	0.125	0.100	0.069	0.625	0.140	0.224
Doubt	0.020	0.102	0.089	0.000	0.167	0.203
Excitement	0.377	0.463	0.474	0.385	0.435	0.452
Exhaustion	0.000	0.308	0.224	0.000	0.077	0.143
Fascination	0.185	0.360	0.309	0.278	0.293	0.314
Fear	0.078	0.216	0.230	0.104	0.253	0.234
Gratitude	0.905	0.877	0.914	0.896	0.867	0.879
Happiness	0.500	0.477	0.506	0.424	0.339	0.401
Hate	0.444	0.535	0.542	0.476	0.543	0.547
Hope	0.508	0.564	0.580	0.476	0.597	0.594
Indifference	0.071	0.100	0.059	0.179	0.050	0.088
Inspiration	0.571	0.352	0.413	0.619	0.352	0.440
Jealousy	0.400	0.379	0.382	0.000	0.448	0.441
Longing	0.146	0.238	0.306	0.024	0.175	0.198
Love	0.444	0.538	0.519	0.397	0.502	0.490
Pride	0.682	0.688	0.676	0.671	0.648	0.676
Regret	0.102	0.308	0.242	0.388	0.394	0.431
Relaxed	0.038	0.216	0.182	0.000	0.176	0.195
Sadness	0.335	0.444	0.461	0.335	0.421	0.468
Shame	0.231	0.281	0.311	0.346	0.422	0.411
Surprise	0.140	0.301	0.278	0.140	0.208	0.222
Sympathy	0.371	0.379	0.426	0.457	0.500	0.495
<b>Macro-avg</b>	<b>0.297</b>	<b>0.366</b>	<b>0.370</b>	<b>0.326</b>	<b>0.351</b>	<b>0.384</b>
<b>Micro-avg</b>	<b>0.400</b>	<b>0.440</b>	<b>0.455</b>	<b>0.376</b>	<b>0.407</b>	<b>0.431</b>

Table 5.7: Recall of basic SMO and BayesNet classifiers across P1, P2 and P1+P2

<b>F1</b>	<b>SMO</b>			<b>BayesNet</b>		
<b>Category</b>	<b>P1</b>	<b>P2</b>	<b>P1+P2</b>	<b>P1</b>	<b>P2</b>	<b>P1 + P2</b>
Admiration	0.261	0.211	0.260	0.054	0.098	0.111
Amusement	0.608	0.728	0.741	0.760	0.828	0.823
Anger	0.238	0.407	0.384	0.131	0.392	0.354
Boredom	0.235	0.526	0.514	0.667	0.566	0.603
Confidence	0.000	0.134	0.140	0.000	0.175	0.128
Curiosity	0.576	0.486	0.590	0.561	0.308	0.701
Desperation	0.222	0.161	0.121	0.714	0.237	0.351
Doubt	0.034	0.146	0.133	0.000	0.248	0.239
Excitement	0.413	0.549	0.550	0.478	0.538	0.545
Exhaustion	0.000	0.429	0.328	0.000	0.133	0.246
Fascination	0.256	0.446	0.396	0.380	0.411	0.427
Fear	0.118	0.311	0.313	0.168	0.390	0.352
Gratitude	0.924	0.895	0.921	0.906	0.903	0.901
Happiness	0.541	0.530	0.558	0.530	0.462	0.526
Hate	0.566	0.645	0.642	0.588	0.676	0.669
Hope	0.574	0.655	0.666	0.610	0.691	0.687
Indifference	0.125	0.151	0.094	0.256	0.093	0.143
Inspiration	0.706	0.475	0.549	0.722	0.507	0.595
Jealousy	0.571	0.524	0.510	0.000	0.605	0.600
Longing	0.231	0.319	0.387	0.039	0.264	0.277
Love	0.514	0.587	0.581	0.520	0.586	0.576
Pride	0.744	0.782	0.758	0.755	0.769	0.780
Regret	0.169	0.400	0.329	0.487	0.485	0.500
Relaxed	0.065	0.310	0.292	0.000	0.281	0.313
Sadness	0.433	0.514	0.539	0.459	0.534	0.569
Shame	0.333	0.371	0.415	0.450	0.574	0.544
Surprise	0.198	0.406	0.371	0.226	0.313	0.322
Sympathy	0.510	0.472	0.531	0.604	0.574	0.541
<b>Macro-avg</b>	<b>0.363</b>	<b>0.449</b>	<b>0.450</b>	<b>0.395</b>	<b>0.452</b>	<b>0.479</b>
<b>Micro-avg</b>	<b>0.474</b>	<b>0.524</b>	<b>0.537</b>	<b>0.493</b>	<b>0.530</b>	<b>0.545</b>

Table 5.8: F1 of basic SMO and BayesNet classifiers across P1, P2 and P1+P2

There are two key takeaways from these experiments. First, using the combined data from P1 and P2 generally yields higher performance than using P1 or P2 data alone. Second, classifiers provided with more training examples usually produce higher overall performance as evidenced by higher macro and micro F1 when larger data sets are used but the results for individual emotion categories shows that it is not guaranteed that more data always leads to

higher performance. The classifiers may behave differently depending on the linguistic characteristics of the category.

### 5.3.2 Comparison across Features

We compare the results using three features groups: corpus-based, lexicon-based and cue-based in this section. The premise is that emotion cues identified by annotators can serve as a good feature subset for fine-grained emotion classification. We start by examining four different term weighting schemes for feature representation. We then present results from our experiments using the one-size-fits-all model. All the binary classifiers in the one-size-fits-all model employ exactly the same set of features. We further explored the potential of creating custom models that allow the features for each binary classifier to vary.

#### 5.3.2.1 Term Weights

Four different term weighting schemes were examined using the basic models to identify the most suitable one for feature representation: no-weight (i.e., binary representation where the feature value is 1 if a term is present and 0 if a term is absent), term frequency (*tf*, where the feature value is the frequency of the term in the tweet), inverse document frequency (*idf*, where the feature value is the logarithmically scaled fraction of the tweets that contain the term<sup>24</sup>) and term frequency-inverse document frequency (*tf-idf*, where the feature value is obtained by multiplying *tf* and *idf*). Results are shown in Table 5.9.

Scheme	SMO	BayesNet	Mean
no-weight	0.482	0.341	0.41
tf	0.474	0.493	0.48
idf	0.482	0.495	0.49
tf-idf	0.473	0.493	0.48

Table 5.9: F1 of basic SMO and BayesNet classifiers based on four weighting schemes

<sup>24</sup> Inverse document frequency (*idf*) is computed by dividing the total number of tweets by the number of tweets containing the term.

Of the four weighting schemes, we opted for *tf* as mean F1 of the basic SMO and BayesNet classifiers using *tf* is higher than *no-weight*. Also, *idf* and *tf-idf* does not seem to provide significant performance gain compared to *tf*.

### 5.3.2.2 One-Size-Fits-All Model

#### A) Overall Performance

In the first part of Task 2, we focus on identifying a common set of features that work well for all 28 emotion categories. We use the term “one-size-fits-all” to refer to this set. The goal is to find a single set of features to represent all 28 emotion categories. Building a model using one-size-fits-all features is more efficient as feature selection is performed only one-time across all 28 emotion categories.

Feature Set	Feature Size	SMO			BayesNet		
		P	R	F1	P	R	F1
Corpus-based							
C1: Corpus unigrams ( $f \geq 3$ )	6526	0.656	0.455	0.537	0.741	0.431	<b>0.545</b>
C2: Corpus bigrams ( $f \geq 3$ )	15812	0.556	0.308	0.396	0.800	0.267	0.400
C3: Corpus unigrams ( $f \geq 3$ ) + POS tags	8603	0.646	0.447	0.528	0.743	0.423	0.539
Lexicon-based							
L1: NRC emotion + non-emotion words	5586	0.694	0.221	0.335	0.778	0.199	0.317
L2: NRC emotion words	2704	0.737	0.205	0.320	0.787	0.190	0.306
L3: NRC categories (10)	10	0	0	0	0.189	0.125	0.150
Cue-based							
E1: Cue unigrams (joined)	1979	0.733	0.417	0.531	0.787	0.374	0.507
E2: Cue unigrams + phrases (joined)	2969	<b>0.750</b>	0.427	<b>0.544</b>	<b>0.810</b>	0.368	0.506
Cross-group							
C5: C1 + L3	6536	0.657	<b>0.459</b>	0.540	0.474	<b>0.493</b>	0.483
E5: E2 + L3	2979	0.748	0.427	0.544	0.485	0.436	0.459

Table 5.10: Precision, recall and F1 of classifiers based on feature sets

Table 5.10 shows the results for different feature groups for both SMO and BayesNet. First, we compare the results within each feature group. In corpus-based features, the basic models using corpus unigram features (C1) remain as the top performing classifiers in the mix (SMO: F1 = 0.537 and BayesNet: F1 = 0.545). Overall classifier performance (F1) deteriorated

slightly when using bigrams and unigrams enriched with POS tags as features. The decrease in performance is more significant for bigrams, a 27% drop for SMO and BayesNet respectively.

In the second feature group, where features consist of only lexicon terms and do not include unigrams, classifier performance using all lexicon terms as features improves by a slight 4% when compared to only using a subset of emotion terms. Lexicon-based features yield the lowest performing classifiers of the three feature groups. For fine-grained emotion classification, it is important to use features that capture vocabulary at a finer-grained level.

The classifiers performed very poorly with very coarse-grained representation of emotion content as evidenced by L3 which only captures the count of words associated with the 10 categories in the NRC lexicon (SMO:  $F1 = 0$  and BayesNet:  $F1 = 0.150$ ). L3 can serve as complementary features to corpus-based as well as cue-based features to boost the recall of the classifier. For instance, performance improved slightly when L3 is combined with corpus unigram features (C1) in SMO. On the other hand, precision significantly decreases while recall only improves slightly when L3 is added to complement other features in BayesNet.

None of the classifiers using lexicon-based features is able to top the performance of the basic models. One possible reason for these results is that the NRC lexicon is created for more coarse-grained analysis of emotions in text and specifically focuses on Plutchik's eight basic emotions. Many words that are important indicators of our set of finer-grained emotions are missing from the lexicon. Also, the lexicon does not include informal words, abbreviations, emoticons and emojis which are commonly used to express emotions in tweet.

SMO using cue unigrams and phrases (E2) as features achieves an F1 of 0.544 and is the overall best one-size-fits-all classifier beating the performance of the basic model (C1) by small margin (1%). Although recall of E2 is slightly lower than C1, the precision of E2 far exceeds C1. Similar findings for precision are observed for the BayesNet classifiers. BayesNet is a little less responsive to cue-based features so we will discuss the results based on SMO.

It is remarkable that cue-based classifiers can achieve performance that is at par with that of the corpus-based classifiers with far fewer features. E1 (SMO:  $F1 = 0.531$ ) contains only one third the number of features as C1 yet it produces a classifier that performs almost as well as C1 (SMO:  $F1 = 0.537$ ). Essentially, C2 and E1 are bag of words representation of the tweets but E1, using only a subset of terms, captures the most relevant predictors of the 28 emotion categories. Cue-based features also outperform lexicon-based features. Even though both contain roughly the same number of features, cue-based features include more relevant terms used to express emotions in tweets.

As evidenced by the best one-size-fits-all SMO classifier which uses the E2 cue-based variant, we can infer that the salient cues humans associate with each emotion category can serve as more compact features for automatic classification of fine-grained emotion in tweets. Classifiers using cue-based features can match the overall performance of the basic models with significantly less number of features.

## **B) Category-specific Performance**

### **B1: Corpus-based Features**

C2 (corpus bigrams) causes a decrease in performance in almost all of the emotion categories except for two categories in SMO (i.e., *pride* and *doubt*) and five categories in BayesNet (i.e., *love*, *pride*, *indifference*, *longing* and *sympathy*). For *pride*, it is interesting to note that the increment in performance happen in both SMO and BayesNet. Similar observation is made for pride in C3 (POS tags). This suggests that *pride* responds positively to features that attempt to capture contextual cues surrounding a term.

As shown in Table 5.11, results are more mixed for C3. Higher F1 scores than C1 are observed in less than half of the emotion categories with *longing* having the most positive effect (i.e., the highest difference in F1 between C3 and C1 in both SMO and BayesNet). For SMO, 12

emotion categories outperform the basic model. Results are similar for BayesNet although the set of 12 emotion categories slightly differs from SMO.

F1	SMO					BayesNet				
Category	C1	C2	C2-C1	C3	C3-C1	C1	C2	C2-C1	C3	C3-C1
Admiration	0.260	0.178	-0.08	<b>0.286</b>	0.03	0.111	0.086	-0.02	<b>0.189</b>	0.08
Amusement	<b>0.741</b>	0.381	-0.36	0.721	-0.02	<b>0.823</b>	0.476	-0.35	0.822	0.00
Anger	<b>0.384</b>	0.209	-0.17	0.342	-0.04	<b>0.354</b>	0.034	-0.32	0.309	-0.04
Boredom	0.514	0.111	-0.40	<b>0.521</b>	0.01	<b>0.603</b>	0.387	-0.22	0.563	-0.04
Confidence	0.140	0.102	-0.04	<b>0.189</b>	0.05	0.128	0.052	-0.08	<b>0.211</b>	0.08
Curiosity	0.590	0.416	-0.17	<b>0.644</b>	0.05	<b>0.701</b>	0.593	-0.11	0.671	-0.03
Desperation	0.121	0.061	-0.06	<b>0.125</b>	0.00	<b>0.351</b>	0.000	-0.35	0.235	-0.12
Doubt	0.133	0.149	0.02	<b>0.177</b>	0.04	0.239	0.162	-0.08	<b>0.240</b>	0.00
Excitement	<b>0.550</b>	0.487	-0.06	0.524	-0.03	<b>0.545</b>	0.474	-0.07	0.504	-0.04
Exhaustion	<b>0.328</b>	0.071	-0.26	0.286	-0.04	<b>0.246</b>	0.111	-0.13	0.218	-0.03
Fascination	0.396	0.186	-0.21	<b>0.455</b>	0.06	0.427	0.194	-0.23	<b>0.444</b>	0.02
Fear	<b>0.313</b>	0.056	-0.26	0.245	-0.07	<b>0.352</b>	0.041	-0.31	0.288	-0.06
Gratitude	<b>0.921</b>	0.792	-0.13	0.914	-0.01	<b>0.901</b>	0.815	-0.09	0.885	-0.02
Happiness	<b>0.558</b>	0.450	-0.11	0.555	0.00	0.526	0.372	-0.15	<b>0.534</b>	0.01
Hate	<b>0.642</b>	0.381	-0.26	0.583	-0.06	<b>0.669</b>	0.438	-0.23	0.656	-0.01
Hope	<b>0.666</b>	0.563	-0.10	0.642	-0.02	0.687	0.619	-0.07	<b>0.698</b>	0.01
Indifference	0.094	0.076	-0.02	<b>0.125</b>	0.03	0.143	0.256	0.11	0.116	-0.03
Inspiration	<b>0.549</b>	0.070	-0.48	0.367	-0.18	<b>0.595</b>	0.125	-0.47	0.528	-0.07
Jealousy	<b>0.510</b>	0.200	-0.31	0.444	-0.07	<b>0.600</b>	0.269	-0.33	0.600	0.00
Longing	0.387	0.382	-0.01	<b>0.485</b>	0.10	0.277	<b>0.430</b>	0.15	0.360	0.08
Love	0.581	0.554	-0.03	<b>0.592</b>	0.01	0.576	<b>0.596</b>	0.02	0.568	-0.01
Pride	0.758	0.778	0.02	<b>0.833</b>	0.08	0.780	0.826	0.05	<b>0.830</b>	0.05
Regret	<b>0.329</b>	0.248	-0.08	0.329	0.00	0.500	0.370	-0.13	<b>0.515</b>	0.02
Relaxed	<b>0.292</b>	0.047	-0.25	0.257	-0.03	<b>0.313</b>	0.000	-0.31	0.255	-0.06
Sadness	<b>0.539</b>	0.258	-0.28	0.535	0.00	<b>0.569</b>	0.207	-0.36	0.551	-0.02
Shame	<b>0.415</b>	0.281	-0.13	0.304	-0.11	<b>0.544</b>	0.372	-0.17	0.477	-0.07
Surprise	<b>0.371</b>	0.230	-0.14	0.353	-0.02	0.322	0.297	-0.03	<b>0.324</b>	0.00
Sympathy	0.531	0.529	0.00	<b>0.561</b>	0.03	0.541	0.545	0.00	<b>0.562</b>	0.02
Macro-F1	<b>0.450</b>	0.294	-0.16	0.443	-0.01	<b>0.479</b>	0.327	-0.15	0.470	-0.01
Micro-F1	<b>0.537</b>	0.396	-0.14	0.528	-0.01	<b>0.545</b>	0.400	-0.14	0.539	-0.01

Table 5.11: F1 of each emotion category for SMO and BayesNet using corpus-based features

The bigram (C2) and POS tag (C3) features were introduced to lower ambiguity present in using single terms as features. However, our experiments show that these enriched corpus-based features mostly led to degradation in classifier performance. This is similar to what was



observed in Lewis (1992) who examined the use of phrases as features as a means to lower ambiguity in unigram features. High dimensionality and low frequency of the enriched features outweigh the advantages these features provide in lowering ambiguity.

## **B2: Lexicon-based Features**

Using solely the lexicon-based features (i.e., L1, L2 and L3), the performance of only 4 emotion categories (i.e., *desperation*, *fear*, *hate* and *jealousy*) is above the performance of the basic SMO classifier. Adding L3 to complement C1 helps boost the classifier performance on 14 emotion categories. On the other hand, BayesNet responds even more negatively to the lexicon-based features. Only a slight increase in performance is seen on only one emotion category (i.e., *confidence*) when only the lexicon-based features are used. Unlike SMO, there is hardly any improvement in the performance in the BayesNet classifier when L3 is added to complement C1 except for *admiration*, *fear*, *love* and *shame*.

As shown in Table 5.12 and Table 5.13, using the lexicon-based features alone cause classifier performance to drop substantially, notably for *gratitude* and *amusement*, two emotion categories in which the basic models are superior at predicting. For *gratitude*, the word “*thank*” which is an important indicator for *gratitude* is not listed in the NRC lexicon. As for *amusement*, many positive examples contain emojis and abbreviated form that are not represented in the feature space because such terms are absent from the lexicon. If a salient emotion term does not exist in the feature space, the classifier will not be able to learn any patterns associated with the term. Overall, the NRC lexicon-based features prove to be less informative for fine-grained emotion detection in tweets. The vocabulary in the NRC lexicon is not exactly the best fit for this problem domain.

F1	SMO								
Category	C1	L1	L1-C1	L2	L2-C1	L3	L3-C1	C5	C5-C1
Admiration	0.260	0.204	-0.06	0.194	-0.07	0	-0.26	<b>0.273</b>	0.01
Amusement	<b>0.741</b>	0.046	-0.70	0.049	-0.69	0	-0.74	0.732	-0.01
Anger	0.384	0.235	-0.15	0.224	-0.16	0	-0.38	<b>0.391</b>	0.01
Boredom	<b>0.514</b>	0.310	-0.20	0.310	-0.20	0	-0.51	0.507	-0.01
Confidence	0.140	0.149	0.01	0.151	0.01	0	-0.14	<b>0.153</b>	0.01
Curiosity	<b>0.590</b>	0.000	-0.59	0.000	-0.59	0	-0.59	0.585	0.00
Desperation	0.121	<b>0.232</b>	0.11	0.209	0.09	0	-0.12	0.121	0.00
Doubt	<b>0.133</b>	0.036	-0.10	0.025	-0.11	0	-0.13	0.114	-0.02
Excitement	<b>0.550</b>	0.469	-0.08	0.471	-0.08	0	-0.55	0.548	0.00
Exhaustion	<b>0.328</b>	0.167	-0.16	0.000	-0.33	0	-0.33	0.328	0.00
Fascination	0.396	0.179	-0.22	0.188	-0.21	0	-0.40	<b>0.410</b>	0.01
Fear	0.313	0.292	-0.02	<b>0.326</b>	0.01	0	-0.31	0.325	0.01
Gratitude	0.921	0.132	-0.79	0.126	-0.79	0	-0.92	<b>0.922</b>	0.00
Happiness	0.558	0.329	-0.23	0.302	-0.26	0	-0.56	<b>0.559</b>	0.00
Hate	0.642	0.669	0.03	<b>0.677</b>	0.04	0	-0.64	0.630	-0.01
Hope	<b>0.666</b>	0.648	-0.02	0.620	-0.05	0	-0.67	0.662	0.00
Indifference	<b>0.094</b>	0.000	-0.09	0.000	-0.09	0	-0.09	0.093	0.00
Inspiration	0.549	0.250	-0.30	0.253	-0.30	0	-0.55	<b>0.557</b>	0.01
Jealousy	0.510	0.588	0.08	<b>0.600</b>	0.09	0	-0.51	0.510	0.00
Longing	0.387	0.418	0.03	0.000	-0.39	0	-0.39	<b>0.454</b>	0.07
Love	0.581	0.572	-0.01	0.581	0.00	0	-0.58	<b>0.597</b>	0.02
Pride	0.758	0.756	0.00	0.763	0.00	0	-0.76	<b>0.766</b>	0.01
Regret	0.329	0.013	-0.32	0.062	-0.27	0	-0.33	<b>0.330</b>	0.00
Relaxed	<b>0.292</b>	0.000	-0.29	0.000	-0.29	0	-0.29	0.286	-0.01
Sadness	0.539	0.271	-0.27	0.245	-0.29	0	-0.54	<b>0.543</b>	0.00
Shame	<b>0.415</b>	0.242	-0.17	0.224	-0.19	0	-0.41	0.414	0.00
Surprise	0.371	0.167	-0.20	0.173	-0.20	0	-0.37	<b>0.378</b>	0.01
Sympathy	0.531	0.298	-0.23	0.000	-0.53	0	-0.53	<b>0.543</b>	0.01
Macro-F1	0.450	0.274	-0.18	0.242	-0.21	0	-0.45	<b>0.455</b>	0.00
Micro-F1	0.537	0.335	-0.20	0.320	-0.22	0	-0.54	<b>0.540</b>	0.00

Table 5.12: F1 of each emotion category based on SMO using lexicon-based features

F1	BayesNet								
Category	C1	L1	L1-C1	L2	L2-C1	L3	L3-C1	C5	L4-C5
Admiration	0.111	0.184	0.07	0.170	0.06	0	-0.11	<b>0.280</b>	0.17
Amusement	<b>0.823</b>	0.033	-0.79	0.033	-0.79	0	-0.82	0.816	-0.01
Anger	<b>0.354</b>	0.165	-0.19	0.158	-0.20	0.219	-0.14	0.345	-0.01
Boredom	<b>0.603</b>	0.310	-0.29	0.310	-0.29	0	-0.60	0.603	0.00
Confidence	0.128	<b>0.151</b>	0.02	0.151	0.02	0	-0.13	0.128	0.00
Curiosity	<b>0.701</b>	0.000	-0.70	0.000	-0.70	0	-0.70	0.701	0.00
Desperation	<b>0.351</b>	0.257	-0.09	0.209	-0.14	0	-0.35	0.351	0.00
Doubt	<b>0.239</b>	0.000	-0.24	0.000	-0.24	0	-0.24	0.239	0.00
Excitement	<b>0.545</b>	0.461	-0.08	0.463	-0.08	0	-0.55	0.515	-0.03
Exhaustion	<b>0.246</b>	0.218	-0.03	0.000	-0.25	0	-0.25	0.246	0.00
Fascination	<b>0.427</b>	0.184	-0.24	0.184	-0.24	0	-0.43	0.427	0.00
Fear	0.352	0.280	-0.07	0.281	-0.07	0	-0.35	<b>0.380</b>	0.03
Gratitude	<b>0.901</b>	0.135	-0.77	0.126	-0.77	0	-0.90	0.895	-0.01
Happiness	<b>0.526</b>	0.319	-0.21	0.294	-0.23	0.283	-0.24	0.504	-0.02
Hate	<b>0.669</b>	0.656	-0.01	0.656	-0.01	0.166	-0.50	0.204	-0.46
Hope	<b>0.687</b>	0.646	-0.04	0.613	-0.07	0.240	-0.45	0.377	-0.31
Indifference	<b>0.143</b>	0.000	-0.14	0.000	-0.14	0	-0.14	0.143	0.00
Inspiration	<b>0.595</b>	0.253	-0.34	0.253	-0.34	0	-0.59	0.595	0.00
Jealousy	<b>0.600</b>	0.600	0.00	0.600	0.00	0	-0.60	0.600	0.00
Longing	<b>0.277</b>	0.000	-0.28	0.000	-0.28	0	-0.28	0.277	0.00
Love	0.576	0.578	0.00	0.578	0.00	0	-0.58	<b>0.606</b>	0.03
Pride	<b>0.780</b>	0.763	-0.02	0.763	-0.02	0.121	-0.66	0.699	-0.08
Regret	<b>0.500</b>	0.088	-0.41	0.088	-0.41	0	-0.50	0.500	0.00
Relaxed	<b>0.313</b>	0.000	-0.31	0.000	-0.31	0	-0.31	0.313	0.00
Sadness	<b>0.569</b>	0.239	-0.33	0.231	-0.34	0	-0.57	0.531	-0.04
Shame	0.544	0.226	-0.32	0.226	-0.32	0	-0.54	<b>0.564</b>	0.02
Surprise	<b>0.322</b>	0.131	-0.19	0.131	-0.19	0	-0.32	0.322	0.00
Sympathy	<b>0.541</b>	0.051	-0.49	0.000	-0.54	0	-0.54	0.536	0.00
Macro-F1	<b>0.479</b>	0.247	-0.23	0.233	-0.25	0.037	-0.44	0.453	-0.03
Micro-F1	<b>0.545</b>	0.317	-0.23	0.306	-0.24	0.150	-0.40	0.483	-0.06

Table 5.13: F1 of each emotion category based on BayesNet using lexicon-based features

### B3: Cue-based Features

SMO using cue-based features produces better results when compared to corpus-based features and lexicon-based features. Note that maximum F1 scores (bold) for each emotion

category in Table 5.14 are not concentrated in the C1 column but are dispersed across the three variants of cue-based features.

