Advancing Fine-Grained Emotion Recognition in Short Text

THÈSE N^{O} 7162 (2016)

PRÉSENTÉE LE 5 OCTOBRE 2016

À LA FACULTÉ INFORMATIQUE ET COMMUNICATIONS
GROUPE SCI IC PFP
PROGRAMME DOCTORAL EN INFORMATIQUE ET COMMUNICATIONS

ÉCOLE POLYTECHNIQUE FÉDÉRALE DE LAUSANNE

POUR L'OBTENTION DU GRADE DE DOCTEUR ÈS SCIENCES

PAR

Valentina SINTSOVA

acceptée sur proposition du jury:

Prof. K. Aberer, président du jury
Dr P. Pu Faltings, Prof. B. Faltings, directeurs de thèse
Dr A. Balahur, rapporteuse
Prof. P. Robinson, rapporteur
Dr A. Popescu-Belis, rapporteur

ÉCOLE POLYTECHNIQUE
FÉDÉRALE DE LAUSANNE

Suisse
2016

To my family and friends…

Acknowledgements

While pursuing my PhD degree I met many wonderful, strong-minded, and inspiring people who helped me to grow both intellectually and personally, supported my ideas, and participated wholeheartedly in my life. I would like to take this moment to express my gratitude to them.

First and foremost, I want to thank my advisors, Dr. Pearl Pu and Prof. Boi Faltings, for their trust in me and for the opportunity to pursue my scientific interests under their supervision. Throughout the years of our collaboration, I have learned from Dr. Pearl Pu a lot about work management and motivation, about related skill development, and about focus and persistence in research. She was always there for me with encouraging comments and right suggestions, either about overcoming challenges in ongoing research projects or about searching for new fascinating topics to investigate. Her guidance towards improving my skills and realizing my potential made me a better researcher. I am also thankful to Prof. Boi Faltings for his helpful feedback on my research ideas and projects.

I am sincerely grateful to the committee members who agreed to participate in my defense and shared that final moment with me: Prof. Peter Robinson, Dr. Andrei Popescu-Belis, Dr. Alexandra Balahur, and Prof. Karl Aberer. Thank you for your helpful comments, thought-provoking questions, and constructive suggestions.

Over these past years, I was lucky to get a chance to work with many talented Master and Bachelor students: Renato Kempter, Simona Traykova, Dana Naous, Margarita Bolívar Jiménez, Sephora Madjiheurem, Julien Marengo, Nataniel Hofer, and Elias Schegg. Sharing their inspiration and motivation for work was an enlightening experience. I am happy that some of these collaborations contributed to this dissertation, as acknowledged in the corresponding sections. I also would like to thank Claudiu Musat for his optimism, his personal example on how to approach research problems effectively, and his significant help in improving my writing skills.

Continuing this wave of thankfulness, I would like to thank all the people who participated in my research progress less directly, but deserve their contributions acknowledged: anonymous reviewers of our paper submissions giving their invaluable opinion and feedback, and people with whom I had fruitful discussions at the conferences and workshops. Many thanks go as well to my friends and colleagues, who helped me with multiple dry-runs and paper revisions, incited interesting and productive discussions, and shared their PhD-related (and not only) wisdom and experience: Onur Yürüten, Yu Chen, Rong Hu, Lionel Martin, George

Popescu, Marina Boia, Igor Kulev, Goran Radanovic, Florent Garcin, Aleksei Triastcyn, Grégoire Devauchelle, Sergii Vozniuk, Andrii Vozniuk, Alexandra Olteanu, Julia Proskurnia, Alevtina Dubovitskaya, Nataliya Gerasimova, Evgeny Petrov, Andrew Becker, Daniel Lupei, Adrian Popescu, Alina Dudeanu, Konstantin Zhurov, Atul Sethi, Immanuel Trummer, Ina Bialova, Irina Prostakova, Eugene Burmako, Ruslan Mavlyutov, Roman Prokofyev, and others. I am particularly grateful to Onur, Yu, Rong, and Lionel. Thank you for the words of encouragement, for your input in managing teaching and within-lab responsibilities, and for the atmosphere of a small HCI family that you created. It was fun and stimulating working side by side with you!

I also thank all my friends for being supportive and motivating. I appreciate very much the many moments of laughter, relaxation, and intellectual growth that you gave me. Thank you for the great time we spent together in many leisure activities, being that hiking, going out, dancing, swimming, snowboarding, coffee-breaks, traveling around, or any other crazy or less crazy activity. I especially want to thank Evgeny Petrov, who always added a note of humour and shifted my vision of perspectives at the most difficult times. Thank you for making me more confident, for pointing out my strengths, and for inspiring me to pursue the PhD degree in the first place. I also want to thank Marina Boia for being a perfect companion while progressing our PhD in parallel and for sharing many research and life interests with me. I would like to thank more personally Andrew Becker for his persistence in joining our swimming evenings, for giving advice when needed, and for his constant readiness to help me with my English writing. I also thank Nataliya Gerasimova, Daniel Lupei, and Adrian Popescu for being wonderful friends who could always cheer me up. I am glad that we shared these years in Lausanne going through this PhD journey together.

In the process of writing this thesis, I was relying so much on Grégoire, who was a vigorous defender of my emotional and mental health. Without his fond love and endless positivity, reaching the finish line of my PhD would have been much more difficult. At last but not least, I want to thank my family who supported my decision to study abroad and who always continued motivating me towards new achievements.

I am sending to you all my gratitude. Thank you for your contributions in this tough yet joyful period of my life.

Abstract

Advanced emotion recognition in text is essential for developing intelligent affective applications, which can recognize, react upon, and analyze users' emotions. Our particular motivation for solving this problem lies in large-scale analysis of social media data, such as those generated by Twitter users. Summarizing users' emotions can enable better understandings of their reactions, interests, and motivations. We thus narrow the problem to emotion recognition in short text, particularly tweets.

Another driving factor of our work is to enable discovering emotional experiences at a detailed, fine-grained level. While many researchers focus on recognizing a small number of basic emotion categories, humans experience a larger variety of distinct emotions. We aim to recognize as many as 20 emotion categories from the Geneva Emotion Wheel. Our goal is to study how to build such fine-grained emotion recognition systems.

We start by surveying prior approaches to building emotion classifiers. The main body of this thesis studies two of them in detail: crowdsourcing and distant supervision. Based on them, we design fine-grained domain-specific systems to recognize users' reactions to sporting events captured on Twitter and address multiple challenges that arise in that process.

Crowdsourcing allows extracting affective commonsense knowledge by asking hundreds of workers for manual annotation. The challenge is in collecting informative and truthful annotations. To address it, we design a human computation task that elicits both emotion category labels and emotion indicators (i.e. words or phrases indicative of labeled emotions). We also develop a methodology to build an emotion lexicon using such data. Our experiments show that the proposed crowdsourcing method can successfully generate a domain-specific emotion lexicon. Additionally, we suggest how to teach and motivate non-expert annotators. We show that including a tutorial and using carefully formulated reward descriptions can effectively improve annotation quality.

Distant supervision consists of building emotion classifiers from data that are automatically labeled using some heuristics. This thesis studies heuristics that apply emotion lexicons of limited quality, for example due to missing or erroneous term-emotion associations. We show the viability of such an approach to obtain domain-specific classifiers having substantially better quality of recognition than the initial lexicon-based ones. Our experiments reveal that treating the emotion imbalance in training data and incorporating pseudo-neutral documents is crucial for such improvement. This method can be applied to building emotion classifiers across different domains using limited input resources and thus requiring minimal effort.

Another challenge for lexicon-based emotion recognition is to reduce the error introduced by linguistic modifiers such as negation and modality. We design a data analysis method that allows modeling the specific effects of the studied modifiers, both in terms of shifting emotion categories and changing confidence in emotion presence. We show that the effects of modifiers vary across the emotion categories, which indicates the importance of treating such effects at a more fine-grained level to improve classification quality.

Finally, the thesis concludes with our recommendations on how to address the examined general challenges of building a fine-grained textual emotion recognition system.

Key words: Emotion Recognition, Emotion Lexicons, Emotion, Affective Computing, Text Classification, Sentiment Analysis, Survey, Crowdsourcing, Human Computation, Distant Supervision, Quality Control, Tutorials, Incentives Framing, Twitter, Sports Events, Modifiers, Data Analysis, Social Media

Résumé

La reconnaissance d'émotions est indispensable au développement d'applications affectives intelligentes capables de reconnaître, analyser et réagir face aux émotions humaines. Notre but précis est d'analyser de larges volumes de données, et plus particulièrement des collections de tweets. Nous réduisons donc notre champ d'étude à la reconnaissance d'émotions dans de courts textes, à savoir des tweets.

Par nos travaux, nous souhaitons également pouvoir décrire plus en détail les expériences émotionnelles détectées. Or, si l'homme est capable de ressentir un nombre important d'émotions, la plupart des chercheurs n'en considèrent qu'un nombre restreint. C'est pourquoi nous avons décidé de considérer les 20 émotions provenant du modèle « Geneva Emotion Wheel ». Notre défi sera de construire des systèmes de reconnaissance d'émotions plus nuancés que ceux construits précédemment.

Nous commençons par une revue des méthodes existantes de construction de classifieurs d'émotions. Nous concentrons nos efforts sur deux méthodes : le *crowdsourcing* et le *distant supervision*, sur lesquelles nous construirons nos systèmes de reconnaissance d'émotions. Ces systèmes visent à classifier des réactions recueillies sur Twitter lors d'événements sportifs.

Nous utilisons tout d'abord le crowdsourcing afin d'obtenir des annotations manuelles des tweets, et d'accéder à une connaissance commune de l'affectif. L'enjeu ici est de collecter des annotations manuelles aussi informatives et fidèles que possible. Pour atteindre cet objectif, nous avons conçu une tâche d'annotation de tweets produisant des labels d'émotions et les indicateurs émotionnels associés. Nous développons également une méthodologie nous permettant de construire des lexiques émotionnels. Notre méthode et la pertinence des lexiques générés pour des domaines spécifiques sont ensuite validées. Enfin, nous proposons une méthode d'éducation et de motivation des annotateurs non-experts. Nous montrons notamment qu'inclure un tutoriel et définir de manière judicieuse une récompense améliorent la qualité des annotations.

Le distant supervision consiste quant à lui à apprendre des classifieurs à partir de données annotées de manière automatique par des heuristiques. Les heuristiques que nous appliquons sont adaptées à des lexiques émotionnels dont la qualité peut être limitée, par exemple du fait d'un nombre limité de termes émotionnels. Nous validons une telle approche et montrons que traiter le déséquilibre des catégories d'émotions et incorporer des données pseudo-neutres permet d'améliorer les performances des classifieurs. Enfin, cette méthode peut être appliquée pour construire des classifieurs d'émotions dans différents domaines et ne nécessite que peu de ressources initiales.

Un autre défi de la reconnaissance d'émotions à partir de lexiques est de minimiser les erreurs introduites par des opérations linguistiques telles que la négation et la modalité. Nous proposons donc une méthode d'analyse de l'impact de ces opérations sur le changement des émotions d'une part, et sur le degré de certitude de la présence des émotions d'autre part. Nous montrons que ces effets varient selon les émotions, ce qui souligne l'importance de traiter les émotions séparément afin d'obtenir des classifieurs de qualité.

Enfin, nous donnons des recommandations sur la démarche à adopter lorsque l'on souhaite construire un système de reconnaissance textuelle d'émotions et sur la résolution des défis que cela soulève.

Mots clefs : Reconnaissance d'émotions, Informatique affective, Emotion, Lexiques émotionnels, Classification de textes, Analyse de textes, Crowdsourcing, Distant supervision, Tutoriels, Contrôle de la qualité, Twitter, Evénements sportifs, Opérations linguistiques, Analyse de données, Réseaux sociaux

Contents

List of Figures

List of Tables

1 Introduction

Emotions lie at the foundation of the human experience. We experience emotions on a daily basis because of events happening around us and with us, interactions we participate in, and our thoughts. We might feel happy while spending time with close people, angry because of unfair treatment, or inspired by an online talk. Emotions motivate our actions and help define our goals. They affect our decisions, behavior, and even our thought processes. Moreover, expressions of emotions are important social signals that help us understand others' feelings, motivations, and intentions and to adapt and improve our interpersonal communications and relationships. Emotions shape our life experience giving it color and meaning.[1]

Because of their crucial role in our day-to-day lives, emotions are an important research topic across multiple disciplines, including philosophy, psychology, sociology, and lately computer science. The field of affective computing arose to model the experience of emotions computationally and give computers the power to relate to emotions [Pic95]. Automatic recognition of human emotions is dreamed to foster multiple affective applications. For example, computer assistants could help us regulate our emotions by increasing self-awareness and projecting positive emotions [KMP99]. Such technology could also enhance human-computer interactions, by making computer agents more empathetic and human-like [Pic95, CDCT$^+$01]. In addition to potential user-facing applications, automatic quantifiable recognition of human emotions would allow us to study emotions in society in general. This would help scientists better understand and model the psychology of emotional experiences [MGP10], as well as provide insights on emotional reactions to specific events, objects, or persons [DNKS10, TBP11].

Recognizing human emotions is not a well-defined problem. Multiple different modalities of emotion expression can be used to detect emotional experience. Humans express their emotions non-verbally via gestures [Wal98], speech tones [KR12], and facial expressions [FL03]. Emotional experiences also affect the physiological state of the human body, for example, by changing electrodermal activity [NALF04] and electrical activity of the brain [Pan98]. Additionally, humans describe their emotional states verbally, e.g. by using emotionally charged

[1] Based on a review of the different functions of emotions [KH99, HMnd].

expressions (such as “Yay! We did it!”) or explicit statements (such as “So happy now”). In this thesis, we focus on recognizing and categorizing such verbalized emotional statements captured in text. Language (spoken or written) transfers the information and meaning in interpersonal communications. Thus, the verbal channel provides access to more cognitive, intended expression of emotions, allows distinguishing more specific fine-grained emotional states, and supports extraction of additional context details of emotional experiences. Furthermore, compared to other non-verbal detection methods, which require physical or visual access to personal emotion expressions, text-based recognition can work directly on top of currently used online communication channels. This allows us to employ freely-available linguistic corpora for a more large-scale analysis of emotional experiences.

To recognize emotions in text, we should conceptualize what we understand by emotions. In this work, we consider “emotions” as fine-grained categories or labels that can categorize different emotional experiences. The objective of a text-based emotion recognition system is to detect which emotion categories a writer expresses in a given text sample. For example, when someone writes “Today was awesome,” the system should conclude that the author is happy.

Many researchers attempted to address this problem in recent years. Machine-learning and rule-based knowledge systems were shown to obtain adequate quality across different types of textual data, including dialogues [LLS03, NPI11a], blog posts [AS07, Mis05], and tweets [Moh12a, RRJ$^+$12]. The related problem of sentiment classification, where the objective is to classify a text sample as positive or negative, attracted even a larger volume of research [PL08, Liu12], especially in reviews [PLV02, HL04, TBT$^+$11]. However, due to the multiplicity and complexity of linguistic expressions, as well as the variety of their different contexts, emotion recognition problem remains a challenging task. Also, the previous works mostly focused on recognizing more pronounced, basic emotions (with up to 8 emotion categories) or simply the polarity (positive or negative), which places the problem of recognition at a coarse-grained level. Yet, humans naturally experience and differentiate multiple, more subtle emotions, which presumably differ based on the patterns of cognitive evaluation of emotion-eliciting events [Sch01]. We believe that only with the fine-grained emotion recognition can we provide enough details for insightful analysis. For example, only fine-grained separation of positive emotions can help us distinguish what makes people feel happy, interested, or in love about the subject of study (being that a person, product, or event). Thus, instead of focusing on a small number of emotion categories, our work addresses an even more challenging problem: recognizing a fine-grained set of more subtle, yet distinguishable emotions. *Our goal is to study how to build a new emotion recognition system for a fine-grained set of emotion categories and to address the challenges that arise in this process.*

In this endeavor, we advance the frontier of the research on textual emotion recognition in four directions. For building emotion lexicons via crowdsourcing, we investigate how to collect more informative annotations of emotional data instead of asking for direct word-emotion associations. For ensuring the quality of such annotations, we study how to elicit more truthful annotations in the context of our subjective task while standard quality control techniques were researched in the context of more objective tasks. For building emotion classifiers automatically via distant supervision, we study how to construct and improve emotion classifiers using less restrictive heuristics than the previously suggested ones. For treating the effects of linguistic modifiers, we suggest how to automatically extract a model of the modifiers' impact on emotional expressions instead of the commonly used hand-coding techniques. This dissertation proposes our solutions to these problems and derives practical recommendations for future developers of textual emotion recognition systems.

1.1 Motivation Scenario

We consider the emotional analysis of data from social media as the main application scenario for our work. Social websites, such as Twitter, Facebook, TripAdvisor, and Livejournal, allow users to share their personal opinions, experiences, and emotions on any subject with their friends, subscribers, and the world. As of January 2015, Internet users are estimated to have an average of 2.8 actively used social media accounts.[2] These shared experience data are easily accessible for analysis and already gave birth to multiple research studies of collective behavior. This allows the application of emotion recognition to summarize reactions to specific events [KSMP14], as well as to analyze differences in emotional experiences across different locations [SEK$^+$13] and produce modeled characteristics of interpersonal conversations [KPV$^+$14].

In this work, we focus on the scenario where we need to study some specific collected data to obtain more details about the expressed emotional reactions. This scenario assumes that we first collect the data for a new domain of study, for example, from a specific platform for analysis and on a specific topic of discussion, such as politics, sports, or brand relations. Our goal is to analyze emotions within these data, for example to understand what emotions are expressed for each specific sub-topic or entity. The available state-of-the-art emotion recognition systems are able to recognize only a fixed set of emotions, while relying on more universal emotional expressions [MT13, KPJD13, SV04, NPI11a]. However, due to the specificity of the dataset, we require a new, application-specific set of emotion categories to distinguish. Another requirement is to have a better coverage of domain-specific expressions in order to achieve better recognition quality. Thus, in our motivation scenario, we need to build a new emotion recognition system capable of recognizing a new set of emotion categories within a given new domain of textual data.

We further narrow this motivation scenario and focus on one specific type of linguistic data—Twitter data. On Twitter, people share short status updates, called tweets, which are limited in length to 140 characters, making it essentially a short-text format. The great benefit of using Twitter data is the highly available human-generated text in large volumes on practically any topic of interest. This made Twitter the *de facto* most popular media for the recent computational social science research [PGS12, DCCH13b, TSSW10]. In this work, we consider

[2] By the estimate from GlobalWebIndex [Man15].

tweets a desired application domain for a novel emotion recognition system. The additional domain specificity comes from the topic of tweets. For example, we consider tweets with reactions to sporting events as a specific domain.

Another particularity of our motivation scenario is the set of emotion categories that would allow to recognize specific details on emotions within a chosen domain. We considered different emotion models and decided on the 20 emotion categories from the Geneva Emotion Wheel, GEW, version 2.0 [Sch05, SSS12].[3] The categories of the GEW were chosen with the goal of summarizing the different emotional experiences reported by participants in the psychological studies. They comprise 10 positive and 10 negative emotions, including both basic emotions, such as Happiness, Sadness, Anger, and Disgust, and more subtle emotions, such as Pride, Pity, Awe, and Contempt.

This is the main application of our work, where we advance fine-grained emotion recognition in short text. Based on these specifications, we formulate the problem of multi-label classification in tweets, with 20 GEW emotion categories chosen as potential answer labels. Our main driving research question is *how can we build a new fine-grained textual emotion recognition system that is able to recognize a specific set of emotion categories within a specific domain of data with a reasonable quality*?

1.2 Research Agenda

Our goal is to study how to build from scratch a novel fine-grained emotion recognition system tailored to a specific domain. This development process commonly goes through six steps, illustrated in Figure 1.1. First, we collect domain data for analysis (step 1) and select an emotion model that is characteristic and of interest for the studied data (step 2). Then, we annotate the subset of the data with the considered emotions (step 3) in order to obtain the input knowledge for our emotion recognition system. We separate two main approaches to annotating data: either to hire people to do it manually, or to annotate the data automatically based on some heuristics (resulting in so called pseudo-labeled data). The next step is to build the system itself based on those annotations (step 4). Afterwards, we evaluate how the built system works (step 5). At last, we investigate how to improve the built emotion recognition further to achieve better quality when applied to the studied data (step 6).