F1	SMO						
Category	C1	E1	E1-C1	E2	E2-C1	E5	E5-C1
Admiration	<b>0.260</b>	0.178	-0.08	0.222	-0.04	0.228	-0.03
Amusement	0.741	<b>0.764</b>	0.02	0.757	0.02	0.757	0.02
Anger	<b>0.384</b>	0.344	-0.04	0.327	-0.06	0.329	-0.05
Boredom	0.514	0.526	0.01	<b>0.556</b>	0.04	0.526	0.01
Confidence	0.140	0.233	0.09	0.238	0.10	<b>0.244</b>	0.10
Curiosity	<b>0.590</b>	0.462	-0.13	0.485	-0.10	0.485	-0.10
Desperation	0.121	<b>0.273</b>	0.15	0.273	0.15	0.273	0.15
Doubt	0.133	0.185	0.05	<b>0.200</b>	0.07	0.200	0.07
Excitement	<b>0.550</b>	0.486	-0.06	0.509	-0.04	0.512	-0.04
Exhaustion	<b>0.328</b>	0.118	-0.21	0.250	-0.08	0.250	-0.08
Fascination	0.396	<b>0.440</b>	0.04	0.435	0.04	0.430	0.03
Fear	0.313	0.309	0.00	<b>0.347</b>	0.03	0.347	0.03
Gratitude	<b>0.921</b>	0.913	-0.01	0.903	-0.02	0.903	-0.02
Happiness	0.558	0.569	0.01	<b>0.580</b>	0.02	0.578	0.02
Hate	<b>0.642</b>	0.605	-0.04	0.628	-0.01	0.591	-0.05
Hope	0.666	0.651	-0.01	0.672	0.01	<b>0.684</b>	0.02
Indifference	<b>0.094</b>	0.000	-0.09	0.087	-0.01	0.091	0.00
Inspiration	<b>0.549</b>	0.457	-0.09	0.457	-0.09	0.457	-0.09
Jealousy	<b>0.510</b>	0.154	-0.36	0.154	-0.36	0.143	-0.37
Longing	0.387	<b>0.491</b>	0.10	0.491	0.10	0.453	0.07
Love	0.581	<b>0.596</b>	0.02	0.593	0.01	0.585	0.00
Pride	0.758	0.782	0.02	<b>0.847</b>	0.09	0.847	0.09
Regret	0.329	0.395	0.07	0.390	0.06	<b>0.400</b>	0.07
Relaxed	<b>0.292</b>	0.154	-0.14	0.154	-0.14	0.154	-0.14
Sadness	0.539	0.516	-0.02	0.540	0.00	<b>0.550</b>	0.01
Shame	<b>0.415</b>	0.242	-0.17	0.188	-0.23	0.182	-0.23
Surprise	0.371	0.308	-0.06	0.436	0.07	<b>0.455</b>	0.08
Sympathy	0.531	0.465	-0.07	<b>0.622</b>	0.09	0.622	0.09
Macro-F1	<b>0.450</b>	0.415	-0.04	0.441	-0.01	0.438	-0.01
Micro-F1	0.537	0.531	-0.01	<b>0.544</b>	0.01	0.544	0.01

Table 5.14: F1 of each emotion category based on SMO using cue-based features

The cue-based SMO classifier exceeds the performance of the basic model for 16 emotion categories, the largest number of categories affected compared to the two other feature groups. On the other hand, cue-based BayesNet only manages to boost the performance of 10

emotion categories as shown in Table 5.15. Regardless of the machine learning algorithm, *desperation* and *longing* both respond favorably to the cue-based features.

F1	BayesNet						
Category	C1	E1	E1-C1	E2	E2-C1	E5	E5-C1
Admiration	0.111	0.048	-0.06	0.048	-0.06	<b>0.194</b>	0.08
Amusement	<b>0.823</b>	0.798	-0.03	0.792	-0.03	0.788	-0.03
Anger	<b>0.354</b>	0.283	-0.07	0.229	-0.13	0.299	-0.06
Boredom	<b>0.603</b>	0.526	-0.08	0.526	-0.08	0.526	-0.08
Confidence	0.128	0.114	-0.01	<b>0.216</b>	0.09	0.216	0.09
Curiosity	<b>0.701</b>	0.564	-0.14	0.333	-0.37	0.333	-0.37
Desperation	0.351	<b>0.435</b>	0.08	0.435	0.08	0.435	0.08
Doubt	<b>0.239</b>	0.075	-0.16	0.075	-0.16	0.075	-0.16
Excitement	<b>0.545</b>	0.516	-0.03	0.513	-0.03	0.539	-0.01
Exhaustion	<b>0.246</b>	0.105	-0.14	0.105	-0.14	0.105	-0.14
Fascination	0.427	<b>0.472</b>	0.05	0.472	0.05	0.472	0.05
Fear	<b>0.352</b>	0.171	-0.18	0.171	-0.18	0.171	-0.18
Gratitude	<b>0.901</b>	0.869	-0.03	0.872	-0.03	0.858	-0.04
Happiness	<b>0.526</b>	0.490	-0.04	0.473	-0.05	0.501	-0.02
Hate	<b>0.669</b>	0.644	-0.03	0.644	-0.03	0.164	-0.50
Hope	<b>0.687</b>	0.651	-0.04	0.684	0.00	0.383	-0.30
Indifference	<b>0.143</b>	0.067	-0.08	0.074	-0.07	0.074	-0.07
Inspiration	<b>0.595</b>	0.500	-0.09	0.500	-0.09	0.500	-0.09
Jealousy	<b>0.600</b>	0.533	-0.07	0.533	-0.07	0.533	-0.07
Longing	0.277	0.353	0.08	<b>0.364</b>	0.09	0.364	0.09
Love	0.576	0.545	-0.03	0.548	-0.03	<b>0.603</b>	0.03
Pride	0.780	0.781	0.00	<b>0.826</b>	0.05	0.758	-0.02
Regret	0.500	0.537	0.04	<b>0.558</b>	0.06	0.558	0.06
Relaxed	<b>0.313</b>	0.160	-0.15	0.160	-0.15	0.160	-0.15
Sadness	<b>0.569</b>	0.481	-0.09	0.483	-0.09	0.440	-0.13
Shame	<b>0.544</b>	0.267	-0.28	0.267	-0.28	0.267	-0.28
Surprise	0.322	0.346	0.02	<b>0.429</b>	0.11	0.429	0.11
Sympathy	0.541	0.526	-0.01	<b>0.564</b>	0.02	0.564	0.02
Macro-F1	<b>0.479</b>	0.423	-0.06	0.425	-0.05	0.404	-0.08
Micro-F1	<b>0.545</b>	0.507	-0.04	0.506	-0.04	0.459	-0.09

Table 5.15: F1 of each emotion category based on BayesNet using cue-based features

Of the two pure variants of cue-based features (E1 and E2), E2 which adds collocations and negated phrases to cue unigrams help improve the performance for a greater number of categories notably for *pride* and *sympathy*. In the case of *pride*, we observe that the term

“honor” is commonly used as a secondary indicator for both *pride* and *admiration*. To distinguish between the two emotion categories, the immediate words surrounding the term provide useful contextual clue to interpret the emotion being expressed by the tweeter. Therefore, the inclusion of phrase features containing the term “honor” such as “honored to” helps both the *pride* and *admiration* classifier better distinguish between positive and negative examples. As for *sympathy*, there are specific phrases (e.g., “thoughts and prayers”, “are with”, “go out to”) that almost seem like templates that tweeters commonly use to express *sympathy*. The inclusion of such phrases improved classification results for *sympathy*.

Our findings reveal that using a large number of bigrams statistically generated from the corpus hurts the performance of the classifier. However, adding only a selective set of bigrams proves to be a more effective approach to reduce the problem of high dimensionality (Tan, Wang, & Lee, 2002).

### **C) Comparison across Features**

Among the various feature sets tested on SMO, the best performing one-size-fits-all classifier across all 28 emotion categories is the one that uses E2 (SMO-E2). We next compare the difference in F1 scores between each binary classifier in SMO-E2 with the best performing binary classifier for each individual category. As a basis for comparison, we first identify the binary classifier with maximum F1 for each emotion category from a pool of our feature-related experiments. This ensemble of 28 binary classifiers with maximum F1 (max-ensemble) is then compared to each binary classifier in the one-size-fits-all model as shown in Table 5.16.

Classifier performance based on the ensemble model (micro-F1 = 0.579) is higher compared to SMO-E2 (micro-F1 = 0.544). It turns out that the best performing binary classifier for each emotion category is generally not SMO-E2. SMO-E2 tops other classifiers for only four emotion categories (i.e., *happiness*, *longing*, *pride* and *sympathy*). It performs well on a majority

of other emotion categories judging from the fairly small differences in F1 observed in Table 5.16 except for *curiosity*, *jealousy* and *shame*.

Category	One-Size-Fits-All (SMO-E2)			Max-Ensemble			Diff(F1)
	Classifier	Features	F1	Classifier	Features	F1	
Admiration	SMO	E2	0.222	SMO	C3	<b>0.286</b>	0.063
Amusement	SMO	E2	0.757	BayesNet	C1	<b>0.823</b>	0.067
Anger	SMO	E2	0.327	SMO	C5	<b>0.391</b>	0.064
Boredom	SMO	E2	0.556	BayesNet	C1	<b>0.603</b>	0.047
Confidence	SMO	E2	0.238	SMO	E5	<b>0.244</b>	0.006
Curiosity	SMO	E2	0.485	BayesNet	C1	<b>0.701</b>	0.216
Desperation	SMO	E2	0.273	BayesNet	E1	<b>0.435</b>	0.162
Doubt	SMO	E2	0.200	BayesNet	C3	<b>0.240</b>	0.040
Excitement	SMO	E2	0.509	SMO	C1	<b>0.550</b>	0.041
Exhaustion	SMO	E2	0.250	SMO	C1	<b>0.328</b>	0.078
Fascination	SMO	E2	0.435	BayesNet	E1	<b>0.472</b>	0.037
Fear	SMO	E2	0.347	BayesNet	C5	<b>0.380</b>	0.033
Gratitude	SMO	E2	0.903	SMO	C5	<b>0.922</b>	0.019
Happiness	SMO	E2	<b>0.580</b>	SMO	E2	0.580	0
Hate	SMO	E2	0.628	SMO	L2	<b>0.677</b>	0.049
Hope	SMO	E2	0.672	BayesNet	C3	<b>0.698</b>	0.027
Indifference	SMO	E2	0.087	BayesNet	C2	<b>0.256</b>	0.169
Inspiration	SMO	E2	0.457	BayesNet	C1	<b>0.595</b>	0.137
Jealousy	SMO	E2	0.154	BayesNet	C5	<b>0.600</b>	0.446
Longing	SMO	E2	<b>0.491</b>	SMO	E2	0.491	0
Love	SMO	E2	0.593	BayesNet	C5	<b>0.606</b>	0.013
Pride	SMO	E2	<b>0.847</b>	SMO	E2	0.847	0
Regret	SMO	E2	0.390	BayesNet	E2	<b>0.558</b>	0.169
Relaxed	SMO	E2	0.154	BayesNet	C1	<b>0.313</b>	0.159
Sadness	SMO	E2	0.540	BayesNet	C1	<b>0.569</b>	0.029
Shame	SMO	E2	0.188	BayesNet	C5	<b>0.564</b>	0.377
Surprise	SMO	E2	0.436	SMO	E5	<b>0.455</b>	0.018
Sympathy	SMO	E2	<b>0.622</b>	SMO	E2	0.622	0
Macro-avg			0.441			<b>0.529</b>	0.088
Micro-avg			0.544			<b>0.579</b>	0.035

Table 5.16: Comparing F1 of each emotion category in the best one-size-fits-all model to the best model per category

Roughly half of the emotion categories respond better to SMO while the other half lean towards BayesNet. Some emotion categories seem to respond well to the same classifier (e.g.,

the best classifier for *amusement*, *boredom*, *curiosity*, *inspiration*, *relaxed* and *sadness* is BayesNet-C1), suggesting that it is possible to use an ensemble of different classifiers on a few subsets of emotion categories to maximize classification performance.

Picking the most appropriate emotion classifier depends on the nature of the problem. If the goal is to obtain a classifier yielding relatively decent performance across all 28 emotion categories, the one-size-fits-all classifier is the best candidate for the task. If it is absolutely necessary to maximize classification for each specific category or a subset of categories, using custom classifiers or an ensemble of different classifiers is more advantageous.

### **5.3.2.3 Custom Model**

#### **A) Overall Performance**

The second part of Task 2 further explores a systematic approach to select custom features for each emotion category. The binary classifier for an emotion category employs only features selected based on emotion cues that are relevant to the particular category. Unlike the one-size-fits-all model (E1 and E2), the number of features varies in each emotion category. We hypothesize that using features custom-selected for each emotion category will increase the performance.

First, selecting custom features tailored to an emotion category can help remove features that are not informative for that category. Second, using custom features also allows us to optimize and refine the features for each specific emotion category without the risk of harming or sacrificing the performance of other emotion categories. We use the terms “custom model” to refer to a set of 28 binary classifiers that are trained with different features but use the same machine learning algorithm.

The results based on six custom model variants and E2 are presented in Table 5.17. For the stemmed variants (E3, E4, E6 and E7), the SMO classifier using E6 as features achieves the highest F1 ( $F1 = 0.564$ ) and exceeds the performance of the best one-size-fits-all classifier



(E2). The BayesNet classifier using only custom cue-based features also outperforms E2. The results show that using custom features tailored to each emotion category can be advantageous for fine-grained emotion classification in tweets. Adding the ten NRC lexicon word count features (L3) to E3 and E4 slightly increase the custom SMO model but significantly decrease the precision and hence the F1 for the custom BayesNet model. The degradation in performance of the BayesNet model containing L3 features suggest that BayesNet does not respond well to the cross-group feature combination (i.e., combining features from cue-based and lexicon-based feature groups).

Feature Set	SMO			BayesNet		
	P	R	F1	P	R	F1
<b>Stemmed</b>						
E2: Cue unigrams + phrases (joined)	0.750	0.427	0.544	<b>0.810</b>	0.368	0.506
E3: Cue unigrams (custom)	0.792	0.404	0.535	0.801	0.375	<b>0.535</b>
E7: E3 + L3	0.795	0.405	0.537	0.479	0.435	0.456
E4: Cue unigrams + phrases (custom)	0.814	0.426	0.559	0.788	0.381	0.514
E6: E4 + L3	<b>0.818</b>	<b>0.430</b>	<b>0.564</b>	0.490	<b>0.443</b>	0.465
<b>Non-Stemmed</b>						
E7a: E3 + L3	0.797	0.387	0.521	0.475	0.409	0.439
E6a: E4 + L3	0.817	0.409	0.545	0.483	0.411	0.444

Table 5.17: Overall precision, recall and F1 based on custom cue features

Basically, E6a is the non-stemmed variant of E6 and the same applies to E7a. We included the non-stemmed variants (E6a and E7a) as part of this set of experiments to test if preserving the morphological variations of the features affects classification performance. Based on the results shown in Table 5.17, classifier performance using E6a and E7a is slightly lower compared to its stemmed counterpart. Of all our feature-related experiments, SMO-E6 produces the highest F1. Hence, our discussion will focus on the SMO custom model.

## B) Emotion Category Performance

Recall that the best one-size-fits-all SMO-E2 classifier shows performance improvements for 16 emotion categories compared to the basic model. Using custom features,

the classifier is able to improve the performance of the basic model for 19 emotion categories as shown in Table 5.18.

F1	SMO								
Category	C1	E7	E7-C1	E6	E6-C1	E7a	E7a-C1	E6a	E6a-C1
Admiration	<b>0.260</b>	0.076	-0.18	0.248	-0.01	0.174	-0.09	0.199	-0.06
Amusement	0.741	0.764	0.02	<b>0.774</b>	0.03	0.769	0.03	0.765	0.02
Anger	<b>0.384</b>	0.333	-0.05	0.334	-0.05	0.248	-0.14	0.264	-0.12
Boredom	0.514	<b>0.526</b>	0.01	0.526	0.01	0.143	-0.37	0.143	-0.37
Confidence	0.140	<b>0.216</b>	0.08	0.216	0.08	0.000	-0.14	0.000	-0.14
Curiosity	0.590	<b>0.615</b>	0.03	0.387	-0.20	0.571	-0.02	0.611	0.02
Desperation	0.121	<b>0.435</b>	0.31	0.435	0.31	0.105	-0.02	0.100	-0.02
Doubt	0.133	0.075	-0.06	0.254	0.12	0.077	-0.06	<b>0.258</b>	0.12
Excitement	0.550	0.531	-0.02	<b>0.579</b>	0.03	0.479	-0.07	0.521	-0.03
Exhaustion	<b>0.328</b>	0	-0.33	0.250	-0.08	0.000	-0.33	0.133	-0.20
Fascination	0.396	<b>0.472</b>	0.08	0.366	-0.03	0.467	0.07	0.400	0.00
Fear	<b>0.313</b>	0.244	-0.07	0.244	-0.07	0.301	-0.01	0.308	-0.01
Gratitude	<b>0.921</b>	0.913	-0.01	0.910	-0.01	0.920	0.00	0.910	-0.01
Happiness	0.558	0.557	0.00	<b>0.592</b>	0.03	0.562	0.00	0.589	0.03
Hate	0.642	<b>0.644</b>	0.00	0.644	0.00	0.636	-0.01	0.615	-0.03
Hope	0.666	0.661	0.00	<b>0.700</b>	0.03	0.656	-0.01	0.696	0.03
Indifference	<b>0.094</b>	0	-0.09	0.087	-0.01	0.000	-0.09	0.087	-0.01
Inspiration	<b>0.549</b>	0.500	-0.05	0.486	-0.06	0.143	-0.41	0.143	-0.41
Jealousy	<b>0.510</b>	0.167	-0.34	0.143	-0.37	0.000	-0.51	0.143	-0.37
Longing	0.387	0	-0.39	<b>0.464</b>	0.08	0.000	-0.39	0.417	0.03
Love	0.581	0.610	0.03	0.592	0.01	0.627	0.05	<b>0.643</b>	0.06
Pride	0.758	0.781	0.02	<b>0.85</b>	0.09	0.817	0.06	0.835	0.08
Regret	0.329	0.466	0.14	<b>0.506</b>	0.18	0.451	0.12	0.500	0.17
Relaxed	<b>0.292</b>	0.16	-0.13	0.160	-0.13	0.083	-0.21	0.083	-0.21
Sadness	0.539	0.519	-0.02	<b>0.567</b>	0.03	0.490	-0.05	0.552	0.01
Shame	0.415	0.267	-0.15	0.207	-0.21	<b>0.424</b>	0.01	0.323	-0.09
Surprise	0.371	0.385	0.01	<b>0.517</b>	0.15	0.237	-0.13	0.404	0.03
Sympathy	0.531	0.545	0.01	<b>0.696</b>	0.16	0.600	0.07	0.600	0.07
Macro-F1	0.450	0.409	-0.04	<b>0.455</b>	0.00	0.356	-0.09	0.401	-0.05
Micro-F1	0.537	0.537	0.00	<b>0.564</b>	0.03	0.521	-0.02	0.545	0.01

Table 5.18: F1 of each emotion category based on SMO custom model

Category	P			R			F1		
	E2	E6	E6-E2	E2	E6	E6-E2	E2	E6	E6-E2
Admiration	0.459	<b>0.500</b>	0.04	0.147	<b>0.165</b>	0.02	0.222	<b>0.248</b>	0.03
Amusement	0.884	<b>0.909</b>	0.03	0.662	<b>0.674</b>	0.01	0.757	<b>0.774</b>	0.02
Anger	0.562	<b>0.752</b>	0.19	<b>0.230</b>	0.215	-0.02	0.327	<b>0.334</b>	0.01
Boredom	<b>1.000</b>	0.833	-0.17	<b>0.385</b>	0.385	0.00	<b>0.556</b>	0.526	-0.03
Confidence	0.556	<b>1.000</b>	0.44	<b>0.152</b>	0.121	-0.03	<b>0.238</b>	0.216	-0.02
Curiosity	<b>1.000</b>	1.000	0.00	<b>0.320</b>	0.240	-0.08	<b>0.485</b>	0.387	-0.10
Desperation	0.750	<b>1.000</b>	0.25	0.167	<b>0.278</b>	0.11	0.273	<b>0.435</b>	0.16
Doubt	0.545	<b>0.571</b>	0.03	0.122	<b>0.163</b>	0.04	0.200	<b>0.254</b>	0.05
Excitement	0.719	<b>0.917</b>	0.20	0.394	<b>0.423</b>	0.03	0.509	<b>0.579</b>	0.07
Exhaustion	<b>1.000</b>	1.000	0.00	0.143	<b>0.143</b>	0.00	<b>0.250</b>	0.250	0.00
Fascination	0.571	<b>0.600</b>	0.03	<b>0.351</b>	0.263	-0.09	<b>0.435</b>	0.366	-0.07
Fear	<b>0.708</b>	0.688	-0.02	<b>0.230</b>	0.149	-0.08	<b>0.347</b>	0.244	-0.10
Gratitude	<b>0.933</b>	0.928	0.00	0.874	<b>0.893</b>	0.02	0.903	<b>0.910</b>	0.01
Happiness	0.730	<b>0.772</b>	0.04	<b>0.481</b>	0.480	0.00	0.580	<b>0.592</b>	0.01
Hate	0.818	<b>0.824</b>	0.01	0.509	<b>0.528</b>	0.02	0.628	<b>0.644</b>	0.02
Hope	0.840	<b>0.901</b>	0.06	0.560	<b>0.572</b>	0.01	0.672	<b>0.700</b>	0.03
Indifference	<b>0.500</b>	0.333	-0.17	0.048	<b>0.050</b>	0.00	<b>0.087</b>	0.087	0.00
Inspiration	<b>0.889</b>	0.818	-0.07	0.308	<b>0.346</b>	0.04	0.457	<b>0.486</b>	0.03
Jealousy	<b>0.333</b>	0.250	-0.08	<b>0.100</b>	0.100	0.00	<b>0.154</b>	0.143	-0.01
Longing	<b>0.722</b>	0.619	-0.10	<b>0.371</b>	0.371	0.00	<b>0.491</b>	0.464	-0.03
Love	0.732	<b>0.837</b>	0.11	<b>0.498</b>	0.458	-0.04	<b>0.593</b>	0.592	0.00
Pride	<b>1.000</b>	0.980	-0.02	0.734	<b>0.750</b>	0.02	0.847	<b>0.850</b>	0.00
Regret	0.536	<b>0.667</b>	0.13	0.306	<b>0.408</b>	0.10	0.390	<b>0.506</b>	0.12
Relaxed	0.667	<b>1.000</b>	0.33	0.087	<b>0.087</b>	0.00	0.154	<b>0.160</b>	0.01
Sadness	0.744	<b>0.859</b>	0.12	0.424	<b>0.424</b>	0.00	0.540	<b>0.567</b>	0.03
Shame	0.429	<b>0.750</b>	0.32	0.120	<b>0.120</b>	0.00	0.188	<b>0.207</b>	0.02
Surprise	<b>0.727</b>	0.705	-0.02	0.312	<b>0.408</b>	0.10	0.436	<b>0.517</b>	0.08
Sympathy	0.700	<b>0.762</b>	0.06	0.560	<b>0.640</b>	0.08	0.622	<b>0.696</b>	0.07
Macro-avg	0.716	<b>0.778</b>	0.06	0.343	<b>0.352</b>	0.01	0.441	<b>0.455</b>	0.01
Micro-avg	0.750	<b>0.818</b>	0.07	0.427	<b>0.430</b>	0.00	0.544	<b>0.564</b>	0.02

Table 5.19: Precision, recall and F1 for each emotion category between E2 and E6

The results shown in Table 5.18 suggest that including features that are only relevant to the emotion category is an effective strategy to maximize the performance of each binary classifier. The greatest improvement in performance is observed in *desperation* and *regret*. Table 5.19 compares the precision, recall and F1 between the E2 (one-size-fits-all) and E6 (custom). E6 outperforms E2 for 18 emotion categories, although the difference is not huge.

There are three categories that respond better to non-stemmed features: *doubt*, *love* and *shame*. For these emotion categories, including specific morphological variants of an emotion-related word serves as a form of word sense disambiguation. The prefixes and affixes attached to the word can change its meaning. For instance, the root word “*love*” and its morphological variants “*loved*” and “*lovely*” are treated as separate features in the non-stemmed feature space. The words “*loved*” and “*lovely*” tend to occur more often in the negative examples of *love*. Therefore, having these two variants in the feature space can assist the classifier in making a better distinction between the positive and negative examples. In the stemmed version, “*loved*” and “*lovely*” are reduced to their root form “*love*”, causing their distinctive meaning to be lost. Similar arguments apply to shame and doubt. In the case of “*shame*”, the root word “*shame*” has multiple senses that can be used to express different emotions but its morphological variant “*shameful*” is consistently used to express shame. Similarly, in doubt, the inclusion of “*confuse*”, “*confusing*” and “*confusion*” as features aids the classifier to better distinguish between positive and negative examples.

For 9 emotion categories, the custom cue-based features do not improve performance beyond that of the corpus unigram features (C1). *Gratitude* is one of these 9 categories. It is possible that *gratitude* with very low lexical diversity (TTR = 0.15) does not require complex features to produce excellent classification results. Note that the performance of our custom *gratitude* binary classifier in E6 (F1 = 0.910) is not significantly worse than C1 (F1 = 0.921). The custom *gratitude* binary classifier contains only 159 features as opposed to C1 which has 6526 features. As for the other 8 emotion categories, it is possible that the underlying linguistic patterns within the category are not adequately captured using the lexical features. Apart from *gratitude*, *anger*, *admiration* and *fear*, the other 5 emotion categories also have very few positive examples in the corpus. Another possible solution is to increase the number of positive examples of these emotion categories for training and testing.

### **5.3.3 Summary: Feature-related Experiments**

We can conclude that customizing features for each emotion category offers great potential for improving the classification performance of the 28 binary classifiers. In most cases, classification performance using cue-based features equaled or bettered corpus-based features. Each emotion category exhibits a set of unique linguistic characteristics that can be leveraged to improve the performance the classifier. We have mainly utilized lexical features in these experiments and have yet to systematically explore the salient syntactic and semantic features that may be useful to improve classification performance. A one-size-fits-all classifier for fine-grained emotion detection produces decent results but gains in performance for some emotion categories may come at the expense of others. Custom models offer us more freedom in maximizing the performance of an emotion category without negatively affecting the others.

## **5.4 Task 3: Sample-related Experiments**

There are two parts to the sample-related experiments. The first part discusses class imbalance experiments and the second part examines if training the classifier with subsets of data based on the four sampling strategies affects performance.

### **5.4.1 Class Imbalance Strategies**

For each emotion category, there are far more negative examples than positive examples. A common notion in machine learning is that unbalanced data can produce unsatisfactory classifiers (Provost, 2000). We examine if artificially rebalancing the data would produce better classifiers. Since we have fewer positive examples in the data, we first downsampled the number of negative examples to match that of the positive examples (i.e., ratio of positive to negative examples is 1:1). The downsampling strategy basically reduces the number of examples from the majority class, which is the negative class in our case.