Different challenges arise during this process of building a new text-based emotion recognition system. We highlight the researched problems on the Figure 1.1. When the process employs manual annotations, *how can we collect an adequate amount of informative annotations? How can we ensure the quality of such manual annotations when they are performed online? How can we build emotion recognition systems without requiring manual annotations? How can we improve the quality of the built classifiers by treating the effects of different linguistic modifiers?* Our research aims to investigate and address these challenges. We detail each of them below.

[3] More details on this decision process can in found in section 3.2 in Common Material chapter.

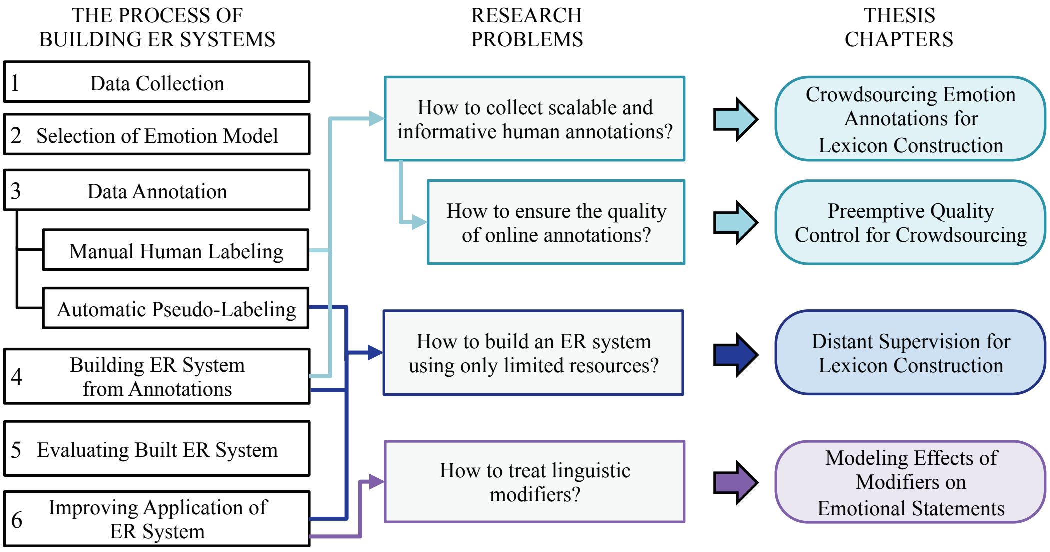

THE PROCESS OF
BUILDING ER SYSTEMS
RESEARCH
PROBLEMS
THESIS
CHAPTERS
1 Data Collection
2 Selection of Emotion Model
3 Data Annotation
Manual Human Labeling
Automatic Pseudo-Labeling
4 Building ER System from Annotations
5 Evaluating Built ER System
6 Improving Application of ER System
How to collect scalable and informative human annotations?
How to ensure the quality of online annotations?
How to build an ER system using only limited resources?
How to treat linguistic modifiers?
Crowdsourcing Emotion Annotations for Lexicon Construction
Preemptive Quality Control for Crowdsourcing
Distant Supervision for Lexicon Construction
Modeling Effects of Modifiers on Emotional Statements

Figure 1.1: An overview of the common process of building emotion recognition (ER) systems aligned with the studied research problems and the corresponding thesis chapters.

Challenge 1: Collecting scalable and informative manual annotations of emotions

Manually annotated data is an indispensable source of knowledge for building computer-based recognition systems. While manual annotation should produce more reliable results than pseudo-labeling, it has its own challenging aspects. When we ask annotators to assess which emotion another person expressed, their answers are subjective to their own appraisal process and past emotional experiences. Their answers also depend on their understanding of the assessed emotional statement and its context. Furthermore, the level of emotion differentiation can vary between people [Bar06]. Thus, every text document should be annotated by several people to account for their differing judgment. Moreover, the more documents are labeled, the more accurate and useful knowledge an emotion classifier will be able to extract. To achieve such scalable annotations, we suggest using paid crowdsourcing.

To construct domain-independent emotion lexicons, previous researchers employed crowdsourcing by requesting direct emotion annotations for a list of predefined dictionary words [MT13, WKB13]. However, such an approach is not effective for generating new domain-specific lexicons, because it would omit multi-word and domain-specific emotional expressions, as well as would involve labeling non-relevant words. In order to obtain more informative emotion annotations, we suggest asking workers to annotate indicators in text [AS07] and to generate additional descriptors for each class. Combining these two ideas, we design a crowdsourcing task for building a fine-grained domain-specific emotion lexicon, and address the following questions. How to formulate and present the task to online workers? How to select documents for annotation? How to aggregate the crowdsourced answers into an emotion lexicon? To what extent do annotators agree with each other when asked to label emotions at a fine-grained level?

While crowdsourcing provides an affordable solution to collect an adequate number of annotations, it also introduces additional challenges for controlling the quality of non-expert workers. We group them as the next sub-challenge 1.1.

Challenge 1.1: Ensuring the quality of the collected manual annotations

Using an online labor marketplace to perform crowdsourcing of data annotations requires a delicate approach. Running a task without unique correct answers, such as emotion annotation, may attract malicious (or lazy) workers, who do not put enough effort to provide good-quality answers. Moreover, workers might misunderstand the requirements of the task, especially in cases when the task requires more attention to separate the nuanced categories. How can we improve the quality of answers?

We suggest to study two quality control mechanisms aiming to preemptively ensure the quality of answers: tutorials and framing of financial incentives. While tutorials are imperative for ensuring that workers perform the task as expected, their use in the context of subjective tasks, such as emotion annotation, raises additional questions. To what extent does the inclusion of the tutorial affect the quality of emotion annotations? How can we validate that workers understand the task specifics, e.g. that we ask for the writer's emotions, not their reactions as readers? Researchers also proposed specifically formulated reward schemes to motivate workers provide good-quality answers [Har11, HSSV15]. However, such schemes usually imply a mathematical computation of reward bonus according to some function of answers' quality. How can we adapt such reward schemes to be employed with more ambiguous, subjective tasks? Which bonus formulation is more suitable for our task of emotion annotation?

Challenge 2: Constructing lexicons from limited resources

Manual annotations are crucial for building accurate recognition systems, but they are expensive and time-consuming to obtain. Even with crowdsourcing, we can afford to collect at most thousands of annotated documents. How can we build emotion recognition systems without manual annotations, by using other more limited sources of affective knowledge?

One solution is to obtain pseudo-annotated data automatically by applying some heuristics. This approach to building recognition systems is referred to as distant learning or supervision [GBH09, MBSJ09]. Previous researchers considered using emoticons and emotional hashtags as input heuristics [Moh12a, DCGC12, YLC07]. However, such restrictive heuristics might not provide adequate amount of pseudo-labeled documents when the input is limited to a within-domain dataset. Therefore, we suggest using more applicable heuristics that label text based on emotion lexicons, either of limited coverage or accuracy, or both. Such distant supervision approach can be considered as a full-fledged method for building an emotion recognition system, or alternatively as a way to further adapt a built lexicon for the specific application. We investigate how to apply the suggested distant supervision framework and answer the following questions. Which initial lexicons lead to better-quality systems? Which methods should be used for learning the emotion classifiers? How to find their parameters?

How to avoid bias towards specific emotions due to the skewed emotion distribution in the pseudo-annotated data? Which heuristics can be used for pseudo-annotation of neutral text documents? Is the inclusion of pseudo-neutral documents helpful for the learning process?

Challenge 3: Treating the effects of different linguistic modifiers

Another challenge of emotion recognition is that even the most explicit emotional terms can relate to another emotion when they occur in the scope of a modifier. For example, the word ‘happy’ refers to Happiness, but can express another emotion when negated, e.g. in the phrase ‘not happy’. The effects of different modifiers on emotions are either ignored or hand-coded only for the most impactful modifiers, such as negations and intensifiers [TBT$^+$11, PZ06]. However, when researchers employ a novel model of emotion categories, the effects of modifiers should be described with respect to that model. Yet, it might be time-consuming to manually derive the fine-grained rules of modifiers’ treatment, while ignoring the modifiers’ effects will damage the quality of recognition.

We suggest to automatically derive a computational model of the modifiers’ effects. To do so, we develop a data analysis method to quantify how different linguistic modifiers, such as negation or modality, change the emotion of emotional terms. Using that method, we investigate the answers to the following questions. How do different modifiers affect emotional statements and, more specifically, how do they change their emotion distribution? Do the emotions shift under a specific modifier and, if yes, towards which emotions? How does the modifiers’ presence change the confidence of an emotional statement? How can we use this information to treat the effects of modifiers in emotion recognition?

Our thesis follows this research agenda and addresses in turn each of the presented challenges. However, they form only a subpart of challenges that can arise in the process of building an emotion recognition system. The other problems that are left out of the scope of our research include among others semantic representation of the feature space, automatic extraction of an appropriate emotion model for the studied domain, unbiased within-domain data collection, and modeling inter-dependency between emotion categories.

1.3 Main Contributions

This thesis makes the following contributions to the field of emotion recognition in text.

Human computation task for crowdsourcing informative annotations and building a fine–grained emotion lexicon We design a human computation task for simultaneous annotation of a textual corpus with emotions and discovery of linguistic emotion indicators. We employ this task to build a domain-specific emotion lexicon suitable for fine-grained analysis of the tweets about the sports events. The built lexicon outperforms the domain-independent baseline. This work shows the potential of using crowdsourcing to build accurate emotion recognition systems. Our manual annotation process indicates that people can clearly separate the emotions at the desired fine-grained level in large proportion of the tweets, while the remaining tweets should be considered as expressing mixed emotional states.

Study of two preemptive quality control mechanisms to ensure the quality of crowdsourced annotations We investigate two mechanisms that can affect the quality and motivation of the crowdsourced workers in emotion annotation. We study the effects of including an obligatory tutorial that aims to ensure workers' task understanding, and show that its inclusion indeed leads to better-quality annotations. We also study different formulations of bonus incentives, aiming to incentivize workers to produce better-quality answers, and show the importance of careful qualitative bonus framing, at least when more difficult data are being labeled. Our research shows the value of properly managing the online crowd by incorporating appropriate preemptive quality control measures.

Framework of distant supervision for building fine-grained emotion recognition systems from limited input lexicons We develop a distant supervision framework that allows building fine-grained emotion recognition systems having only the emotion lexicons of limited quality and unlabeled within-domain data. Using that framework, we build several systems for the domain of tweets about sports events, starting from three different initial lexicons and employing different training methods. The majority of them outperform the baseline initial lexicons. We discover how the behavior of initial input lexicons and the choice of the supervised learner affects the quality of the built emotion classifiers. We also show the positive effects of rebalancing emotion categories within the pseudo-annotated data. And we reveal the importance of including pseudo-neutral tweets during the learning stage.

Computational modeling of the modifiers' effects on fine-grained emotional statements We design a data analysis method for deriving a computational model of the different modifiers' effects from the usage of modifiers and emotional expressions in Twitter. With this method, we study the effects of six detectable modifier types that affect the emotional terms in their scope: negation, intensification, modality, interrogation, past tense, and conditionality. Our analysis shows that the effects of all modifiers vary across emotion categories, and reveals that negation is not the only modifier type that can have a large impact on emotions. In addition, the extracted model specifies how each emotion can change under each modifier and suggests how the effects of different modifiers could be treated in classification.

Finally, based on the findings from our research, we generate general recommendations for building emotion recognition systems in text.

1.4 Thesis Structure

We continue this dissertation by presenting the background for our research in chapter 2. It discusses in more detail what are emotions in general and how they can be conceptualized for text-based emotion recognition. We will review different psychological and computational emotion models, and argue for the application-specific or study-specific models. The second part of the background chapter reviews the state of the art in text-based emotion recognition. We present the modeling of different aspects of recognition, including different emotion models, recognition models and methods. The same chapter also describes different potential applications of text-based emotion recognition systems, as well as links our problem to other related ones.

Chapter 3 “Common Material" formulates our classification problem, justifies the choice of the Geneva Emotion Wheel [Sch05] as our emotion model, enumerates the input affective linguistic resources, and presents the data used in this dissertation.

Chapter 4 “Crowdsourcing Emotion Annotations for Lexicon Construction" presents our work on using crowdsourcing for collecting manual emotion annotations. It describes important aspects of task design, presents how to aggregate annotations into an emotion lexicon, and evaluates the quality of the resultant lexicon.

Chapter 5 “Preemptive Quality Control for Crowdsourcing" investigates two different approaches to preemptive quality control in crowdsourcing online emotion annotations. The first part investigates the effects of including an obligatory tutorial in the given online task. The second part describes a crowdsourcing experiment studying the effects of different qualitative framing of conditions for obtaining an additional reward.

Chapter 6 “Distant Supervision for Lexicon Construction" presents a framework of distant supervision for building emotion recognition systems out from the emotion lexicons of limited quality (either because of limited coverage or accuracy). We describe our method, present the additional heuristics for detecting pseudo-neutral tweets, and investigate the quality of the built systems within the distant learning framework.

Chapter 7 “The Impact of Modifiers on Emotional Statements" describes our effort to automatically analyze the effects of six different linguistic modifiers on emotional statements. It presents the data analysis method for extracting and modeling such effects computationally based on comparison of the corresponding emotion distributions.

Finally, chapter 8 concludes this dissertation by reviewing its contributions, presenting our recommendations on building a new fine-grained textual emotion recognition system, and suggesting the perspectives of future work to further advance and apply text-based emotion recognition.

2 Background

2.1 Foundations of Emotion Modeling

2.1.1 A Concept of Emotion

Emotion is a widely studied psychological concept. Emotions can explain humans' behavior and motivation, and can affect our memories and thoughts. However, there is no consensus for their definition. Plutchik estimated that more than 90 definitions of emotions were suggested in the 20th century [Plu01]. We report two exemplary operational definitions of emotion, which were compiled based on findings and arguments from multiple researchers:

> "Emotions are episodes of coordinated changes in several components (including at least neurophysiological activation, motor expression, and subjective feeling but possibly also action tendencies and cognitive processes) in response to external or internal events of major significance to the organism." [Sch00]
>
> "Emotion is a complex chain of loosely connected events that begins with a stimulus and includes feelings, psychological changes, impulses to actions and specific, goal-directed behavior." [Plu01]

These definitions describe two generally agreed-upon properties of emotions.

First, emotions are activated as a response to some event important to an individual, i.e. relevant to his or her needs, goals, and concerns. Scherer separates such events into external and internal ones, where external events include behavior of other people, change of situation or a novel situation, and where internal events include thoughts, memories, or sensations [Sch00]. For example, we can feel happy when meeting an old friend and worried when thinking about the future interview.

Second, emotions are experienced as interrelated changes within the different organismic components, including physiological, subjective, behavioral, and cognitive ones [Sch00]. Physiological (or neurophysiological) component describes bodily symptoms. These symptoms include the response of autonomic nervous system with a change of the arousal level (which can be measured by heart rate, blood pressure, respiration, and skin conductivity) [NALF04] and neurological response in the brain (which can be measured by electroencephalography or more precisely by magnetic resonance imaging) [Pan98]. Another component are motor expressions, such as facial and vocal expressions [BDY+04]. For example, we frown when we are angry and smile when happy. Similarly, our voice can change when we experience an emotion: we can have a different speech rate, intensity, and tone [EAKK11]. Calvo and D'Mello [CD10] and Cowie et al. [CDCT+01] provide a comprehensive overview of different methods of emotion detection from neurophysiological and motor expressions. The subjective component consists of the corresponding subjective feeling or emotional experience, which can be reported by a person. The behavioral (or motivational) component presents the action tendencies induced by the emotion, for example, preparing the body to fight when Anger is experienced [Plu01]. Finally, the cognitive component presents the process of appraisal of a stimulus event. It refers to the cognitive evaluation of different properties of the event, such as relevance, pleasantness, or novelty [Sch01]. Researchers generally agree upon the emotion presence in the first three components (physiological, motor expressions, and subjective). However, there is less agreement on whether or not the last two components (behavioral and cognitive) are essential for the emotion experience [Sch00].

Emotions can be further characterized as having high intensity, low duration, rapid onset, high focus on event, and ability to affect person's behavior [Sch05]. Based on these characteristics, psychologists can differentiate full-blown emotions from other affective concepts, such as moods, interpersonal stances, sentiments (attitudes or preferences), and personality traits. While emotions have an episodic nature and can last from seconds to hours, all other mentioned affective states usually have prolonged duration: from days and weeks for moods to years and decades for personality traits [CDCT+01]. Additionally, moods and personality traits differ from emotions by having less focus on a specific event, person, or other stimuli [Sch05].

Regardless of relatively clear boundaries of the given emotion definition as a psychological concept, the same words can describe qualitatively different emotional states in English language. For example, the word *happy* can be used to describe an emotion ("This makes me happy"), a mood ("I feel happy these days"), and a personality trait ("I am a happy person"). Because of that, emotion recognition in text rather considers all possible affective terms as potential descriptors of a relevant emotional state. The emotion is then considered as expressions of an experienced affective state *in situ*, whether it is an actual emotion, mood, or even a personality trait.

2.1.2 Psychological Models of Emotion

Psychologists developed multiple emotion models to explain the nature of emotions and to characterize them. Surveys generally distinguish such models based on how they conceptualize emotions [Sch00, CD10, Iza13]. Instead of reviewing the full psychological argument on the emotions' nature, we focus here only on the main approaches to represent emotions. We distinguish two families of the emotion representation: dimensional and categorical models. This section describes their differences and enumerates the example models.

Dimensional models

Dimensional models imply the representation of the emotional experience within a certain continuous dimensional space. Such emotional spaces can be modeled by one to four dimensions. The simple one-dimensional models can involve two alternative dimensions: valence (also called pleasure or pleasantness) describes how positive or negative is the feeling and arousal (also called activation) describes how aroused-activated or calm-sleepy a person is. These two dimensions were combined into a model of "core affect" argued for by Russel [Rus03] and adapted by many scientists. Three-dimensional models include again the dimensions of valence and arousal, and add the dimension of dominance (also called control or power), which represents how capable the person is to change or cope with the experienced emotion [OMM75]. Pleasure-Arousal-Dominance model (PAD) [Meh96] is widely used in linguistic community [BL99]. More recently, a fourth dimension of unpredictability was added to better separate different reported affective states in the emotional space [FSRE07].

The dimensional representation is argued to be more psychologically primitive than the categorical one (described below) meaning that it can describe any subjective feeling, even across cultures [Rus03]. Also, the practitioners of dimensional models argue that any real-world or language concept can be represented as a point in this space to capture its connotative emotional meaning [OMM75]. Analogously, any emotional term, including the names of emotion categories, can be described as a point in the corresponding dimensional space.

Categorical models

The ability of humans to categorize their emotional experience in terms of specific emotion names or labels founds the basis for the categorical representation of emotions. While the categorical emotion models are mostly represented as the finite sets of emotion categories, they can also model the hierarchical or inter-linked componential structure of emotion categories.

Discrete models of basic emotions A large group of researchers assumes the existence of the small set of discrete basic or fundamental emotions, which are considered to be universal across humans. They usually enumerate a short list of 7 to 14 emotions that have distinct eliciting conditions, body expressions, and resultant action tendencies. Different principles of distinguishing such basic or fundamental emotions were suggested. For example, Plutchik [Plu80] based his classification on the distinctive patterns of behavioral action tendencies (e.g. flight behavior induced by Fear vs. fight behavior — by Anger). Ekman's six basic emotions of Happiness, Sadness, Fear, Disgust, Anger, and Surprise were motivated by Darwin's research on facial expressions of emotions, and are argued to have distinctive universal patterns of expression, antecedent, and behaviors [Ekm92]. The separation of fundamental emotions can be also based on the evolutionary development of a specific neural circuit for each emotion [Pan82]. The variety of ways to conceptualize basic emotions explains the lack of agreement on which emotions should form the basic set.