<b>SMO</b>	<b>Cross Validation: 10-fold</b>					<b>Train/Test Split</b>				
<b>Category</b>	<b># Train</b>	<b># Test</b>	<b>P</b>	<b>R</b>	<b>F1</b>	<b># Train</b>	<b># Test</b>	<b>P</b>	<b>R</b>	<b>F1</b>
Admiration	761	85	0.744	0.757	0.750	340	4661	0.072	0.672	0.131
Amusement	1247	139	0.800	0.774	0.787	573	4661	0.124	0.728	0.212
Anger	2270	252	0.722	0.688	0.704	1010	4661	0.167	0.663	0.266
Boredom	90	10	0.822	0.771	0.796	42	4661	0.005	0.769	0.010
Confidence	208	23	0.682	0.682	0.682	90	4661	0.017	0.636	0.033
Curiosity	176	20	0.857	0.774	0.814	77	4661	0.032	0.680	0.062
Desperation	109	12	0.574	0.466	0.514	50	4661	0.008	0.611	0.016
Doubt	298	33	0.741	0.690	0.715	119	4661	0.028	0.735	0.053
Excitement	1296	144	0.799	0.805	0.802	575	4661	0.161	0.721	0.264
Exhaustion	92	10	0.706	0.735	0.720	37	4661	0.003	0.571	0.007
Fascination	385	43	0.709	0.691	0.700	170	4661	0.033	0.719	0.063
Fear	451	50	0.738	0.661	0.698	189	4661	0.031	0.676	0.058
Gratitude	985	109	0.966	0.927	0.946	438	4661	0.310	0.849	0.454
Happiness	3377	375	0.754	0.738	0.746	1488	4661	0.310	0.717	0.433
Hate	363	40	0.871	0.807	0.838	168	4661	0.054	0.698	0.100
Hope	986	110	0.800	0.797	0.798	455	4661	0.126	0.811	0.218
Indifference	128	14	0.769	0.735	0.752	48	4661	0.015	0.667	0.029
Inspiration	141	16	0.732	0.693	0.712	60	4661	0.023	0.577	0.045
Jealousy	64	7	0.793	0.676	0.730	33	4661	0.006	0.700	0.012
Longing	229	25	0.841	0.744	0.789	94	4661	0.025	0.800	0.049
Love	1285	143	0.837	0.825	0.831	543	4661	0.181	0.769	0.293
Pride	402	45	0.929	0.854	0.890	191	4661	0.178	0.797	0.291
Regret	289	32	0.800	0.680	0.735	113	4661	0.054	0.714	0.101
Relaxed	145	16	0.732	0.675	0.703	67	4661	0.010	0.652	0.020
Sadness	985	109	0.798	0.766	0.782	436	4661	0.101	0.806	0.180
Shame	170	19	0.762	0.711	0.736	81	4661	0.010	0.520	0.020
Surprise	502	56	0.727	0.680	0.703	228	4661	0.042	0.740	0.080
Sympathy	191	21	0.891	0.812	0.850	90	4661	0.038	0.680	0.072
<b>Macro-avg</b>			<b>0.782</b>	<b>0.736</b>	<b>0.758</b>			<b>0.077</b>	<b>0.703</b>	<b>0.128</b>
<b>Micro-avg</b>			<b>0.785</b>	<b>0.757</b>	<b>0.771</b>			<b>0.066</b>	<b>0.727</b>	<b>0.121</b>

Table 5.20: Precision, recall and F1 for SMO classifier based downsampled data evaluated using cross validation and train/test split

First, we tested a downsampling strategy similar to that employed in Brooks et al. (2013). The positive and negative examples in the P1 corpus are downsampled to equal proportions (i.e., ratio of 1:1) for each emotion category. Since 10-fold cross validation is used to evaluate the classifiers, we are essentially applying downsampling on both the training and test

data sets. We then conducted a second set of experiments by having the classifiers learn only on downsampled training data but evaluated using the same held out test set across all emotion categories.

<b>BayesNet</b>	<b>Cross Validation: 10-fold</b>					<b>Train/Test Split</b>				
<b>Category</b>	<b># Train</b>	<b># Test</b>	<b>P</b>	<b>R</b>	<b>F1</b>	<b># Train</b>	<b># Test</b>	<b>P</b>	<b>R</b>	<b>F1</b>
Admiration	761	85	0.659	0.658	0.658	340	4661	0.431	0.190	0.263
Amusement	1247	139	0.887	0.800	0.841	573	4661	0.895	0.703	0.787
Anger	2270	252	0.705	0.612	0.655	1010	4661	0.184	0.581	0.280
Boredom	90	10	1.000	0.458	0.629	42	4661	0.833	0.385	0.526
Confidence	208	23	0.891	0.373	0.526	90	4661	0.000	0.000	0.000
Curiosity	176	20	0.889	0.860	0.874	77	4661	0.059	0.520	0.105
Desperation	109	12	1.000	0.224	0.366	50	4661	0.000	0.000	0.000
Doubt	298	33	0.738	0.589	0.655	119	4661	0.034	0.347	0.063
Excitement	1296	144	0.809	0.730	0.768	575	4661	0.167	0.779	0.276
Exhaustion	92	10	1.000	0.245	0.393	37	4661	0.000	0.000	0.000
Fascination	385	43	0.969	0.461	0.625	170	4661	0.656	0.368	0.472
Fear	451	50	0.727	0.657	0.690	189	4661	0.032	0.757	0.062
Gratitude	985	109	0.952	0.908	0.929	438	4661	0.408	0.818	0.544
Happiness	3377	375	0.732	0.696	0.714	1488	4661	0.272	0.662	0.386
Hate	363	40	0.835	0.740	0.785	168	4661	0.041	0.736	0.078
Hope	986	110	0.768	0.705	0.735	455	4661	0.097	0.723	0.171
Indifference	128	14	0.963	0.382	0.547	48	4661	0.008	0.381	0.016
Inspiration	141	16	0.825	0.440	0.574	60	4661	0.900	0.346	0.500
Jealousy	64	7	0.792	0.559	0.655	33	4661	0.000	0.000	0.000
Longing	229	25	0.970	0.537	0.691	94	4661	0.323	0.571	0.412
Love	1285	143	0.852	0.731	0.787	543	4661	0.314	0.636	0.421
Pride	402	45	0.798	0.798	0.798	191	4661	0.046	0.766	0.086
Regret	289	32	0.812	0.621	0.704	113	4661	0.120	0.592	0.199
Relaxed	145	16	1.000	0.039	0.075	67	4661	0.000	0.000	0.000
Sadness	985	109	0.772	0.714	0.742	436	4661	0.091	0.535	0.156
Shame	170	19	0.757	0.311	0.441	81	4661	1.000	0.080	0.148
Surprise	502	56	0.836	0.440	0.576	228	4661	0.609	0.182	0.280
Sympathy	191	21	0.930	0.653	0.767	90	4661	0.134	0.800	0.230
<b>Macro-avg</b>			<b>0.852</b>	<b>0.569</b>	<b>0.650</b>			<b>0.273</b>	<b>0.445</b>	<b>0.231</b>
<b>Micro-avg</b>			<b>0.789</b>	<b>0.671</b>	<b>0.725</b>			<b>0.133</b>	<b>0.592</b>	<b>0.218</b>

Table 5.21: Precision, recall and F1 for BayesNet classifier based downsampled data evaluated using cross validation and train/test split

Cross validation results in Table 5.20 (SMO) and Table 5.21 (BayesNet) seem to suggest that the downsampling strategy significantly increases the performance of the classifiers (i.e., high micro precision and recall). It is not clear, however if these improvements are real or if they are an artifact of the evaluation methodology; it is possible that we are not really producing better classifiers but, rather, have made the classification problem easier by using balanced data sets. A huge contrast in F1 is observed when the cross validation results are compared to the results generated using the train/test split.

F1 in the train/test split is significantly lower than cross validation. The drop in micro F1 in the train/test split is due to low precision. This suggests that the apparent increase in performance based on the cross validation results is not the result of better classifiers. Therefore, care must be taken when interpreting the inflated cross validation results. The train/test split offers a more reliable portrayal of how downsampling affects the behavior of the classifiers.

Based on the train/test split results, overall F1 for both SMO and BayesNet trained on downsampled data is relatively low compared to the classifiers trained on actual distribution of the data. Contrary to conventional wisdom, we found that downsampling does not lead to improvement in overall classifier performance. Even for our top performing emotion category, *gratitude*, the classifier trained on downsampled data only manages to achieve  $F1 = 0.454$  for SMO and  $F1 = 0.544$  for BayesNet. This pales in comparison to the performance of the basic models used in our feature-related experiments which achieve  $F1 = 0.921$  for SMO and  $F1 = 0.901$  for BayesNet.

One other conclusion we can draw from Table 5.20 and Table 5.21 is that downsampling the training data generally produces classifiers with higher recall. Based on this observation, we then examine if we can increase the recall of the classifier without significantly reducing precision using our basic SMO and BayesNet models. Since we have a limited amount of positive examples, one approach is to use all of the positive examples and hold it constant but



gradually increase the number of negative examples in the training data in subsequent rounds. We subsample data from the training set to get a 1:1 ratio of positive and negative examples in the first round. We then incrementally increase the number of negative examples by two times, three times and so forth up till the ratio of 1:10. For each classifier, we added an extra round with a 1:20 ratio of positive and negative examples. The final round uses all of the training data without downsampling. The classifiers from all rounds are evaluated on the same test set. The overall precision, recall and F1 trend lines are shown in Figure 5.2.

SMO and BayesNet exhibit similar trends in the results. Figure 5.2 shows that a classifier trained with a perfectly balanced ratio (1:1) of positive and negative examples starts out with very high recall but very low precision. As the number of negative examples increases in subsequent rounds, recall starts to fall and precision starts to climb up more steeply at first and then more gradually at the end. It is interesting to note that precision and recall coincide at a certain point on the chart. This is the point where the classifier achieves balanced precision and recall. After this point, precision continues to climb uphill as more negative examples are added while recall keeps heading downhill.

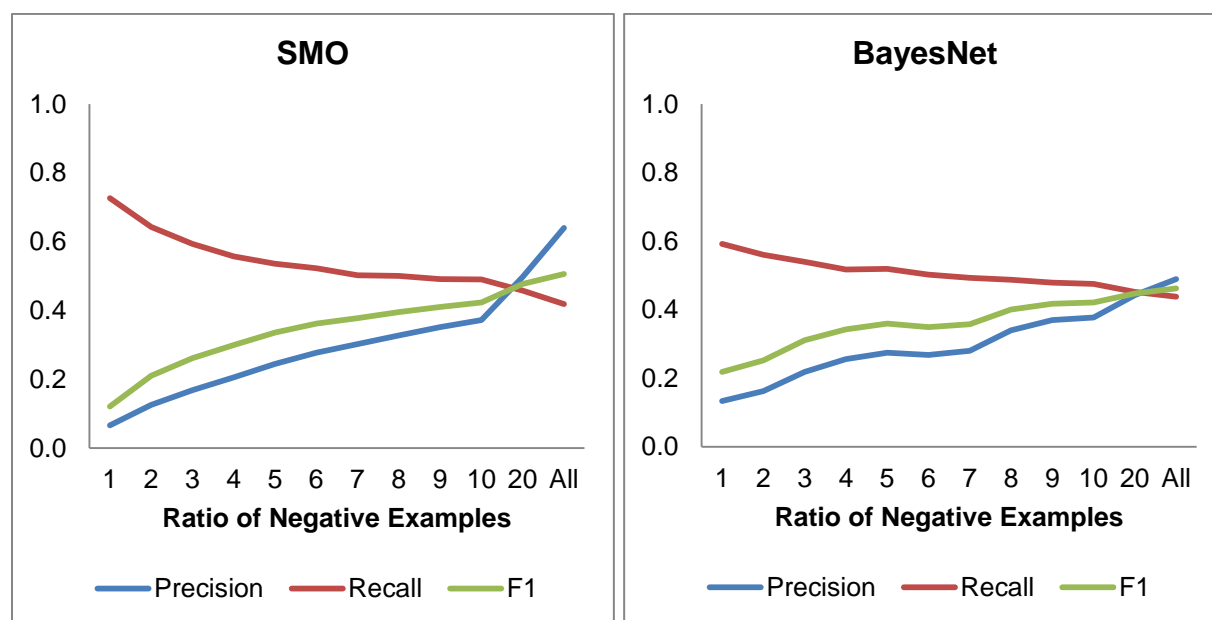


Figure 5.2: Precision, recall and F1 based on 12 iterations of downsampled training data

Several conclusions about the classifier behavior can be drawn from the results of this experiment. First, we can obtain a classifier with higher recall by training the classifier with a more balanced ratio of positive and negative examples but at the expense of precision. Second, adding negative examples to train a classifier is not entirely useless. This strategy can help increase precision but usually results in a lower recall. Third, it is also possible to obtain a classifier that yields roughly the same precision and recall by adjusting the ratio of positive and negative examples in the training data.

The decision on what proportion of positive and negative examples to be used as training data in a binary classifier depends on the application. For example, a classifier with high precision is more useful in a system used to detect occurrences of natural disasters by recognizing the expression of fear and sadness in tweets. As long as the system is able to correctly identify one or a few instances of fear expressed in the event of a disaster, the system can then set off the right alarm to warn relevant authorities. In this particular application, a system with low precision will set off many false alarms. In contrast, in an automatic qualitative content analysis system, if the goal of a researcher is to automatically identify all instances of happiness and sadness in a large corpus of tweets, recall then becomes more important as the cost of missing an example is higher than having the classifier make an incorrect prediction. It is harder to find examples that the classifier misses as opposed to correcting an incorrect prediction.

#### **5.4.2 Effects of Sampling Strategies on Classifier Performance**

Tweets in the corpus are sampled using four different sampling strategies. We hypothesize that a classifier trained on combined data from the four sampling strategies will perform better than a classifier trained only on data from a single sampling strategy. Using the basic SMO and BayesNet models, we first train the classifier on the full set of training data. We then split the training set into four subsamples based on sampling strategy, and train four other

separate classifiers using the four subsamples respectively. All five classifiers are evaluated on the same test set.

Train Sample	SMO			BN		
	P	R	F1	P	R	F1
<b>ALL</b>	0.640	0.419	<b>0.506</b>	0.489	0.438	<b>0.462</b>
<b>RANDOM</b>	0.585	0.242	0.343	0.567	0.295	0.388
<b>TOPIC</b>	0.572	0.260	0.357	0.547	0.343	0.421
<b>AVG-USER</b>	0.514	0.295	0.375	0.369	0.324	0.345
<b>SEN-USER</b>	0.611	0.206	0.308	0.638	0.235	0.343

Table 5.22: Precision, recall and F1 of SMO and BayesNet trained with different samples

Table 5.22 shows that a classifier utilizing data from all four sampling strategies (ALL) performs better than RANDOM, TOPIC, AVG-USER and SEN-USER respectively. ALL achieves  $F1 = 0.506$  for SMO and  $F1 = 0.462$  for BayesNet. Note that the performance values reported in Table 5.22 differ from the ones reported in the feature-related experiments as the classifier in this section is evaluated using the train/test split. Cross validation uses the full corpus for training in the feature-related experiments.

We also examined the outcome of testing a classifier trained using data from a sampling strategy on a subsample of the test set containing data from the same sampling strategy as well as on three other subsamples from different sampling strategies. We expect classifiers trained on a subsample to perform well on test data retrieved using the same sampling strategy but we hypothesize that the ALL classifier trained with data from all four sampling strategies will perform equally well or better when tested on subsamples from each sampling strategy.

First, we split the test set into four subsamples based on sampling strategy. The five classifiers, each trained on ALL, RANDOM, TOPIC, AVG-USER and SEN-USER, are then evaluated on each of the four test subsamples. We compare the test results based on the four test subsamples to ALL (i.e., evaluation of the classifiers were done on the full test set) in Table 5.23 (SMO) and Table 5.24 (BayesNet).

The ALL classifier generally outperforms classifiers trained on only a single subsample even when tested on data from the same subsample as evidenced by the results in Table 5.23 and Table 5.24. Classifiers trained on only a single subsample exhibit worse performance when evaluated on test data from other subsamples. The ALL classifier shows consistent performance well when tested on samples from a single sampling strategy and as well as combined data from all four sampling strategies. This shows that the ALL classifier is more generalizable compared to classifiers utilizing training data from only a single sampling strategy.

Train Sample	Test Sample				
	ALL	RANDOM	TOPIC	AVG-USER	SEN-USER
ALL	<b>0.506</b>	<b>0.513</b>	<b>0.492</b>	<b>0.438</b>	<b>0.605</b>
RANDOM	0.343	0.407	0.306	0.291	0.386
TOPIC	0.357	0.372	0.418	0.270	0.356
AVG-USER	0.375	0.374	0.350	0.407	0.370
SEN-USER	0.308	0.271	0.249	0.172	0.567

Table 5.23: F1 of SMO based on testing classifiers trained and tested based on different subsamples

Train Sample	Test Sample				
	ALL	RANDOM	TOPIC	AVG-USER	SEN-USER
ALL	<b>0.462</b>	<b>0.486</b>	<b>0.448</b>	<b>0.415</b>	0.510
RANDOM	0.388	0.462	0.348	0.359	0.394
TOPIC	0.421	0.468	0.434	0.373	0.408
AVG-USER	0.345	0.366	0.321	0.346	0.353
SEN-USER	0.343	0.310	0.288	0.196	<b>0.608</b>

Table 5.24: F1 of BayesNet based on testing classifiers trained and tested based on different subsamples

Sampling tweets using various sampling strategies increases the diversity of training data. As a result, the classifier is less biased to a particular topic and is more generalizable to tweets generated by the population on Twitter.

### 5.4.3 Summary: Sample-related Experiments

From the experiments described in this section, we observed three sample-related factors that affect the classifier performance: 1) size of training and test data, 2) proportion of positive and negative examples, and 2) diversity of training examples. First, downsampling the data in our skewed data set did not improve classifier performance. Training data downsampled to a 1:1 ratio of positive and negative examples can produce a classifier with high recall but very low precision. We also identified an important methodological issue wherein cross validation when used with a downsampling strategy tends to inflate the classifier performance measures. Thus, the train/test split offers a more accurate portrayal of the classifier performance when training data is downsampled. We also show that the ratio of positive and negative examples used in the training set affects the classifier behavior. Finally, training a classifier with the diversity of tweets collected using different sampling strategy reduces bias and increase the generalizability of the classifier.

## 5.5 Discussion

This section discusses the results from the pool of experiments we ran and compares machine learning performance to human performance at recognizing the 28 emotion categories.

### 5.5.1 Comparing Machine Learning Classification Performance

Overall classifier performance across all 28 emotion categories peaks at 0.564 in terms of micro-F1. However, a wide range in F1 scores is observed at the level of each individual emotion category. Classifiers have shown to perform remarkably well for some emotion categories achieving F1 as high as 0.9 but there are also a handful of emotion categories where F1 remains as low as 0.2. Table 5.25 compares the performance of the best overall custom SMO model using E6 (*custom*) on each emotion category to the best performing binary

classifier for each of the 28 individual categories (*max-ensemble*). *Max-ensemble* contains a series of binary classifiers that represent the “*cream of the crop*” in each emotion category.

Category	Custom (SMO-E6)					Max-Ensemble					Diff (F1)
	CL	FT	P	R	F1	CL	FT	P	R	F1	
Admiration	SMO	E6	0.500	0.165	0.248	SMO	C3	0.396	0.223	<b>0.286</b>	0.037
Amusement	SMO	E6	0.909	0.674	0.774	BN	C1	0.899	0.759	<b>0.823</b>	0.050
Anger	SMO	E6	0.752	0.215	0.334	SMO	C5	0.481	0.329	<b>0.391</b>	0.056
Boredom	SMO	E6	0.833	0.385	0.526	BN	C1	0.880	0.458	<b>0.603</b>	0.076
Confidence	SMO	E6	1.000	0.121	0.216	SMO	E5	0.625	0.152	<b>0.244</b>	0.028
Curiosity	SMO	E6	1.000	0.240	0.387	BN	C1	0.859	0.591	<b>0.701</b>	0.314
Desperation	SMO	E6	1.000	0.278	<b>0.435</b>	BN	E1	1.000	0.278	0.435	0
Doubt	SMO	E6	0.571	0.163	0.254	SMO	E6a	0.615	0.163	<b>0.258</b>	0.004
Excitement	SMO	E6	0.917	0.423	0.579	SMO	E4	0.926	0.423	<b>0.581</b>	0.002
Exhaustion	SMO	E6	1.000	0.143	0.250	SMO	C1	0.611	0.224	<b>0.328</b>	0.078
Fascination	SMO	E6	0.600	0.263	0.366	BN	E6	0.656	0.368	<b>0.472</b>	0.106
Fear	SMO	E6	0.688	0.149	0.244	BN	C5	0.653	0.268	<b>0.380</b>	0.135
Gratitude	SMO	E6	0.928	0.893	0.910	SMO	C5	0.930	0.914	<b>0.922</b>	0.011
Happiness	SMO	E6	0.772	0.480	<b>0.592</b>	SMO	E6	0.772	0.480	0.592	0
Hate	SMO	E6	0.824	0.528	0.644	SMO	L2	0.876	0.552	<b>0.677</b>	0.034
Hope	SMO	E6	0.901	0.572	<b>0.700</b>	SMO	E6	0.901	0.572	0.700	0
Indifference	SMO	E6	0.333	0.050	0.087	BN	C2	0.611	0.162	<b>0.256</b>	0.169
Inspiration	SMO	E6	0.818	0.346	0.486	BN	C1	0.917	0.440	<b>0.595</b>	0.108
Jealousy	SMO	E6	0.250	0.100	0.143	BN	C5	0.938	0.441	<b>0.600</b>	0.457
Longing	SMO	E6	0.619	0.371	0.464	SMO	E2	0.722	0.371	<b>0.491</b>	0.026
Love	SMO	E6	0.837	0.458	0.592	SMO	E6a	0.821	0.529	<b>0.643</b>	0.051
Pride	SMO	E6	0.980	0.750	<b>0.850</b>	SMO	E6	0.980	0.750	0.850	0
Regret	SMO	E6	0.667	0.408	0.506	BN	E2	0.649	0.490	<b>0.558</b>	0.052
Relaxed	SMO	E6	1.000	0.087	0.160	BN	C1	0.789	0.195	<b>0.313</b>	0.153
Sadness	SMO	E6	0.859	0.424	0.567	BN	C1	0.724	0.468	<b>0.569</b>	0.001
Shame	SMO	E6	0.750	0.120	0.207	BN	C5	0.667	0.489	<b>0.564</b>	0.357
Surprise	SMO	E6	0.705	0.408	<b>0.517</b>	SMO	E6	0.705	0.408	0.517	0
Sympathy	SMO	E6	0.762	0.640	<b>0.696</b>	SMO	E6	0.762	0.640	0.696	0
Macro-avg			0.778	0.352	0.455			0.763	0.434	<b>0.537</b>	0.082
Micro-avg			0.818	0.430	0.564			0.732	0.484	<b>0.582</b>	0.018

Table 5.25: Comparing F1 of each emotion category between SMO-E6 and the best performing binary classifier per category (CL: Classifier, FT: Features and the highest F1 between SMO-E6 and Max-Ensemble for each emotion category is in bold)

For five emotion categories in particular (i.e., *gratitude*, *pride*, *amusement*, *hope* and *sympathy*), performance remains consistently high across the two sets of results. Gratitude remains as the best performing category with F1 above 0.9. Results of the top performing emotion categories are promising and show that machine learning classifiers can be relied upon to produce high quality predictions in fine-grained emotion classification. Interestingly, three emotion categories (i.e., *curiosity*, *jealousy* and *shame*) fall into the opposite end: low performing categories in *custom* with significantly better performing counterparts in *max-ensemble*. The largest difference is observed in *jealousy*, where the classifier in *custom* achieves F1 of 0.1 but the counterpart in *max-ensemble* achieves F1 of 0.6. The emotion categories that remain at the bottom in both *custom* and *max-ensemble* are *relaxed*, *admiration*, *doubt*, *indifference* and *confidence*.

F1	Custom: SMO-E6		Max-Ensemble	
	Emotion Categories	Count	Emotion Categories	Count
High	Gratitude, Pride, Amusement, Hope, Sympathy, Hate	6	Gratitude, Pride, Amusement, Curiosity, Hope, Sympathy, Hate, Love, Boredom, Jealousy	10
Medium	Love, Happiness, Excitement, Sadness, Boredom, Surprise, Regret, Inspiration, Longing, Desperation	10	Inspiration, Happiness, Excitement, Sadness, Shame, Regret, Surprise, Longing, Fascination, Desperation	10
Low	Curiosity, Fascination, Anger, Doubt, Exhaustion, Admiration, Fear, Confidence, Shame, Relaxed, Jealousy, Indifference	12	Anger, Fear, Exhaustion, Relaxed, Admiration, Doubt, Indifference, Confidence	8

Table 5.26: Emotion categories with high, medium and low performance

Based on the F1 scores presented in Table 5.25, we divide the emotion categories into three groups based on level of performance. Table 5.26 shows the three groups of emotion categories with high ( $F1 \geq 0.6$ ), moderate ( $0.4 \leq F1 < 0.6$ ) and low ( $F1 < 0.4$ ) performance. Each group roughly contains one third of the emotion categories. *Max-ensemble* shows high performance for 10 emotion categories as opposed to *custom* with only 6 high performing categories. Apparently, *curiosity*, *boredom* and *jealousy* respond more favorably to BayesNet than SMO. By comparing *custom* and *max-ensemble*, we can infer that an ideal supervised

learning solution for fine-grained emotion classification is one that allows flexibility not only in feature selection but also in the choice of classifier per category.

### 5.5.2 Comparing Human and Machine Classification Performance

We computed the correlation coefficients between the human annotation performance measures (i.e., *distinctiveness*, *percentage of full agreement* and *intuitiveness*) and the machine learning classifier performance (i.e., F1 based on Custom: SMO-E6). There is a strong positive correlation between F1 and the distinctiveness ( $r(26) = 0.55$ ), full agreement ( $r(26) = 0.64$ ), and intuitiveness ( $r(26) = 0.66$ ) respectively. The results suggest that classifier performance tends to be high for the emotion categories that annotators are able to detect with high reliability. Lexical diversity is negatively correlated with F1 ( $r(26) = 0.67$ ), indicating that classifier performance tends to be high for the emotion categories with low lexical diversity. All four correlations are statistically significant ( $p < 0.01$ ).

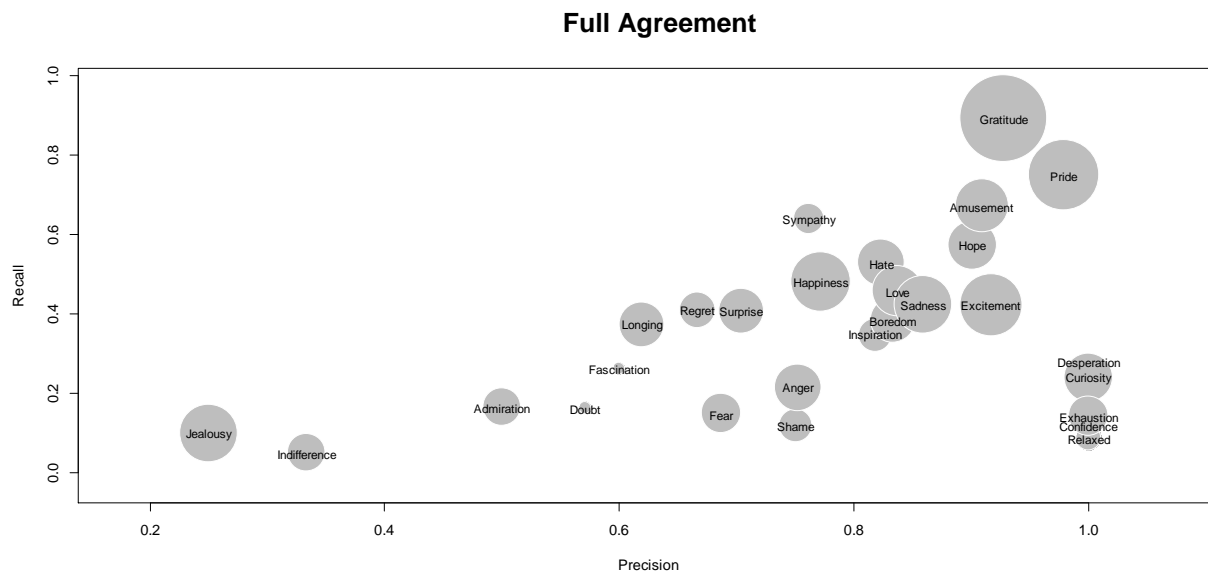


Figure 5.3: Bubble chart with three data dimensions (x: precision, y: recall and bubble size: percentage of full agreement)



To gain more insights of the relationship between human and machine classification, we map the precision and recall of the Custom model into a bubble chart shown in Figure 5.3. The size of the bubble with a category name represents percentage of full agreement among human annotators, the bigger the bubble size, the higher the percentage.

It can be noted from Figure 5.3 that the larger bubbles are generally concentrated in the upper right position of the chart. Indeed, we can verify that the machine learning classifier tends to yield better performance for the emotion categories with higher frequency of full agreement. There are two notable exceptions. *Jealousy* has a relatively high percentage of full agreement among annotators but the machine learning classifier produces very low precision and recall. *Jealousy* has the least number of positive examples in the corpus as shown in Figure 5.4 so the lack of training data may explain the low classifier performance. Also, jealousy is a rather complex emotion. Unless explicit emotion words such as “*jealous*” or “*envy*” are present in the tweet, the detection of jealousy may require a better understanding between the relationships of the actors mentioned in the tweet as illustrated by Example 5.1 and Example 5.2.