Furthermore, humans by nature can differentiate larger variety of emotional states than those studied as basic emotions. To explain the presence of other emotional states that humans can differentiate, the theorists of discrete basic emotions postulate a mechanism of emotion mixing or blending, where other non-basic emotions are generated as a certain mixture of the basic emotions. For example, Plutchik [Plu01] suggested to model Love as the mixture of Joy and Acceptance. Damasio [Dam94] suggested to distinguish between primary and secondary emotions. The emotions that scientists suggest to be basic and fundamental are rather more primitive, universal emotions called primary. Other more subtle emotions can be described as secondary and their generation is considered to involve cognitive processes.

Cognitive-based Models Psychologists design cognitive modeling of emotions. Such models describe the process of evaluation (or appraisal) of the emotion-eliciting event (or antecedent) and the derivation of the emotional state based on the appraisal patterns. Each emotion is assumed to be determined by the specific parameters of evaluation dimensions, such as novelty, desirability, or relevance of an event. For example, the OCC model [OCC88] forms the structure of the emotional concepts based on the evaluative analysis of their causes, while distinguishing events, persons, and objects as potential causes of emotions. Many psychological cognitive-based models correspond to componential models, which further specify the mechanisms of the appraisals within the organism [Sch00]. Such models can differ in terms of what emotional states they describe and what appraisal variables are used. For example, Lazarus argues for the limited number of major emotions produced via specific patterns of appraisals [Laz91]. At the same time, Scherer assumes that the multitude of emotions can be potentially experienced, each being elicited by a specific combination of appraisal variable outputs [Sch84]. However, he also argues for the limited number of commonly repeated patterns that generate the *modal emotions*, corresponding to more primary emotions, such as Fear, Disgust, or Joy.

Lexical and Self-Report Models The distinctive categories of emotions were also generated as clusters of affective terms that are used in similar situations and that have similar characteristics of usage in language. Such clusters can be referred to as semantic fields, and their characteristics — as semantic properties. Manual cluster analysis of linguistic emotion terms can reveal emotion taxonomies, or structural trees of emotion concepts [SSKO87]. The reduction of self-reported emotion words can also generate a categorical model of emotions. For instance, the 11 affective scales of PANAS-X were derived by the factor analysis of the self-reports on experience of 60 specific moods [WC99]. Scherer manually mapped the lexicon of affective words in several languages into the smaller number of emotion categories, resulting in the Geneva Affect Label Coder (GALC) [Sch05]. For English, GALC includes 36 emotion categories that are frequently used in self-reports and distinguished as separate states by psychologists.

2.1.3 Applied Models of Emotion

While psychologists focus on revealing the underlying nature and mechanisms of emotions (as well as their reproducibility and universality), researchers in computer-mediated emotion recognition address emotion modeling from more practical, applied perspective. Applied emotion models aim to describe the states and behaviors of artificial agents as modulated by emotions, to differentiate specific affective states of users expressed in text or during human-computer interactions, and to investigate the impact of particular relevant emotions within the specific domain of study.

Computational Emotion Models Various models were developed in order to allow computers to reason about emotions as computationally modeled concepts [MGP10, LSZ12]. Their internal structure mostly relies on the psychologically inspired cognitive representations of emotions, such as appraisal models of Frijda [Fri87], Lazarus [Laz91], and Scherer [Sch84] or cognitive OCC model [OCC88]. Such computational models provide the architecture for cognitive processing of emotions: they model evaluation process of experienced situations (“appraisal derivation”), derivation of the evoked emotions out of them (“affect derivation”), and regulation of actions and cognitive states as outcomes of the experienced emotion ("affect consequent"). These formalized system-structure descriptions are suitable for modeling the affective cognition of artificial intelligent agents, such as robots and virtual characters [Geb05, MG09]. However, at the current stage of their development, the derivation of the appraisal attributes values (or “appraisal derivation”) is limited to the prototypical manually coded situations.

Emotion Models for Text Analysis For the goal of text understanding and analysis, different psychological emotion models were applied. Many researchers adapted basic sets of emotion categories from Ekman [Ekm92] (6 emotions, including Happiness, Sadness, Disgust, Fear, Surprise, and Anger) [KPJD13, Moh12a] and Plutchik [Plu01] (8 emotions, including 6 Ekman’s ones plus Anticipation and Acceptance) [MT13, SI13]. Among other used models are 9 emotions of Izard [Iza71, NPI11a] and the 11 categories from PANAS-X [WC99, DCGC12]. The set of emotion categories to recognize can be also induced automatically from the collected within-domain data based on clustering principles [BB13]. This approach helps to increase the separability of emotions as well as to produce their taxonomy [LK06]. Alternatively, the researchers can employ the dimensional models, such as Valence-Arousal [Rus03] or Pleasantness-Arousal-Dominance models [Meh96], to investigate the expressed emotions [SDB$^+$15]. Researchers also adapted the four-dimensional model, e.g. the Hourglass of Emotions [CLH12], inspired by the Plutchik’s emotion structure. The dimensional representation allows extracting the emotion labels based on the assignments of emotion regions to specific emotion categories [KVC10].

The more fine-grained emotion models were adapted to capture more subtle emotions. The OCC model with 22 emotion categories [OCC88] was also employed to discern the differences in emotions in text [SPI09]. The Emotion Annotation and Representation Language (EARL) from HUMAINE project distinguishes 48 emotions, combined into 10 higher-level classes [HUM06]. The fine-grained models of emotions were also imposed by the provided tools to express own emotion on the web platforms. For example, researchers built emotion recognition systems aiming to separate the 132 mood labels (40 of which are frequently appearing) assigned to blog posts in LiveJournal [Mis05], 40 emoticons from Yahoo! Kimo Blog [YLC07], and 12 emotional reactions from the TED platform [PPB13]. With the growth of the available emotional data, differentiation of even a more fine-grained set of all affective terms was envisioned [MK15].

Emotion Models for Specific Application Domains In the area of emotion studies and affective applications, modeling of emotions depends on what aspects of emotional experience are relevant to the application or the studied domain. Researchers derived specific categorical sets of relevant emotions evoked by food, music, pictures, movies, and visual interfaces. For example, Desmet derived a set of 25 positive emotion categories evoked by consumer products, based on the analysis of survey answers [Des12]. He suggests that defining specific emotions, such as Amusement, Relaxation, and Inspiration, as explicit design goals can guide the designers to produce more successful products. To understand how emotions play role in learning, Confusion, Boredom, and Interest are included in the studied emotions [CGSG04, AR10]. In the studies of achievements and failures, e.g. in sport, the corresponding emotions of Pride and Shame can be of relevance. While the impact of Love and Tenderness can be crucial to study in the formation of romantic or friendship relationships, these emotions are less important in the studies of business, scientific, or technical communications. The research of professional communications does not study emotions per se, but related interpersonal stances, such as Aggressiveness or Dependence [VD11]. Personal assistants that monitor user's stress and attention (for example, to ensure safe driving experience [GYT14]) can focus on detecting the related emotions, such as Fear or Anger. Therefore, the set of categories to study is rather an application-specific choice.

Overall, multiple different emotion models were developed both by psychologists and researchers in affective computing. To unify the descriptions of different emotion models in the computer systems, researchers develop emotion representation languages [SRI16, SBB$^+$11, HUM06] and ontologies [Gra09, HCSM11, LGG$^+$08], which aim to represent different modalities of detection and detected emotions themselves with the fixed structures. This short survey of applied emotion models supports the relevance of our motivation scenario, where we assumed the need for developing an emotion recognition system that uses a new set of emotion categories suitable for the within-domain data. The choice of the model for a specific application rather depends on the goals of the application. In our case, the goal is to empower an analytical tool to discern specific emotional reactions, for which we adapt a fine-grained model of 20 emotion categories from the Geneva Emotion Wheel, version 2.0 [Sch05].[1]

2.2 State of the Art for Emotion Recognition in Text

Emotion recognition in text is an increasingly popular sub-topic in sentiment analysis, which aims to extract personal opinions, sentiments, and feelings expressed in text [PL08, Liu12]. While borrowing many methods from polarity and multi-category text classification problems, emotion recognition has evolved into a distinct field of research due to the multiplicity of ways to express and discern emotions in language. Recently, several papers surveyed this relatively new field [KLY$^+$09, BP12, Moh16]. We first summarize different aspects of emotion recognition in text that can differentiate recognition systems, and then we describe the prominent existing systems.

2.2.1 Different Aspects of Emotion Recognition Systems

Systems for emotion recognition in text vary along multiple aspects, for example, which domain of the data is in focus, whose emotions are being recognized and at which segment granularity level, how emotions are modeled, which recognition model is used and how it was obtained. We review the main options across this manifold variation. While describing these differing aspects, we classify our work accordingly.

Domain Emotions can appear in many domains, with emotion distribution and their expressions depending on the communication style (e.g. formal vs. informal), sharing channel (e.g. Twitter vs. emails), and topic of discussions (e.g. politics vs. personal events). Different source of data were analyzed, including social status updates (on Twitter [Moh12a, WCTS12] and Facebook [PPSP$^+$16]), blog posts (from LiveJournal [Mis05]), news [SG14] and news headlines [SM08], love and suicide letters [Moh12b, DH13], literature (fairy tales [ARS05]), emails [HLH11], and instant messages [NPI10b, KPJD13]. Different topics of discussions were in the focus of emotional studies, including politics [MZKM15], sports events [KSMP14], crisis events [CS14, BJJW14], and mental disorders (e.g. depression [DCCH13b]). Researchers also aimed at developing domain-independent (also called universal or general-purpose) systems, which represent and use general affective commonsense knowledge on textual expressions and indicators of emotional experiences. For example, universal affective lexicons are generally applicable across the domains [SV04, MT13]. As was described in the motivation scenario (section 1.1), this thesis aims at building domain-specific emotion recognition systems from scratch. We focus on the domain of tweets (essentially informal short documents), with further specificity on the topic of reactions towards sports events.

[1] We present more details and arguments on this choice in section 3.2.

Perspective The perspective of emotion recognition specifies whose emotions a system aims to recognize from text: writer's or reader's. Generally, the user-generated content, such as blog posts and status updates, is treated from the writer's perspective, because researchers are interested in what emotions the users express [Mis05, AS07, WCTS12]. In contrast, more reader-oriented media, such as news and literature, are analyzed from the reader's perspective, because researchers focus on the reactions that the content evokes [SM08, LYC08, SG14]. Interpersonal conversations, such as emails, chats, and online discussions, could be studied from both of these perspectives, but up to now the focus is on the writer's emotion [NPI10b]. In fact, a third perspective can be distinguished—to detect the emotions of referenced persons, such as Happiness of the third person in the phrase "He is happy". This perspective is less-studied and rather relates to detection of the writer's emotions attributed to a different emotion experiencer. Our work focuses uniquely on recognizing emotions from the writer's perspective.

Scope Different sizes of text segments for emotion recognition result in different scopes of recognition. Researchers recognized emotions of words, sentences, short documents, paragraphs and long documents, and even collection of documents. When the emotions of words are recognized, the goal is to build a reusable affective lexicon, describing word-emotion associations [PGH$^+$13, MT13]. Recognition of emotions in sentences [KPJD13, NPI11a] and short documents [WCTS12, Moh12a] allows for the precise modeling of each separate emotional statement. Longer scope (paragraph or long document levels) can summarize the entire emotional experience [Mis05, SG14]. Recognizing emotions in a collection of documents is useful for summarizing and modeling all reactions towards a specific event or product (e.g. detecting an emotional reaction towards a video based on the user's comments [PPB14]). Our goal is to recognize emotions in short text documents, such as tweets. This can help to analyze the expressed emotional reactions on a specific topic.

Language Most emotion recognition works, including this thesis, focus on English language because of the availability of many linguistic resources and corpora. Other frequently studied languages are Chinese [QR10, RQ12, ZDWX12] and Japanese [TIM08, MKSR11]. Multilingual systems were built for the problem of polarity classification [Den08, BT14], but left understudied for emotion recognition.

Emotion Model As reviewed in the previous section 2.1.2, emotions can be modeled in different ways. To recognize them in text, researchers adapted both dimensional and categorical models. Among the dimensional models, the most widely studied ones are unidimensional valence scale (which is reminiscent of polarity classification along with intensity), two-dimensional Valence-Arousal space [Rus03], and three-dimensional Pleasure-Arousal-Dominance model [Meh96]. There is less agreement on the choice of a categorical model to recognize. Researchers adapted various models with different emotion granularity (i.e. with different number of categories to recognize): from two (Happiness vs. Unhappiness) [WCL06] to 40 top mood labels from the LiveJournal website [Mis05]. The most used categorical models are the 6 basic emotions of Ekman [Ekm92] and 8 primary emotions of Plutchik [Plu01]. In our work, we adapt a more fine-grained model of 20 emotion categories from the Geneva Emotion Wheel, version 2.0 [Sch05] (see section 3.2 for our argument on this choice).

Recognition Model We differentiate the systems based on the type of recognition model they employ, that is how the affective knowledge is represented and how it is applied to derive the emotions from the text. Researchers identified different classes of emotion recognition models: lexicon-based, keyword-spotting, rule-based, knowledge-based, statistical approaches, machine-learning, learning-based, corpus-based, and lexical affinity methods. We find the boundaries between these classes to be either ambiguous or too generic. Thus, in our classification, we attempted to ensure the clear separation between the suggested categories. Based on our analysis of different systems, we distinguish five general classes of emotion recognition models:

- *Lexicon-based* These systems rely on representing the direct associations of linguistic terms (words or phrases) with emotions, stored essentially in the form of emotion or affective lexicons [SV04, MT13, BL99]. The text is classified either by spotting and aggregating the words from the lexicons [Ell92] (the *keyword-spotting* systems), or, for more accurate results, incorporating the rules of syntactic relations between words to derive the final emotion of a statement [NPI11a] (the *lexicon rule-based* systems).

- *Statistical Feature-based* Such systems perform emotion recognition using a trained machine-learning classifier, deriving the emotions of the text from multiple text-level features, e.g. the presence of *n*-grams or specific emotional cues [Moh12a, RRJ$^+$12, AS07].

- *Prototype-based* These systems represent each emotion category as one or multiple prototype objects and classify the text by finding closely matched prototype objects. Two sub-types of recognition models follow this description. One type is the model of *dimensional affinity*, where both documents and emotions are represented as vectors in multidimensional space and the vector similarity metrics are employed to find the closest emotions of the text [KVC10, DA08]. Another type is the models employing *databases of emotional experiences*, which use as prototypes the collections of labeled textual emotion descriptions (raw [TIM08] or in reduced representations [BHM12]).

- *Appraisal-based* These systems employ psychologically-inspired cognitive appraisal models to derive the emotion in text. They first detect from the text the values of appraisal component variables, and then employ a corresponding theoretical model to derive an emotion category [SPI09, UH15].

- *Hybrid or Other* Hybrid systems essentially combine together several recognition models, while a class of other emotion recognition models contains those models that did not fit to any of the presented types.

In the described classification, our work belongs to the class of lexicon-based models. The next section 2.2.2 presents in detail different existing systems corresponding to each of these models.

Construction Principle The principles of building emotion recognition systems differ depending on how the knowledge about emotions and their expressions was obtained. We distinguish the following five types of construction principles:

- *Manual Coding* In this approach, expert human raters manually build emotion recognition systems, e.g. by annotating words with the associated emotions [SDSO68, TP10]. The rules for application of the lexicons and derivation of emotion label are usually hand-coded as well.
- *Crowdsourcing or Human Computation* These methods build systems based on aggregation of answers from multiple people, not necessarily experts. It can be performed in the form of offline or online surveys [BL99, KPJD13], using paid marketplace platforms [MT13], or as games with a purpose [PS10].
- *Supervised* These methods extract recognition models automatically from the annotated data. Both statistical feature-based [RRJ$^+$12] and prototype-based [BHM12] recognition models can be built in this way. The input data can be annotated manually [AS07, ARS05] or collected from online resources with user-provided emotional labels [Mis05, SG14].
- *Semi-Supervised* These methods aim to build emotion recognition models using only a limited input knowledge about emotional expressions, e.g. a list of emotional seed words or a small subset of annotated data. One subcategory of semi-supervised methods is *semi-supervised lexicon extension* based on computed similarities between words. Another subcategory is *distant supervision* methods, which adapt supervised techniques but train classifiers on data that are generated automatically based on some heuristics.
- *Unsupervised* These methods first build a reduced data representation, e.g. vector space model or data clustering. The emotion recognition is then performed based on the emotion prototypes defined in the same representation [KVC10]. Because the affective knowledge is added at the later stage, we refer to these methods as *practically unsupervised.*

We separate the aspects of construction principle and recognition model because our survey of the existing systems reveals that the same recognition models can be built using different construction principles, while the same construction principle can result in different recognition models. Yet, there are some strong relations between them, for example statistical feature-based models are usually extracted via supervised techniques. Thus, we do not summarize separately each of these principles, but describe instead the specific methods of construction of different systems in the next section. Our work concerns two of the presented construction principles: crowdsourcing (chapter 4) and distant supervision (chapter 6). We describe the detailed related works on those topics in the corresponding chapters.

2.2.2 Different Systems for Textual Emotion Recognition

Emotions can be expressed in language in different ways, for example by emotionally charged expressions (such as *yay!* or *what a jerk*), or by explicit words (such as *that was funny*), or by describing emotion-eliciting situations (e.g. *I passed my exam*). From any of those emotional statements, humans arguably can infer the emotional state of the writer. In order to derive the emotion expressed in the text, a system of emotion recognition should model and incorporate the knowledge about associations of linguistic expressions with emotions. We separate five classes for different recognition models: lexicon-based, statistical feature-based, prototype-based, appraisal-based, and hybrid/other. Below we describe existing emotion recognition systems, separated into these five classes of recognition models.

Lexicon-Based Systems

Lexicon-based systems rely on representing the direct associations of linguistic terms (words or phrases) with emotions. Essentially they form emotion (or affective) lexicons, which list terms that bear emotions with their corresponding emotion association. The emotion lexicons can classify each linguistic term to one or several emotion categories, or associate it with the weight for each emotion.

Existing Affective Lexical Resources With the increase of attention towards emotion recognition, the number of available affective lexical resources grows. The first developed emotion (or affective) lexicons enumerate all terms directly expressing an emotion, such as “happy”, “angry”, “frustrated”, etc. An example of such lexicons for English language is the list of approximately 500 explicit affective terms studied by Ortony et al. [OCF87]. The explicit terms can be clustered into a smaller number of categories, based on the semantic clustering analysis [SSKO87] or based on manual categorization [Sch05]. For example, words “happy”, “joyous”, “elated” are all assigned to category *Happiness* in the GALC lexicon [Sch05]. Other emotion lexicons additionally contain terms indicative of an emotional experience, thus more indirectly expressing an emotion. Examples of indirect terms linked to *Happiness* are “approval” in WordNetAffect [SV04], “entertain” in NRC [MT13], and “visit friend” in EmoSenticNet [PGH$^+$13]. All of these lexicons associate linguistic terms to specific sets of chosen emotion categories. Other similar categorical lexical resources are available, including Synesketch [KPJD13], DepecheMood [SG14], AffectDatabase [NPI07]. The underlying emotion representation model differs from one emotion lexicon to another. For instance, Plutchik’s basic categories are used by NRC lexicon [MT13], Ekman’s categories – by WordNet-Affect [SV04], Synesketch [KPJD13], and EmoSenticNet [PGH$^+$13], and Izard’s – by the AffectDatabase [NPI07]. The most widely used lexicon with dimensional representation is ANEW (Affective Norms of English

Words), which contains connotative associations of words with pleasure, arousal, and dominance scores [BL99]. In the social-linguistic studies, investigation of the affective text content commonly applies the Linguistic Inquiry and Word Count (LIWC), which contains limited manually annotated words for some negative emotional categories, such as Sadness, Anxiety, and Anger, as well as positive and negative words [TP10].

The emotion lexicons are similar in nature to sentiment lexicons, which store terms' polarities for polarity classification and opinion mining, such as positive *good*, *great*, and *awesome*. Commonly used examples of polarity lexicons include GeneralInquirer (GI) [SDSO68], Bing Liu's lexicon [HL04], OpinionFinder [WWH09], and SentiStrength [TBP12]. For example, GeneralInquirer (GI) categorizes English words into multiple categories, including subjective ones, such as Positive, Negative, and Emotional [SDSO68].

The presented affective lexicons mostly contain universal knowledge about emotional expressions and emotional connotations of terms, and thus can be applied across different domains.

Keyword-Spotting The direct spotting of words from the emotion lexicons allows extracting intuitively the emotions of the text by counting the appearing words or aggregating their emotional weights. This idea was applied by multiple systems for emotion recognition [Ell92, FG13, SH01], with the minimal refinement to treat the effects of negations in some cases. It was also employed for summarization of stylistic features of the text, e.g. to analyze differences in emotion references across domains, times, or authors [Moh12b, DD10, CDH14]. The keyword-spotting techniques were applied in the context of multi-modal emotion recognition to capture the linguistic modality features [CW04]. Even though such simple approaches are intuitive and comprehensive, they neglect the structure of the sentences and the contextual meaning of words.

Rule-based Lexicon Application Rule-based algorithms go beyond simple keyword-spotting by taking into account syntactic structures and semantic composition of terms, e.g. by modeling the presence of negations, intensity modifiers, and conjunctions [KPJD13, MPI05]. We give a detailed review of different possible strategies for modeling the modifiers' effects in chapter 7, where we suggest how to extract a model of such effects automatically. More complex rules can describe how to derive an emotion of term compositions, e.g. of verb-noun phrases, from the emotion associations of individual terms [NPI11a]. The rule-based systems are also employed in the sentiment analysis research, where multiple lexicon-based systems adapt compositional rules to increase the application quality of the built lexicons and compute the final sentiment intensity [TBT^{+}11, TBP12, NPI11b].

Such rules for lexicon applications rely on knowledge about the grammar and semantic composition, as well as about other syntactic and semantic properties of the words. Thus, they are usually hand-coded by experts based on the exploration of linguistic data.

Lexicon Construction Methods As lexicon-based recognition models are popular among researchers, almost all construction principles were used to obtain them. Many sentiment and emotion lexicons, especially older ones, were built based on manual expert annotations. In such *manual coding*, the words are assigned to specific emotion or affective categories by a small number of linguistic experts. The previously mentioned lexicons GI [SDSO68] and LIWC [TP10] are example lexicons generated by manual coding.

More recently, *crowdsourcing* techniques allowed to scale the annotations to be performed by multiple non-expert human raters. Offline surveys for collecting annotations can be considered as a preliminary form of crowdsourcing. Researchers used it to directly annotate words with the associated emotions in order to have a word-emotion association lexicon. The NRC lexicon [MT13] was built using online crowdsourcing, where workers from Amazon Mechanical Turk were asked to rate the strength of association of a given word sense with each of 8 Plutchik's emotions. The extended version of the ANEW lexicon was also built using paid crowdsourcing [WKB13], while the original ANEW used an offline survey approach [BL99]. More details on the prior crowdsourcing techniques and human-encoded lexicons can be found in section 4.2, whereas the related approaches for ensuring the quality of the annotations in crowdsourcing are enumerated in section 5.2.

The emotion lexicons were also build using *semi-supervised lexicon extension* methods, where a small-sized input affective lexicon of the seed words is extended with new terms. The new words are classified into specific emotions based on their similarity to the given input terms, where similarity is computed either based on their semantic relations (e.g. synonymy) [SV04] or based on their co-occurrences within linguistic data, e.g. web *n*-grams [PIMK13]. For the latter, the approaches based on Pointwise Mutual Information (PMI) [TL03] are often adapted to estimate correlation between words. WordNet-Affect lexicon [SV04], Synesketch [KPJD13], and EmoSenticNet [PGC$^+$12, PGH$^+$13] were built using one of such semi-supervised techniques.

Supervised approaches to lexicon construction use some input annotated data. The weights of emotional terms are derived based on their statistical correlation with the emotion classes. One common way to derive such weights is to compute Pointwise Mutual Information (PMI) scores between terms and emotions [YLC07, Moh12a]. Another method, which was deployed to build DepecheMood [SG14], is to build term-emotion matrix out of term-document and document-emotion matrices. Furthermore, some machine-learning algorithms can also output lexicon-format recognition models, e.g. SVM in binary setting [YLC07], but we rather attribute them to the class of statistical feature-based methods.

Limitations While the above-mentioned lexicon-based methods can be applied to any textual data, due to their term-level nature they are unlikely to capture the full variety of emotional expressions used in language. Also, without rigorous modeling of context and semantic composition that would cover all potential expressions variations, such approaches are error-prone due to, for example, their failure to model different word senses (e.g. 'like' is emotional in the sentence “i really like this job”, while neutral in the sentence “do it like this”) or change in word order [KLY$^+$09]. As our system belongs to the class of lexicon-based models, it will be subject to these limitations too. This leaves room for more advanced recognition models.

Statistical Feature-Based Systems

Statistical feature-based systems perform emotion recognition using a trained machine-learning classifier, deriving the emotions of the text from multiple text-level features. Various linguistic, stylistic, and syntactic features are used in the context of emotion recognition. Those include not only features typical for text classification, such as *n*-grams, punctuation marks, length of text, part-of-speech (POS) tags, but also topics, syntactic dependencies, word clusters, concepts, and known affective terms. Such systems were developed for different domains, including web-logs [AS07, Mis05], fairy tales [ARS05], news headlines [SM08], and tweets [Moh12a, RRJ$^+$12].

Many statistical feature-based systems are built via *supervised methods* that derive the emotion recognition models from the annotated data. These data can come from manual annotations [RRJ$^+$12, AS07, ARS05] or from crawling the websites where users can provide emotion labels to text documents (e.g. LiveJournal, where users can self-label their blog posts with moods [Mis05], or Rappler, where users can react with emotional labels to a news article [SG14]). Alternatively, the data can be labeled or collected automatically based on some heuristics, following the *distant supervision* approach. Among the explored heuristics are the use of emoticons [PB12, YLC07] and emotional hashtags [DCGC12, Moh12a, WCTS12], as well as of the explicit patterns of emotional sharing (as in “i feel happy”) [KH11]. Using the data from a specific website or domain results in building within-domain emotion classifiers.

The process of building the statistical feature-based models is essentially the process of training a given classifier on the annotated dataset. Different classifiers, suitable for application for text classification tasks, were adapted for emotion recognition, including SVM [Moh12a], Multinomial Naïve Bayes [WCTS12], and Logistic Regression [DCGC12]. Their learning and application schema differ depending on how the classification problem is formulated. The frequent formulation for the categorical emotion models is one-vs.-rest classification, where a separate binary classifier is built for each emotion category and its goal is to identify positive samples for the corresponding emotion separately versus all other emotion classes [Moh12a, RRJ$^+$12]. Alternatively, some machine-learning classifiers support solving a multi-class classification problem with one class output directly (e.g. Logistic Regression).

Feature engineering process often defines the quality of a built model. Currently used features are mostly shallow linguistic features, such as *n*-grams, emoticons, hashtags, and punctuation marks. Adding word clusters, intensification patterns (e.g. word elongations), and negations can help cope with informal text and its structure [KZM14]. Adding aggregation statistics on terms from the known affective and sentiment lexicons as classification features was shown to improve the quality of recognition [WCTS12, Moh12a]. Domain-specific features, e.g. modeled topics of discussion, were also adapted [RRJ+12].

Instead of using such shallow features, it is desirable to generate features that are able to better capture the semantic representation of the text. The first step in this direction is done via extracting the concepts from the text [GCHP11, PGH+13]. In the area of sentiment analysis, learning word embeddings via deep learning was shown to achieve better quality of recognition because of better semantic space representation of the text [SPW+13, MSC+13]. The deep learning approach was also adapted for reader's perspective emotion classification of emotional statements from the Experience Project [SPH+11]. However, in contrast with using word-level features, the dimensions of word embeddings are less interpretable.

Limitations Using supervised techniques for training such feature-based classifiers requires substantial annotated data, which are expensive to obtain. Even when large annotated data are directly available, they come from a specific domain. It is not clear how generalizable such approaches are to be applicable across different domains. Finally, the models trained on one specific dataset are likely to be biased towards the emotions and expressions within that dataset.

Prototype-Based Systems

We distinguish a class of prototype-based systems, where the recognition model represents each emotion category as one or multiple prototype objects and classifies the text by finding closely matched prototype objects. We separate this class in order to underline the difference of such systems from other approaches. This group is further split into two subgroups: the methods based on dimensional affinity and the methods that use database of emotional experiences.

Dimensional Affinity This group combines systems that represent both documents and emotions as vectors in multidimensional space and classify documents based on their affinity (or similarity) to emotion prototypes in that space. The construction methods of such dimensional representation can be unsupervised: for example, they can be based on Latent Semantic Analysis (LSA) or other dimensionality reduction methods of the term-document matrix for the studied linguistic corpus [KVC10, SM08]. Prototype emotion vectors in this case are built along with the document representation by adding pseudo-documents containing known emotional words as prototypes. The text is then classified based on finding the closest emotion vector in the dimensional space, e.g. based on Cosine similarity between the extracted vector representations of the text and emotions in the same space. Such methods were deployed with coarse-grained emotion categories of up to six emotions. It is not clear whether they would work with fine-grained emotion models.

Using a Database of Emotional Experiences This group of systems combines those that represent emotion categories as collections of example (or prototype) emotional statements in the form of a database of emotional experiences and classify new text by finding the closest examples in that database. Such databases are generated either based on self-labeled descriptions of emotional expressions or via pseudo-labeling of the data. The example of the first is the ISEAR dataset (International Survey On Emotion Antecedents And Reactions) [SW94], which collects human descriptions of events that caused specific emotion. The example of the second is the repository of the emotional events extracted based on the grammatical relations of the sentence clauses with explicit emotion words [TIM08]. The database entries can be stored as raw text, with defined word-based cosine similarity between two text documents. In this case, the kNN-style classifier can be employed directly to derive the emotion of the text [TIM08]. Another approach is to use more structured representation of emotional events and situations in the database. For example, EmotiNet represents descriptions of emotional events as a sequence of atomic events in [subject, verb, object] structure, which are detected using available semantic parsing tools [BHM12]. Similarly, split of sentences into nouns, actions, and other transition operators was suggested to represent the emotion association rules for derivation [WCL06]. More recently, researchers suggested to store the emotional entries in the database in the semantically pruned form, where only words that convey emotional meaning are included [SEHHE14]. This direction of research is promising, especially when combined with more explicit emotional features. However, application of semantic parsers to short informal text from social media is likely to lead to errors due to poor grammar of this text genre.

Appraisal-Based Systems

Appraisal-based systems directly employ the theoretical rules of cognitive emotion generation based on the appraisal of a situation, interacted person or object. They automatically extract the corresponding appraisal variables from the text, and then apply the appraisal-based or cognitive theory to derive an emotion label. Examples of such systems are those that implement the OCC model [OCC88], by Shaikh et al. [SPI09] and Udochukwu and He [UH15]. The authors suggested to model different appraisal variables via automatic evaluation of polarity of related concepts, i.e. happening situation, interacted person or object, as well as via hand-coded rules. For example, the pleasantness of the events or attraction of the object an be estimated based on the polarity classification using sentiment lexicons. Whereas, identification of the events' attribution (e.g. self vs. others) can be coded based on analyzing the subject of the event. Such appraisal-based systems theoretically can achieve significant results, because of their grounding in emotion theory. Yet, the quality and applicability of such approaches are not studied enough. Also, using such appraisal-based approach restricts the choice of emotion categories to those that are modeled within the respective appraisal theory. Thus, its potential adaptation to new emotion categories requires further investigation. Finally, while the rules themselves are based on the emotion theories, the detection of the appraisal components (variables) is currently coded manually.

Hybrid and Other Approaches

Hybrid models rely on several sources of information and its representation for recognition. For example, Liu et al. [LLS03] develop a hybrid system that relies both on the lexical-level associations of words with emotions and polarities and on the prototypes of emotional situations (both extracted from the commonsense knowledge database). Lexicon-based approaches can be incorporated into the statistical feature-based systems by including aggregation of words from sentiment and emotion lexicons as features of classifiers. Such approach has shown promise in improving the quality of emotion classification [Moh12a, WCTS12].

We also assign to this category other approaches that did not fit to the described types of emotion recognition models. Kim et al. [KBO12] suggested to assign emotions based on the topics discovered by Latent Dirichlet Allocation (LDA). First, the topics would be extracted, and then automatically labeled based on the presence of seed keywords. The classification proceeds using the labeled topics. Another method of emotion classification was suggested by Wang and Pal [WP15]. They apply a constraint optimization approach, which allows to incorporate within the same classification framework the lexical features and topic similarity constraints, as well as other constraints on emotion co-existence and bias towards frequent terms.

2.3 Applications of Textual Emotion Recognition

Textual emotion recognition can support multiple affective applications. We enumerate here the main envisioned applications across different areas. We compiled this list by reviewing the stated motivation in the papers on emotion recognition [CDCT$^+$01, Pic95, Moh16] and by summarizing the papers that describe the affective applications themselves.

Applications Oriented towards End Users

Enhancing Human-Computer Interaction Automatic emotion recognition will help computerized conversational agents to become more empathetic and human-like by responding properly to users' emotions. Machines able to recognize the emotion of a current user could adapt their behavior and appearance accordingly [Pic95, CDCT$^+$01]. Research suggests that users perceive agents who react to their emotions as more intelligent and adequate [RN96, LLS03]. Interactive chat bots, robots, and personal assistants could benefit from this ability. Novel affective technology already marches into our lives (A Pepper robot is one example [Pep14]). Such conversational agents could play a role of active listeners for lonely people who need to share their daily experiences with somebody. They could also be personal psychological assistants that help people to cope with their everyday emotional overload [KMP99]. For example, they could show a funny picture in a right moment or provide an appropriate inspiring citation.

Enhancing Interpersonal Online Communications A computer can play a role of an emotional mediator making other people aware of someone's emotion. Better awareness of another person's emotions can enhance the social group experience [Che15]. For example, a computer could share the emotion of a user with another trusted user when the first user needs social help. Automatic visualization techniques can help to express and share the experienced emotion with other users. For example, EmoHeart system visualizes a detected user's emotional state by changing his or her avatar correspondingly [NPI10a]. Similarly, Synesketch enhances the instant messaging experience by artistically visualizing the shared emotions in the adjacent window of the chat [KPJD13]. Automatic emotion recognition could improve online support for companies by reacting quickly and appropriately to angry or frustrated customers.

Increasing Personal Self-Awareness Retrospective analysis of own emotions can help revive the past memories and re-evaluate own experiences [LSH$^+$06]. A machine equipped with a reliable emotion recognition mechanism is appropriate for saving a history of our emotions within an affective diary and presenting them later for analysis. Prototype systems in this direction were already designed. One example is the Muse system that allows users to visualize the emotional topics in their personal email archives [HLH11]. Another example is AffectAura—a multi-modal emotion tracking and visualization system that allows users to retrospectively analyze their emotional experiences [MKK$^+$12].

Helping Children to Develop Emotional Intelligence People with low emotional intelligence might lack understanding of the social interaction rules, including the emotional component of such interactions. Children with autism can exhibit such behavior. Automatic tools of emotion recognition could help them learn how to identify emotions themselves and help reacting to the expressed emotions [AR11]. Developing emotional intelligence can also help children to learn how to process and react to their own emotions.

Emotion-Aware Recommender Systems and Search Engines Knowing the emotional state of a user, a recommender system could adapt the proposed items correspondingly [TBK10]. For example, a movie recommender system could avoid suggesting a movie with immensely sad elements to a person tending to depression. Analogously, activity recommender systems could suggest some relaxing activity to a stressed or frustrated user. Knowing affective properties of operated items (e.g. movies, music, news articles, or blog posts) either based on their content or on users' comments can also help find and recommend more appropriate items [AMPI07, KNS$^+$11, SKCL09, PPB13]. For example, users could search for people with alike emotional experiences or for a news article or blog post that evoke a desired emotion.

Applications for Studying Aggregated Emotional Reactions

Studying and Modeling Interpersonal Communications Emotions play an important role in regulating our interpersonal communications and relationships. Automatic emotion recognition can help analyze and model the impact of an emotional constituent in our conversations. Researchers have studied the patterns of emotion transition in online conversations [KBO12, KPV+14], how emotional expressions can influence the sharing behavior in social media [PGS12], as well as how emotions impact collaboration, e.g. among the website contributors [ILC+14]. Other studies investigated whether the phenomenon of emotion contagion (or assortativity) is present in defining our social networks [BGRM11, CSK+14]. Emotion recognition can also help to characterize our social roles, e.g. by separating supporters from opponents [LYTL14], and identifying strong social ties, e.g. close friends.

Understanding Collective Emotional Experiences Quantifying the emotional experience across populations, e.g. for different location, gender, age, and other demographic parameters, can help characterize their differences. For example, it can help compare happiness index and well-being across different locations [SEK+13]. Visualization of emotions shared in geo-located social media can help understand the emotional flow of the cities [GAAES14, GLHG16]. Systems that visualize the time-based emotion flow can help capture the global emotional trends [MdR06, BMP11]. Such systems provide the field of digital humanities an unprecedented way to capture and reflect upon the ongoing history of the global emotions. Automatic emotion recognition can also empower psychology researchers to build and validate theoretical models of emotions by studying shared self-revealed emotional experiences [MGP10].

Affective Analysis of Multiple Linguistic Corpora Understanding which emotions are expressed in the text becomes a ubiquitous component of linguistic analysis. Researchers quantified affective content in news articles, books, movie plots, user-generated blog posts, and other large linguistic corpora [Kle11, Moh12b, Har13]. Patterns of used emotional expressions are also indicative of human personality [MK15], mental disorders (e.g. depression [DCCH13b]), and change in behavior (e.g. after child birth [DCCH13a]). As such, they are important features for mining those characteristics from the user-generated content.

Affective Analysis of Social Media Social media allow people to share their experiences, feelings, thoughts, and emotions. Besides the previously mentioned global understanding of the collective emotional experiences, analysis of the shared emotional reactions can help understand the collective experience and opinion regarding a specific topic, towards a person or a product, or during an event. For example, EmotionWatch was designed to visualize and summarize the emotional experiences evoked by specific global events [KSMP14]. Change in the emotion flow can help analyze the global events and trends [MdR06, BMP11]. Quantifying emotions expressed regarding products and brands quickly becomes an important indicator of the success or failure of the marketing strategies and products themselves [RQ12]. Researchers also employed detection of the shared emotions to predict the stock market value [BMZ11] and the outcome of political elections [TSSW10], as well as to discern public opinions on controversial topics [GMC+14].

2.4 Related Recognition Problems

Text-based emotion recognition is similar in nature to many other problems of text classification and analysis. We enumerate here the main related text recognition problems, as well as other formulations of the emotion recognition problem.

First of all, textual emotion recognition is a sub-task of sentiment analysis. This area of research concerns extraction of opinions, feelings, and sentiments expressed in text [PL08, Liu12]. The researchers in sentiment analysis mostly address the problem of polarity classification [PLV02, Tur02, GBH09], which classifies whether the text is positive, negative, or neutral. Another classical problem is subjectivity detection [WWB$^+$04, WR05], which classifies whether the text is subjective (opinionated) or objective (factual, neutral). Both of these problems share the recognition models and construction principles presented above for textual emotion recognition, especially lexicon-based and statistical feature-based models. In fact, emotion classification can follow the hierarchical recognition model, where the text is first classified as whether neutral or emotional, second, as whether positive or negative for emotional texts, and then only categorized into specific emotion categories [GIS10].

The sentiment analysis field also formulates multiple other problems of analyzing subjective text. One of them is sentiment-aspect mining [TM08, JO11, BE10, PPB14], which aims at detecting the sentiment expressed in the text towards specific aspects of a described product, particularly in the context of online product reviews. Such methods model both the sentiment categories and product aspects (e.g. cleanliness of the rooms for hotels, or quality of battery for the electronic devices). Following the same principle, the problem of emotion cause detection was formulated for emotion recognition [CLLH10, RCR$^+$11, NA13], where the goal is to identify a trigger that evoked the expressed emotion.