**Example 5.1:** My boyfriend isn't allowed to hug other chicks, you better pet that hoe on the head like a dog. [**Jealousy**]

**Example 5.2:** I would kill to have the body of the inzano twins ok [**Jealousy**]

On the other hand, sympathy has a relatively low percentage of full agreement but the classifier manages to produce high precision and recall. Interestingly, the way sympathy is expressed in tweets is to a large extent dictated by social conventions. Tweeters tend to repetitively use similar phrases when expressing sympathy towards the misfortune of others (see Example 5.3 and Example 5.4). These repetitive patterns can be picked up easily by the machine learning classifier.

**Example 5.3:** My prayers are w/ families of Ambassador Stevens & three other Americans killed in this appalling & completely unacceptable attack [**Sympathy**]

**Example 5.4:** My prayers are with the loved ones of the Bramlage family of Junction City. They were community role models and will be greatly missed. **[Sympathy]**

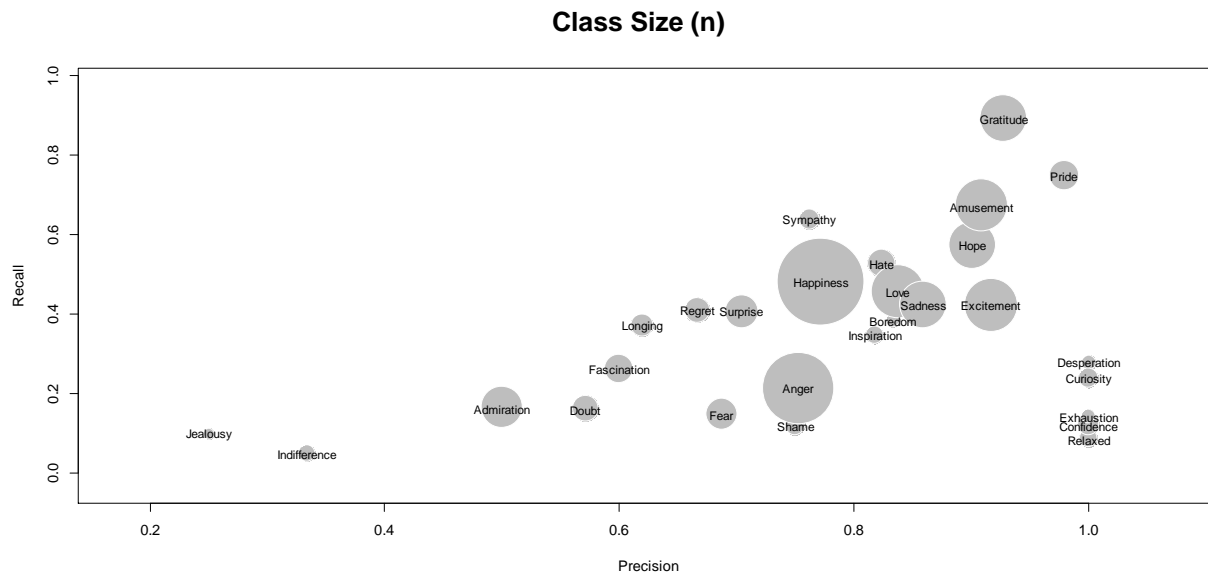


Figure 5.4: Bubble chart with three data dimensions (x: precision, y: recall and bubble size: class size)

We also examine if there is any relationship between the class size and the performance of the machine learning classifier. We use the same bubble chart to represent precision and recall on the x-axis and y-axis. The bubble size in Figure 5.4 represents the number of positive examples for each emotion category. Note that the larger bubbles are not on the far upper right position in the chart, suggesting that classifier performance is not necessarily related to class size. For instance, the classifier performs remarkably well for *pride* and *gratitude* although the size of both classes is smaller than *happiness*. Based on our observation, as long as the class size reaches a critical mass, the classifier should be able to yield decent performance. Classifier performance is likely to be low for the emotion categories with very few positive examples as in the case of *jealousy* and *indifference*.

Table 5.27 summarizes human and machine classification performance for each of the 28 emotion categories. The three emotion categories in which the annotators and machine

learning classifiers achieve consistently high performance are *gratitude*, *pride* and *amusement*. The real strength of the machine learning classifier lies in its capability to reliably detect some of the emotion categories that are difficult for the annotators to recognize, notably for *sympathy*, *boredom*, *curiosity*, *love*, *jealousy*, *hate* and *hope*.

Category	Human			Machine		
	Distinctiveness	Full Agreement	Intuitiveness	P	R	F1
Amusement	H	H	H	H	H	H
Excitement	H	H	H	H	M	M
Gratitude	H	H	H	H	H	H
Happiness	H	H	H	H	M	M
Pride	H	H	H	H	H	H
Sadness	H	H	H	H	M	M
Jealousy	H	H	M	H	M	H
Love	H	H	M	H	M	H
Anger	M	M	M	M	L	L
Curiosity	M	M	H	H	M	H
Exhaustion	M	M	H	H	L	L
Admiration	M	M	M	L	L	L
Boredom	M	M	M	H	M	H
Fear	M	M	M	H	L	L
Hate	M	M	M	H	M	H
Hope	M	M	M	H	M	H
Indifference	M	M	M	H	L	L
Longing	M	M	M	H	L	M
Surprise	M	M	M	H	M	M
Regret	M	M	M	H	M	M
Shame	M	L	M	H	M	M
Inspiration	L	L	M	H	M	M
Sympathy	L	L	M	H	H	H
Confidence	L	L	L	H	L	L
Desperation	L	L	L	H	L	M
Doubt	L	L	L	H	L	L
Fascination	L	L	L	H	L	M
Relaxed	L	L	L	H	L	L

Table 5.27: Comparing human and machine annotation performance

## 5.6 Conclusion

In this chapter, we have addressed R3 by showing empirical evidence that the salient cues humans associate with each emotion category (i.e., cue-based features) serve as better features for machine learning classification of fine-grained emotion in text (Section 5.3). In general, cue-based features equaled or bettered corpus-based and lexicon-based features with far fewer features. A classifier using cue-based features is also more advantageous as it can be trained much more quickly than corpus-based features, which allows it to scale to large data sets. In addition, leveraging the emotion cues to select custom features for each emotion category further improves overall classifier performance (Section 5.3.2.3). As for the individual emotion categories, half the emotion categories respond most favorably to the cue-based features.

We addressed R4 by showing that it is feasible to apply current machine learning techniques to fine-grained emotion classification through the series of machine learning experiments we ran using EmoTweet-28. First, we have identified two machine learning algorithms that perform well in this problem domain: SMO and BayesNet (Section 5.2). We have also demonstrated that classifier performance on 28 emotion categories is not far worse than having a classifier perform classification based on emotion valence (positive, negative, neutral and no emotion) and emotion presence (has emotion and no emotion) (Section 5.2.2.1).

Classifier performance varies for each of the 28 emotion categories. Of the 28 emotion categories, the classifier achieves high performance in detecting 10 emotion categories as shown in Table 5.27. Only 8 categories suffer from low performance. Interestingly, these 8 emotion categories do not exactly match the 8 that humans have the most difficulty in recognizing, suggesting that machine learning can be leveraged to detect some emotion categories that humans cannot reliably recognize in tweets. In our sample-related experiments, we have demonstrated that downsampling the training data did not improve classifier

performance (Section 5.4.1) whereas training the classifier with data retrieved from four different sampling strategies improves its generalizability to detect emotions on Twitter (Section 5.4.2).

Emotion Category	Ang	Dis	Fea	Joy	Sad	Sur	Lov	Gra	# Cat: Macro-avg
<b>F1</b>									
Mohammad (2012)	0.28	0.19	0.51	0.62	0.39	0.45	-	-	6: 0.41
Roberts et. al (2012)	0.64	0.67	0.74	0.68	0.69	0.61	0.66	-	7: 0.67
Wang et. al (2012)	0.72	-	0.44	0.72	0.65	0.14	0.52	0.57	7: 0.54
<b>EmoTweet-28 (Custom)</b>	<b>0.33</b>	<b>0.64</b>	<b>0.24</b>	<b>0.59</b>	<b>0.57</b>	<b>0.52</b>	<b>0.59</b>	<b>0.91</b>	<b>8: 0.55</b> <b>28: 0.45</b>
<b>EmoTweet-28 (Ensemble)</b>	<b>0.39</b>	<b>0.68</b>	<b>0.38</b>	<b>0.59</b>	<b>0.57</b>	<b>0.52</b>	<b>0.64</b>	<b>0.92</b>	<b>8: 0.59</b> <b>28: 0.54</b>
<b>Precision</b>									
Mohammad (2012)	0.37	0.31	0.6	0.65	0.42	0.51	-	-	6: 0.48
Roberts et. al (2012)	0.67	0.72	0.9	0.66	0.75	0.63	0.73	-	7: 0.72
Wang et. al (2012)	0.7	-	0.6	0.68	0.63	0.45	0.58	0.67	7: 0.62
<b>EmoTweet-28 (Custom)</b>	<b>0.75</b>	<b>0.82</b>	<b>0.69</b>	<b>0.77</b>	<b>0.86</b>	<b>0.71</b>	<b>0.84</b>	<b>0.93</b>	<b>8: 0.80</b> <b>28: 0.78</b>
<b>EmoTweet-28 (Ensemble)</b>	<b>0.48</b>	<b>0.88</b>	<b>0.65</b>	<b>0.77</b>	<b>0.72</b>	<b>0.70</b>	<b>0.82</b>	<b>0.93</b>	<b>8: 0.75</b> <b>28: 0.76</b>
<b>Recall</b>									
Mohammad (2012)	0.22	0.13	0.44	0.6	0.36	0.41	-	-	6: 0.36
Roberts et. al (2012)	0.62	0.62	0.63	0.7	0.64	0.59	0.6	-	7: 0.63
Wang et. al (2012)	0.73	-	0.35	0.77	0.67	0.08	0.46	0.50	7: 0.51
<b>EmoTweet-28 (Custom)</b>	<b>0.22</b>	<b>0.53</b>	<b>0.15</b>	<b>0.48</b>	<b>0.42</b>	<b>0.41</b>	<b>0.46</b>	<b>0.89</b>	<b>8: 0.44</b> <b>28: 0.35</b>
<b>EmoTweet-28 (Ensemble)</b>	<b>0.33</b>	<b>0.55</b>	<b>0.27</b>	<b>0.48</b>	<b>0.47</b>	<b>0.41</b>	<b>0.53</b>	<b>0.91</b>	<b>8: 0.49</b> <b>28: 0.43</b>

Table 5.28: Classifier performance in the state-of-the-art of automatic emotion detection in tweets (Ang: Anger, Dis: Disgust/Hate, Fea: Fear, Joy: Joy/Happiness, Sad: Sadness, Sur:

Surprise, Lov: Love, Gra: Gratitude/Thankfulness)

Automatic emotion detection in text, regardless of the level of granularity, is a challenging task. Table 5.28 shows our results in the context of other related work on coarse-grained emotion classification in tweets. We report average results from our *custom* and *max-ensemble* models based on the full set of 28 emotion categories as well as on only the 8 emotion categories found in the related work. The purpose is not to make a direct comparison between our results and the state-of-the-art since the data, features and classifier setup vary

across the different studies. Rather, we want to highlight that it is possible to extend the capability of machine learning classifiers to handle fine-grained emotion detection in tweets while achieving current expected standards.

State-of-the-art machine learning classifiers achieve only moderate performance in detecting emotions in tweets. Training the classifier with a significant amount of data collected using distant supervision (i.e., retrieving labeled data using emotion hashtags) as seen in the study by Mohammad (2012) and Wang et al. (2012) also yields similar outcome. Overall performance of the classifiers in our study is comparable to related work shown in Table 5.28 even with significantly more number of categories. Our study adds to the current discourse on automatic classification of emotion categories that are common (*anger, fear, joy, sadness* and *surprise*) as well as those that are less common (*disgust, love* and *gratitude*). More importantly, our study provides a baseline of expected performance for new emotion categories (e.g., *amusement, jealousy, sympathy*, etc.) yet to be explored in this problem domain.

## Chapter 6: Conclusion and Future Work

In this final chapter, we first review the major contributions of the research. We then discuss the limitations of the research and describe future challenges to address as well as directions to pursue.

### 6.1 Contributions

This research narrows the gap between our understanding of the linguistic characteristics of a fine-grained set of emotions expressed in tweets and the computational linguistic approaches that can be leveraged to automatically recognize this set of emotion categories. Our research has important theoretical, language resource and methodological contributions.

#### 6.1.1 Theoretical

We have identified 28 discrete emotion categories that are representative of the range of emotions expressed in tweets, an extension to the six or eight basic emotion categories commonly used in the state-of-the-art. The 28 emotion categories are derived from actual data based on human knowledge of emotion. These categories are more relevant to the content of microblog text than the categories adopted from existing emotion theories in psychology which are mainly based on other manifestations of emotion such as facial or physiological expressions.

This research takes a first step in creating a framework or taxonomy of emotion categories based on text. Since there is currently no unifying theory of emotion in text, we adopted a more pragmatic and integrative view by drawing from various existing emotion theories in defining emotion in text. We treat this as a starting point to advance an emotion

theory based on textual expressions especially in the context of more informal types of text. Our findings further extend and enrich the current discourse on emotion not only in the computational linguistic community but also in psychology and linguistics in general.

We have also developed a detailed annotation scheme for the 28 emotion categories that can be adopted by other researchers to extend the size of the current corpus as well as to extend corpus development to in other domains such as customer reviews, blogs, chat logs, etc. The annotation scheme provides clear definitions and linguistic specifications of each emotion category that is aimed at encouraging the development of additional corpora to be shared and reused in the research community. As demonstrated in the first two phases of this research, data annotated by different annotators but using the same annotation scheme can be merged to expand the size of training and test data for machine learning classification. This reduces redundant effort to develop multiple small isolated emotion corpora that are difficult to merge into a single data set due to conflicts in the semantic representation of concepts with the same name.

### **6.1.2 Language Resources**

We have created a carefully hand-crafted emotion corpus containing 15,553 tweets. EmoTweet-28 is currently the largest emotion corpus annotated with 28 emotion categories. The corpus also comes with polarity and intensity ratings so it is useful for other kinds of research in sentiment analysis. We have developed detailed annotator agreement statistics for the corpus. The corpus contains a diversity of examples for an automated classifier to learn from. The corpus is not restricted to any single topic in particular and is generated with the goal of capturing emotions that reflect the actual range of emotions expressed by the general population on Twitter.

We produced an emotion lexicon containing emotion cues associated with each emotion category. We did not pose a limit to length of the lexical items included in the lexicon. Therefore,



the emotion lexicon contains emotion words, phrases, punctuation marks, emoticons, emojis and hashtags associated with each emotion category. We note that emojis only started to appear in tweets recently but are popular as a means to express emotions in microblog posts. Emojis often refer to slang – not the classic definition of the picture. Our emotion lexicon represents one of the first attempts to map emojis to emotion categories based on how these pictograms are actually used by tweeters.

### **6.1.3 Machine Learning**


We have built and tested a series of computational models to identify 28 emotion categories in tweets. We have demonstrated the feasibility of applying machine learning to fine-grained emotion detection in tweets and show that the classifiers can achieve performance that is comparable to the state-of-the-art emotion classification based on six to eight basic emotions. We identified through an extensive series of experiments a set of classifier and feature combinations that are effective for each of the 28 emotion categories. We also implemented our classifiers to handle the multiple-emotion-per-tweet problem, which many earlier studies avoid. Our binary classifiers achieve high precision but only moderate recall.

We developed a novel approach utilizing the emotion cues identified by human annotators to select informative features for each emotion category. Classifiers using cue-based features use more compact features but are able to yield performance comparable to or slightly better than that achieved with conventional corpus-based and lexicon-based features. Reducing the number of features to a set that is most relevant to each emotion category decreases training time, thus increasing the efficiency of the classifiers.

This research represents a step towards the development of accurate automatic emotion detectors and opens up new territories to be explored in sentiment analysis beyond the scope of identifying the semantic orientation in text. The automatic emotion detector utilizing the full set of 28 emotion categories can be applied to enhance sentiment analysis in a variety of applications

(e.g., identifying customer sentiment, personality detection, threat detections, recommender systems, etc.) and to advance the development of affective systems (e.g., building more expressive agents or avatars). For those who are interested in the detection of only a subset of the emotion categories, our classification setup also provides the flexibility to integrate the predictions from only a subset of the binary classifiers.

## 6.2 Limitations

*“Sometimes I have to keep my feelings to myself,  because I could find no language to describe them.”*

The tweet above highlights one limitation of automatic emotion detection. It is possible that some emotions can be communicated non-verbally or indirectly but cannot be directly expressed using words. Current automatic classifiers depend solely on textual features and detect only emotions that can be communicated in written form. Also, classifiers are trained to detect specific emotion categories and may miss ambiguous emotion signals where tweeters do not describe their emotion in definite terms.

The taxonomy of emotion categories we have developed is not exhaustive. First, the emotion categories are obtained from a corpus of 15,553 tweets. It is likely that we have captured the emotion categories commonly expressed in tweets but it is possible for some rare, but important, emotion categories to slip through the cracks. Second, we imposed a limit on the number of categories as a means to achieve a balance based on the ability of humans to distinguish emotions in text. If too many categories are used, humans cannot reliably distinguish between them. If too few categories are used, important distinctions are lost. The emotion categories are conceived on a level of granularity that the annotators are able to intuitively recognize.

We have also not tested how generalizable the set of 28 emotion categories across different social media platforms and various types of informal text. The set of 28 emotion

categories is derived from our content analysis of tweets. Tweeters may have adopted certain slang or lingo in expressing their emotions in tweets. Therefore, we cannot assume generalizability of the emotions categories to other media. There may be differences in the use of language to express emotions on different online platforms due to the affordances of the technology as well as the nature in the communities formed on these platforms.

From the empirical evidence we have presented on the human and machine learning performance on the 28 emotion categories, we think this is the right base set to represent emotions that are expressed in tweets. This set offers a good balance between too few and too many categories to represent the range of emotions expressed by tweeters. Both humans and machine learning achieve low performance for three emotion categories: *confidence*, *doubt* and *relaxed*. These categories offer limited use if both humans and machine learning are not able to recognize them in tweets. We will examine in our future work if these emotion categories can be better defined to boost both human and machine learning performance.

It is possible that labeling these categories as “emotion” rather than “affect” may leave space for contention as there is yet to be clear definition of what an emotion is in psychology. However, the categories lend themselves credibility as they were developed collectively based on lay people’s knowledge on emotion.

We also acknowledge that bias may be introduced into the ground truth by expert annotators who reviewed the annotator labels in order to determine the gold labels associated with each tweet. This is particularly true in Phase 2 where the adjudication was done by the primary researcher. The goal of the manual review effort is to ensure consistent annotation across Phase 1 and Phase 2 since AMT annotators in Phase 2 received less training than the annotators in Phase 1.

Our machine learning experiments are also not exhaustive. It is not practical to test every possible combination of features, classifiers and parameters. We started by examining the basic features and classification algorithms used in related work and pursued directions that

show promising results. Finally, we focused on lexical features in our feature-related experiments, and have not thoroughly explored syntactic and semantic features.

## 6.3 Challenges and Future Work

We have accomplished the goals established in this thesis and have presented the range of emotions humans and computers can detect in microblog text. Fine-grained emotion classification proves to be a challenging task. It is not a task that can be accomplished solely through the detection of emotion words. Through careful analysis of tweets with high disagreement and those that annotators struggle with, we have identified challenges that both annotators and automatic emotion detectors can be trained to better tackle going forward:

- The same emotion words can sometimes be used to express different emotions.
  - **Example 6.1:** Your sad devotion to Keynes has not helped you conjure up more jobs, or given you clairvoyance enough to predict... #StarWarsFiscalCliff **[Anger]**
  - **Example 6.2:** ugh the hymn of proof is making me sad again go away tox 2 **[Sadness]**
- The meaning of an emotion word can be modified by its surrounding words (e.g., negations and conditionals).
  - **Example 6.3:** I spent the majority of my weekend sitting in bed, editing & watching the entire 1st season of Dexter. 'Twas divine. I have no regrets. **[Happiness]**
  - **Example 6.4:** my confidence would be sky high if i wasn't just skin and bones **[Shame]**
- An emotion word can form a part of a proper name (e.g., movie, song or band name).
  - **Example 6.5:** ...off to Milan...c ya on the 19th of August in Novi Sad with Silicon Soul and Fakir!!! **[Excitement]**

- **Example 6.6:** 2009: Fruit Ninja 2010: Cut The Rope 2011: Temple Run 2012: Angry Birds 2013: Candy Crush 2014: Now THIS <http://t.co/YFzE5Mk7rH> **[No emotion]**
- Emotions can be embodied in subtle literary devices such as sarcasm or irony in which the literal meaning of the words is not the intended meaning.
  - **Example 6.7:** Too many shows going on tonight in Philly. GOOD THING I CAN GO TO NONE OF THEM THANKS SPRING BREAK **[Anger]**
  - **Example 6.8:** #americanairlines thanks for canceling my flight and rebooking it a day later. You book a specific return time and day for a reason! #fail **[Anger]**
- Emotions can be expressed through descriptions of behaviors, actions, relationships between actors and even physiological reactions (e.g., pounding heart or blood rushing to the face).
  - **Example 6.9:** @cvnvr i remembered his time table at my school so i would know which classes he would come out of so i could walk past him im a mess **[Love]**
- Tweeters have shown creativity in using figures of speech (e.g., similes, metaphors and metonyms) to convey emotions in text, the most challenging being the use of allusions and references to popular culture.
  - **Example 6.10:** Being asked to take a note to another teacher and feeling like you had just been honoured with the task of taking the ring to Mordor **[Happiness]**
  - **Example 6.11:** If a typical fast food worker put 5% the effort into their work as Spongebob, production would go up 100000%. But no, they're all Squidwards **[Anger]**
  - **Example 6.12:** Drove the bike today, about 40 miles. Felt like Jim carrey on me myself and Irene! **[Exhaustion]**

Part of our future work is to better formalize these problems so that appropriate natural language processing techniques can be used to address them. Some of these challenges such as sarcasm and irony detection (González-Ibáñez, Muresan, & Wacholder, 2011) as well as metaphor detection (Mohler, Bracewell, Hinote, & Tomlinson, 2013) have received attention in the computational linguistics community so it will be interesting to apply the techniques in the context of fine-grained emotion detection in tweets.

We view our set of 28 emotion categories as a starting point to advance a theory or taxonomy for emotion in text. Currently, the emotion categories are not tied together by any structure but our analysis suggests that some of the emotion categories share underlying semantic ties. For example, *sympathy* can be treated as a close cousin of *sadness* as *sympathy* is a form of *sadness* but is expressed towards the misfortune of others. Two particular emotion categories, *confidence* and *doubt* have exhibited properties indicating that they are polar opposites. Therefore, one potential next step is to group the emotion categories based on the notion of emotion families proposed by Ekman (1992) and examine how the categories are related to one another. If the discrete emotion categories naturally form a hierarchy, it would be possible to further explore hierarchical classification techniques for this task. We also plan to test the annotation scheme on other types of text to evaluate its utility and compare the resulting emotion taxonomies with the taxonomy based on tweets.

We will also continue our efforts to expand the emotion corpus used to train and test the computational models to increase their robustness in handling diverse emotion expressions and descriptions in text. Language resources must be updated frequently to reflect the changing nature of data for real world applications. One approach we intend to explore is the use of purposeful gaming to collect emotion annotation. This is potentially an economical approach to obtain emotion annotations. However, annotators or players have different expectations in such environments. It is crucial to make the annotation task fun while maintaining the quality of data. The annotated data obtained through purposeful gaming could be compared to the data

generated from paid annotators recruited from AMT in order to better understand the character of the two annotation sources. Another potential approach is to use a semi-automatic method to scale the size of the corpus. We can first tune classifiers for high recall and use them as a filtering mechanism to identify the most probable instances of an emotion category from an unannotated data set. Annotators can then correct the machine predictions instead of performing annotation on the entire unannotated data set.

We are also particularly interested in studying the role that emojis play in emotion expressions. The use of emoticons and emojis need further investigation as our study reveals that actual use of these pictorial symbols does not necessarily follow their prescribed use. Also, we found conventional tokenizers and part-of-speech taggers are not well-equipped yet to handle these symbols appropriately, thus there is a need to build more specialized natural language processing tools for tweet processing.

Another potential hypothesis to test relates to non-standard spellings in tweets (e.g., */ooool* or *yesssss*). Baldwin & Chai (2011) posit that non-standard word forms are not merely misspellings but can contain extra pragmatic information not found in standard word forms. Using our corpus, we can examine if there is a correlation between non-standard words and the expression of emotions in tweets. If non-standard words contain information useful for emotion classification, normalizing the non-standard words in the features could negatively affect the performance of classifiers. Findings from this follow-up study can also help us identify how to define features to capture the pragmatic information embedded within the non-standard word forms.

In this thesis, we framed the machine classification problem as “for tweet  $x$ , how accurately can we predict whether or not it contains an expression of emotion  $y$ ”. Using our current framing, we first evaluate a binary classifier for each emotion category separately and then average over all the categories. Essentially, we are measuring how accurately the classifier set performs in predicting each of the 28 emotion categories given a tweet. The average

accuracy we obtained from the 28 binary classifiers is above 90%. This does not, however, mean that the combined classifier is 90% accurate in predicting the set of emotions expressed in a tweet. In future work, we will address the related question: for tweet  $x$ , how accurately can we predict the set of emotions that are expressed in it. In the multi-label setup, we will collect the set of predicted labels from the 28 binary classifiers and compare the predicted set to the set of gold labels for each tweet. A completely correct instance means that all the predicted labels match all the gold labels for a tweet while a completely incorrect instance means that none of the predicted labels match any of the gold labels. We will also encounter partial matches where some of the predicted labels match the gold labels. We can then compute the accuracy of the combined model using exact-match-ratio, Hamming Loss (Sorower, 2010) or a formula that exercises more leniency towards partial matches.

On the machine learning end, we plan further efforts to test other types of features to improve classifier performance. First, we have only adopted the NRC emotion lexicon to implement the lexicon-based features. We are interested to incorporate other sentiment and emotion lexicons such as the ANEW lexicon (Bradley & Lang, 1999), WordNet-Affect (Strapparava & Valitutti, 2004), General Inquirer (Stone, Dunphy, Smith, & Ogilvie, 1966), AFINN (Nielsen, 2011) and more Twitter-specific sentiment lexicons<sup>25</sup> (Davies & Ghahramani, 2011; Tang, Wei, Qin, Zhou, & Liu, 2014) into our future machine learning experiments. Our experimentation with syntactic (e.g., unigrams with POS information) and semantic features (e.g., word frequencies based on the NRC emotion classes) are limited in this thesis. Other more sophisticated features worth exploring include named entities, semantic roles, position of emotion cues in the tweet, user metadata and hashtag patterns.

Our machine learning experiments also revealed that building classifiers with feature sets that are tailored to each category is a promising direction. We also have evidence that some emotion categories respond better to some classification algorithms than others. We will

---

<sup>25</sup> Sentiment Analysis Word List: <http://alex-davies-4lq6.squarespace.com/twitter-sentiment-analysis/>



continue our efforts to select “good” classifier and feature combinations for each emotion category. Part of future work includes more in-depth exploration on the role of context and culture in this problem domain. Not only are we interested to study how context and culture affect the performance of the machine learning models, we also hope to find context-specific features that can be leveraged to build more robust automatic emotion classifiers. Our ultimate vision is to build a generalizable model to detect emotion signals regardless of topic, domain and context on Twitter. This emotion model can then be specialized to meet the needs of more specific topics and domains and can be generalized for other material types.