Another related problem is detecting an opinion holder [KH05, KJM07], i.e. a person whose sentiment is expressed in the text. For the area of emotion recognition, this problem is referred to as experiencer detection, i.e. detecting a person who experienced a stated emotion. This problem is especially relevant for the literature analysis, where the emotions are frequently referenced from the third-person perspective rather than stated in first person. For the domain of tweets about politics, Mohammad et al. [MZKM15] showed that the emotion experiencer is almost always the author of the tweet (however, this finding might be specific to the domain of political discussions).

The textual multi-category emotion recognition in nature is essentially a text classification task. As such, it is related to the classical problem of topic categorization of documents [Joa98, NMTM00], where the goal is to categorize textual documents, such as news articles or web pages, into the specific topics, such as sports, politics, or art. Regardless the high granularity of classes such topic classification methods can achieve high accuracy, as the topics can be distinguished at the word-level due to a large number of topic-specific words. Emotion recognition is a more challenging problem because even explicit emotional words can frequently appear across multiple emotion categories due to the different in their contextual usage. Thus, more advanced text understanding models are required for its successful solution.

Emotion recognition can also help to identify multiple characteristics of text and its author. For example, the use of emotional expressions is predictive of the personality, gender, and age of a writer, and thus can help to solve the related classification problems. For example, extracting features using emotion lexicons has shown promise for predicting reviews' helpfulness [MP14], personality detection [PGC$^+$14, MK15], and polarity classification [BMMP13, CdAP13].

Finally, emotion recognition from text is conceptually related to other emotion recognition problems from alternative modalities. Those include emotion recognition from facial expressions [FL03], gestures and body postures [Wal98], behavior patterns in computer interaction (e.g. typing and mouse movements) [Koł13], physiological changes (e.g. galvanic skin response or heart rate) [NALF04], brain activity [Ado02], and voice [KR12]. The extraction of the evoked emotions was also addressed for music [KSM$^+$10], images [MFL$^+$05], and videos [SPP12, HX05].

3 Common Material

3.1 Problem Statement

The objective of writer-side emotion recognition is to detect which emotions are expressed in a given text document. As Twitter data are considered as the main domain of application in this thesis, this problem can be translated into detecting emotions of the author of a tweet based on its text. In this thesis, we model "emotions" as belonging to a finite number of categories and formulate the problem of emotion recognition as a multi-label classification task. Given the set of emotion categories $E = \left\{e_1, e_2, ..., e_{|E|}\right\}$, the system detects which emotion categories are expressed in a given document d—in our case, a tweet— and produces a label set $Y_d = \left\{e_{i_k}\right\} \subseteq E$ containing the detected emotions. In order to separate neutral documents from emotional ones, we use the extended set of categories $E^0 = \{e_0\} \cup E$, where e_0 represents the *Neutral* or *No emotion* label. If e_0 is within the multi-label output Y_d (i.e. $e_0 \in Y_d$) or if no emotion is present (i.e. $Y_d = \varnothing$), we consider that a *Neutral* category e_0 is assigned alone (forcing $Y_d = \{e_0\}$).

This formulation implies using the categorical representation of emotions where an emotional state is described in terms of a discrete set of emotion names. Compared to dimensional models where emotional states are described as points in space (a more detailed description of this alternative can be found in section 2.1.2), categorical modeling has an advantage of allowing for a more fine-grained analysis. For example, it allows us to separate nuanced emotional states differing insignificantly within the dimensional space, such as the widespread Pleasure-Arousal-Dominance or PAD model [Meh96]. It is also more natural for humans, because in daily life we use emotion names to describe specific feelings rather than give numerical evaluations or specify polarity. So far, the multi-item emotion classification problem has received much less attention than polarity or valence classification.

Alternatively, we could also formulate this problem as a multi-class classification task, where only one category is assigned per document (tweet). However, previous research showed that even for the small number of basic emotions the text can often express several emotions [AS07]. Psychologists also pointed out that the level of emotion differentiation differs across people even from the same cultural background [Bar06], and thus people can describe the same emotional experience with different emotion labels, choosing between more general and specific categories. With the more fine-grained emotion categorization, the chances for the presence of emotion mixture, confusion, or overlap are higher. Therefore, multi-label classification corresponds better to the specifics of emotional experiences.

Another option would be to return as an output of the system an intensity or weight for each studied emotion, as in soft or fuzzy classification problems. Strapparava and Mihalcea [SM08] suggested this formulation for the reader-perspective emotion recognition in news headlines with six basic emotions. However, the more emotion categories are under investigation, the more effort is required to extract reliable data annotation in this format. Thus, even though such format can describe more precisely the expressed emotions, its direct evaluation is more complicated.

Yet, we suggest to use this format for the intermediate, more precise weighted representation of emotions expressed in text. To model the presence of the given emotions from E^0 within a text segment, we define the *emotionality* – emotion distribution represented as a tuple in the probability space

$$\mathbb{P} = \left\{ \vec{p} = (p_0, \ldots, p_{|E|}),\ \sum_{i=0}^{|E|} p_i = 1,\ \forall i\ p_i \geq 0 \right\} \tag{3.1}$$

where p_i represents the percentage of ith emotion in felt emotion mixture, or, in other words, the weight of ith emotion. That is we will estimate for each text segment (whether it is an entire tweet or a lexical term) what emotions it expresses by providing a weight for each emotion. Note that in this formulation of emotionality we assume that the emotional states are independent but can be present together. Also, the total weight of the emotional presence can be computed as $1 - p_0$ (corresponding to the sum of weights for all emotions, excluding neutral state). Using this intermediate emotionality representation, we will detect the emotions of a tweet in the multi-label format as the emotion categories having the highest weights within a detected emotionality of the tweet.

3.2 Fine-Grained Emotion Model (GEW)

In this thesis, we use the emotion categories from the Geneva Emotion Wheel (GEW, version 2.0) [Sch05, SSS12] as the basis for recognition. The GEW was developed as a tool for obtaining self-reports of emotional experience with a goal to structure the exhaustive list of possible emotion names used in free-format self-reports with minimal loss in expressibility. The visual representation of the GEW is shown in Figure 3.1. It presents 20 (10 positive/10 negative) emotion categories that frequently appear in free-format self-reports and are studied as categories of interest in psychological research. Each emotion category is represented by two common emotion names to emphasize its family nature (e.g. *Happiness/Joy*). Throughout the thesis, the first names will be used for a shorter reference. These categories are arranged

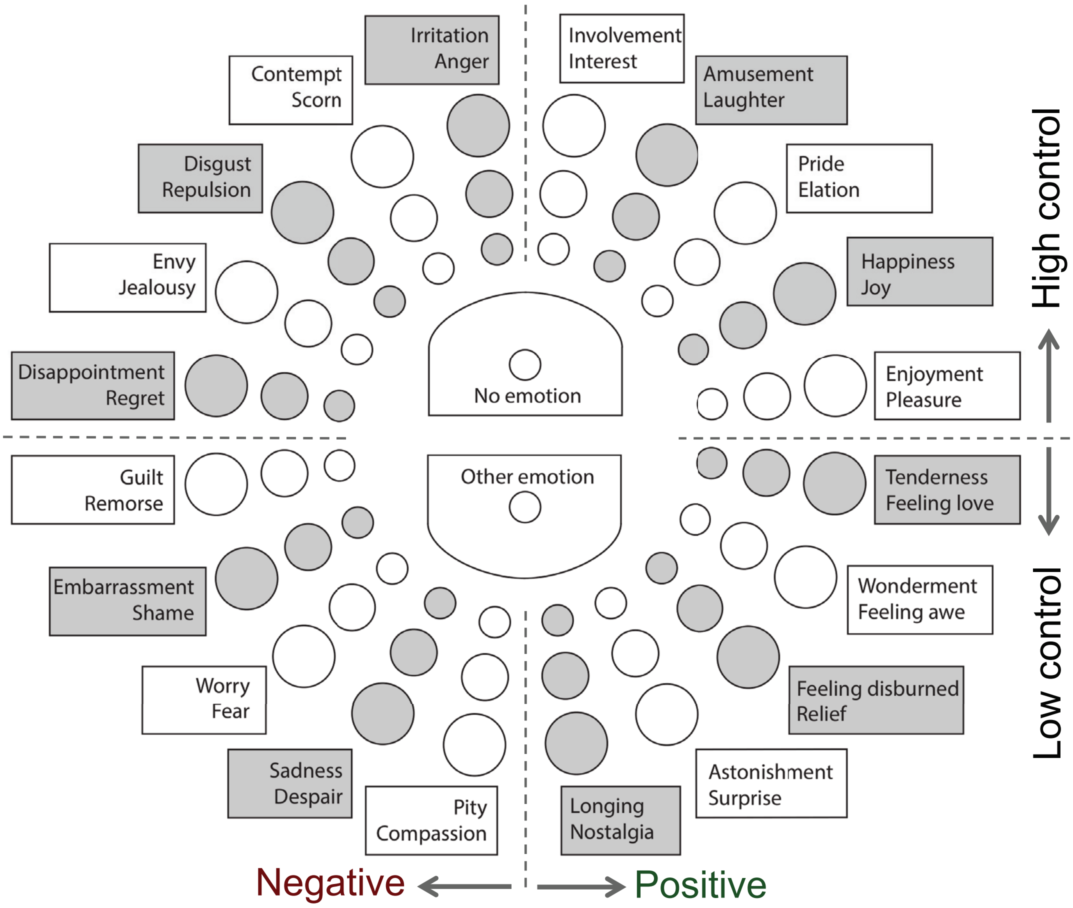


Figure 3.1: The Geneva Emotion Wheel, version 2.0, with 20 emotions placed the valence-control space, and two additional categories of response for No Emotion and Other Emotion.

on a circle following the underlying two-dimensional space of valence (positive-negative) and control (high-low). Several levels of intensity for each emotion category are presented as answer options. Also, two alternative answers are possible: *No emotion* and *Other emotion* with free-format input in the latter case. For the classification purposes, we use the 20 main emotions and *No emotion* category only, ignoring *Other emotion* as it can aggregate multiple ambiguous categories.

The GEW has multiple advantages. Whereas common sets of basic emotions, such as from Ekman [Ekm92] or Plutchik [Plu01], contain up to 8 categories, the GEW's 20 categories provide a more accurate approximation of the full range of emotions that humans are capable of experiencing. The most commonly distinguished emotions are included in GEW, for example, Happiness, Fear, Anger, Sadness, Surprise, Disgust, and Love. However, humans feel not only these strong emotions, but also other, more subtle emotions. Such a fine-grained model allows us to discover more insightful details about emotional reactions. The GEW is advantageous to other models of fine-grained emotions. For example, to the OCC model [OCC88] contains 22 categories differentiated based on cognitive attribution of factors evoking emotion, such as how desirable or relevant for the person are the consequences of an event. Compared to this model, we believe that the GEW emotions are more likely to be distinguished correctly based solely on lexical terms (e.g. it can be difficult to distinguish *Gratification* from *Satisfaction* without proper context modeling). Another alternative were the 24 primary emotions from the Plutchik's circumplex model [Plu01], adapted later into the Hourglass of Emotions [CLH12], distinguishing 4 affective dimensions and specifying 6 levels of emotion in each. However, that model lacks cognitive-based emotions such as *Pride*, *Envy*, or *Pity*, thus precluding the analysis of potentially different context of their experience. Even more fine-grained models could be adapted, e.g. all 36 categories suggested by Scherer as summarizing personal experience [Sch05] or 48 categories from HUMAINE Emotion Annotation and Representation Language (EARL) designed to describe the full variety of emotional experience [SPL06]. Using them could be insightful as well, but requires stronger discriminating power of the classifiers, as well as leads to higher cognitive load for separating these emotions during the annotation. An advantage of GEW as a well-designed emotion elicitation tool is in its structured way to allocate the emotions in the wheel structure for the annotation. This helps overcome the possible difficulties in distinguishing the fine-grained emotions. The desired, but limited emotion granularity also reduces the cognitive load of analyzing the output of emotion recognition, e.g. by avoiding a visual clutter within emotion summarization.

Another question one might want to ask is why do we aim at the fine-grained emotion analysis at all? We believe that only with the fine-grained emotion recognition can we draw later the insightful details for making decisions out of human behavior. Different emotions imply different appraisal values and different elicitation situations, as suggested by appraisal emotion theories (enumerated in the background section 2.1.2). Thus, by separating them, we can extract more information on those different experiences and make the correspondingly adjusted decisions. For example, Desmet [Des12] argues for a fine-grained, but limited set of positive emotions, which can help product designers to focus on the specific goals for evoking emotions. Another argument for the fine-grained models is the fine granularity of computational emotion models designed to model the behavior of intelligent agents. The preference to use more categories for defining the behavior of such agents suggests that more categories would also be needed to understand and distinguish the behavior of humans. Finally, with the more fine-grained model, we can identify the emotions specific to the studied novel domain among the full list of considered emotions.

3.3 Used Affective Lexical Resources

Lexicon of Explicit Emotional Terms (GALC) A list of explicit emotional terms is associated with the chosen GEW model—the Geneva Affect Label Coder (GALC) [Sch05]. It is a domain-independent affective lexicon that enumerates for each emotion category the stemmed words that explicitly express the corresponding emotion, for example, *happ*∗ for *Happiness*. Some examples of the terms are given in Table 3.1. It was developed along with the GEW, for automatically classifying free-format survey responses into given emotion categories.

Interest/Enthusiasm	absor*, alert, animat*, ardor*, attenti*, curi*, eager*, enrapt*, engross*, enthusias*, ferv*, interes*, zeal*
Happiness	cheer*, bliss*, delect*, delight*, enchant*, enjoy*, felicit*, happ*, merr*
Joy	ecstat*, elat*, euphor*, exalt*, exhilar*, exult*, flush*, glee*, joy*, jubil*, overjoyed, ravish*, rejoic*
Surprise	amaze*, astonish*, dumbfound*, startl*, stunn*, surpris*, aback, thunderstruck, wonder*
Sadness	chagrin*, deject*, dole*, gloom*, glum*, grie*, hopeles*, melancho*, mourn*, sad*, sorrow*, tear*, weep*
Fear	afraid*, aghast*, alarm*, dread*, fear*, fright*, horr*, panic*, scare*, terror*
Disappointment	comedown, disappoint*, discontent*, disenchant*, disgruntl*, disillusion*, frustrat*, jilt*, letdown, resign*, sour*, thwart*
Disgust	abhor*, avers*, detest*, disgust*, dislik*, disrelish, distast*, loath*, nause*, queas*, repugn*, repuls*, revolt*, sicken*
Anger	anger, angr*, cross*, enrag*, furious, fury, incens*, infuriat*, irate, ire*, mad*, rag*, resent*, temper, wrath*, wrought*

Table 3.1: Excerpt of explicit emotional terms from the GALC affective lexicon.

GALC contains 279 stemmed terms for 36 emotion categories, covering the variety of emotion categories extracted from self-reports on emotional experience. From these, we use 212 stems associated with 20 GEW, 2.0 categories, leaving us with 10.9 in average per each emotion category. However, we discovered that using stems with a wild token $*$ at the end is undesirable, as sometimes non-related terms would be also matched. For instance, one of the most frequently mapped instances of the GALC stemmed term *happ*$*$ (*Happiness* emotion) is *happy*, which is the correct association, but the instance *happen* is also frequent while it does not correspond to this emotion category. Other common mismatched examples include *made* for *mad*$*$ and *please* for *pleas*$*$. Thus, we instantiate the stemmed words into actual linguistic tokens by matching the GALC stems in the dataset of 15 millions of sampled general tweets.[1,2] Then, we manually discovered correctly matched emotional terms among the most frequent instances. Based on this investigation, we also moved the terms matched to *regret*$*$ into *Regret* category instead of the original *Nostalgia*. The new revised lexicon GALC-R is composed of 1026 terms, 52.9 in average per emotion category. Note that some terms among them correspond to two emotion categories: 18 terms to Pleasure and Happiness, and 13 terms to Awe and Surprise. The full list of the terms can be found in the appendix, section A.1.

Emotional Hashtags We specify the list of 167 emotional hashtags assigned to the 20 GEW categories based on the previously introduced GALC lexicon [Sch05]. This list of hashtags was aimed for crawling the tweets using Twitter API, and thus it could include only a limited

[1] These tweets were collected using Twitter Sample API, between 1st November and 10th December of 2014, without any keyword filtering, restricted to English. We used only non-retweets and non-replies.

[2] Alternatively, we could instantiate the GALC stems using a word dictionary, such as WordNet. Yet, instantiating the stems based on their mapped entries from the actual tweets allows us to capture additional word variations used in informal communications. For example, we could associate the stem *happ*$*$ not only to tokens *happy* and *happiness*, but also to *happygirl*, *happyy*, and *happi*.

number of instantiated terms. Also, we aimed at direct explicit emotional hashtags, to avoid disputable associations of emotions. To extract such a list, we matched the hashtags from one million of collected tweets against GALC terms and discovered the top of frequent GALC-based hashtags. Then, we manually excluded the confusing hashtags that could be used to describe non-emotional concepts (e.g. the word *glee* is also the name of popular TV series) and non-frequent hashtags. This resulted in the list of 167 manually chosen hashtags. It includes such hashtags as *#happy*, *#elated*, *#proud*, *#love*, *#loveit*, etc. The full list is enumerated in the appendix, section A.2. These emotional hashtags will be used to collect and detect tweets reliably expressing a specific emotion, as required for building, refining, and evaluating emotion recognition systems.

3.4 Used Datasets of Tweets

Twitter is used in this thesis as a source of data, as it contains a large amount of emotion-bearing tweets that are easy to collect, and because tweets are short enough to assume that appeared emotional statements can summarize the whole emotional reaction of a tweet's author.

General Tweets Labeled by Emotional Hashtags In order to obtain the dataset of tweets labeled with the chosen emotion categories, we follow the distant learning idea of using the emotional hashtags appearing at the end of the text as a self-reported emotion label for the tweet [WCTS12, DCGC12, Moh12a]. The constraint on the position of a hashtag helps ensure that the hashtag is used as a label of the tweet and not as a part of the content text. This automatic data collection eliminated the requirement of using subjective and time-consuming manual annotation, as well as allowed collecting a dataset of significant size for the subsequent analysis. Concerning the quality, we rely on the previous evaluations of similar hashtag-based labeling, which showed that the emotion of the hashtag correctly corresponded to the tweet content in 83% of tweets for a large set of emotional and mood-descriptive hashtags [DCCG12].

17.6 millions of English tweets with the emotional hashtags were collected via Twitter Streaming API between 27th February and 26th May of 2014. Among them, we extracted $1,729,980$ tweets that had those hashtags at the end of the text, were not repeated, were no retweets, did not contain URLs, and were assigned to only one emotion category. We refer to this dataset as EMHASH_DATASET. The emotion category associated with an emotional hashtag is considered to be an emotion label for the full tweet text. Figure 3.2 presents the distribution of the associated categories. We can observe that Love is the dominant category present in 27% of tweets, followed by Happiness (18%), Sadness (12%), and Anger (11%). Note that we do not have in this dataset any neutral tweets. Some example tweets are "My group of friends are such good-hearted intelligent kids #lovethem", "Yay for new tires!!! (: #happytweet" and "Last party , last in college , last day with friends ,#sad".

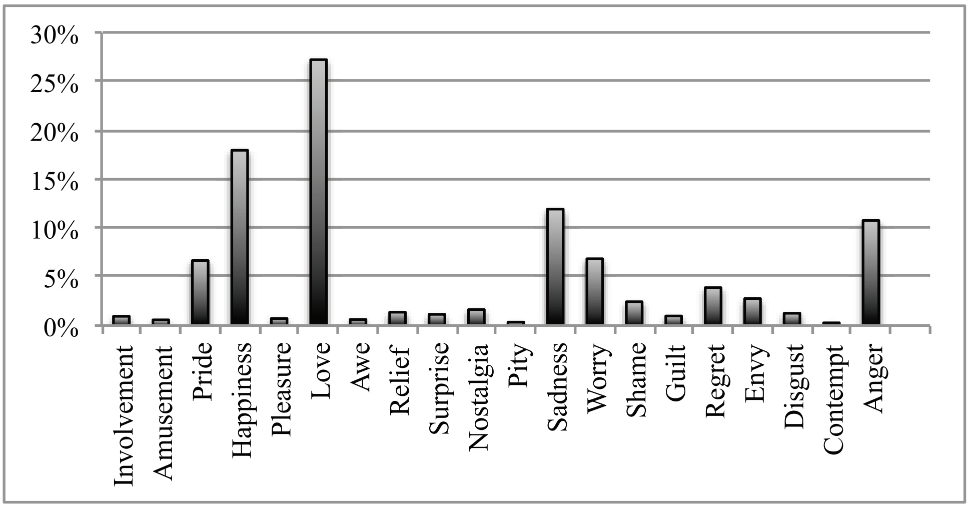


Figure 3.2: Distribution of emotion categories in the collected dataset of general tweets with emotional hashtags. The category is assigned to a tweet based on the category of the hashtag.

Tweets about the Olympic Games In the motivation scenario (section 1.1), we suggested to investigate how to build an emotion recognition system for the new set of emotion categories and for the new domain of application. Our particular application domain that we consider are tweets with reactions towards sports events. This domain was chosen because it contains various emotions with domain-specific emotional expressions, thus requiring the development of a tailored emotion recognition system.

We focus on analyzing the emotions of the spectators of the Olympic games. Our data consist of 33.2 million English Twitter posts collected over two weeks during the 2012 Olympic Games by querying Olympic-related keywords using Twitter Streaming API. The list of keywords was the following: “london2012", “olympics", “olympic", "olympicgames", "OG2012", included with and without a hashtag sign ‘#’ in front. Table 3.2 enumerates example tweets for each GEW emotion category, as labeled by the human annotators (the details of this annotation process can be found in section 6.5.1).[3] We refer to this dataset as OLYMP_DATASET.

[3] We preserve the exact writers’ punctuation and style in the provided tweet examples, except for replacing the appeared usernames. The author of this thesis does not share any opinion or emotion expressed in the given examples.

Emotion	Example tweet
Involvement	Always wondering what <username>listens to before his races... #GoTeamUSA
Amusement	i watch the olympics thats enough sport for meeeeeee
Pride	YEEESSS! Holly Bleasdale gets it right on 3rd attempt at first height.
Happiness	just managed to get athletics tickets
Pleasure	Wine olympics goodcompany =goodtimes
Love	Nathan Adrian, i wish you knew my existence. Congrats on the gold also
Awe	It is insane how smoothly and quickly these track women can run..
Relief	Thank goodness all 204 flags are coming in at once !!! #closingceremony
Surprise	wow pll over the Olympics all night.
Nostalgia	Wishing I would have stayed in gymnastics. . .
Pity	Unlucky Tom Stalker! fault he nicked it! Judges are terrible at this Olympics..
Sadness	Please don't tell me the Olympics is ACTUALLY ending.
Worry	Olympics on but I have no tv. What's happening!!!???
Shame	Too much cheating going on in this Olympics. Those asian badminton players, Cameron Van der Burgh.
Guilt	Watchin olympic instead of doin project
Regret	Work be crampin my style with these Olympic games! I always miss the good stuff!
Envy	You live the dream life Jack snap out of it! #sojealous
Disgust	Olympic girls look like men...
Contempt	Easily amused British public. Makes me sick. Crack on Olympics, let's get it over with.
Anger	Why is water polo always on when I want to watch the Olympics?
Involvement + Pleasure	Olympic tickets arrived. Absolute win.
Amusement + Pleasure	Love the hurdles! Love seeing them falling over hahahaha.
Happiness + Love + Surprise	<username> OMFG. I think Kate Bush is going to be at the Olympic closing ceremony! #Babooshka
Involvement + Worry	Every time I watch the Olympics it gives me chills.
Involvement + Regret	Its too damn hard to do anything productive when the Olympics are on
Nostalgia + Sadness	Can't believe the Olympics is over tomorrow already
Shame + Contempt	If you have to choose between a house and food, or gymnastics, then your priorities are wrong if you choose gymnastics.
Regret + Anger	Aish!!! AGD is cancelled this week due to london olympic...
Disgust + Anger	Omg olympics are so long and boring #saidnobody

Table 3.2: Examples of manually annotated Olympic-related tweets.

4 Crowdsourcing Emotion Annotations for Lexicon Construction

4.1 Introduction[1]

Our goal is to build an emotion recognition system for analyzing the fine-grained emotions within short text documents, such as tweets. One way to achieve it is to allow a computer to learn from human commonsense knowledge. The data collections manually annotated with emotions are a valuable source of extracting such knowledge based on which emotion recognition system can make its decisions. Traditional approaches to label textual data involve offline labeling by expert raters, which is an expensive and time-consuming process, mainly due to the valuable, but limited time resource of experts. Crowdsourcing presents an alternative annotation method, which can scale easily the data annotation process, but requires increased attention to the task design and quality control, due to the less trustworthy and less experienced annotators.

With the multi-class fine-grained classification, an additional annotation challenge arises: if emotion representation is not carefully designed, the annotator agreement can be very low. The higher the number of considered emotions is, the more difficult it is for humans to agree on a label for a given text. Low quality of labeling can lead to difficulties in extracting powerful classification features. This problem is further compounded in parsimonious environments, like Twitter, where the short text leads to smaller number of present emotional cues. Moreover, with tweets having problems with grammar and being short, their sense is not always clear to annotators without additional context, which may cause additional problems in annotations. All this presents challenges in collecting high-quality training corpora for building emotion recognition systems operating with a fine-grained emotion category set in short text. The question remains open: How to reliably annotate data with fine-grained emotions, while obtaining maximum useful information for building emotion classification system?

In this chapter, we show how to tackle the above challenges while designing a novel human computation task for emotion annotation, which we run using an online labor market, the

[1] The presented annotation task was developed in collaboration with Claudiu Musat.

Amazon Mechanical Turk or AMT (www.mturk.com). We directly employ the Geneva Emotion Wheel (GEW) [Sch05]—a well-designed emotion assessment tool introduced in the previous chapter,— which helps us to overcome the possible difficulties in annotation. In a given task, we show to the annotators the tweets and ask them to classify the tweets' emotional content into one of the 20 provided emotion categories. To simplify the task, we ask the annotators for one emotion label per tweet, while the aggregation of their answers will generate the multi-label annotation of each tweet. In order to increase the usefulness of annotations, the action sequence requires raters to additionally specify the textual constructs that support their decision. We view the selected textual constructs as probable classification features and refer to them as emotion indicators. We also ask workers to provide new analogous emotion indicators that could be used as a replacement for the selected ones. The proposed method thus simultaneously produces *an emotion annotated corpus* of tweets and creates an *emotion lexicon*. The resultant emotion lexicon is a list of phrases indicative of emotion presence with the weights of associated emotions. It consists solely of the phrases selected by respondents, while their weights are learnt based on their occurrence in the annotated corpus.

Following our motivation scenario presented in the thesis introduction, we aim at building a fine-grained emotion lexicon for a specific domain — the tweets with reactions towards the sports events. The emotion recognition within this domain opens great opportunities to understand the fans' reactions and to summarize their emotions for enhanced collective experience of emotion. For annotation, we thus focus on the tweets related to a fixed sports event, the Olympics 2012.

We show that the human-constructed lexicon, *OlympLex*, is well-suited for the particularities of the chosen domain, and also for an emotion model with a high number of categories. We show that domain specificity of the lexicon matters, and that non-specialists, using their common sense, can extract features that are useful in classification. We use the resultant lexicon in a binary polarity classification problem on the within-domain data and show that it outperforms several traditional lexicons. In multi-label emotion classification, we show that it is highly accurate in classifying tweets into 20 emotion categories of the Geneva Emotion Wheel (GEW), version 2.0 [Sch05]. As a baseline for comparison, we use the GEW compatible lexicon, the Geneva Affect Label Coder (GALC) [Sch05]. The experiments show that *OlympLex* significantly outperforms this baseline.

In short, the contribution of this chapter research is two-fold. First, we propose and investigate the properties of the annotation task for simultaneous collection of data labels and indicative linguistic features. Second, we describe and validate the human computation method for creating a domain-specific emotion lexicon.

4.2 Related Work

We review the main ideas related to crowdsourcing tasks, while focusing on collecting linguistic annotations, extracting common-sense knowledge, and generating affective information. We also discuss different approaches to use annotations for building affective resources.

Crowdsourcing Human Annotations Online data annotation via crowdsourcing became available with the development of the web and its economy. It provides a way to outsource expensive and time-consuming manual data annotation to “an undefined, generally large group of people in a form of an open call” [QB11]. The process of crowdsourcing data annotation usually implies that multiple human raters each perform a small task and their answers are aggregated in the resultant data annotation. In the most common case of crowdsourcing, annotations are conducted via online labor markets, such as Amazon Mechanical Turk (or AMT), where annotators get paid for each performed micro-task. An annotation micro-task can be to label one image, text document, or audio file, according to the desired classification taxonomy. Crowdsourcing was used to create ground-truth corpora for many linguistic tasks, including word sense disambiguation [Rum11], named entity detection [FMK+10], and document-query relevance assessment [ABY11]. Snow et al. [SOJN08] collected crowdsourced annotations for five linguistic tasks, including reader-side emotion classification, and showed that with the aggregated crowdsourced answers (some required redundancy) are comparable in quality with the expert annotations. In the field of affect recognition, paid crowdsourcing was successfully applied to collect emotion and sentiment annotations of images (including landscapes [QOC14] and facial expressions [TMM13]), audio (including speech utterances [STCD12] and music songs [SCS+13]), and text documents (e.g. tweets [NKR+13]). Morris and McDuff [MM14] surveyed different approaches to collect annotated affective data using crowdsourcing. They discuss not only annotation of the provided data, but also generation of the original content, e.g. by acting, or finding specific affective content within a given collection. In this work, we use paid crowdsourcing to collect fine-grained emotion annotations of selected short text documents (tweets).

An alternative to paid crowdsourcing are games with a purpose (GWAPs) that use fun as intrinsic motivation for attracting and engaging users [VAD04, VAD08]. The examples of designed GWAPs in the linguistic domain include ESP Game, asking users to generate labels for images until agreement is achieved [VAD04], Verbosity, asking one user to describe attribute relations for a given word so that another user could guess the word [VAKB06], and Concept Game, using gamification principles for validating the extracted commonsense relations between concepts [HB12]. For emotion recognition, another GWAP was designed for generating the emotionally annotated dataset [PS10]: it asks one user to generate a sentence expressing a specific emotion, while another user should guess what emotion it is. In this work, we use more predictable paid crowdsourcing, because it allows to ensure the required number of annotators is attracted and the full corpus is labeled within a desired time frame.

Building Affective Lexical Resources using Human Annotations In order to build emotion recognition systems, researchers manually annotated with emotions (not always via crowdsourcing) different text documents, including tweets [RRJ$^+$12], fairy tales [ARS05], blog posts' sentences [AS07], and news headlines [SM08]. Emotion recognition systems are then built in a supervised manner using the collected annotated corpora for training the machine-learning classifiers over extracted linguistic features.

An alternative approach for building emotion recognition and sentiment analysis systems is to annotate emotions and sentiment of separate words, by requesting their direct annotation. An example of such approach for multi-category emotion recognition is the construction of the NRC lexicon, which was also extracted via crowdsourcing on AMT [MT13]. The authors developed a task where, for a given word from WordNet, annotators rated to what extent a given word is associated to each of the 8 Plutchik's basic emotions [Plu80]. Affect Database was created in a similar way, but annotating terms into 9 emotion categories by three offline annotators and including emoticons, slang terms, and interjections for annotation [NPI07]. For binary polarity classification, manual annotation of words was used to construct various sentiment lexicons, including ANEW [BL99], its enlarged version [WKB13], and VADER [HG14]. In these annotations, raters were asked to label the intensity of polarity (positive or negative) for context-free potentially emotional terms. Such lexicons, generated by labeling context-free words, form a general-purpose knowledge, which can be applied within any domain. However, these lexicons might miss the concepts and emotional expressions used in the context of a specific domain.

In contrast to such context-free approaches, in this work, we harvest emotional labels of the potential features in context, taking into account the studied domain. The terms are associated with emotions in the context of the tweet they appear in. We use the approach suggested by Aman and Szpakowicz[AS07] where humans are asked to select an excerpt of the text that expresses emotion. The similar approach for polarity classification was suggested as well, where workers in the crowdsourcing task are asked to find and label sentiment-related features in the review text [BMF14]. Asking users to directly select indicative features in the text was also shown to be beneficial for sentiment analysis in the GWAP setup [MTGF12]. In this work, in addition to selecting indicative features from the text, we ask the annotators for additional interchangeable, emotional expressions that can be used in the described situation. Such additional indicators will extend the set of annotated emotional expressions. Therefore, our human computation task produces simultaneously document-level labels and various potential indicators of the stated emotion.

To summarize, the main differences of our emotion lexicon compared to its predecessors lie in the usage of a new fine-grained emotion set, new methods of human computation employed in its construction, and its specificity to the context of Twitter posts and sport-related events.

4.3 Domain Data for Annotation: Tweets about Gymnastics

Social media platforms such as Twitter have become a common way for people to share opinions and emotions. Sports events are traditionally accompanied by strong emotions and the 2012 summer Olympic Games in London were not an exception. Our motivation scenario is to analyze the emotions shared by the spectators of the Olympic games. We consider the tweets about the Olympics posted during the 2012 Olympic games as a data source for this analysis. However, we also assume that the same emotions are expressed in similar manner for all the sports disciplines. We thus narrow the scope of our annotation to a single sport – gymnastics.

Traditionally, the gymnastics teams from the USA make strong bid for victory. Thus, we assume that a large group of English-speaking nations may be interested in it. Then, gymnastics is a dynamic type of sport where each moment of performance can play a crucial role in final results, enhancing the emotional experience in audience. Also, it is less commonly practiced individually than, for instance, running or swimming. Thus, the occurrence of the term ‘gymnastics’ in tweets from the Olympics period will rather signal a reference to the Olympic gymnasts than description of a personal exercise session.

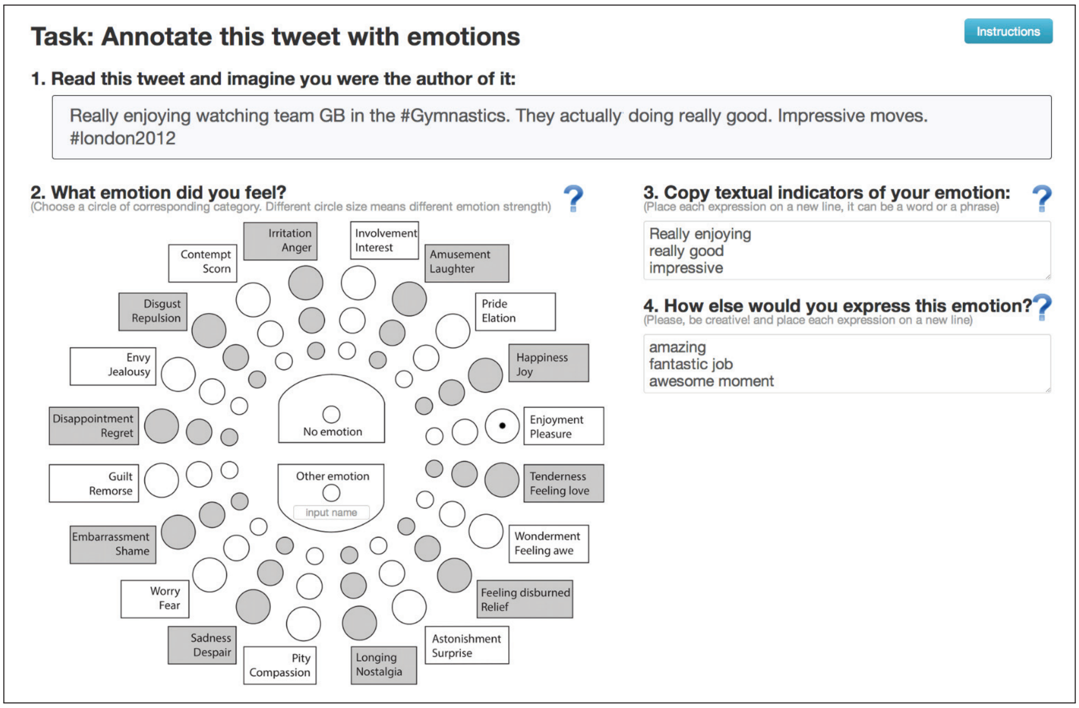
We used the hashtag *#gymnastics* (hashtags represent topics in tweets) to obtain the tweets related to the gymnastic competitions during the Olympics 2012 time, between July 26th and August 14th 2012. This resulted in $199,730$ such tweets. An emotional example is *“Well done #gymnastics we have a SILVER yeayyyyyyyyy!!!! Wohoooo”*. A subset of these tweets will be used for the annotation and evaluation of the proposed method.

4.4 Human Computation Task for Emotional Labeling and Emotion Feature Elicitation

We create a Human Computation task, using the online labor market (Amazon Mechanical Turk or AMT) to simultaneously accomplish two goals. The first one is to have a reliable, human annotation of the emotion categories within a text corpus. That is we aim to collect the most probable emotion labels corresponding to the tweet as felt by its author. The second goal is to enable the workers[2] to provide us with the lexical features needed to construct an emotion lexicon. That is we aim to collect emotional cues (called indicators) representative for each emotion. These can be words or word sequences (n-grams).

Both goals are incorporated into the designed AMT Human-Intelligence Task, or HIT. This approach guarantees that the emotion cues and emotion labels are tied together by the same person in the same context. We developed the task for annotating a subset of the collected tweets with emotion-related information. In this section we describe the design of this task, iterations of the annotation launches, as well as provide the statistical description of the collected data.

[2] The users of AMT are called *workers*, because they receive micro-payments for doing the HITs.



Task: Annotate this tweet with emotions

1. Read this tweet and imagine you were the author of it:

Really enjoying watching team GB in the #Gymnastics. They actually doing really good. Impressive moves. #london2012

2. What emotion did you feel?
(Choose a circle of corresponding category. Different circle size means different emotion strength)

3. Copy textual indicators of your emotion:
(Place each expression on a new line, it can be a word or a phrase)

Really enjoying
really good
impressive

4. How else would you express this emotion?
(Please, be creative! and place each expression on a new line)

amazing
fantastic job
awesome moment

Figure 4.1: The screenshot of the emotion annotation task interface. Possible answers are included for overview.

4.4.1 Task Description

Our task comprises the annotation of one tweet presented to a worker. The task interface for labeling one tweet is shown on Figure 4.1. For simplicity, the work flow is explicitly divided into small action steps with a given order.

Action 1. A worker is asked to read a tweet text and imagine that he or she was the author of it. This is to ensure that the most probable writer's side emotion is being labeled, instead of the workers personal reaction as a reader of the tweet. All the following actions (2, 3, and 4) should be performed following this idea. To assure it, the mentions of the author of the tweet in the task instructions are replaced with a direct address to the worker as "You". We chose this presentation style because it was believed to help workers put themselves in the author's place, and also to make the task more socially appealing.

Action 2. For the given tweet, a worker should identify what dominant emotion the author of the tweet felt when writing it (***emotion label***) and how strong it was (***emotion strength***). Even though an emotion mixture could be felt, a worker had to choose one emotion that prevailed all others. This kept him focused on one main emotion in the next actions asking for the emotional indicators. To elicit this information, we directly employ the Geneva Emotion Wheel (GEW, version 2.0 [Sch05]), designed to elicit and categorize emotional responses. As was described in section 3.2, the GEW presents 20 emotion categories, each represented by two names, in the structure of the wheel. Three circles are shown for each category standing for different strength labels (low, medium, or high).[3] Thus, each circle corresponds to a specific combination of emotion label and strength. A worker should click on one of the circles to answer. Two additional answer options are available: *No emotion* and *Other emotion*. We required workers to type the emotion name if *Other emotion* circle was selected. Additionally, the GEW layout allows the extraction of ***polarity label*** based on the chosen option, because the right side of the wheel corresponds to the positive emotions and left side corresponds to the negative ones.