## **6.4 Conclusion**

In this thesis, we have identified a set of 28 emotion categories representative of the range of emotions expressed in tweets and have identified a set of lexical items (i.e., words, phrases, emoticons, emojis and hashtags) that characterize each emotion category. Of the 28 emotion categories, annotators can recognize 8 with high reliability, 13 with moderate reliability and 7 with low reliability. We found only a handful of single terms in the lexical items that can serve as unique primary indicators of an emotion category. It is common for tweeters to use the same terms but with different surrounding terms or in different contexts to express distinct emotions in tweets. As a result, emotion-related lexicons that contain only single word lexical items are of limited utility in the detection of fine-grained emotions in tweets. We developed an emotion lexicon that contains not only single terms but multi-word terms or phrases that serve as salient indicators for each of the 28 emotion categories.

We then built automatic emotion detectors to perform classification on all 28 emotion categories. We tested a novel approach utilizing the emotion cues identified by annotators to select features for machine learning in this problem domain. Our experimental results demonstrate that the classifier using cue-based features slightly outperforms the conventional corpus-based and lexicon-based features in fine-grained emotion classification. Most

importantly, the performance gain by the cue-based features is achieved using far fewer features. Thus, cue-based features offer great advantages in terms of increasing the efficiency of classifiers through the use of more compact feature sets without any loss in performance. Of the 28 emotion categories, the classifier using custom cue features per category achieves high performance ( $F1 \geq 0.6$ ) for 6 categories, moderate performance ( $0.4 \leq F1 < 0.6$ ) for 10 categories and low performance ( $F1 < 0.4$ ) for 12 categories.

We also discovered that each emotion category has unique linguistic properties and achieves maximum performance using different combination of classifier and features. The overall machine classification results can be improved through the use of an ensemble model encompassing the best performing binary classifier for each category. The ensemble model encompasses a mix of SMO and BayesNet binary classifiers with 14 using corpus-based features, 13 using cue-based features and 1 using lexicon-based features. Using the ensemble classifier yields high performance for 10 categories, moderate performance for 10 categories and low performance for only 8 categories.

We have shown from our experimental results that it is feasible to extend machine learning classification to fine-grained emotion detection in tweets (i.e., as many as 28 emotion categories) with results that are comparable to state-of-the-art classifiers that detect six to eight basic emotions in text. In fact, the machine learning classifier proves to be highly effective in the detection of a specific set of emotion categories (i.e., gratitude, pride and amusement) that annotators can reliably recognize in tweets. The real strength of the machine learning classifier lies in its ability to perform well even in some of the categories that annotators find difficult to recognize. Our findings thus open up new possibilities for the development of more sensitive automatic emotion detectors that can be applied in sentiment analysis to help augment the ability of humans to better recognize emotion signals in massive amounts of text.

# Appendix A

The instrument used in the emotion word rating task (Phase 1: Task 3) is shown in Figure A.1, Figure A.2, Figure A.3 and Figure A.4.

The screenshot shows a HIT preview page with a light blue header. The header contains the following information: "Rate emotion words for valence and arousal", "Requester: Jasy Liew Suet Yan", "Reward: \$0.20 per HIT", "HITs available: 0", and "Duration: 1 Hours". Below the header, there is a section titled "Qualifications Required:" which states: "HIT Approval Rate (%) for all Requesters' HITs greater than or equal to 95, Number of HITs Approved greater than or equal to 1000, Location is one of AU, CA, NZ, GB, US, UM". The main content area is titled "HIT Preview" and contains an "Instructions" box. The instructions box has a blue header and contains the following text: "We are conducting a study about words people use to describe their emotions. You will make a valence (pleased - displeased) rating and an arousal (aroused or calm) rating for 50 emotion words. Select the link below to complete the word rating task. At the end of the task, you will receive a validation code to paste into the box below to receive credit for rating 50 emotion words. **Make sure to leave this window open as you complete the task.** When you are finished, you will return to this page to paste the code into the box." Below the instructions box, there is a "Study link:" label followed by the URL "https://syracuseuniversity.qualtrics.com/SE/?SID=SV\_0U1Iz9OQ8zyfyqF". Below the link, there is a "Provide the validation code here:" label followed by a text input field containing "e.g. 123456". At the bottom right of the form, there is a blue "Submit" button.

Figure A.1: HIT describing the emotion word rating task on AMT

The screenshot shows a survey interface with a dark grey background. At the top, there is a Syracuse University logo. Below the logo, there is a white box containing the following text: "You will be shown 50 words people use to describe their emotions. You will make a valence (pleased or displeased) rating and an arousal (aroused or calm) rating for all 50 words. Make your ratings based on your first and immediate reaction as you read each word." At the bottom right of the white box, there is a small grey button with the text ">>". Below the white box, there is a small text that says "Survey Powered By Qualtrics".

Figure A.2: Instructions on the emotion word rating task

Valence rates the emotion word on a displeased - pleased scale. Choose the point at the far left of this scale if the emotion word describes someone feeling extremely displeased. Choose the point at the far right of this scale if the emotion word describes someone feeling extremely pleased. If the word does not describe someone feeling pleased or displeased, choose the point in the middle.

Rate the degree of valence for each of the following 10 emotion words.

	Extremely Displeased	Quite Displeased	Moderately Displeased	Mildly Displeased	Neither Displeased nor Pleased	Mildly Pleased	Moderately Pleased	Quite Pleased	Extremely Pleased	Unsure
Admiration	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Amazement	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Amivalence	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Amusement	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Anger	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Annoyance	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Anticipation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Anxious	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Boredom	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Confidence	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Rate the degree of valence for each of the following 10 emotion words.

	Extremely Displeased	Quite Displeased	Moderately Displeased	Mildly Displeased	Neither Displeased nor Pleased	Mildly Pleased	Moderately Pleased	Quite Pleased	Extremely Pleased	Unsure
Confusion	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Curiosity	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Desire	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Desperation	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Disappointment	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Disgust	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Displeased	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Doubt	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Dread	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Empathy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Rate the degree of valence for each of the following 10 emotion words.

	Extremely Displeased	Quite Displeased	Moderately Displeased	Mildly Displeased	Neither Displeased nor Pleased	Mildly Pleased	Moderately Pleased	Quite Pleased	Extremely Pleased	Unsure
Enchantment	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Exhaustion	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Fascination	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Fear	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Gratitude	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Guilt	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Happiness	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hate	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Hope	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Indifference	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Rate the degree of valence for each of the following 10 emotion words.

	Extremely Displeased	Quite Displeased	Moderately Displeased	Mildly Displeased	Neither Displeased nor Pleased	Mildly Pleased	Moderately Pleased	Quite Pleased	Extremely Pleased	Unsure
Inspiration	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Jealousy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Lie	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Longing	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Love	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Lost	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Nostalgia	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Pleased	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Pride	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Regret	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Rate the degree of valence for each of the following 10 emotion words.

	Extremely Displeased	Quite Displeased	Moderately Displeased	Mildly Displeased	Neither Displeased nor Pleased	Mildly Pleased	Moderately Pleased	Quite Pleased	Extremely Pleased	Unsure
Relaxed	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Relief	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sadness	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Shame	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Shock	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Surprise	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Sympathy	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Torn	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Worry	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>
Yearning	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>	<input type="radio"/>

Survey Powered by [Qualtrics](#)

Figure A.3: Valence rating for 50 emotion words



# Appendix B

List of negators is shown below.

Negative words:

- No
- Not
- None
- No one
- Nobody
- Nothing
- Neither
- Nowhere
- Never

Negative Adverbs:

- Hardly
- Scarcely
- Barely

Negative verbs:

- Doesn't
- Doesnt
- Isn't
- Wasn't
- Wasnt
- Shouldn't
- Wouldn't
- Wouldnt
- Couldn't
- Won't
- Wont
- Can't
- Cant
- Cannot
- Don't
- Dont
- Didn't
- Didnt
- Weren't

## Appendix C

Table C.1 shows the frequency distribution of emotion categories in the AVG-USER, RANDOM, SEN-USER and TOPIC samples.

Category	AVG-USER	RANDOM	SEN-USER	TOPIC
Admiration	108	112	113	70
Amusement	309	180	10	161
Anger	398	267	137	399
Boredom	28	9	0	11
Confidence	28	32	16	34
Curiosity	36	25	4	28
Desperation	22	12	5	19
Doubt	65	32	12	49
Excitement	114	85	163	324
Exhaustion	20	10	4	15
Fascination	58	64	37	45
Fear	61	48	57	73
Gratitude	78	114	263	66
Happiness	398	302	604	483
Hate	90	54	4	44
Hope	104	102	108	208
Indifference	36	18	0	14
Inspiration	8	21	29	17
Jealousy	6	25	0	3
Longing	43	42	6	30
Love	184	248	31	217
Pride	19	42	123	29
Regret	59	43	9	42
Relaxed	23	14	11	29
Sadness	146	138	61	176
Shame	43	16	7	24
Surprise	84	57	26	99
Sympathy	8	13	44	36
Total	<b>2576</b>	<b>2125</b>	<b>1884</b>	<b>2745</b>

Table C.1: Frequency distribution of emotion categories for each sample

# Appendix D

Annotation scheme (code book) developed in Phase 1 is presented in Appendix D. This annotation scheme describes the initial set of 48 emotion categories and includes the heuristics collectively developed by the expert annotators for emotion valence, intensity and category.

## Phase 1 Code Book: Emotion Annotation

### Goal

The goal of the emotion annotation is to manually detect emotions expressed by the tweeter and other emotional phenomena described in tweets collected from Twitter. This manually-annotated data can be used to identify linguistic patterns associated with each emotion, and to train computer systems to automatically detect emotions in text.

### Coding Description

Emotion in text is defined as a subset of particularly visible and identifiable feelings that are expressed in written form. A tweet is emotional if it contains one of the aspects below:

- 1) Is the tweeter expressing an emotional reaction towards any stimuli? Emotions typically arise as reactions to situational events in an individual's environment that are appraised to be relevant to his/her needs, goals or concerns. It may be helpful to first identify the stimuli causing the tweeter to express a particular emotional reaction. *Example: I am excited to watch the game tonight.*
- 2) Is the tweeter describing the emotional experience of others? *Example: She is upset that her boyfriend did not remember their first month anniversary.* Flag these tweets as “Not Self: Other” in Notes.
- 3) Is the tweeter describing an emotional phenomenon? *Example: You should be glad to be given this opportunity.* Flag these tweets as “Not Self” in Notes.

Annotators are required to read each tweet, and provide annotations for 4 facets of emotion:

- 1) Polarity
- 2) Intensity
- 3) Emotion Tag
- 4) Emotional Cues



## Polarity

Polarity measures whether an emotion is pleasant or unpleasant. Is the tweeter expressing positive, negative, neutral or no emotion?

Code	Description
Positive	Positive emotions are evoked by positive events, objects or situations. Expressing pleasure (e.g., happy, relaxed, fascination, love).
Negative	Negative emotions are evoked by negative events, objects or situations. Expressing displeasure (e.g., anger, fear, sad).
Neutral	Emotion expressed is neither positive nor negative (e.g., surprise).
None	No emotion expressed.

## Instructions

- 1) Select only one code for each tweet.

## Heuristics

- 1) For tweets with quotations from the tweeter (self-quotations), consider it emotional if the quote contains emotional cues.

ID	Text	Polarity	Explanation
108250	"Our victory proves neither corporations nor billionaires can buy Montana." Jon Tester <a href="http://t.co/QeQQtzGs">http://t.co/QeQQtzGs</a> #MTSen	Positive	Key word is "our victory". Conveys positive emotion about the victory.
114554	@MariaCantwell: "It is a great night for Senate Democrats and a repudiation of stalemate in Washington, DC" <a href="http://t.co/ZHxbABwN">http://t.co/ZHxbABwN</a>	Positive	

- 2) Look carefully at phrases like "must-read" and "check it out". These are usually suggestions for readers to read an article or watch a video. Most often, these tweets contain the title of an article, and are followed by a URL to the article. Code call-for-actions as "None" unless the tweet contains emotional cues.

ID	Text	Polarity	Explanation
110486	I hope you'll check out my new piece up at @HuffingtonPost: "A Historic Election For #Women" <a href="http://t.co/MjnkjlHt">http://t.co/MjnkjlHt</a> #offthesidelines	Positive	The cue "I hope" indicates hope/anticipation, which is a positive emotion.
100398	READ: Bank bailout opponent Shelby praised in Neil Barofsky's new book: <a href="http://t.co/MkjcbY30">http://t.co/MkjcbY30</a> #TARP	None	
108247	Watch Jon's victory speech from yesterday morning: <a href="http://t.co/69KA5sLj">http://t.co/69KA5sLj</a> #MTSen	None	The word "victory" describes the type of speech given by the tweeter but does not describe the emotion of the tweeter.
106626	Be sure to tune in tonight at 9:30pm to watch my sitcom debut on @ParksandRecNBC! Here's a preview: <a href="http://t.co/ofkQjA8V">http://t.co/ofkQjA8V</a>	Positive	

- 3) When a tweeter describes someone opposing him/her, code as "Negative".

ID	Text	Polarity	Explanation
104265	@BarackObama's admin <b>opposed our</b> Iran sanctions. Take a look <a href="http://t.co/dAPtBwCQ">http://t.co/dAPtBwCQ</a> ; then tried to water them down again <a href="http://t.co/zG0emLzu">http://t.co/zG0emLzu</a>	Negative	Implicit expression of emotion. The tweeter is expressing displeasure/dissatisfaction because his/her proposal has been rejected.

- 4) If tweeter mentions injustice or unfairness, code as Negative.

ID	Text	Polarity	Explanation
101356	<b>California women make 85 cents for every dollar made by men.</b> We need to pass the Paycheck Fairness Act #RU4fairpay <a href="http://t.co/eCi1r0Xh">http://t.co/eCi1r0Xh</a>	Negative	Expressing unfairness

- 5) When a tweeter describes that his/her expectations are not met, code as "Negative".

ID	Text	Polarity	Explanation
104970	sent ltr to Pres. today: <b>His failure</b> to aid those in Benghazi has caused servicemembers to question bond of helping those in distress.	Negative	Expression of disappointment. Someone has failed to meet tweeter's expectations.

- 6) Code tweets that merely ask someone to do something or mention that someone should do something as "None".

ID	Text	Polarity	Explanation
109515	Urged Pres Obama and @USDA to expand crucial food aid programs for #NJ families in all 21 counties impacted by #Sandy <a href="http://t.co/hf3e7A1S">http://t.co/hf3e7A1S</a>	None	
103160	Isakson & @SenBobCorker Continue to Press State Department for Disclosure of Communications from Benghazi Attacks <a href="http://t.co/401M0kT5">http://t.co/401M0kT5</a>	None	
103161	Isakson, @SenBobCorker Call on President to "Come Clean" and Release Communications from Benghazi Attack <a href="http://t.co/SNLalOns">http://t.co/SNLalOns</a>	None	
104969	my letter to Obama urging him to set the record straight on Benghazi: <a href="http://t.co/lrTbaTJn">http://t.co/lrTbaTJn</a>	None	

- 7) Tweeter expressing concerns or worries about a topic/entity should be coded as Negative.

ID	Text	Polarity	Explanation
104972	My concerns about new USDA National School Breakfast and Lunch program rules: <a href="http://t.co/zObGChvd">http://t.co/zObGChvd</a>	Negative	Tweeter is expressing his/her concerns about an issue.

- 8) For tweets depicting someone else's emotions (not the tweeter's emotion), code the tweet but flag them in Notes as "Not Self: Other".

ID	Text	Polarity	Explanation
100998	<b>Lots of enthusiasm</b> here in Lynchburg - we already hit Virginia Beach and Charlottesville, Roanoke up next #Virginia <a href="http://t.co/p5CiploV">http://t.co/p5CiploV</a>	Positive	"Lots of enthusiasm" is an emotional cue describing the emotion of others (not necessarily the emotion of the tweeter). This is an example of a "Not Self".
105932	<b>great crowd</b> of phoners/doorknockers in Sylvania, OH today! <b>Lots of energy out there</b> for the President <a href="http://t.co/KhOM8O72">http://t.co/KhOM8O72</a>	Positive	Contains emotional cues describing the emotions of others.

- 9) Tweets merely reporting events or activities of the tweeter without any indication of how the tweeter feels about them should be coded as "None".

ID	Text	Polarity	Explanation
101715	Today I met w/ women business leaders in #COSprings to hear their thoughts on fiscal responsibility, small business & the #waldocanyonfire.	None	
103530	Senator Inouye addressing the #Hawaii Farm Bureau Federation <a href="http://t.co/bwH8bq3g">http://t.co/bwH8bq3g</a>	None	

- 10) For tweets that are opinions (tweeter's belief/judgment on a topic), we should flag them in Notes as "Opinion". It is possible for tweeter to state a positive/negative opinion about a topic with no expression of emotion (pleasure/displeasure).

ID	Text	Polarity	Explanation
106242	Good #MDSandy advice from @fema on staying safe during storm, conserving power and communicating with loved ones after outages	None	Opinion but no emotion
101550	Op-Ed: Hundreds of workers in Brighton, Windsor, & Pueblo would still have their jobs had Congress passed the Wind PTC. <a href="http://t.co/41yqkx1q">http://t.co/41yqkx1q</a>	None	Opinion but no emotion
101712	We must make smarter investments in #education, so students can have access to #PellGrants & the opportunity of higher education.	None	Opinion but no emotion
111752	Cheating by China's #solar industry sets a <b>dangerous</b> precedent. Read my @washingtonpost Letter to the Editor: <a href="http://t.co/P2x7w1qV">http://t.co/P2x7w1qV</a>	Negative	Use of the word "dangerous" indicates fear.

11) Statements should be coded as None.

ID	Text	Polarity	Explanation
101357	The California Desert Protection Act (18 years old today) preserved 7 million acres of desert, largest-ever designation in continental U.S.	None	Statement
101189	The steady leadership of President Obama was reflected in this 31st straight month of private-sector job creation. <a href="http://t.co/QzbFkcCf">http://t.co/QzbFkcCf</a>	None	Statement

12) If tweeter talks about winning, tweeter is usually expressing Positive emotion. If tweeter talks about losing (e.g., sad about losing the election), tweeter is usually expressing Negative emotion. Do check the context of the tweet. Tweeter may say something about not giving up hope after losing, and such a case will be coded as Positive.

ID	Text	Polarity	Explanation
91230278244237300	All of the "real" Democrats won their respective races & will move on 2 face the 6 recalled Republican state senators on August 9 #YesWeCan	Positive	

13) If tweeter talks about being chilled, chilling out, being restful or relaxing, tweeter is expressing Positive emotion.

ID	Text	Polarity	Explanation
100018512445186000	excuse me while i spend my last saturday of the summer making cds, watching netflix, and chillin in the woman cave #legit #crazy #woah	Positive	Expression of relaxed.
100033738754363000	A night to just lay up & watch movies on Netflix.	Positive	Expression of relaxed.

14) If tweeter expresses that he/she likes something, is an expression of passion. If tweeter says that he/she likes an object, person or an entity, then code it as Positive. If tweeter use "like" in a simile (e.g., big like elephant), then "like" is not an emotional cue.

ID	Text	Polarity	Explanation
100001160882176000	@QuietusCyn @ShatteredYuuki Yeah, It's on netflix I think now. I still like the futuristic and action concepts even today.	Positive	

15) If more than one emotion with the SAME POLARITY are expressed in a tweet, code the tweet with the appropriate polarity (i.e., Positive, Negative or Neutral). Identify all the multiple emotions in Emotion Tag. Flag in Notes as "Multiple Emotions".

ID	Text	Polarity	Explanation
103000	Thanks! RT @AMGravitt: @SaxbyChambliss loved hearing you on @Talkmaster Boortz today broadcasting on #klbj.	Positive	Tweeter is expressing gratitude to @AMGravitt. @AMGravitt is expressing happiness.

- 16) If more than one emotion with DIFFERENT POLARITY are expressed in a tweet, then code as “Neutral”. For example, if a tweeter expresses a positive emotion towards a stimulus and also a negative emotion towards a different stimulus in the same tweet, code as “Neutral”. Identify all the multiple emotions in Emotion Tag. Flag in Notes as “Multiple Polarity” and also all the polarity codes identified in the tweet.

ID	Text	Polarity	Explanation
100004037872713000	@theN5er lol I been in Oz 6months now on my 4th & final stint. Love it but do miss London, prob coz I can't get 2d Emirates no more. #afc	Neutral	Tweeter is expressing both happiness (emotional cues: “lol”, “Love it”) and sadness (emotional cues: “miss London”). You will put in Notes “Multiple Polarity: Positive, Negative”.
100001779579748000	Niagara Falls today. Fairly spectacular but too touristy for me. Prefer Vic Falls, thus one of the 7 wonders. Detroit tomorrow, hopefully.	Neutral	Tweeter expresses disappointment in the second and third sentences. Fourth sentence contains expression of hope. You will put in Notes “Multiple Polarity: Positive, Negative”.

- 17) Expressing support/stance towards a topic/entity (e.g., I support Obama) is considered to be non-emotional. Code tweet as “None” only if tweeter is merely expressing his/her support/stance towards a topic/entity in a tweet without any emotion.

ID	Text	Polarity	Explanation
111364	@MittRomney provides the right tone and leadership in response to the attacks on our embassies. Obama's appeasement is wrong for U.S.	None	Tweeter is expressing support for @MittRomney and stating that he is not supporting Obama.
104266	Appreciate @Schneider4IL10's past support but I didn't work with you;I've worked with @RobertDold closely and he has my strong support #IL10	Positive	Emotion is positive because tweeter is expressing “gratitude”. The part where the tweeter expresses that he/she supports @RobertDold instead of @Schneider4IL is not emotional.

- 18) Tweets to promote/discourage votes for a political candidate are considered to be non-emotional. These tweets usually appear in the form of “a tweet for person A is a vote for issue B”.

ID	Text	Polarity	Explanation
112290	A vote for "Bailout Joe" Donnelly is a vote for Harry Reid, more bailouts & more reckless debt.	None	
112291	A vote for @RichardMourdock is a vote to repeal Obamacare, to stop job destroying tax hikes, and fiscal sanity.	None	

19) If the tweet appears in the format of the tweeter's response towards another tweeter (i.e., tweeter's response is usually in front of or at the back of the RT), code for both the emotion of the tweeter as well as the emotion expressed by the quoted tweeter. Focus of the annotation task should be on identifying linguistic units in the tweet that serve as emotional cues. Follow the heuristics below to code these tweets:

- If both tweeter and quoted tweeter are not expressing emotion, code as "None" regardless if the tweeter agrees or disagrees with quoted tweeter.
- If tweeter expresses an emotional response towards a non-emotional retweet, code based on the emotion of the tweeter.
- If tweeter is not expressing any emotion but quoted tweeter expresses emotion, code tweet based on emotion expressed by quoted tweeter. You have to first make sure that @anotheruser or retweet mentioned in the message expresses emotion(s), and not merely a fact, statement or opinion. If quoted tweeter expresses emotion, flag as "Not Self: Other" in Notes.
- If both tweeter and quoted tweeter express different emotions, identify both the emotions expressed by the tweeter and the quoted tweeter.

ID	Text	Polarity	Explanation
102998	Agree- trying for open hearings @marti6619:The <b>American people need answers</b> to the <b>Benghazi attack, they cannot have died in vain!</b>	Negative	Emotion expressed by @marti6619 in the second sentence "@marti6619:The American people need answers to the Benghazi attack, they cannot have died in vain!". Tweeter did not express any emotion but agrees with the quoted tweeter.
107417	Agree! "@danschoen54a: A <b>big thanks</b> 2 volunteers out in crummy weather today in So. St. Paul @amyklobuchar @Obermueller2012 @katiesieben"	Positive	Emotion expressed by @danschoen54a. Tweeter did not express any emotion but agrees with the quoted tweeter.
100019165590597000	Yep and that might b my move in Dallas tonight RT @AntwannetteBond: True! Tht would b <b>fun..</b> @Breedlove_08	Positive	Emotion expressed by @AntwannetteBond. Tweeter did not express any emotion but agrees with the quoted tweeter.
100032137759166000	Netflix?? RT @Pink_Dagger: I'm watching the ORIGINAL X-Men cartoons I used to watch on Saturday Mornings....#throwback <b>#iamhappyagain</b>	Positive	Emotion is expressed by quoted tweeter, @Pink_Dagger. Tweeter did not express any emotions.

## Intensity

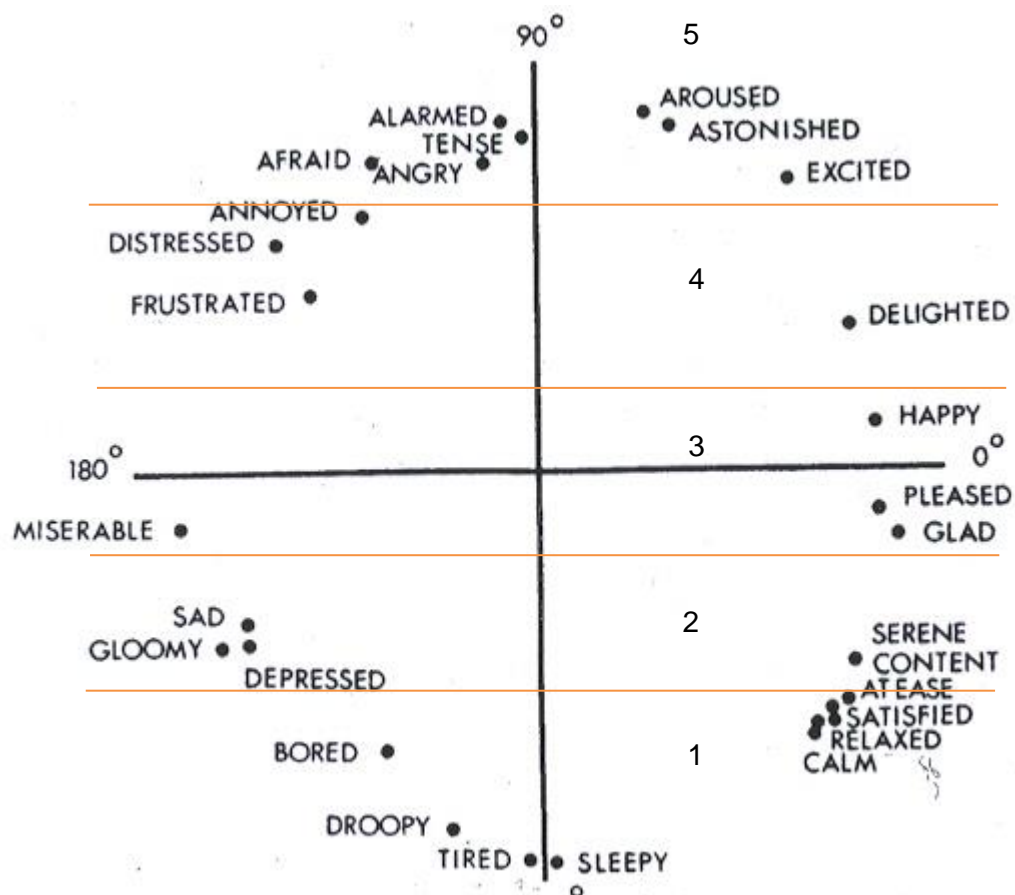
Intensity measures the degree of emotion arousal, which can range from being calm to excited. How intense is the tweeter's emotional reaction to the stimuli? Intensity is linked to the tweeter's level of arousal/activation.

Code	Description
1	Very low intensity
2	Low intensity
3	Moderate intensity
4	High intensity
5	Very high intensity

## Instructions

- 1) Select only one code for each tweet.

## Guidelines<sup>26</sup>



<sup>26</sup> Adapted from Russell's circumplex model of affect (Russell, 1980).

## **Emotion Tag**

### **Stage 1: Discover emotion categories**

This is an open coding exercise to identify discrete categories of emotions. Based on your knowledge about emotion, what is the emotion expressed by the tweeter?

#### **Instructions**

- 1) Code each tweet with the best emotion tag that describes the main emotion being expressed.
- 2) If you identify multiple emotions being expressed in the tweet, you can code the tweet with more than one emotion tag. If you are coding the tweet with multiple emotion tags, put the main emotion tag as the first in your list of emotion tags.

### **Stage 2: Test emotion categories**

You are given a set of emotion categories identified from the card sorting activity to group semantically-related emotion tags together into categories. Annotate emotion tag of a tweet by selecting an emotion category from the given set below.

#### **Instructions**

- 1) Code each tweet by selecting the best emotion category describing the main emotion being expressed.
- 2) If you identify multiple emotions being expressed in the tweet, you can code the tweet with more than one emotion category. If you are coding the tweet with multiple emotion categories, put the main emotion category as the first in your list of emotion categories.
- 3) Suggest a new emotion tag only if none of the emotion categories in the given set fit the emotion(s) described in the tweet.



### Emotion Categories: List

Fascination	Gratitude	Admiration	Love	Like	Pride
Fascination Interest Interested	Gratitude Grateful Thank Thankfulness Thankful Appreciate Appreciative Blessed	Impressed Veneration Admiration Appreciation Honorable Honored Admire Respect Respectful	Love Loved Loving Obsessed Affection Bonding	Like Liking Fond Affinity	Pride Proud Proudness Accomplished Praiseful

Pleased	Happiness	Excitement	Amusement	Relaxed	Relief
Pleased Pleasure Glad Satisfied Satisfaction Content Contented	Happiness Happy Cheerful Cheering Joy Joyful Delight Delighted Elated Blessed Enjoyment Beatific Congratulation Congratulations Congrats Celebrate Celebratory Gratification	Excitement Excited Exciting Enthusiastic Energetic Enthused Aroused	Amused Amusement Fun Funny Humorous Humored Humor Teasing	Relaxed Relax Comfortable Calm Serene At ease	Relief

Hope	Anticipation	Confidence	Inspiration
Hope Hopeful Optimistic Optimism	Anticipation Anticipated Expected Expect Expecting Eager Keen	Confidence Confident Loyalty	Inspiration Inspired Inspiring Moved Motivation Motivated Encouragement Encouraged Supportive Touched

Sadness	Yearning	Annoyance	Anger	Fear	Worry
Sad Saddened Sadness Sorrow Distress Distressed Depressed Depression Grief Dejected Miserable Pain Gloomy Cry Hurt Somber Lonely Loneliness Sick Resigned Commemoration	Yearning Longing Miss Missing	Frustration Frustrated Annoy Annoyed Annoying Annoyance Irritated Irritation Aggravated Miffed Upset Unhappy Offended	Anger Angry Blame Blamed Outraged Aggressive Pissed Indignation Resentment	Fear Scary Frightened Afraid Scare Scared Crazy Craze Alarm Caution Danger	Worried Anxious Anxiety Worry Concerned Concern Concerns Urgent Tense Mad Restless Agitated Stress Worrisome

Guilt	Doubt	Dread	Desperation	Boredom	Jealousy
Guilt Guilty	Doubtful Doubt Pessimism Negative	Dread	Desperation Desperate Hopeless	Boring Bored	Jealousy Jealous

Shame	Regret	Sympathy	Disgust	Hate	Awkward
Shame Shameful Embarrassed Embarrassment Shy	Regret Regretful Sorry Remorse Remorseful	Sympathy Sorry Sympathetic Pity	Disgust Disgusted Degradation	Dislike Hate Hatred Aversion Disdain Averse Self-loathing Revenge Vengeance Contempt Condescension Scorn	Awkward Uncomfortable Weird Strange

Disappointment	Displeased
Disappointment Disappointing Disappointed	Displeased Dissatisfaction Dissatisfied Unsatisfied Disapproval Discontent

Surprise	Shock	Amazement	Empathy	Curiosity	Confusion
Surprise Astonish Astonished Surprised Unexpected Unbelievable Disbelief	Shock Shocked Dismayed	Amazed Amazing Amazement Awed	Empathetic Compassion Compassionate	Curiosity Curious	Confusion Confused Confuse

Exhaustion	Indifference	Nostalgia	Ambivalence	Desire	Lust
Exhausted Tired	Indifference Indifferent	Nostalgia Nostalgic Reminiscent Reminiscing	Torn Conflicted	Desire Ambitious Wish Wishes Wishful Wishing	Lust

## Emotion Categories: Description

### Positive Emotions

Emotion Category	Related Emotion Tags	Description
Pleased	Pleased Pleasure Glad Satisfied Satisfaction Content Contented	Feeling pleased or satisfied about a desirable event
Happiness	Happiness Happy Cheerful Cheering Joy Joyful Delight Delighted Elated Blessed Enjoyment Beatific Congratulation Congratulations Congrats Celebrate	Feeling happy, joyful or delighted about a desirable event
Excitement	Excitement Excited Exciting Enthusiastic Energetic Enthused Aroused	Feeling of great enthusiasm and eagerness
Amusement	Amused Amusement Fun Funny Humorous Humored	State or experience of finding something funny
Pride	Pride Proud Proudness Accomplished Praiseful	Feeling of deep pleasure or satisfaction derived from one's own achievements, the achievements of those with whom one is closely associated, or from qualities or possessions that are widely admired

Emotion Category	Related Emotion Tags	Description
Inspiration	Inspiration Inspired Inspiring Moved Motivation Encouragement Encouraged	Feeling that makes someone want to do something or that gives someone an idea about what to do or create
Gratitude	Gratitude Grateful Thank Thankfulness Thankful Appreciate Appreciative Blessed	State of being thankful or readiness to show appreciation for and to return kindness
Confidence	Confidence Confident	Feeling of self-assurance arising from one's appreciation/trust of one's own abilities or qualities
Hope	Hope Hopeful Optimistic Optimism	Feeling of expectation and desire for a certain event to happen
Fascination	Fascination Interest Interested	State of being fascinated or interested in something
Anticipation	Anticipation Anticipated Expected Expect Expecting Eager Keen	Emotion involving pleasure in considering some expected or longed-for good event
Love	Love Loved Loving Obsessed Affection Bonding	Feeling of affection towards another person
Like	Like Liking Fond	Natural liking for an object or event
Admiration	Impressed Veneration Admiration Appreciation Honorable Honored Admire Respect Respectful	Feeling of respect towards another person or state of being impressed

Emotion Category	Related Emotion Tags	Description
Relaxed	Relaxed Relax Comfortable Calm Serene At ease	Feeling calm or at ease
Relief	Relief	Feeling of reassurance and relaxation following release from anxiety or distress

### **Negative Emotions**

Emotion Category	Related Emotion Tags	Description
Boredom	Boring Bored	State experienced when an individual is left without anything in particular to do, and not interested in their surroundings
Awkward	Awkward Uncomfortable Weird Strange	Feeling uncomfortable or strange in a situation
Sadness	Sad Saddened Sadness Sorrow Distress Distressed Depressed Grief Dejected Miserable Pain Gloomy Cry Hurt Somber	Feeling of loss, helplessness or sorrow for own misfortune
Sympathy	Sympathy Sorry Sympathetic Pity	Feeling of pity and sorrow for someone else's misfortune
Guilt	Guilt Guilty	Feeling of having done wrong or failed in an obligation
Regret	Regret Regretful Sorry Remorse Remorseful	Feeling of remorse or repentance over something that has happened or been done
Doubt	Doubtful Doubt Pessimism	Feeling doubtful or having the tendency to see the worst aspect of things or believe that the worst will happen, a lack of hope or confidence in the future

Emotion Category	Related Emotion Tags	Description
Yearning	Yearning Longing Miss Missing	Feeling of longing for someone or for something that one cannot have or cannot get easily
Dread	Dread	Anticipate with great apprehension or fear
Desperation	Desperation Desperate	State of despair, typically one that results in rash or extreme behavior
Jealousy	Jealousy Jealous	Feeling or showing envy of someone or their achievements and advantages
Shame	Shame Shameful Embarrassed Embarrassment	Painful feeling of humiliation or distress caused by the consciousness of wrong or foolish behavior
Disappointment	Disappointment Disappointing Disappointed	Feeling of sadness or displeasure caused by the nonfulfillment of one's hopes or expectations
Displeased	Displeased Dissatisfaction Dissatisfied Unsatisfied Disapproval Discontent	Displeased or dissatisfied about an undesirable event
Annoyance	Frustration Frustrated Annoy Annoyed Annoying Annoyance Irritated Irritation Aggravated Miffed Upset Unhappy Offended	State of being annoyed, upset or frustrated
Anger	Anger Angry Blame Blamed Outraged Aggressive Pissed	Strong feeling of annoyance, displeasure or hostility

Emotion Category	Related Emotion Tags	Description
Worry	Worried Anxious Anxiety Worry Concerned Concern Concerns Urgent Tense Mad Restless	State of anxiety and uncertainty over actual or potential problems
Fear	Fear Scary Frightened Afraid Scare Scared Crazy Craze Alarm Caution Danger	Unpleasant emotion caused by the belief that someone or something is dangerous, likely to cause pain, or a threat
Hate	Dislike Hate Hatred Aversion Disdain Averse Self-loathing Revenge Vengeance	Intense or passionate dislike for someone, something or some event
Disgust	Disgust Disgusted Degradation	Feeling of revulsion or profound disapproval aroused by something unpleasant or offensive



### **Neutral Emotions**

<b>Emotion Category</b>	<b>Related Emotion Tags</b>	<b>Description</b>
Curiosity	Curiosity Curious	Strong desire to know or learn something
Confusion	Confusion Confused Confuse	State of being bewildered or unclear in one's mind about something
Exhaustion	Exhausted Tired	State of physical or mental fatigue or being tired
Indifference	Indifference Indifferent	Lack of interest, concern, or sympathy
Surprise	Surprise Astonish Astonished Surprised Unexpected	Unexpected or astonishing event, fact, or thing
Shock	Shock Shocked	Sudden upsetting and surprising event or experience (negatively surprised)
Amazement	Amazed Amazing Amazement Awed	Feeling of great wonder/awe and surprise (positively surprised)
Desire	Desire Ambitious Wish Wishes Wishful Wishing	Strong feeling of wanting to have something or wishing to have something
Lust	Lust	Strong sexual desire for someone
Empathy	Empathetic Compassion Compassionate	Expressing the ability to understand and share the feelings of another
Nostalgia	Nostalgia Nostalgic	Sentimental longing or wistful affection for the past, typically for a period or place with personal associations
Ambivalence	Torn Conflicted	State of having mixed feelings or contradictory ideas about something or someone

## Emotion Cues

What is/are the phrase(s) that influenced you to annotate the tweet with a particular emotion tag?

### Instructions

- 1) Copy and paste the portion of text you have identified as emotion cues into the Emotion Cues column in the coding sheet.
- 2) There is no restriction to the length of each emotion cue. Emotion cues can be symbols, characters, words or phrases.
- 3) There can be multiple emotion cues associated with an emotion in a tweet. Identify all the cues you deem relevant to each emotion in a tweet. You can separate each cue for the same tweet with a comma (,).
- 4) If you have annotated a tweet with multiple emotion tags, make sure you specify which emotion tag the emotion cues are associated with.
  - Example:       emo\_tag\_1: emo\_cue\_1, emo\_cue\_2  
                  emo\_tag\_2: emo\_cue\_3

# Appendix E

Instructions and annotation scheme given to AMT workers in Phase 2 are shown in Appendix E.

This annotation scheme contains emotion specification for 28 emotion categories.

## Instructions

In this task, you will annotate a batch of 30 tweets with various aspects of emotion (i.e., polarity, arousal, tag, cues, and source). First, read the task and code descriptions for polarity, arousal, emotion tag, emotion cues, emotion source, and multiple emotions in the codebook below. Click the “start annotation” button at the bottom of the page to begin your annotation task. Read each tweet and provide your annotations. You will be compensated \$ 0.50 upon completion of the task.

## Codebook

### Q1: Polarity

Polarity measures whether an emotion is pleasant or unpleasant. Is the tweeter expressing positive, negative, neutral or no emotion?

Code	Description	Examples
Positive	Positive emotions are evoked by positive events, objects or situations. Emotions of pleasure (e.g., happiness, relaxed, fascination, love).	<ul style="list-style-type: none"><li>• <i>Getting one of these bad boys in your cereal box and feeling like your day simply couldn't get any better</i> <a href="http://t.co/Fae9EjyN61">http://t.co/Fae9EjyN61</a></li><li>• <i>Thank you to all of our local first responders, police, and volunteers who have helped out during #Sandy</i></li></ul>
Negative	Negative emotions are evoked by negative events, objects or situations. Emotions of displeasure (e.g., anger, fear, sadness).	<ul style="list-style-type: none"><li>• <i>Shocked, saddened by deaths of ambassador, staffers. Praying for families of fallen and those still on front lines.</i> <a href="http://t.co/rVYFEZvJ">http://t.co/rVYFEZvJ</a></li><li>• <i>but i cant play tales of vesperia becasue they decided not to localise the ps3 version and chose xbox instead &gt;:(</i></li></ul>
Neutral	Emotion expressed is neither positive nor negative (e.g., surprise).	<ul style="list-style-type: none"><li>• <i>was surprised to run into Mitt this afternoon in Sylvania, OH</i> <a href="http://t.co/l9Pked2J">http://t.co/l9Pked2J</a></li><li>• <i>@ZeddRebel secular era in the ME is over. pretending otherwise is wishful thinking. as long as oil remains under \$150, I really don't care.</i></li></ul>
No Emotion	No emotion is expressed.	<ul style="list-style-type: none"><li>• <i>CT residents who suffered damage in disaster declared counties should register online at</i> <a href="http://t.co/lwZU8vCf">http://t.co/lwZU8vCf</a></li><li>• <i>READ: Shelby in Press-Register op-ed: We want #RESTORE Act money to go to Alabama, not the federal government</i> <a href="http://t.co/gl2uXqUz">http://t.co/gl2uXqUz</a></li></ul>

## Q2: Arousal

Arousal measures the degree an emotion causes someone to be in a state of activation or arousal, which can range from being calm to excited.

Code	Description	Examples
1	Very low arousal	Bored, Tired, Sleepy, Calm, Relaxed
2	Low arousal	Sad, Gloomy, Depressed, Serene, Content, Satisfied
3	Moderate arousal	Miserable, Happy, Pleased, Glad
4	High arousal	Annoyed, Distressed, Frustrated, Delighted
5	Very high arousal	Alarmed, Tense, Afraid, Angry, Aroused, Astonished, Excited

## Q3: Emotion Tag

Emotion tag represents an emotion category with a set of distinctive features. You can annotate a tweet with more than one emotion tag if the tweet contains multiple emotions. For tweets with multiple emotions, identify the primary emotion tag first, and then followed by the others.

Code	Description	Examples
Admiration	Someone or something regarded as impressive or worthy of respect. Honoring or looking up to someone.	<ul style="list-style-type: none"><li><i>I have respect for the boys who will stand up for their girls.</i></li><li><i>Women of PG County are educated, empowered, engaged &amp; are working 3 shifts - at job, at home &amp; in the community. You are my champions</i></li></ul>
Amusement	State of finding something funny or entertaining.	<ul style="list-style-type: none"><li><i>Hilarious intro! well written: "How NOT To Go Up In Flames During A Social Media Crisis" via @AndyVale <a href="http://t.co/uy2gerrD9A">http://t.co/uy2gerrD9A</a></i></li><li><i>@whatisrightt hahahaaha omfg yes</i></li></ul>
Anger	Feeling of disappointment, displeasure, dissatisfaction, annoyance, frustration, hostility or rage caused by the non-fulfillment of one's hopes/expectations or about an undesirable event.	<ul style="list-style-type: none"><li><i>Give me one reason why i shouldnt run up in mcdonalds wit a ak47 &amp; get a killstreak for this disrespect @mcdonalds <a href="http://t.co/N5TDK00ItV">http://t.co/N5TDK00ItV</a></i></li><li><i>I intro'd a bill that would create longterm, sustainable jobs for vets &amp; not add to deficit. But Majority wont allow vote. Why? Politics.</i></li></ul>

Code	Description	Examples
Boredom	Feeling dull, uninterested or left without anything in particular to do.	<ul style="list-style-type: none"> <li>• <i>@FilozofA ok bored of playing with little slavonic barbarian, #scumblock</i></li> <li>• <i>@danielclifford6 @LovelyLee_G I don't read any marvel now, or DC for that matter. Tired of corporate superheroes, my money goes to creators.</i></li> </ul>
Confidence	Feeling of self-assurance arising from one's appreciation of one's own abilities or qualities. Feeling one can trust or rely on someone or something.	<ul style="list-style-type: none"> <li>• <i>I'm confident it won't be long now before human footprints follow in the path of the Mars rover #Curiosity.</i></li> <li>• <i>Headng to a Dem Rally dinner in. Hagerstown for Obama Cardin n Delaney. Forward Together. We Can Do It !!!!!</i></li> </ul>
Curiosity	Strong desire to know or learn something.	<ul style="list-style-type: none"> <li>• <i>I wonder if @cartoonnetwork would be interested in Fudge Lord <a href="http://t.co/tY5jA5UsAB">http://t.co/tY5jA5UsAB</a></i></li> <li>• <i>I curious to know how many of you will be tweeting #ChurchFlow on Sunday after the NFL season starts</i></li> </ul>
Desperation	Feeling complete loss of hope or despair, typically one that results in rash or extreme behavior. Suffering or driven by great need or distress.	<ul style="list-style-type: none"> <li>• <i>@marzy08 You're good with computers - how do i get microsoft word back on my computer!? may have deleted :( pleaseee i need your help!</i></li> <li>• <i>"#callmatterface I want to talk about how desperate Luca Modric is to leave Spurs he ransacked Tottenham last night."</i></li> </ul>
Doubt	State of being bewildered, confused, uncertain or unclear about something. Having mixed feelings about someone or something. Feeling of distrust, suspicion or one cannot rely on someone or something.	<ul style="list-style-type: none"> <li>• <i>i dont even know why im watching it tbh</i></li> <li>• <i>@BaneXelphir @jeffcannata @Humin's the app I mentioned I'm using- I just didn't call it out by name because I'm still deciding how I like it</i></li> </ul>
Excitement	Feeling great enthusiasm and anticipation in considering some expected or longed-for good event.	<ul style="list-style-type: none"> <li>• <i>Go Giants! @SenFeinstein and I bet Sen. @Stabenow &amp; @SenCarlLevin that we beat the Tigers. The stakes: <a href="http://t.co/dyw1beaj">http://t.co/dyw1beaj</a> #WorldSeries</i></li> <li>• <i>I can't wait until I'm at the point of my life where I get to see something like this everyday in my home. <a href="http://t.co/OlvRgQMyTp">http://t.co/OlvRgQMyTp</a></i></li> </ul>
Exhaustion	State of physical or mental fatigue or feeling tired.	<ul style="list-style-type: none"> <li>• <i>I'm already exhausted from everything I'm going to avoid getting done today.</i></li> <li>• <i>On my way to the airport with 50 min sleep all night...eyes burning like I'm on True Blood..WTF</i></li> </ul>

Code	Description	Examples
Fascination	State of being fascinated, amazed or interested in something. Feeling of great wonder or awe.	<ul style="list-style-type: none"> <li>• <i>Amazing middle students @SalkSchool in ElkRiver... learned about their #STEM program &amp; saw projects firsthand. We should double stem schools!</i></li> <li>• <i>@artofbaz Sounds intriguing, what's it for?</i></li> </ul>
Fear	Feeling caused by the belief that someone or something is dangerous, likely to cause pain, or a threat. Feeling dread or anticipate with great apprehension or fear. Feeling anxious or worried over actual or potential problems.	<ul style="list-style-type: none"> <li>• <i>PAX panic is starting to hit me. I can't find my backpack. : </i></li> <li>• <i>My concerns about new USDA National School Breakfast and Lunch program rules: <a href="http://t.co/zObGChvd">http://t.co/zObGChvd</a></i></li> </ul>
Gratitude	State of being thankful or readiness to show appreciation for and to return kindness.	<ul style="list-style-type: none"> <li>• <i>Thank you to everyone who came out to the Worcester rally to show their support. #peopleoverparty #masen <a href="http://t.co/oFZkdbOE">http://t.co/oFZkdbOE</a></i></li> <li>• <i>U guys tweet out my video every week, the least I can do is follow a butt load of ya. So ima do daaat :) #JcsNewVideo <a href="http://t.co/3JjQxcdqb3">http://t.co/3JjQxcdqb3</a></i></li> </ul>
Happiness	Feeling pleased, satisfied, happy or delighted about a desirable event.	<ul style="list-style-type: none"> <li>• <i>Fun to sit w/Megan Rybak &amp; son while @MayorRTrybak gave his gr8 speech at the Dem. National Convention</i></li> <li>• <i>Excellent news for Marylanders! MT @baltimoresun: BGE: Power has been restored to all customers affected by #MDSandy.</i></li> </ul>
Hate	Feeling of dislike, distaste or aversion towards a person, event or thing. Feeling of disgust or profound disapproval aroused by something unpleasant or offensive.	<ul style="list-style-type: none"> <li>• <i>and i really didnt like the character designs for cheria/sophie</i></li> <li>• <i>@noahmittman @Gibbomadness I hate everything about it</i></li> </ul>
Hope	Feeling of expectation and desire for a certain event to happen or grounds for believing something good will happen.	<ul style="list-style-type: none"> <li>• <i>Hoping the promised new direction for Doctor Who 'materialises' tonight, no more running, shouting, magic wand waving please.</i></li> <li>• <i>@peterfacinelli Good luck tonight @ The Teen Choice Awards :) i know Twilight is gonna take them all!! Ill b watching Xoxo</i></li> </ul>
Indifference	Lack of interest, concern, or sympathy.	<ul style="list-style-type: none"> <li>• <i>Trying to get the boys to watch last nights Doctor Who, but apart from the T Rex they're not that bothered. Not for them I guess!</i></li> <li>• <i>I know retweeting praise for yourself is frowned on, but what the hell... Just this once I don't care!</i></li> </ul>

Code	Description	Examples
Inspiration	Feeling that makes someone want to do something or that gives someone an idea about what to do or create.	<ul style="list-style-type: none"> <li>• <i>Inspirational kindness from #CT's @NestleUSA donating half a million bottles of water for #Sandy victims (that's 12 truckloads!)</i></li> <li>• <i>Women power! As a dad to 3 girls, inspired to see over 2000 strong leaders at the Women's Success Forum #wvfwf</i></li> </ul>
Jealousy	Feeling or showing envy of someone or their achievements and advantages. Feeling or showing suspicion of someone's unfaithfulness in a relationship.	<ul style="list-style-type: none"> <li>• <i>I would kill to have the body of the inzano twins ok</i></li> <li>• <i>DO YOU EVER JUST GET JEALOUS SO EASILY LIKE NO THAT PERSON IS MINE DON'T BREATHE AROUND THEM PLEASE AND THANK YOU</i></li> </ul>
Longing	Yearning for or missing someone or something that one cannot have or cannot get easily. Feeling nostalgic, sentimental longing or wistful affection for the past, typically for a period or place with personal associations.	<ul style="list-style-type: none"> <li>• <i>Looking at a photograph and wishing you could re-live that moment over and over again.</i></li> <li>• <i>Miss that Louis <a href="https://t.co/rVyQ2Lby1">https://t.co/rVyQ2Lby1</a></i></li> </ul>
Love	Feeling of affection or natural liking towards another person, event or thing.	<ul style="list-style-type: none"> <li>• <i>I am in love 😊 <a href="http://t.co/vuSeaVUB2b">http://t.co/vuSeaVUB2b</a></i></li> <li>• <i>Doing my speech on the Olympics because I'm obsessed #Olympics #Sochi2014 #USAUSAUSA us🇺🇸</i></li> </ul>
Pride	Deep pleasure derived from one's own achievements, the achievements of those with whom one is closely associated, or from qualities or possessions that are widely admired.	<ul style="list-style-type: none"> <li>• <i>Proud to get a 100% score on hunger and nutrition issues from @FPAction: <a href="http://t.co/UZB4I3Jy">http://t.co/UZB4I3Jy</a></i></li> <li>• <i>Honored to visit w/some of our evacuated VA hospital patients &amp; staff at #FortHamilton in #Brooklyn today. #Sandy <a href="http://t.co/da7Po7Y3">http://t.co/da7Po7Y3</a></i></li> </ul>
Regret	Feeling remorse or repentance over something that has happened or has been done. Feeling guilty of having done wrong or failed in an obligation.	<ul style="list-style-type: none"> <li>• <i>im so sorry i will never say that again omfg</i></li> <li>• <i>@AlfieGallagher Sometimes I wish I did less to achieve more.</i></li> </ul>
Relaxed	Feeling calm, at ease. Relief following release from anxiety or distress.	<ul style="list-style-type: none"> <li>• <i>A night to just lay up &amp; watch movies on Netflix.</i></li> <li>• <i>Lucky i didnt go to the spurs game last night</i></li> </ul>
Sadness	Feeling of loss, helplessness or sorrow for own misfortune.	<ul style="list-style-type: none"> <li>• <i>It's sad that ppl try 2 hurt ppl they love because they refuse to be truthful. So they'd rather hurt you with "kill shots" and nasty fights</i></li> <li>• <i>whAT THE IMCRYIBG SO HARD <a href="http://t.co/8bqRF6iLMk">http://t.co/8bqRF6iLMk</a></i></li> </ul>

Code	Description	Examples
Shame	Humiliation or embarrassment caused by the consciousness of wrong or foolish behavior. Feeling uncomfortable or awkward in a situation.	<ul style="list-style-type: none"> <li>Yesterday is the day the atomic bomb in Japan. I embarrassed as a Japanese, was a moment of silence for the first time.</li> <li>That awkward moment when your stalking someones instagram and you like a pic from 8238.3 weeks ago</li> </ul>
Surprise	Unexpected or astonishing event, fact or thing. Sudden shocking event or experience.	<ul style="list-style-type: none"> <li>I CANT BELIEVE AT THE END BECAUSE LUDGER WAS IN FULL CHROMATUS WE COULDNT SEE HIS FACE</li> <li>@whatisrightt a week :O actually im surprised by how quick it went</li> </ul>
Sympathy	Feeling of pity and sorrow for someone else's misfortune. Feeling empathy and expressing the ability to understand and share the feelings of another.	<ul style="list-style-type: none"> <li>My thoughts and prayers are with all those affected by hurricane #Sandy.</li> <li>From Minnesota to Massachusetts: our hearts go out to the victims of today's tragedy. #prayforboston</li> </ul>

#### Q4: Emotion Cues

Identify the linguistic cues that influenced you to annotate the tweet with a particular emotion tag.

- Highlight the portions of text in the tweet you have identified as emotion cues, and the highlighted text will appear in the Emotion Cues text box.
- There is no restriction to the length of each emotion cue. Emotion cue can be symbols, characters, words or phrases.
- There can be multiple emotion cues associated with a single emotion tag. Identify all the cues you deem relevant to each emotion tag in a tweet.

#### Q5: Emotion Source

Identify whose emotion is being expressed or described.