Action 3. In case an emotion was present, a worker is asked to choose the excerpts of the tweet indicating its presence, the ***tweet emotion indicators***. These indicators play the role of emotional cues given in the tweet text and are linked to the chosen emotion label via the text context. A worker is asked to find the expressions of the chosen emotion present in the tweet text. They can be either one word, emoticon, or a sequence of the tweet words (n-grams). We ask to aim for shorter expressions and to also include the words modifying the strength of emotion (e.g. to choose *so excited* instead of *excited*).

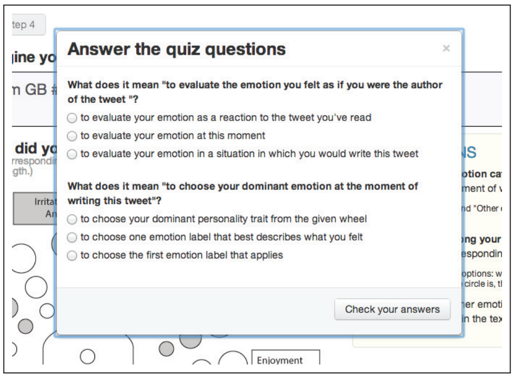
Action 4. As the last action, a worker is asked to input ***additional emotion indicators*** of chosen emotion. Similarly to the tweet emotion indicators, a worker should input the textual expressions of the chosen emotion. However, in this case the expressions have to be not from the tweet text, but generated based on the personal experience. The provided emotion indicators are also emotional cues, but outside of the tweet scope. For example, a worker could state that s/he uses *poor thing* to express *Pity*. Having this question allows us to collect a wider variety of possible emotional expressions that are similar to and inspired by those appeared in the tweet text. This question is also directed as additional validation of workers engagement and qualifications, as it would be more difficult for non-native speakers to produce other emotion indicators.

In the presented task interface, the visible instructions are concise and describe only the main actions to remember. Detailed instructions are shown under a mouse hover over the large question marks near each action to explain.[4] They describe what should be the output and explain the main concepts. For example, in the instructions for action 3 we explain what can be an emotion indicator and how it should be stated in the answer. The clue words in the instructions text are highlighted and the text is structured hierarchically to make it more legible.

Alternatively, we could split our human computation task into three separate subtasks: label emotion in the tweet, select tweet indicators, and provide additional analogous indicators. This would follow the common guideline for designing crowdsourcing tasks that suggests to split the task into small chunks [CTIB15]. However, by combining the three subtasks together, we ask annotators to read and understand each tweet only once. It makes the annotation more efficient in this sense. Furthermore, this task design allows us to use the subtasks about

[3] We use three strength options instead of the five in the original version to simplify the choice for workers.

[4] These detailed instructions are given in the appendix, section A.4.



Answer the quiz questions

What does it mean "to evaluate the emotion you felt as if you were the author of the tweet "?

○ to evaluate your emotion as a reaction to the tweet you've read
○ to evaluate your emotion at this moment
○ to evaluate your emotion in a situation in which you would write this tweet

What does it mean "to choose your dominant emotion at the moment of writing this tweet"?

○ to choose your dominant personality trait from the given wheel
○ to choose one emotion label that best describes what you felt
○ to choose the first emotion label that applies

Figure 4.2: The screenshot of the comprehension quiz questions shown after the first step of the tutorial for the annotation task.

emotion indicators as validation of workers' engagement as well as additional effort barrier for malicious workers. Last but not the least, this will allow us to connect the provided indicators directly to the specified emotion label.

4.4.2 Preemptive Quality Control: Tutorial

We also include an obligatory tutorial into the described task. It forces the workers to review the detailed instructions and validate their understanding. It also provides examples of answers. Another role of the tutorial is creating an additional barrier to start the task, which can discourage poorly motivated workers. As such, the tutorial plays a role of a preemptive quality control measure: it helps to select workers who are willing to put more efforts in the task and teaches them how to perform the task better. We describe below the design of our tutorial for the task.[5]

As the task flow itself is divided into actions, we separate analogously the tutorial into manageable steps. We present the detailed instructions for each action alongside the same interface sub-part for this action. This is done to simplify the learning process for workers. Moreover, an example of appropriate answer is shown for each answer-requiring action (2,3, and 4) to exemplify the required output. Providing the examples in the tutorial is assumed to give workers better understanding of what are good-quality answers and what answers they should provide.

We use the same tweet as an example during the explanation of all task actions to limit the memory load of workers. However, the annotated tweet examples are valuable for overview and repetition of learned instructions. Thus, we include an additional step showing the full unchanged task interface with the included answers for another tweet example. No other examples are given in order to limit the number of steps in the tutorial and preserve its simplicity.

[5] The detailed steps of the tutorial and its quiz questions can be found in the appendix, section A.4.

In addition to the detailed instructions and examples of fulfilled answers, the tutorial should also verify that workers indeed read and understood the provided information. We include quiz questions to avoid not-attentive behavior and to enforce additional thinking about the instructions. The quiz consisted of four multiple-choice comprehension questions, one for each task action (e.g. "What is the dominant emotion?" for action 2). The Figure 4.2 shows an example of quiz questions appearance. The answer options were carefully chosen based on the common pitfalls.For instance, in the pilot launches we discovered that some workers input the names of other, non-selected emotions as additional emotion indicators (for action 4). Thus, to inform the workers that such answers are undesirable, we include this misconception as an incorrect option in the related multiple-choice quiz question. This allows workers who misunderstood the instructions and selected an incorrect answer to get feedback on its incorrectness and to avoid making this error in the annotation process.

An alternative could be asking workers to practice the actions on the examples with provided feedback. We refrained from it: Due to the high subjectivity and ambiguity of the task, it is problematic to define an appropriate feedback for every situation. For instance, input of additional emotion indicators (action 4) is not even entry-based, and thus cannot be predicted in advance.

To summarize, our tutorial contains four instruction steps, interrupted by three quiz steps. The first three instruction steps describe each task action in details, and the last one reviews a full task interface with the example answers given. During the first three steps, workers are presented with one positive tweet and the example answers for each task action are shown for this tweet. To balance the perspective, the example tweet at the last step is negative. The quiz questions are shown as intermediate tasks between tutorial steps. To proceed to the next step, a worker is asked to answer multiple-choice questions testing his understanding of the task actions and related concepts described on this tutorial step. Until the worker provides a correct answer for each question stated on the step, he will not have access to the next step. However, we do not eliminate any worker based on their quiz performance. We believe that incorrect answers induce more thoughtful reconsideration of the instructions.

4.4.3 Task Design Development and Annotation Launches

The design of the annotation schema and the corresponding instructions as well as the search for the optimal HIT parameters took several iterations. Table 4.1 contains the statistics on each iteration of annotations.

For the labels, we evaluate the quality of the provided answers using inter-annotator agreement metrics. We do not have the ground-truth labels to evaluate the correctness of annotations, because the goal was to collect such ground-truth labels. The inter-annotator agreement is computed separately for each tweet as the percentage of agreed label pairs, and then averaged across all labeled tweets. Beside emotion agreement, we also consider polarity agreement. The polarity label of an answer is defined as the polarity of its emotion label. *No emotion* implies a

Neutral polarity. For answers with *Other Emotion*, we manually detected their polarity based on provided emotion name if applicable, or set *Neutral* polarity otherwise. Strength agreement considers four strength labels: *No Emotion*, *Low*, *Medium*, and *High*, and all answers of *Other Emotion* are assigned to *Low* strength.

For analyzing the provided indicators, we compute the average number of tweet and additional emotion indicators, Jaccard-based inter-annotator agreement of tweet indicators, and average acceptability of indicators per answer. For all of these metrics, we consider only the answers where non-neutral emotion label was assigned. To compute the agreement between the two lists of tweet indicators returned by two workers for a same tweet, we first extract for each indicators' list an unordered set of linguistic tokens that appeared in it. Then, we compute the agreement as the Jaccard similarity between these two sets of tokens, computed as $\frac{|T_1 \cap T_2|}{|T_1 \cup T_2|}$ (it is an alternative for inter-annotator agreement described in [AP08] and adapted to the specifics of our task input). For each tweet we compute the average Jaccard-based agreement across all answers' pairs as the inter-annotator indicator agreement. The reported agreement of tweet indicators is the average per-tweet inter-annotator indicator agreement. We also suggest to evaluate the acceptability of the indicators. It validates the indicators' compliance with the answer requirements. In our case, to determine whether a tweet indicator is acceptable, we validate that it appears in the text of the tweet (while ignoring the differences in main punctuation signs, case, and elongation patterns). For additional indicators, it is more difficult to derive a simple heuristic for measuring their acceptability. We noticed that even though we asked for new emotional expressions, many users returned the given names of emotions as additional indicators. We believe that users returned them as an alternative emotion label for a tweet, and as such they should not be accepted as alternative emotional expressions for the main stated emotion. We report the average percentage of indicators (tweet or additional) per answer.

We started our research from the preliminary within-lab and crowdsourced annotations for another set of 14 emotion categories (iterations 1–3). Only later our task matured to the presented task interface and annotations in terms of the described GEW categories (iterations 4–5). We present here all iterations and subsequent decisions in chronological order.

Iteration 1 Firstly, we annotated 200 tweets (set $\mathscr{S}_1$), using respondents within our laboratory, into a set of 14 emotion categories that we considered to be representative for the emotions incited by sport events: *Love, Pride, Excitement, Positive Surprise, Joy, Like, Other Positive, Anger/Hate, Shame, Anxiety, Shock, Sadness, Dislike, Other Negative.* For each tweet an annotator gave the *emotion label* and chose corresponding *tweet emotion indicators*. The tweets of $\mathscr{S}_1$ included both tweets with predefined explicit emotional words and without. Emotion agreement of the obtained annotation is 38.5%, with Fleiss Kappa [Fle71] of 0.32. Considering the difficulty of classifying emotions into multiple categories and the expert-like nature of this annotation iteration, such annotation quality can be regarded as acceptable for the multi-label classification. The data from this annotation allowed us to extract new emotion indicators used in the domain of sports events, such as 'congrats' and 'goteamgb'.

Metrics	Iterations						
	1	**2** (B_{en})	**2** (B_{all})	**3**	**4**	**5**	**4+5**
	14 emotions (preliminary)				20 GEW emotions		
Polarity agreement	78.5	68	33.3	66.7	73.9	75.9	75.7
Emotion agreement	38.5	24.7	13.3	29.3	25.8	29.7	29.3
Strength agreement	-	-	-	-	38.2	44.4	43.8
Agreement of tweet indicators	44.8	44.8	0	23.8	17.9	43.7	41.0
# of tweet indicators	1.6	1.28	0.48	1.26	1.20	1.72	1.67
# of additional indicators	-	0.24	0.27	1.39	1.31	2.06	1.99
% of accepted tweet indicators	81	89	15	65	65	92	89
% of accepted additional indicators	-	95	100	94	56	95	91

Table 4.1: Descriptive statistics on the data collected over the emotion annotation iterations.

Iteration 2 We launched two batches of HITs on AMT: B_{all} and B_{en}. One HIT asked for the full annotation of one tweet. A HIT batch is defined by a set of tweets to label, with some parameters specific for AMT, such as the number of different workers for each tweet (we used 4 in all our launches), the payment for one HIT, or specific worker requirements (e.g. for B_{en} we also required that workers should be from the United States). We grouped 25 tweets from $\mathscr{S}_1$ with payment of \$0.05 per HIT in B_{en}, whereas for B_{all} we included only 10 tweets with payment of \$0.03 per HIT. The annotation schema used again the same 14 emotions. For each tweet an annotator gave the *emotion label* and provided *tweet* emotion indicators. The field for *additional* emotion indicators input was presented as optional.

We discovered that the answers in B_{all} had an unacceptable quality, with the low agreement both in labels (13.3%) and indicators (0%). This can be explained either by lower understanding of English or less reliability of workers from all around the world compared to the U.S. workers. Consequently, all our next iterations had the requirement on workers to be from the U.S.

Iteration 3 We launched a new annotation batch to annotate the full set $\mathscr{S}_1$ with the same 14 emotions. Starting with this iteration, the payment was fixed to \$0.04 per HIT, i.e. per labeling one tweet. The *additional emotion indicators* field was shown as compulsory. The experiment showed that AMT workers generally followed the instructions achieving emotion agreement only slightly worse than the agreement achieved in our within-laboratory offline annotation (iteration 1).

Iteration 4 At that moment, we decided to use the more fine-grained and well-researched GEW emotion categories. Thus, we launched another HIT batch to annotate $\mathscr{S}_1$ again, in terms of GEW emotion categories (with an interface and task model described in section 4.4.1 on Task Description). The details of the HIT description and final parameters are given in the appendix, section A.3. Particularly, the payment for labeling one tweet was set to \$0.04 and we requested 4 workers' answers for each tweet. Even though a new task contained more answer options, emotion agreement stayed in the same range between 0.25 and 0.3.

Iteration 5 We launched a final batch with the described GEW schema to annotate more tweets. We selected Olympics related tweets that had a high likelihood of being emotional. We first selected tweets using the emotion indicators obtained during the previous iterations and found more than 5 times in the collected corpus (418 terms). For each keyword in this list we extracted up to 3 tweets containing this term (1244 tweets). In addition, we added the tweets without keywords from the list, but posted by the users who used these emotional keywords in their other tweets, supposing that these users are more likely to express their emotions. Overall, 1800 tweets were selected, but 13 were excluded later because they were not written in English.

The results from the iteration 4 were promising, but we discovered possible misunderstanding of the task instructions which could be addressed by inclusion of the tutorial. Therefore, we designed the tutorial described in section 4.4.2 and included it into the next, 5th iteration. The tutorial was shown to a worker before labeling the first tweet, and the worker could not submit that HIT until he passed the tutorial.

Further analysis of the data from iteration 4 showed that half of the workers who submitted high number of answers (3 out of 6 workers who labeled more than 50 tweets) actually submitted poor-quality answers, with either high percentage of random and neutral emotion labels or with repeated misconception problems. To avoid such accumulation of non-usable answers, we set a limit on the number of assignments submitted per worker in iteration 5, which could be surpassed only after manual authorization. The round number of 50 assignments was chosen as a limit to afford conclusive decisions on worker's quality and yet bound the number of possible incorrect answers to the acceptable amount. Only 10% of workers exceeded this limit in launch 4. This limitation worked as follows: After workers accepted their 51st assignment, they were shown a message about the impossibility to submit it until we review their previous work. If no quality-related problems were revealed during the manual review of their answers, we authorized such workers to continue performing the task. Even though no notification message was sent, 24.4% of approved workers returned and continued performing the HITs without any further limitations.

Hence, the iteration 5 differed from the iteration 4 in two ways (besides having new tweets to label). It included 1) the tutorial and 2) the limitation on number of answers per worker. We can observe that the quality metrics are better in iteration 5, potentially because of these two changes. We analyze to what extent the tutorial impacted these improvements in the next chapter 5 on the quality control.

During this largest and final iteration, we attracted 674 different workers for annotating emotions in the selected tweets. On average (when aggregating at worker's level), each worker spent 93.8 seconds on one tweet and annotated 10.7 tweets, with 3 being the median and 724 – the maximum. When averaging at the answers' level, annotation of one tweet took 68.7 seconds (it is smaller than the per-worker average because the annotation of a first tweet in average took longer: 108.9 seconds vs. 64.7 for all other tweets).

The resulting corpus contains the data gathered during the iterations 4 and 5. It consists of 1987 tweets each annotated by 4 workers with emotion label, emotion strength, and related emotion indicators. The Fleiss Kappa [Fle71] for emotion labels is 0.24 which is considered to be fair by Landis and Koch [LK77], but quite low compared to usual kappa values in other tasks (e.g. polarity annotation usually has Kappa in a range of 0.7–0.8). We conclude that the annotation in terms of multi-category emotions is highly subjective and ambiguous task, confirming our assumptions on existence of emotion mixtures and providing further argument for treating emotion recognition problem as multi-label classification. While in the context of one-class annotations the values of Kappa less than 0.3 can raise concerns for the reliability of annotated data, in this work we alleviate this concern by directly aggregating all the provided tweet's labels into a multi-label.

We further note that the strength agreement is low: its Fleiss Kappa is 0.13, which corresponds to only a slight agreement. Thus, we suggest to ignore such strength labeling in the current version. To improve it, more related examples and explanations should be provided in the tutorial. Another alternative could be to simplify the task and ask workers to provide only the labels without considering the strength of the emotion.

4.4.4 Posterior Quality Control via Answer Filtering

The results of crowdsourcing usually require additional refinement. The workers who give malicious answers intentionally or due to lack of understanding worsen the data quality. We detect such workers automatically using the following two criteria:

Average Polarity Conformity A worker's answer has a *polarity conformity* of 1 if at least one worker indicated the same polarity for the same tweet (0 otherwise). A worker's average polarity conformity is computed from all his answers. This criterion aims to detect the workers who repeatedly disagree with all other workers on polarity.

Dominant Emotion Frequency The dominant emotion of a worker is the one which appears most frequently in his answers. A worker's dominant emotion frequency is the percentage of the dominant emotion among the worker's answers. This criterion aims to detect workers biased towards a specific emotion.

A worker who has the average polarity conformity below a predefined threshold or the dominant emotion frequency above a certain threshold is considered to have an insufficient quality and all his answers are excluded from the corpus. The threshold for each criterion is computed as a percentile of an approximated normal distribution of workers criterion values for probability limit of 0.01.

To increase the confidence in the computed criteria values, we establish a minimum number of tweets T_{min} any worker should annotate to be subjected to the criteria. To establish this number for each criterion, we use the following algorithm:

Let $X_n(w)$ be the criterion value computed using only first n answers of worker w in order of their submission. For each worker we detect $N_{min}(w)$ – the minimum number of answers after which the criterion value stops varying greatly:

$$|X_n(w) - X_{n-1}(w)| \leq 0.05,\ \forall n \geq N_{min}(w) \tag{4.1}$$

We then compute T_{min} as the ceiling of the average value of of $N_{min}(w)$ among workers who annotated at least 20 tweets. It equals to 9 for the criterion on average polarity conformity, resulting in 0.68 for its cut-off threshold. And it equals to 12 for dominant emotion frequency criterion, resulting in the threshold of 0.42 for it.

The described procedure on detection of bad workers allowed the analysis of 83% of the answers. Using it, we excluded 8 workers, with their corresponding 260 answers.

In addition to removing these workers, we also excluded potentially incorrect answers: 736 answers that had a polarity conformity of 0. This additional filter was applied to all the remaining answers from the previous method. We also excluded the 121 answers with *Other emotion* and the answers for 12 tweets, that were left with only 1 answer by this stage.

As a result of the quality control step, 14.2% of initial answers were excluded. Overall, 1957 tweets with 6819 corresponding annotations remained. After the posterior filtering, the Fleiss Kappa [Fle71] increased to 0.33 for emotion labels and to 0.18 for strength labels.

4.4.5 Analysis of the Collected Data (SREC)

The pre-filtered answers compose the final *Sport-Related Emotion Corpus (SREC)*. It contains 1957 tweets, with 3.48 answers per tweet in average.

Label Distribution To provide a glimpse of these data, we present the distribution of emotion categories among all answers in SREC in Figure 4.3. The most frequently answered emotion category was *Pride*, followed by *Involvement*. Even though the large proportion of the labeled tweets was selected using certain keywords, this annotation distribution is likely to be reminiscent of the emotions in the gymnastics dataset, because we did not put any criteria on the emotions of those keywords. The discovered emotions are natural in the context of sports events, however more coarse-grained emotion models could not distinguish them. For example, using Ekman's set of six basic emotions [Ekm92], most of the positive tweets would likely to be labeled with *Happiness*. This highlights the advantage of fine-grained GEW emotion set to express the subtleties of the domain.

We also analyze the distribution of polarity labels in the tweets that have perfect polarity agreement (comprising 92.7% of the annotated tweets).[6] This polarity distribution is skewed as well: 63.3% of tweets are positive and only 26.3% of tweets are negative. The neutral (or non-

[6] Note that the polarity agreement is high due to the removal of answers with disagreeing polarity during the filtering process.