Code	Description
Tweeter	Tweeter is expressing his or her own emotion.
Other Person	Tweeter is describing another person's emotion.
No One	Description of an emotion-related phenomenon (e.g., general description of a particular emotion, trying to make someone feel certain emotion or talking about how someone should feel in a particular situation).

#### Q6: Multiple Emotions

Indicate if a tweet contains another emotion other than the one you have specified. If you select "Yes", you will be asked to specify another emotion for the same tweet.



# References

- Abbasi, A., Chen, H., Thoms, S., & Fu, T. (2008). Affect analysis of web forums and blogs using correlation ensembles. *IEEE Transactions on Knowledge and Data Engineering*, 20(9), 1168–1180.
- Agarwal, A., Xie, B., Vovsha, I., Rambow, O., & Passonneau, R. (2011). Sentiment analysis of Twitter data. In *Proceedings of the Workshop on Languages in Social Media* (pp. 30–38). Stroudsburg, PA, USA.
- Ahmad, S. N., & Laroche, M. (2015). How do expressed emotions affect the helpfulness of a product review? Evidence from reviews using latent semantic analysis. *International Journal of Electronic Commerce*, 20(1), 76–111.
- Alm, C. O., Roth, D., & Sproat, R. (2005). Emotions from text: Machine learning for text-based emotion prediction. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing* (pp. 579–586). Stroudsburg, PA, USA.
- Aman, S., & Szpakowicz, S. (2007). Identifying expressions of emotion in text. In *Text, Speech and Dialogue* (pp. 196–205).
- Aman, S., & Szpakowicz, S. (2008). Using Roget's thesaurus for fine-grained emotion recognition. In *Proceedings of the Third International Joint Conference on Natural Language Processing* (pp. 296–302).
- Antoine, J.-Y., Villaneau, J., & Lefeuve, A. (2014). Weighted Krippendorff's alpha is a more reliable metrics for multi- coders ordinal annotations: Experimental studies on emotion, opinion and coreference annotation. In *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2014)* (pp. 550–559). Gotenborg, Sweden.
- Arnold, M. B. (1960). *Emotion and personality*. New York, NY, US: Columbia University Press.
- Averill, J. R. (1980a). A constructivist view of emotion. In *Emotion: Theory, research, and experience* (Vol. 1, pp. 305–339). New York: Academic Press.
- Averill, J. R. (1980c). The emotions. In *Personality: Basic Aspects and Current Research* (pp. 134–199). Engelwood Cliffs, NJ: Prentice-Hall.
- Balahur, A., & Hermida, J. M. (2012). Affect detection from social contexts using commonsense knowledge representations. In *International Conference on Social Computing (SocialCom)* (pp. 884 –892).
- Balahur, A., Hermida, J. M., & Montoyo, A. (2012a). Building and exploiting EmotiNet, a knowledge base for emotion detection based on the appraisal theory model. *IEEE Transactions on Affective Computing*, 3(1), 88 –101.
- Balahur, A., Hermida, J. M., & Montoyo, A. (2012b). Detecting implicit expressions of emotion in text: A comparative analysis. *Decision Support Systems*, 53(4), 742–753.

- Balahur, A., Hermida, J. M., Montoyo, A., & Muñoz, R. (2011). EmotiNet: A knowledge base for emotion detection in text built on the appraisal theories. In *Natural Language Processing and Information Systems* (pp. 27–39).
- Baldwin, T., & Chai, J. Y. (2011). Beyond normalization: Pragmatics of word form in text messages. In *Proceedings of the 5th International Joint Conference on Natural Language Processing* (pp. 1437–1441). Chiang Mai, Thailand.
- Beck, H., & Kumar, B. (1998). Clustering lexical patterns obtained from a text corpus. In *Proceedings of the 11th International FLAIRS Conference* (pp. 304–308).
- Bellezza, F. S., Greenwald, A. G., & Banaji, M. R. (1986). Words high and low in pleasantness as rated by male and female college students. *Behavior Research Methods, Instruments, & Computers*, 18(3), 299–303.
- Benamara, F., Cesarano, C., Picariello, A., Recupero, D. R., & Subrahmanian, V. S. (2007). Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*.
- Besnier, N. (1990). Language and affect. *Annual Review of Anthropology*, 19, 419–451.
- Bollen, J., Mao, H., & Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1), 1–8.
- Bollen, J., Pepe, A., & Mao, H. (2011). Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media* (pp. 450–453).
- Bouckaert, R. R. (1967). *Bayesian belief networks: From construction to inference*. Universiteit Utrecht, Faculteit Wiskunde en Informatica.
- Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. *Journal of Behavior Therapy and Experimental Psychiatry*, 25(1), 49–59.
- Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Instruction manual and affective ratings*. University of Florida: Technical Report C-1, The Center for Research in Psychophysiology.
- Brave, S., & Nass, C. (2009). Emotion in human-computer interaction. In *The Human-Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications* (pp. 53–68). CRC Press.
- Breiman, L. (1996). Bagging predictors. *Machine Learning*, 24(2), 123–140.
- Brooks, M., Kuksenok, K., Torkildson, M. K., Perry, D., Robinson, J. J., Scott, T. J., Anicello, O., Zukowski, A., Harris, P., & Aragon, C. R. (2013). Statistical affect detection in collaborative chat. Presented at the Conference on Computer Supported Cooperative Work and Social Computing, San Antonio, TX.
- Byrt, T., Bishop, J., & Carlin, J. B. (1993). Bias, prevalence and kappa. *Journal of Clinical Epidemiology*, 46(5), 423–429.

- Calvo, R. A., & D'Mello, S. (2010). Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1), 18–37.
- Calvo, R. A., & Mac Kim, S. (2012). Emotions in text: Dimensional and categorical models. *Computational Intelligence*, 29(3), 527–543.
- Chaffar, S., & Inkpen, D. (2011). Using a heterogeneous dataset for emotion analysis in text. In C. Butz & P. Lingras (Eds.), *Advances in Artificial Intelligence* (pp. 62–67).
- Chaumartin, F. R. (2007). UPAR7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)* (pp. 422–425).
- Cherry, C., Mohammad, S. M., & de Bruijn, B. (2012). Binary classifiers and latent sequence models for emotion detection in suicide notes. *Biomedical Informatics Insights*, 5, 147–154.
- Chuang, Z.-J., & Wu, C.-H. (2004). Multi-modal emotion recognition from speech and text. *Computational Linguistics and Chinese Language Processing*, 9(2), 45–62.
- Church, K. W., & Hanks, P. (1990). Word association norms, mutual information, and lexicography. *Comput. Linguist.*, 16(1), 22–29.
- Clore, G. L., & Ortony, A. (1988). The semantics of the affective lexicon. In V. Hamilton, G. H. Bower, & N. H. Frijda (Eds.), *Cognitive Perspectives on Emotion and Motivation* (pp. 367–397). Springer Netherlands.
- Clore, G. L., & Schnall, S. (2005). The influence of affect on attitude. In D. Albarrac, B. T. Johnson, & M. P. Zanna (Eds.), *The Handbook of Attitudes* (pp. 437–489). Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.
- Clore, G. L., Wyer, R. S., Dienes, B., Gasper, K., Gohm, C., & Isbell, L. (2001). Affective feelings as feedback: Some cognitive consequences. *Theories of Mood and Cognition: A User's Guidebook*, 27–62.
- Corbin, J., & Strauss, A. (2008). *Basics of qualitative research: Techniques and procedures for developing grounded theory*. Sage.
- Cornelius, R. R. (1996). *The science of emotion: Research and tradition in the psychology of emotions*. Upper Saddle River, New Jersey: Prentice Hall.
- Cowie, R. (2009). Perceiving emotion: Towards a realistic understanding of the task. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, 364(1535), 3515–3525.
- Cui, A., Zhang, M., Liu, Y., & Ma, S. (2011). Emotion tokens: Bridging the gap among multilingual Twitter sentiment analysis. In M. V. M. Salem, K. Shaalan, F. Oroumchian, A. Shakery, & H. Khelalfa (Eds.), *Information Retrieval Technology* (pp. 238–249). Springer Berlin Heidelberg.
- Darwin, C. (1872). *The expression of the emotions in man and animals*. Chicago: University of Chicago Press.

- Davies, A., & Ghahramani, Z. (2011). Language-independent Bayesian sentiment mining of Twitter. In *Workshop on Social Network Mining and Analysis*.
- De Choudhury, M., Counts, S., & Horvitz, E. (2013). Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (pp. 3267–3276). New York, NY, USA.
- De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the Seventh International AAI Conference on Weblogs and Social Media* (pp. 128–137).
- Derczynski, L., Ritter, A., Clark, S., & Bontcheva, K. (2013). Twitter part-of-speech tagging for all: Overcoming sparse and noisy data. In *Proceedings of Recent Advances in Natural Language Processing (RANLP)* (pp. 198–206). Hissar, Bulgaria.
- DeRose, S. J. (2005). The Compass DeRose guide to emotion words.
- Dixon, T. (2012). “Emotion”: The history of a keyword in crisis. *Emotion Review*.
- Doan, S., Vo, B.-K. H., & Collier, N. (2012). An analysis of Twitter messages in the 2011 Tohoku earthquake. In P. Kostkova, M. Szomszor, & D. Fowler (Eds.), *Electronic Healthcare* (pp. 58–66). Springer Berlin Heidelberg.
- Dodds, P. S., & Danforth, C. M. (2010). Measuring the happiness of large-scale written expression: Songs, blogs, and Presidents. *Journal of Happiness Studies*, 11(4), 441–456.
- Donath, J., Karahalios, K., & Viégas, F. (1999). Visualizing conversation. *Journal of Computer-Mediated Communication*, 4(4), 0–0.
- Dong, Y., Chen, H., Qian, W., & Zhou, A. (2015). Micro-blog social moods and Chinese stock market: The influence of emotional valence and arousal on Shanghai Composite Index volume. *International Journal of Embedded Systems*, 7(2), 148–155.
- Ekman, P. (1971). Universals and cultural differences in facial expressions of emotion. *Nebraska Symposium on Motivation*, 19, 207–283.
- Ekman, P. (1973). *Darwin and facial expression: A century of research in review*. New York: Academic Press.
- Ekman, P. (1977). Biological and cultural contributions to body and facial movement. In *The Anthropology of the Body: A. S. A. Monograph 15* (pp. 34–84). New York: Academic Press.
- Ekman, P. (1992). An argument for basic emotions. *Cognition & Emotion*, 6(3-4), 169–200.
- Ekman, P. (1999). Basic emotions. In *Handbook of Cognition and Emotion* (pp. 45–60). John Wiley & Sons, Ltd.
- Ekman, P., & Friesen, W. V. (1971). Constants across cultures in the face and emotion. *Journal of Personality and Social Psychology*, 17(2), 124–129.
- Ekman, P., Friesen, W. V., O’Sullivan, M., Chan, A., Diacoyanni-Tarlatzis, I., Heider, K., Rainer, K., LeCompte, W. A., Pitcairn, T., Ricci-Bitti, P. E., Scherer, K., Tomita, M., & Tzavaras,

- A. (1987). Universals and cultural differences in the judgments of facial expressions of emotion. *Journal of Personality and Social Psychology*, 53(4), 712–717.
- Ekman, P., & Rosenberg, E. L. (1997). *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*. Oxford University Press.
- Elliott, C. D. (1992). *The affective reasoner: A process model of emotions in a multi-agent system*. Northwestern University, Evanston, IL, USA.
- Farzindar, A., & Inkpen, D. (2015). Natural Language Processing for Social Media. *Synthesis Lectures on Human Language Technologies*, 8(2), 1–166.
- Fehr, B., & Russell, J. A. (1984). Concept of emotion viewed from a prototype perspective. *Journal of Experimental Psychology: General*, 113(3), 464–486.
- Fleiss, J. L. (1971). Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5), 378–382.
- Fleiss, J. L., Levin, B., & Paik, M. C. (2013). The measurement of interrater agreement. In *Statistical methods for rates and proportions*. John Wiley & Sons.
- Freund, Y., & Schapire, R. E. (1996). Experiments with a new boosting algorithm. In *International Conference on Machine Learning (ICML)* (pp. 148–156).
- Frijda, N. H. (1986). *The emotions*. Cambridge University Press.
- Frijda, N. H. (1987). Emotion, cognitive structure, and action tendency. *Cognition & Emotion*, 1(2), 115–143.
- Fussell, S. R. (2002). *The verbal communication of emotions: Interdisciplinary perspectives*. Lawrence Erlbaum Associates, Inc.
- Garcia, D., & Schweitzer, F. (2011). Emotions in product reviews—Empirics and models. In *Proceedings of the 2011 IEEE Third International Conference on Social Computing* (pp. 483–488).
- Ghazi, D., Inkpen, D., & Szpakowicz, S. (2010). Hierarchical approach to emotion recognition and classification in texts. In A. Farzindar & V. Kešelj (Eds.), *Advances in Artificial Intelligence* (pp. 40–50).
- Glaser, B. G., & Strauss, A. L. (1967). *The discovery of grounded theory: Strategies for qualitative research*. Chicago: Aldine.
- González-Ibáñez, R., Muresan, S., & Wacholder, N. (2011). Identifying sarcasm in Twitter: A closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies* (Vol. 2, pp. 581–586). Stroudsburg, PA, USA.
- Grassi, M. (2009). Developing HEO Human Emotions Ontology. In J. Fierrez, J. Ortega-Garcia, A. Esposito, A. Drygajlo, & M. Faundez-Zanuy (Eds.), *Biometric ID Management and Multimodal Communication* (pp. 244–251).

- Grefenstette, G., Qu, Y., Shanahan, J. G., & Evans, D. A. (2004). Coupling niche browsers and affect analysis for an opinion mining application. In *Proceedings of the 12th International Conference on Recherche d'Information Assistée par Ordinateur (RIA'O '04)* (pp. 186–194).
- Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing? *International Journal of Human-Computer Studies*, 43(5–6), 907–928.
- Gupta, N., Gilbert, M., & Di Fabrizio, G. (2010). Emotion detection in email customer care. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 10–16).
- Halliday, M. A. K., Cermáková, A., Teubert, W., & Yallop, C. (2004). *Lexicology and corpus linguistics*. A&C Black.
- Hall, M., Frank, E., Holmes, G., Pfahringer, B., Reutemann, P., & Witten, I. H. (2009). The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1), 10–18.
- Hasan, M., Agu, E., & Rundensteiner, E. (2014). Using hashtags as labels for supervised learning of emotions in Twitter messages.
- Hasan, M., Rundensteiner, E., & Agu, E. (2014). EMOTEX: Detecting emotions in Twitter messages. Presented at the 2014 ASE BIGDATA/SOCIALCOM/CYBERSECURITY Conference, Stanford University.
- Hatzivassiloglou, V., & McKeown, K. R. (1997). Predicting the semantic orientation of adjectives. In *Proceedings of the Eighth Conference on European Chapter of the Association for Computational Linguistics* (pp. 174–181). Stroudsburg, PA, USA.
- Holzman, L. E., & Pottenger, W. (2003). Classification of emotions in internet chat: An application of machine learning using speech phonemes.
- HUMAINE. (2013). International Survey On Emotion Antecedents And Reactions (ISEAR).
- Isbister, K., & Nass, C. (2000). Consistency of personality in interactive characters: verbal cues, non-verbal cues, and user characteristics. *International Journal of Human-Computer Studies*, 53(2), 251–267.
- Izard, C. E. (1971). *The face of emotion* (Vol. xii). East Norwalk, CT, US: Appleton-Century-Crofts.
- Izard, C. E. (1977). *Human emotions*. New York: Plenum Press.
- Izard, C. E. (1994). Innate and universal facial expressions: Evidence from developmental and cross-cultural research. *Psychological Bulletin*, 115(2), 288–299.
- Izard, C. E. (2010). The many meanings/aspects of emotion: Definitions, functions, activation, and regulation. *Emotion Review*, 2(4), 363–370.
- James, W. (1884). What is an emotion? *Mind*, 13(IX(34)), 188–205.

- Jia, L., Yu, C., & Meng, W. (2009). The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management* (pp. 1827–1830). New York, NY, USA.
- Joachims, T. (1998). Text categorization with support vector machines: Learning with many relevant features. In *European Conference on Machine Learning (ECML)* (pp. 137–142).
- Johnson-Laird, P. N., & Oatley, K. (1992). Basic emotions, rationality, and folk theory. *Cognition and Emotion*, 6(3-4), 201–223.
- Kagan, J. (1978). On emotion and its development: A working paper. In M. Lewis & L. A. Rosenblum (Eds.), *The Development of Affect* (pp. 11–41).
- Kalra, A., & Karahalios, K. (2005). TextTone: Expressing emotion through text. *Human-Computer Interaction-INTERACT*, 966–969.
- Kao, E. C. C., Liu, C. C., Yang, T. H., Hsieh, C. T., & Soo, V. W. (2009). Towards text-based emotion detection a survey and possible improvements. In *International Conference on Information Management and Engineering, 2009. ICIME '09* (pp. 70 –74).
- Karlgren, J., Sahlgren, M., Olsson, F., Espinoza, F., & Hamfors, O. (2012). Usefulness of sentiment analysis. In *Advances in Information Retrieval* (pp. 426–435).
- Kennedy, A., & Inkpen, D. (2006). Sentiment classification of movie reviews using contextual valence shifters. *Computational Intelligence*, 22(2), 110–125.
- Kennedy, A., Kazantseva, A., Mohammad, S., Copeck, T., Inkpen, D., & Szpakowicz, S. (2011). Getting emotional about news. In *Proceedings of the Text Analysis Conference*. Gaithersburg, MD.
- Keshtkar, F., & Inkpen, D. (2012). A hierarchical approach to mood classification in blogs. *Natural Language Engineering*, 18(01), 61–81.
- Kim, E., Gilbert, S., Edwards, M. J., & Graeff, E. (2009). *Detecting sadness in 140 characters: Sentiment analysis and mourning Michael Jackson on Twitter* (Web Ecology Project) (pp. 1–15).
- Kim, S., Bak, J., & Oh, A. H. (2012). Do you feel what I feel? Social aspects of emotions in Twitter conversations. In *Proceedings of the 6th International AAI Conference on Weblogs and Social Media (ICWSM)*. (pp. 495-498).
- Kleinginna Jr, P. R., & Kleinginna, A. M. (1981). A categorized list of emotion definitions, with suggestions for a consensual definition. *Motivation and Emotion*, 5(4), 345–379.
- Kouloumpis, E., Wilson, T., & Moore, J. (2011). Twitter sentiment analysis: The good the bad and the omg! In *Proceedings of the 5th International AAI Conference on Weblogs and Social Media (ICWSM)* (pp. 538–541).
- Kövecses, Z. (1990). *Emotion concepts*. New York, NY, US: Springer-Verlag Publishing.
- Kövecses, Z. (2007). *Metaphor and emotion: Language, culture, and body in human feeling*. New York: Cambridge University Press.

- Kövecses, Z., & Palmer, G. B. (1999). Language and emotion concepts: What experimentalists and social constructionists have in common. In *Languages of sentiment: Cultural constructions of emotional substrates* (Vol. 18). John Benjamins Publishing.
- Krippendorff, K. (2004). *Content analysis: An introduction to its methodology*. Newbury Park, CA: Sage.
- Lange, C. G., & James, W. (1922). *The emotions*. Williams & Wilkins.
- Lazarus, R. S. (1991). *Emotion and adaptation*. New York: Oxford University Press.
- Lewis, D. D. (1992). An evaluation of phrasal and clustered representations on a text categorization task. In *Proceedings of the 15th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 37–50). New York, NY, USA.
- Liu, H., Lieberman, H., & Selker, T. (2003). A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th International Conference on Intelligent User Interfaces* (pp. 125–132). New York, USA.
- Liu, H., & Singh, P. (2004). ConceptNet — A practical commonsense reasoning tool-kit. *BT Technology Journal*, 22(4), 211–226.
- Ma, C., Prendinger, H., & Ishizuka, M. (2005). Emotion estimation and reasoning based on affective textual interaction. In J. Tao, T. Tan, & R. W. Picard (Eds.), *Affective Computing and Intelligent Interaction* (pp. 622–628).
- Mehrabian, A., & Russell, J. A. (1974). *An approach to environmental psychology* (Vol. xii). Cambridge, MA, US: The MIT Press.
- Merriam-Webster. (2013). Merriam-Webster online: Dictionary and thesaurus. Retrieved from <http://www.merriam-webster.com/>
- Miestamo, M. (2007). Negation – An overview of typological research. *Language and Linguistics Compass*, 1(5), 552–570.
- Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11), 39–41.
- Mishne, G. (2005). Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access* (Vol. 19, pp. 321–327).
- Mishne, G., & De Rijke, M. (2006). Capturing global mood levels using blog posts. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* (pp. 145–152).
- Mohammad, S. M. (2011). From once upon a time to happily ever after: Tracking emotions in novels and fairy tales. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities* (pp. 105–114). Stroudsburg, PA, USA.
- Mohammad, S. M. (2012a). #Emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics* (pp. 246–255). Montreal, QC.



- Mohammad, S. M. (2012b). From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4), 730–741.
- Mohammad, S. M. (2012c). Portable features for classifying emotional text. In *Proceedings of The 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL: HLT-2012)* (pp. 587–591). Montreal, QC.
- Mohammad, S. M., & Kiritchenko, S. (2013). Using nuances of emotion to identify personality. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media* (pp. 27–30).
- Mohammad, S. M., & Kiritchenko, S. (2015). Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2), 301–326.
- Mohammad, S. M., & Turney, P. D. (2008). Crowdsourcing the creation of a word–emotion association lexicon. *Computational Intelligence*, 59.
- Mohammad, S. M., & Turney, P. D. (2010). Emotions evoked by common words and phrases: Using Mechanical Turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text* (pp. 26–34).
- Mohammad, S. M., & Yang, T. W. (2011). Tracking sentiment in mail: How genders differ on emotional axes. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis* (pp. 70–79). Portland, OR.
- Mohammad, S. M., Zhu, X., & Martin, J. (2014). Semantic role labeling of emotions in tweets. In *Proceedings of the ACL 2014 Workshop on Computational Approaches to Subjectivity, Sentiment, and Social Media (WASSA)* (pp. 32–41). Baltimore, MD.
- Mohler, M., Bracewell, D., Hinote, D., & Tomlinson, M. (2013). Semantic signatures for example-based linguistic metaphor detection. In *Proceedings of the First Workshop on Metaphor in NLP* (pp. 27–35).
- Moore, B. S., & Isen, A. M. (1990). *Affect and social behavior*. Cambridge: Cambridge University Press.
- Mulcrone, K. (2012). Detecting emotion in text. Presented at the UMM CSci Senior Seminar Conference, Morris, MN.
- Nakov, P., Kozareva, Z., Ritter, A., Rosenthal, S., Stoyanov, V., & Wilson, T. (2013). SemEval-2013 task 2: Sentiment analysis in Twitter. In *Proceedings of the 7th International Workshop on Semantic Evaluation (SemEval 2013)* (pp. 312–320).
- Narducci, F., de Gemmis, M., & Lops, P. (2015). A general architecture for an emotion-aware content-based recommender system. In *Proceedings of the 3rd Workshop on Emotions and Personality in Personalized Systems 2015* (pp. 3–6). New York, NY, USA.
- Neuendorf, K. A. (2002). *The content analysis guidebook*. SAGE.
- Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2007a). Analysis of affect expressed through the evolving language of online communication. In *Proceedings of the 12th International Conference on Intelligent User Interfaces* (pp. 278–281). New York, NY, USA.

- Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2007b). Narrowing the social gap among people involved in global dialog: Automatic emotion detection in blog posts. In *Proceedings of the International Conference on Weblogs and Social Media* (pp. 293–294).
- Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2007c). Recognition of affect conveyed by text messaging in online communication. In *Online Communities and Social Computing* (pp. 141–150).
- Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2007d). Textual affect sensing for sociable and expressive online communication. In *Affective Computing and Intelligent Interaction* (pp. 218–229).
- Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2011a). Affect Analysis Model: Novel rule-based approach to affect sensing from text. *Natural Language Engineering*, 17(01), 95–135.
- Neviarouskaya, A., Prendinger, H., & Ishizuka, M. (2011b). SentiFul: A Lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 2(1), 22–36.
- Nielsen, F. Å. (2011). A new ANEW: Evaluation of a word list for sentiment analysis in microblogs. *arXiv:1103.2903*.
- Nikfarjam, A., Emadzadeh, E., & Gonzalez, G. (2012). A hybrid system for emotion extraction from suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1), 165–174.
- Nissenbaum, H. F. (1985). *Emotion and focus*. Center for the Study of Language and Information.
- Ochs, E., & Schieffelin, B. (1989). Language has a heart. *Text*, 9(1), 7–25.
- Ortony, A., Clore, G. L., & Collins, A. (1988). *The cognitive structure of emotions*. Cambridge University Press.
- Ortony, A., & Turner, T. J. (1990). What's basic about basic emotions? *Psychological Review*, 97(3), 315–331.
- Pajupuu, H., Kerge, K., & Altrov, R. (2012). Lexicon-based detection of emotion in different types of texts: Preliminary remarks. *Eesti Rakenduslingvistika Ühingu Aastaraamat*, (8), 171–184.
- Pak, A., & Paroubek, P. (2010). Twitter as a corpus for sentiment analysis and opinion mining. In *Seventh International Conference on Language Resources and Evaluation (LREC)*. (pp. 1320-1326).
- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1–135.
- Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing* (Vol. 10, pp. 79–86). Stroudsburg, PA, USA.

- Park, M., Cha, C., & Cha, M. (2012). Depressive moods of users portrayed in Twitter. In *Proceedings of the ACM SIGKDD Workshop on Healthcare Informatics (HI-KDD)* (pp. 1–8).
- Parrott, W. G. (2001). *Emotions in social psychology: Essential readings* (Vol. xiv). New York, NY, US: Psychology Press.
- Passonneau, R. (2004). *Computing reliability for coreference annotation*.
- Passonneau, R. (2006). Measuring agreement on set-valued items (MASI) for semantic and pragmatic annotation. In *Proceedings of the Fifth International Conference on Language Resources and Evaluation (LREC)* (pp. 831–836).
- Pennebaker, J. W., Chung, C. K., Ireland, M., Gonzales, A., & Booth, R. J. (2007). The development and psychometric properties of LIWC2007. *Austin, TX, LIWC. Net*.
- Pennebaker, J. W., Mehl, M. R., & Niederhoffer, K. G. (2003). Psychological aspects of natural language use: Our words, our selves. *Annual Review of Psychology*, 54(1), 547–577.
- Pestian, J. P., Matykiewicz, P., Linn-Gust, M., South, B., Uzuner, O., Wiebe, J., Cohen, K. B., Hurdle, J., Brew, C. (2012). Sentiment analysis of suicide notes: A shared task. *Biomedical Informatics Insights*, 5(Suppl. 1), 3–16.
- Picard, R. W. (1998). *Affective computing*. MIT Press.
- Platt, J. C. (1998). Fast Training of Support Vector Machines Using Sequential Minimal Optimization - Microsoft Research. In *Advances in Kernel Methods - Support Vector Learning* (pp. 41–65). MIT Press.
- Plutchik, R. (1962). *The Emotions: Facts, theories, and a new model*. New York: Random House.
- Plutchik, R. (1980). *Emotion: A psychoevolutionary synthesis*. New York: Harper & Row.
- Plutchik, R. (1984). Emotions: A general psychoevolutionary theory. *Approaches to Emotion*, 1984, 197–219.
- Potter, W. J., & Levine-Donnerstein, D. (1999). Rethinking validity and reliability in content analysis. *Journal of Applied Communication Research*, 27(3), 258–284.
- Provost, F. (2000). Machine learning from imbalanced data sets 101. In *Proceedings of the AAAI 2000 Workshop on Imbalanced Data Sets* (pp. 1–3).
- Purver, M., & Battersby, S. (2012). Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 482–491). Stroudsburg, PA, USA: Association for Computational Linguistics.
- Quercia, D., Ellis, J., Capra, L., & Crowcroft, J. (2012). Tracking “Gross Community Happiness” from tweets. In *Proceedings of the ACM 2012 Conference on Computer Supported Cooperative Work* (pp. 965–968). New York, NY, USA.
- Qu, Y., Huang, C., Zhang, P., & Zhang, J. (2011). Microblogging after a major disaster in China: A case study of the 2010 Yushu earthquake. In *Proceedings of the ACM 2011*

- Conference on Computer Supported Cooperative Work (pp. 25–34). New York, NY, USA.
- Read, J. (2004). *Recognising affect in text using pointwise-mutual information* (M. Sc. Dissertation). University of Sussex.
- Rezabeck, L. L., & Cochenour, J. J. (1995). Emoticons: Visual cues for computer-mediated communication. In *Imagery and Visual Literacy: Selected Readings from the Annual Conference of the International Visual Literacy Association*. Tempe, Arizona.
- Roberts, K., Roach, M. A., Johnson, J., Guthrie, J., & Harabagiu, S. M. (2012). EmpaTweet: Annotating and detecting emotions on Twitter. In *8th International Conference on Language Resources and Evaluation (LREC)* (pp. 3806–3813).
- Rosenthal, S., Nakov, P., Ritter, A., & Stoyanov, V. (2014). Semeval-2014 task 9: Sentiment analysis in Twitter. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)* (pp. 73–80). Dublin, Ireland.
- Rubin, V. L., Stanton, J. M., & Liddy, E. D. (2004). Discerning emotions in texts. In *The AAAI Symposium on Exploring Attitude and Affect in Text (AAAI-EAAT)*.
- Russell, J. A. (1980). A circumplex model of affect. *Journal of Personality and Social Psychology*, 39(6), 1161–1178.
- Russell, J. A. (2003). Core affect and the psychological construction of emotion. *Psychological Review*, 110(1), 145–172.
- Russell, J. A. (2012). From a psychological constructionist perspective. In R. D. Ellis & P. Zachar (Eds.), *Categorical versus dimensional models of affect: A seminar on the theories of Panksepp and Russell* (pp. 79–118). John Benjamins Publishing.
- Russell, J. A., & Mehrabian, A. (1977). Evidence for a three-factor theory of emotions. *Journal of Research in Personality*, 11(3), 273–294.
- Russell, J. A., & Pratt, G. (1980). A description of the affective quality attributed to environments. *Journal of Personality and Social Psychology*, 38(2), 311–322.
- Russ, S. W. (1993). *Affect and creativity: The role of affect and play in the creative process*. Hillsdale, New Jersey: Lawrence Erlbaum Associates.
- Salvetti, F., Reichenbach, C., & Lewis, S. (2006). Opinion polarity identification of movie reviews. In J. G. Shanahan, Y. Qu, & J. Wiebe (Eds.), *Computing Attitude and Affect in Text: Theory and Applications* (pp. 303–316). Springer Netherlands.
- Scherer, K. R. (1993). Studying the emotion-antecedent appraisal process: An expert system approach. *Cognition & Emotion*, 7(3-4), 325–355.
- Scherer, K. R. (1999). Appraisal theory. In T. Dalgleish & M. J. Power (Eds.), *Handbook of Cognition and Emotion* (pp. 637–663). New York: John Wiley & Sons.
- Scherer, K. R. (2005). What are emotions? And how can they be measured? *Social Science Information*, 44(4), 695–729.

- Scott, T. J., Kuksenok, K., Perry, D., Brooks, M., Anicello, O., & Aragon, C. (2012). Adapting grounded theory to construct a taxonomy of affect in collaborative online chat. In *Proceedings of the 30th ACM International Conference on Design of Communication* (pp. 197–204). New York, USA.
- Shaikh, M. A. M., Helmut, P., & Ishizuka, M. (2006). A cognitively based approach to affect sensing from text. In *Proceedings of the 11th International Conference on Intelligent User Interfaces* (pp. 303–305). New York, NY, USA.
- Shaikh, M. A. M., Islam, M. T., & Ishizuka, M. (2006). ASNA: An intelligent agent for retrieving and classifying news on the basis of emotion-affinity. In *Proceedings of the International Conference on Intelligent Agents, Web Technologies and Internet Commerce* (pp. 133–138). Sydney.
- Shaikh, M. A. M., Prendinger, H., & Ishizuka, M. (2007a). Emotion sensitive news agent: An approach towards user centric emotion sensing from the news. In *IEEE/WIC/ACM International Conference on Web Intelligence* (pp. 614–620).
- Shaikh, M. A. M., Prendinger, H., & Ishizuka, M. (2007b). SenseNet: A linguistic tool to visualize numerical-valence based sentiment of textual data. In *Proceedings of the International Conference on Natural Language Processing (ICON)* (pp. 147–152).
- Shaikh, M. A. M., Prendinger, H., & Ishizuka, M. (2009). A linguistic interpretation of the OCC emotion model for affect sensing from text. In *Affective Information Processing* (pp. 45–73).
- Shaikh, M. A. M., Prendinger, H., & Mitsuru, I. (2007). Rules of emotions: A linguistic interpretation of an emotion model for affect sensing from texts. In *Affective Computing and Intelligent Interaction* (Vol. 4738, pp. 737–738).
- Shaver, P., Schwartz, J., Kirson, D., & O'Connor, C. (2001). Emotion knowledge: Further exploration of a prototype approach. In *Emotions in Social Psychology* (pp. 26–56). Psychology Press.
- Shivhare, S. N., & Khethawat, S. (2012). Emotion detection from text. *arXiv:1205.4944*.
- Short, J., Williams, E., & Christie, B. (1976). *The social psychology of telecommunications*. John Wiley and Sons.
- Siegle, G. (1994). The original Balanced Affective Word List project.
- Sim, J., & Wright, C. C. (2005). The kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy*, 85(3), 257–268.
- Sintsova, V., Musat, C.-C., & Pu Faltings, P. (2013). Fine-grained emotion recognition in Olympic tweets based on human computation. Presented at the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, Atlanta, Georgia.
- Sohn, S., Torii, M., Li, D., Waghlikar, K., Wu, S., & Liu, H. (2012). A hybrid approach to sentiment sentence classification in suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1), 43–50.

- Sokolova, M., & Lapalme, G. (2009). A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4), 427–437.
- Sorower, M. S. (2010). *A literature survey on algorithms for multi-label learning*. Oregon State University, Corvallis.
- Staiano, J., & Guerini, M. (2014). DepecheMood: A lexicon for emotion analysis from crowd-annotated news. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics* (pp. 427–433). Baltimore, Maryland, USA.
- Stone, P. J., Dunphy, D. C., Smith, M. S., & Ogilvie, D. M. (1966). *The general inquirer: A computer approach to content analysis* (Vol. 8). Cambridge, MA: MIT Press.
- Strapparava, C., & Mihalcea, R. (2007). Semeval-2007 task 14: Affective text. In *Proceedings of the 4th International Workshop on Semantic Evaluations* (pp. 70–74). Prague.
- Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM Symposium on Applied Computing* (pp. 1556–1560). New York, USA.
- Strapparava, C., & Valitutti, A. (2004). WordNet-Affect: An affective extension of WordNet. In *Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC)* (pp. 1083–1086).
- Subasic, P., & Huettnner, A. (2001). Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems*, 9(4), 483–496.
- Suttles, J., & Ide, N. (2013). Distant supervision for emotion classification with discrete binary values. In A. Gelbukh (Ed.), *Computational Linguistics and Intelligent Text Processing* (pp. 121–136).
- Taboada, M., Anthony, C., & Voll, K. (2006). Methods for creating semantic orientation dictionaries. In *Proceedings of the 5th Conference on Language Resources and Evaluation (LREC'06)* (pp. 427–432).
- Taboada, M., Brooke, J., Tofiloski, M., Voll, K., & Stede, M. (2011). Lexicon-based methods for sentiment analysis. *Computational Linguistics*, 37(2), 267–307.
- Tan, C.-M., Wang, Y.-F., & Lee, C.-D. (2002). The use of bigrams to enhance text categorization. *Information Processing & Management*, 38(4), 529–546.
- Tang, D., Wei, F., Qin, B., Zhou, M., & Liu, T. (2014). Building large-scale Twitter-specific sentiment lexicon: A representation learning approach. In *Proceedings of COLING 2014, the 25th International Conference on Computational Linguistics: Technical Papers* (pp. 172–182). Dublin, Ireland.
- Tao, J. (2004). Context based emotion detection from text input. Presented at the International Conference on Spoken Language Processing (ICSLP2004), Jeju, Korea.
- Teubert, W. (2004). Units of meaning, parallel corpora, and their implications for language teaching. *Language and Computers*, 52(1), 171–189.
- Teufel, S., Siddharthan, A., & Tidhar, D. (2006). An annotation scheme for citation function. In *Proceedings of the 7th SIGdial Workshop on Discourse and Dialogue* (pp. 80–87). Stroudsburg, PA, USA: Association for Computational Linguistics.

- Turney, P. (2001). Mining the web for synonyms: PMI-IR versus LSA on TOEFL. In *Machine Learning: ECML 2001* (pp. 491–502).
- Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: The kappa statistic. *Family Medicine*, 37(5), 360–363.
- Vo, B.-K. H., & Collier, N. (2013). Twitter emotion analysis in earthquake situations. *International Journal of Computational Linguistics and Applications*, 4(1), 159–173.
- Walther, J. B., Loh, T., & Granka, L. (2005). Let me count the ways: The interchange of verbal and nonverbal cues in computer-mediated and face-to-face affinity. *Journal of Language and Social Psychology*, 24(1), 36–65.
- Wang, W., Chen, L., Thirunarayan, K., & Sheth, A. P. (2012). Harnessing Twitter “big data” for automatic emotion identification. In *2012 International Conference on Privacy, Security, Risk and Trust (PASSAT), and 2012 International Conference on Social Computing (SocialCom)* (pp. 587–592).
- Watson, D., Clark, L. A., & Tellegen, A. (1988). Development and validation of brief measures of positive and negative affect: The PANAS scales. *Journal of Personality and Social Psychology*, 54(6), 1063–1070.
- Wiebe, J. M., Wilson, T., Bruce, R., Bell, M., & Martin, M. (2004). Learning subjective language. *Computational Linguistics*, 30(3), 277–308.
- Wiebe, J. M., Wilson, T., & Cardie, C. (2005). Annotating expressions of opinions and emotions in language. *Language Resources and Evaluation*, 39(2-3), 165–210.
- Wiener, M., & Mehrabian, A. (1968). *Language within language: Immediacy, a channel in verbal communication*. Ardent Media.
- Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E., & Patwardhan, S. (2005). OpinionFinder: A system for subjectivity analysis. In *Proceedings of HLT/EMNLP on Interactive Demonstrations* (pp. 34–35). Stroudsburg, PA, USA.
- Witten, I. H., & Frank, E. (2005). *Data mining: Practical machine learning tools and techniques* (2nd ed.). Morgan Kaufmann.
- Yang, H., Willis, A., de Roeck, A., & Nuseibeh, B. (2012). A hybrid model for automatic emotion recognition in suicide notes. *Biomedical Informatics Insights*, 5(Suppl. 1), 17–30.
- Yang, Y., & Liu, X. (1999). A re-examination of text categorization methods. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 42–49). New York, NY, USA.
- Zachar, P., & Ellis, R. D. (2012). *Categorical versus dimensional models of affect: A seminar on the theories of Panksepp and Russell* (Vol. 7). John Benjamins Publishing Company.
- Zhang, L. (2013). Contextual and active learning-based affect-sensing from virtual drama improvisation. *ACM Transactions on Speech and Language Processing (TSLP)*, 9(4), 8:1–8:25.

- Zhang, L., & Barnden, J. (2012). Affect sensing using linguistic, semantic and cognitive cues in multi-threaded improvisational dialogue. *Cognitive Computation*, 4(4), 436–459.
- Zhang, P. (2013). The Affective Response Model: A theoretical framework of affective concepts and their relationships in the ICT context. *Management Information Systems Quarterly*, 37(1), 247–274.
- Zhe, X., & Boucouvalas, A. C. (2002). Text-to-emotion engine for real time internet communication. In *Proceedings of International Symposium on Communication Systems, Networks and Digital Signal Processing* (pp. 164–168).
- Zhu, X., Guo, H., Mohammad, S., & Kiritchenko, S. (2014). An empirical study on the effect of negation words on sentiment. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, ACL* (pp. 304-313).



# JASY LIEW SUET YAN

245 Moore Ave Apt 1A, Syracuse, NY 13210, USA  
(315) 706-5060 | jliewsue@syr.edu

## PROFILE

I aspire to become a researcher and educator who can help bridge the communication gap between people and computers through the creation of more emotion-sensitive computers. As humans become ever-more emotive in our use of social media, my research investigates how the use of various natural language processing (NLP) techniques enables computers to better detect the emotional cues expressed in texts. Being able to automatically detect emotion in texts on a large scale allows me to study various emotion-related phenomena online, and to improve the design of emotion-sensitive systems.

## EDUCATION

2011 – 2016: **SYRACUSE UNIVERSITY** Syracuse, NY, U.S.A.

Ph.D. in Information Science and Technology

School of Information Studies (iSchool)

Co-advisors: Dr. Elizabeth Liddy & Dr. Howard Turtle

Dissertation: *Fine-grained Emotion Detection in Microblog Text*

2009 – 2011: **SYRACUSE UNIVERSITY** Syracuse, NY, U.S.A.

M.S. in Information Management, GPA: 4.0 / 4.0

School of Information Studies (iSchool)

Projects: *Smart Dressing Room, Annotating Political Tweets for Sentiments Analysis*

2006 – 2008: **UNIVERSITY OF SCIENCE MALAYSIA (USM)** Malaysia

Bachelor in Computer Science (Honors)

Major: Software Engineering, Minor: Management

Graduated with First Class Honours, GPA: 3.95 / 4.0

Projects: *Crowd Density Estimation System, Malay Date Pattern Recognizer, Improving the Tuanku Fauziah Museum and Gallery Web Site, Virtual Paper, Virtual Health Connect*

2003 – 2005: **UNIVERSITY OF TECHNOLOGY MALAYSIA (UTM)** Malaysia

Diploma in Computer Science (IT)

Graduated with First Class Diploma, GPA : 4.0 / 4.0

1998 – 2002: **METHODIST GIRLS SCHOOL (MGS) IPOH**, Perak, Malaysia

## RESEARCH EXPERIENCES

### Research Assistant

NSF Grant: Qualitative Data Repository

PI: Howard Turtle

Fall 2011 – Fall 2013

Projects: *Conducted interviews to identify tools social scientists use or need, nature of their data and data management practices*

**Research Assistant**

NSF Grant: SOCS – Socially Intelligent Computing for Coding of Qualitative Data

PI: Nancy McCracken

Fall 2011 – Present

*Projects: Conducted research to identify features for machine learning, run experiments to improve the performance of machine learning models for qualitative content analysis, design and evaluate the user interfaces (UIs)*

**Research Intern**

Technology for Emerging Markets (TEM)

Microsoft Research Lab India (MSRI), Bangalore, India

Summer 2010

*Project: Conducted a field study evaluating the effects of Visual Syntactic Formatting (VSF) on reading comprehension for non-native English speakers in rural schools*

**Faculty Assistant**

School of Information Studies, Syracuse University

Fall 2009, Fall 2010

*Projects: Conducted research on consumer's perceived value and attitude towards digital advertisements*

**Research Officer**

Grid Computing Research, University of Science Malaysia

November 1, 2008 – July 31, 2009

National IPv6 Advanced Centre of Excellence, University of Science Malaysia

28 July 2008 – 31 October 2008

*Projects: Conducted research on automatic crowd density estimation*

**RESEARCH AREAS**

- Sentiment Analysis
- Natural Language Processing, Computational Linguistics, Text Mining
- Affective Computing, Human Computer Interaction
- Information Science and Technology

**TEACHING EXPERIENCES****Instructor**

School of Information Studies, Syracuse University

Spring 2016

Course: Natural Language Processing (IST 664)

**Teaching Mentor**

Graduate School, Syracuse University

Summer 2012 - 2015

Teaching Assistant Orientation Program

**Discussion Leader**

School of Information Studies, Syracuse University

Fall 2011

Course: Introduction to Information Management (IST 621)

**Teaching Assistant**

School of Information Studies, Syracuse University

Spring 2014

Course: Basics of Information Retrieval Systems (IST 657)

Spring 2010

Course: Information Systems Analysis and Design (IST 552)

Fall 2010

Course: Information Analysis of Organizational Systems (IST 352)

**Lab Tutor**

University of Science Malaysia

Session 2007/2008 – Semester 2

January 2008 – April 2008

Course: Microsoft Access

**TEACHING INTERESTS**

- Natural Language Processing
- Human Computer Interaction
- Systems Analysis and Design
- Information Retrieval

**PROFESSIONAL EXPERIENCES****Part-time Web Developer**

Tuanku Fauziah Museum and Gallery, University of Science Malaysia

January 2, 2009 – July 15, 2009

**Internship**

Great Eastern Life Assurance Malaysia

May 3, 2007 – September 14, 2007

- Agency Administration Department: Conduct an analysis on the business models for Agency Rating System and Agent Rank Movement System
- IT Department: Maintain and update the Agency Web Portal using Web scripting languages, JSP and Flash, run unit testing for eAOM software system, and develop a web site for School Adoption Program

**JOURNAL PUBLICATIONS****2013**

- Zhang, P., Liew, J. S. Y., & Hassman, K. D. (2013). The intellectual characteristics of the information field: Heritage and substance. *Journal of the American Society for Information Science and Technology*, 64(12), 2468–2491.

**2011**

- Hussain, N., Yatim, H. S. M., Hussain, N. L., Liew, J. S. Y., & Haron, F. (2011). CDES: A pixel-based crowd density estimation system for Masjid al-Haram. *Safety Science*, 49(6), 824–833.

**2010**

- Liew, J. S. Y., & Haron, F. (2010). A survey of crowd density estimation methods for high density crowds. *Journal of Advanced Computing and Application Volume*, 1(1), 8–25.

## CONFERENCE PROCEEDINGS

### 2015

- Liew, J. S. Y. (2015). Discovering Emotions in the Wild: An Inductive Method to Identify Fine-Grained Emotion Categories in Tweets. In *Proceedings of the Twenty-Eighth International Florida Artificial Intelligence Research Society Conference* (pp. 317-322). Hollywood, Florida, USA.

### 2014

- McCracken, N., Liew, J. S. Y., & Crowston, K. (2014). Design of an active learning system with human correction for content analysis. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces* (pp. 59-62). Baltimore, Maryland, USA.
- Liew, J. S. Y., McCracken, N., Zhou, S., & Crowston, K. (2014). Optimizing features in active machine learning for complex qualitative content analysis. In *Proceedings of the ACL 2014 Workshop on Language Technologies and Computational Social Science* (pp. 44-48). Baltimore, Maryland, USA.
- Liew, J. S. Y. (2014). Expanding the range of automatic emotion detection in microblogging text. In *Proceedings of the Student Research Workshop at the 14th Conference of the European Chapter of the Association for Computational Linguistics* (pp. 38-44). Gothenburg, Sweden.
- Liew, J. S. Y., McCracken, N., & Crowston, K. (2014). Semi-automatic content analysis of qualitative data. In *iConference 2014 Proceedings* (pp. 1128-1132). Berlin, Germany.

### 2012

- Liew, J. S. Y., & Kaziunas, E. (2012). What is a tweet worth? Measuring the value of social media for an academic institution. In *Proceedings of the 2012 iConference* (pp. 565-566). Toronto, Canada.
- Liew, J. S. Y., Haron, F., Alginahi, Y., & Kabir, M. (2012). A preliminary crowd monitoring framework for Al-Masjid Al-Haram. In *Proceedings of the 1st Taibah University International Conference on Computing and Information Technology (ICCIT 2012)*, Vol. 1 & 2 (pp. 111-116). Al-Madinah Al-Munawwarah, Saudi Arabia.

### 2011

- Liew, J. S. Y., Kaziunas, E., Liu, J., & Zhuo, S. (2011). Socially-interactive dressing room: An iterative evaluation on interface design. In *Proceedings of the 2011 Annual Conference Extended Abstracts on Human Factors in Computing Systems, CHI EA 2011* (pp. 2023-2028). New York, USA.

### 2008

- Ranaivo-Malançon, B., Liew, J. S. Y., Wong, D. C. P., & Fam, Y. K. (2008). Malay date identifier. In *Second International MALINDO Workshop* (pp. 74-79). Cyberjaya, Selangor, Malaysia.

## CONFERENCE POSTERS/PRESENTATIONS

- Prestopnik, N., Liew, J. S. Y. (2014). Obscuring the task: Story and theme as motivators in an emotion annotation game. Presented a poster at Collective Intelligence Conference 2014, June 10 – 12, 2014, Massachusetts Institute of Technology (MIT), Cambridge, Massachusetts, USA.
- Liew, J. S. Y. (2014). Expanding the range of automatic emotion detection in microblogging text. Presented a poster at CRA-W Graduate Cohort Workshop 2014, April 11 – 12, 2014, Santa Clara, California, USA.

- Liew, J. S. Y., Hassman, K., Zhang, P. (2013). Conceptualizations of technology in the information field. Presented a poster at American Society for Information Science and Technology Annual Meeting, November 1 – 5, 2013, Montreal, Canada.
- Liew, J. S. Y. (2012). Emotion polarity and its effects on a leader's influence on social media. Presented a poster at #Influence12: Symposium & Workshop on Measuring Influence on Social Media, September 28 - 29, 2012, Dalhousie University, Halifax, Nova Scotia, Canada.
- Liew, J. S. Y., Santoso, S. (2010). Twittering to Congress: Mining sentiments from tweets. Presented a poster at New York Celebration of Women in Computing (NYCWIC), April 8 – 9, 2010, Albany, NY, USA.
- Liew, J. S. Y., Kaziunas, E., Liu, J., Zhuo, S. (2010). Smart dressing room: A formative evaluation on design. Presented a poster at Ontario Celebration of Women in Computing (ONCWIC), October 22 – 23, 2010, Kingston, Ontario, Canada.

## **AWARDS/FUNDING**

2014 : ACL 2014 Workshop on Language Technologies and Computational Social Science Travel Award  
 2013 : NEASIS&T Student Travel Award  
 2012 : Graduate School Organization Travel Grant  
 2012 : Elizabeth D. Liddy Summer Fellowship  
 2011 : IPTA Training Scheme Fellowship – SLAB/SLAI (Ministry of Higher Education Malaysia)  
 2011 : IBM Destination Z Enterprise Computing Scholarship  
 2011 : 2011 International Student Leadership Award (International Center of Syracuse)  
 2010 : Graduate School Organization Travel Award  
 2009 : University of Science Malaysia Fellowship  
 2008 : University of Science Malaysia Chancellor's Gold Medal Award  
 2008 : Royal Education Award for Non-Bumiputera by Conference of Rulers (Majlis Raja-Raja)  
 2008 : Best Final Year Student for Computer Science Gold Medal Award  
 2008 : Best Final Year Female Undergraduate Gold Medal Award  
 2006 - 2008 : University of Science Malaysia School of Computer Sciences Dean's List Award  
 2007 : Great Eastern Supremacy Local Scholarship Award  
 2007 : Best Second Year Student for Computer Science  
 2006 : University of Technology Malaysia Vice Chancellor Best Student Award 2005/2006  
 2006 : University of Technology Malaysia KL Best Potential Student Award  
 2006 : Second College Best Student Award  
 2003 – 2006 : University of Technology Malaysia KL Dean's List Award

## **SERVICE**

### **Committees**

- Teaching Mentor Selection Committee, Graduate School, Syracuse University, 2014.
- Excellence in Graduate Education Faculty Recognition Award Selection Committee, Syracuse University, 2013.
- Doctoral Committee, iSchool, Syracuse University, 2011 – 2012.

### **Conferences/Workshops**

- Student Volunteer, 52nd Annual Meeting of the Association for Computational Linguistics (ACL 2014), June 22 – 27, 2014, Baltimore, Maryland, USA.

- Student Volunteer, American Society for Information Science and Technology Annual Meeting (ASIS&T 2013), November 1 – 5, 2013, Montreal, Canada.
- Student Volunteer, Sentiment Analysis Symposium, May 7 – 8, 2013, New York, USA.
- Student Volunteer, ACM CHI Conference on Human Factors in Computing Systems (CHI 2012), May 5 – 10, 2012, Austin, Texas, USA.
- Web Administrator for Second International MALINDO Workshop, 2008, Cyberjaya, Selangor, Malaysia.

#### **Others**

- Women in Science and Engineering-Future Professionals Program (WiSE-FPP) Associate, Syracuse University, 2014-2015.
- Project Manager, Women in Technology (WIT) Girls are IT Workshops, Syracuse University, 2011.
- Graduate Mentor, iSchool, Syracuse University, 2010.
- Director of International Students, iSchool Graduate Student Organization (iSGO), Syracuse University, 2010.
- Conflict Resolution Fellow, Orange Dialogue for Peace Program, 2010 – 2011.
- BLIST Leadership Fellow – Volunteered for a high school leadership outreach program, 2010.
- English Tutor for University of Technology Malaysia KL Student Development Program, 2003 – 2007.

#### **PROFESSIONAL MEMBERSHIPS**

- American Society for Information Science and Technology (ASIS&T)
- Association of Computational Linguistics (ACL)
- Association for Computing Machinery (ACM)
- Women in Technology (WIT), Syracuse University (2010 – 2011)

#### **OTHER ACHIEVEMENTS**

##### **2009**

- Team Leader representing Malaysia in Microsoft Imagine Cup Worldwide Finals Design for Development Award 2009 in Cairo, Egypt (Winner – International Level)
- Team Leader for Microsoft Imagine Cup Software Design Category 2009 (First Runner-up – National Level)
- Team Leader representing Malaysia in Hong Kong PolyU Global Student Challenge 2009 (Finalist – International Level)
- HSBC Young Entrepreneur Awards 2008-09 (Semi-Finals – National Level)
- Team Leader for MSC-IHL Business Plan Competition 2008/09 – Business Plan Category (First Runner-up – National Level)

##### **2008**

- Team Leader for MSC-USM Business Plan Competition 2008 – Business Plan Category (Winner – University Level)
- Student Leader in Microsoft Education Leadership Forum (ELF) in UNESCO, Paris
- Team Leader representing Malaysia in Microsoft Imagine Cup Worldwide Finals in Paris (Quarter Finals – International Level)
- Team Leader for Microsoft Imagine Cup Software Design Category 2008 (Winner – National Level)
- Student Ambassador for Infosys Foundation Program 2008 in Mysore, India

## SKILLS

### Computing

- Programming: PHP, SQL, Python
- Software Applications: SQL Server 2008, Weka

### Languages

- English, Malay

## PERSONAL INTERESTS

- Traveling, learning new cultures, and reading

## REFEREES

- 1) Dr. Howard Turtle  
Director of Center for Natural Language Processing  
316 Hinds Hall, Syracuse, NY 13244  
(315) 443-4061  
[turtle@syr.edu](mailto:turtle@syr.edu)
- 2) Dr. Elizabeth Liddy  
Trustee Professor, Interim Vice Chancellor & Provost  
500 Crouse-Hinds Hall, Syracuse, NY 13244  
(315) 443-1728  
[liddy@syr.edu](mailto:liddy@syr.edu)