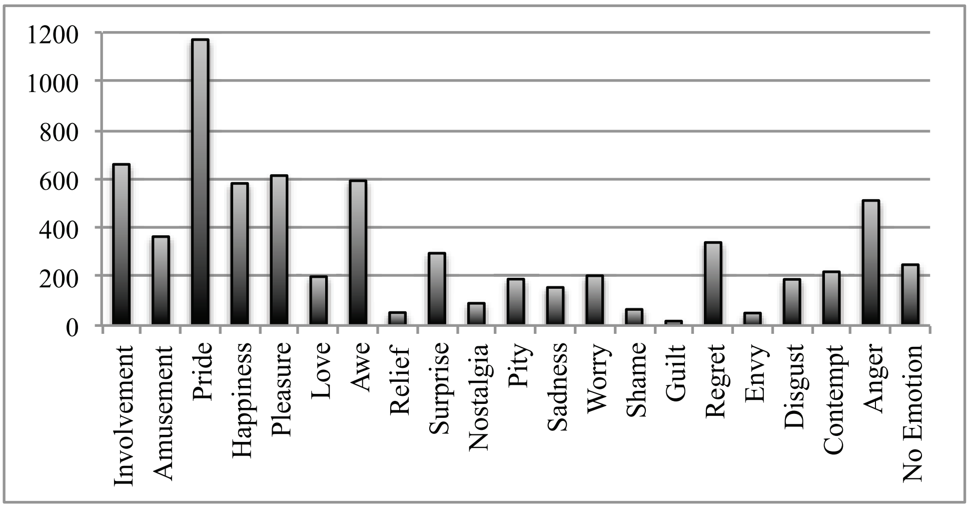


Figure 4.3: Distribution of emotion labels in crowdsourced workers' answers comprising the SREC data (i.e. after the application of posterior quality control).

emotional) tweets comprise 3.2% of all labeled tweets, while polarity disagreement appears in 7.3% of tweets. The small presence of neutral tweets confirms the success of our tweet selection process to find mostly emotional tweets. We aimed for emotional tweets in the annotation in order to obtain larger number of different emotion indicators for the lexicon.

Case Analysis of Label Agreement and Disagreement Out of 1,753 tweets with agreed positive or negative polarity, 65.7% have no definite majority label for emotion (we consider majority label as definite if either all or three out of four workers returned the same label for a tweet). This shows that for some tweets, it was easy to state the polarity, positive or negative, but harder to agree on a specific emotion. For example, the tweet "YES YES YES! Congrats boys!! #teamGB #gymnastics #olympic2012" was labeled as positive by all four workers, but two of them labeled it with *Pride* and two other with *Happiness*. Similar negative example without definite emotion is "referees on #gymnastics disgrace again", which workers labeled differently as *Contempt*, *Regret*, *Disgust*, and *Anger*. Moreover, the average number of different emotion labels per tweet is 2.24. These facts confirm the difficulty of fine-grained emotion labeling and the need for modeling the problem of emotion recognition as a multi-label classification problem instead of the multi-class classification, where only one class is assigned to the document.

The other 34.3% of tweets with agreed positive or negative polarity have a definite dominant emotion. The examples include "Oh, how i miss #gymnastics): <URL>" labeled three times as *Nostalgia* and once as *Love*, "Way to go fab five!!! :) #takinhomethegold #gymnastics #woooo" labeled by all four workers as *Pride*. This shows that specific emotions are definite for some tweets, potentially due to the use of more explicit emotional expressions.

We further analyze the 7.3% of tweets where workers did not agree on the polarity of the tweet. The most common case of such disagreement is simultaneous presence of positive and negative labels (71.8%). Such tweets can represent a mixture of emotions, e.g. during tense moments: "bloody hell tell me someone saw that guy fall off the bar!! OUCH.. #gymnastics #Olympics" is labeled with *Involvement*, *Surprise*, *Pity*, and *Worry*. They can also express several different emotions towards different subjects: "Congrats 2 Aly Raisman and Gabby Douglas for making it 2 the #gymnastics all-around. Def feel badly 4 poor Jordyn Wieber. :(@NBCOlympics" is labeled twice with *Pity*, and once with *Love* and *Pride*. Some of this polarity disagreement also appears to be due to the difficulties and differences in interpretation of some tweets: e.g. some writers interpreted the tweet "Nothing I'm interested in on all day now all the cool stuff on at once #Olympics2012 #gymnastics #swimming" as *Involvement* and *Happiness*, and others as *Anger*. One potential task improvement could be to add a possibility to flag the tweets that are difficult for understanding or contain several contradicting emotions, as well as the option to skip them.

The disagreement on whether the tweet is neutral explains the remaining 28.2% of polarity disagreement cases. The neutral category is mostly confused with some positive emotions (89.6%), as in the tweet "#TeamGB 2nd at the end of round 1 of the #Gymnastics #London2012" labeled twice as *No emotion*, once as *Involvement*, and once as *Pride*. Such tweets rather describe the emotion-provoking situations without explicit reference to emotions. To further improve agreement, we could explicitly state whether to recognize emotions in such factual tweets or label them as neutral. Asking workers to flag such tweets could also be beneficial by allowing their separate treatment.

Co-Occurrence of Emotion Labels To understand which specific emotions appear together in the annotations, we compute a confusion matrix between emotion category labels returned by workers. Such annotation-based confusion matrices can show to what extent annotators agree on each category. In this work, we compute a normalized symmetric confusion matrix as follows. For each returned label (an answer label), we detect which emotion categories are labeled together with it in the same tweet (paired labels). Note that all labels participate in both sides of the computation, making the raw confusion matrix symmetric, with each pair of returned labels counted twice (once per each direction). To obtain a normalized matrix, we divide each cell count by the sum of the row counts (i.e. we normalize based on answer labels). Without this normalization step, the counts would be difficult to compare across emotion categories because of their unequal presence in the annotation. Figure 4.4 presents the resultant confusion matrix as a heat map. The values in each row provide an estimate of how likely a specific paired label would be provided given that the answer label is known.

We can observe a visually clear diagonal in this confusion matrix, meaning that emotion categories co-appear with themselves relatively frequently. However, the average of such diagonal values is far from perfect agreement (0.34). The highest diagonal value corresponds to *No Emotion* (0.725), while the smallest one – to *Guilt* (0), for which the answers did not match any paired labels. As expected, we can observe that some emotions are often labeled

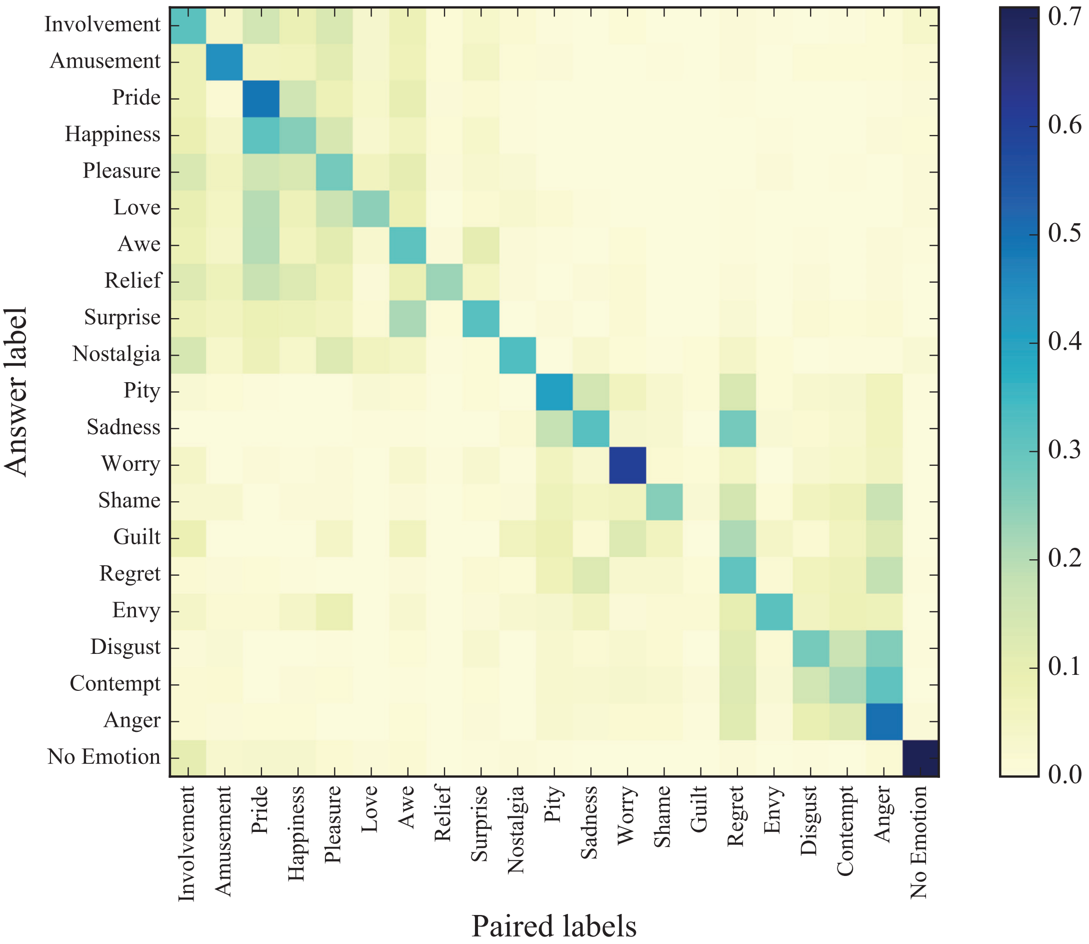


Figure 4.4: The normalized confusion matrix between the emotion category labels returned in the annotations for the SREC data.

together, e.g. *Happiness* has high co-occurrence with *Pride* and *Pleasure*, *Surprise* – with *Awe*, *Disgust* and *Contempt* – with *Anger*. One can observe that *Involvement*, *Pride*, and *Pleasure* frequently co-appear with other positive emotions, while *Regret* and *Anger* frequently co-appear with other negative emotions.[7] There is also an interesting confusion pattern between *Sadness*, *Pity*, and *Regret*: *Pity* often appears with *Sadness* and *Regret*, *Sadness* – with *Regret* and *Pity*, whereas *Regret* often appears with *Sadness* and *Anger*, but less with *Pity*. It is worth mentioning that some emotion categories seem to be more distinguishable than others, e.g. *Amusement*, *Pride*, *Worry*, *Anger*, and *No Emotion* frequently co-appear with themselves (their diagonal values are ≥ 0.45). Knowing the confusion matrix of annotated labels, it could be interesting to derive from it an emotion similarity function in order to take into account the emotion inter-dependence in the classification process.

[7] Note that this is partially because of their higher frequency among all the answers.

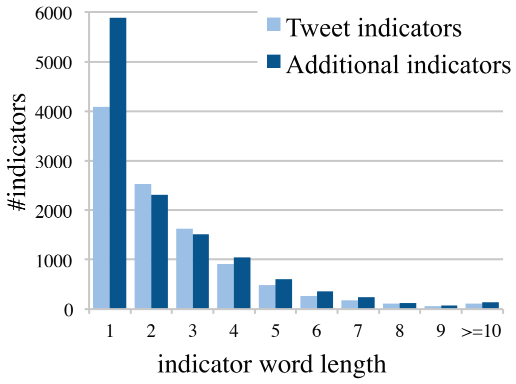


Figure 4.5: Distribution of the length of the collected indicators in the SREC corpus in terms of word number.

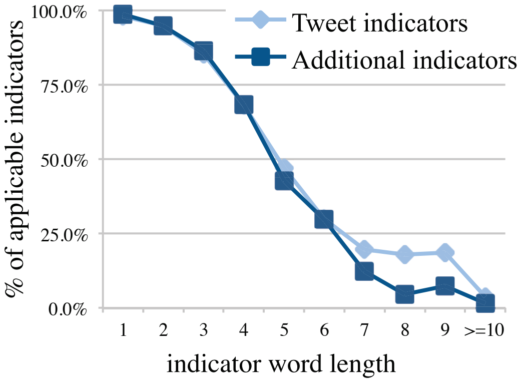


Figure 4.6: Dependency of indicators' applicability on the indicators' length, estimated in the SREC corpus.

Analysis of Provided Indicators We further analyze the collected SREC data to understand the properties of annotated emotion indicators. In the beginning of section 4.4.3, we introduced the concept of indicators' acceptability: tweet indicators should be from the tweet text to be accepted, while additional indicators should not refer to the given GEW emotion names. Only 94% of tweet indicators were accepted after the posterior data filtering. To discover why some tweet indicators are not accepted, we manually explore 100 non-matched tweet indicators. We find that 43 of them correspond to the GEW emotion names, and other 5 are adjectives related to the given emotions, e.g. '*happy*'. This reveals the misconception the workers commonly have about what they should input as the tweet indicators. Another common case of non-accepted tweet indicators (32 out of 100 reviewed cases) are misspelled (e.g. '*lovley*' instead of '*lovely*') or changed closely to the original indicators (e.g. '*can't wait*' instead of '*cant wait*'). These cases show that misspelled and almost identical words should be treated as similar in order to increase the robustness of the emotion recognition. Among other non-accepted cases are indicators with missing words (e.g. 'immense on bar' instead of 'immense on the horizontal bar') and new emotional expressions not from the tweet text.

To understand the properties of the accepted emotion indicators, we compute the distribution of their length in terms of word number. Figure 4.5 shows how many of accepted indicators (tweet and additional) have a specific word number. We can observe the monotonous decrease of indicators number with the increase of word length for both tweet and additional indicators. Also, the number of unigrams is higher for additional indicators, while the number of longer indicators is comparable. We additionally investigate the *applicability* of the provided indicators: we consider the indicator as applicable if it appears at least once within the full dataset of Olympic tweets OLYMP_DATASET (excluding the labeled ones). Overall, 85% of tweet and additional indicators are applicable. The dependency of the applicability on the word length of the indicators is shown on Figure 4.6. It shows that unigrams have almost perfect applicability. The same plot also shows that the greater is the number of words in the indicator, the less applicable it is. This is easily comprehensible when we consider that the exact combination of words in the longer indicator is less likely to appear. To increase the applicability of provided tweet indicators, one strategy could be to cautiously select the tweets with more repeatable expressions, while ensuring that they are likely to be emotional. Fixing the spelling errors within the provided indicators could also be helpful. In the next section, we suggest using only the repeatably provided emotional indicators, which can imply their higher applicability (as well as provide more confident information on their emotion association).

4.5 Building an Emotion Lexicon from Annotations

Our emotion recognition aims to detect expressed emotions, in terms of the emotion set E^0 containing 20 GEW emotion categories $e_1, \dots, e_{20}$ and *No Emotion* (or *Neutral*) category e_0 (as presented in section 3.1). To model the presence of these emotions within a text, we use the *emotionality* – emotion distribution represented as a tuple $\vec{p} = (p_0, \dots, p_{|E|})$ in the probability space $\mathbb{P}$, defined in equation (3.1).

4.5.1 Lexicon Construction

An emotion lexicon is traditionally defined as a list of terms that are indicative of emotion presence, along with their specific emotion associations. In our case, we will use as the lexicon terms the words and word sequences (n-grams) that were returned by workers as *emotion indicators*. The term' emotion association is represented as the term's emotionality $\vec{w}(t) \in \mathbb{P}$. The process of constructing the lexicon consists of finding the appropriate emotion indicators and assigning them emotionalities based on the collected annotations.

We use an annotated corpus for building our lexicon. As described in the previous sections, it has the following format: each annotation answer specifies an emotion label for a tweet and provides tweet-based and additional emotion indicators. The lexicon construction process consists of the following steps:

1) Among all *tweet* and *additional* emotion indicators provided by workers, we select those that were suggested more than once.

2) For each tweet d, we aggregate the emotion labels corresponding to that tweet in the annotated data. We extract its emotionality $\vec{p}(d) \in \mathbb{P}$ by computing the frequency of each allocated emotion label.

3) Each time a term t is returned as an *additional* emotion indicator, we construct a link between this indicator and the corresponding answer's emotion label. This link is represented as an emotionality $\vec{l} \in \mathbb{P}$ with weight 1 for the linked emotion category.

4) Then, for each selected emotion indicator t, we compute its final emotionality by averaging all the emotionalities associated with it. This includes the emotionalities of the tweets

where this indicator occurred without a negation and the emotionalities of the corresponding indicator-emotion links from additional emotion indicators' field.

5) We define an indicator to be ambiguous if its dominant polarity (i.e. polarity having the highest sum of the weights for the corresponding emotions) has summary weight smaller than 0.75. All such terms are removed from the resultant lexicon. In this way, we eliminate emotionally ambiguous terms that appear almost evenly in both positive and negative contexts, e.g.'man', 'results', 'live', and 'lot'.[8]

4.5.2 Resultant Lexicon: OlympLex

Following the specified process over the full SREC data, we computed an emotion lexicon, *OlympLex*, that contains $3,193$ terms. The ratio of positive terms to negative ones is 7:3 (term polarity is defined as dominant polarity of term emotion distribution). Unigrams compose 37.5% of the lexicon, bigrams – 30.5%, all other terms are n-grams of a higher order (up to 5). Table 4.2 presents the top terms within *OlympLex* associated with each emotion category. For each category, it enumerates the terms whose weights for the corresponding emotion are the highest among all terms.

We can observe that in general the top emotion associations are reasonable and rather express the corresponding emotions explicitly. However, for some emotion categories, the top associated terms are also associated with other emotion categories (i.e. the highest weight within the term's emotionality corresponds to another emotion). Such terms are marked with * in the table. This is especially characteristic for the emotion categories that appear less often in the annotation, such as Guilt, Relief, Nostalgia, and Envy. Thus, we hypothesize that this behavior is due to the lack of the explicit terms for those emotions in our annotated dataset, or, when such terms are present, a weaker evidence of their association with a corresponding emotion.

4.5.3 Lexicon-Based Emotion Recognition

Using the constructed lexicon, we compute the result emotionality for a text d $\vec{p}(d) \in \mathbb{P}$ as follows. We sum up all the emotionalities of the lexicon terms $\vec{w}(t)$ found within this text with the number of their occurrences $n_t(d)$, and normalize the result to obtain an emotionality:

$$\vec{p}(d) = \frac{\sum_{t \in d} n_t(d)\ \vec{w}(t)}{\sum_{t \in d} n_t(d)} \tag{4.2}$$

If no indicators are present in the text, a full weight is given to *No emotion* category ($p_0 = 1$). We also ignore all negated indicators occurrences detected by the negation words (*no*, *not*, **n't*, *never*) placed one word ahead of an indicator.

[8] We note that this simple procedure also removes some ambiguous emotional terms, e.g. 'jeez', 'weeping', and 'miss', which could be potentially beneficial for restricting the set of possible emotions present in text. This behavior may be avoided by further tuning of the threshold or by adapting more advanced feature selection methods, e.g. selecting terms having high PMI score with positive or negative emotions as we do in chapter 6.

Emotion	Top 20 *OlympLex* terms per category
Involvement	very interesting, bring it home, encouraging, ready, entertaining, encouragement, cant wait, c'mon, let's do this, looking forward to, comeonteamgb, curious, keep it up*, kill it tonight, pretty fly*, curiosity, neat, this is crazy, got your back, this is exciting*
Amusement	lolol, hysterical, amusing, lmfao, hilarious, joking, funny, amused, hahahahahaha, humorous, lmao, haha, hahahaha, humor, lol, hahaha, freaken, laughing, f'n, rofl
Pride	standing ovation, atta girl, we did it, whoop whoop, conquered, they did it, we were incredible, yesssss, she's the best, proudtobeanamerican, stoked, awesome job, wohoooo, get it girls, killing it, we won, pumped up, usa usa usa, shazam, steady as a rock
Happiness	hurray, woo hoo, jubilant, sohappy, elation, all right, smiling, yes yes yes yes yes, winner*, yippee*, excellent work*, job well done*, love this girl*, hell yeah, yes baby, lifted my spirits, amazingly perfect, woiii, gogabby, yahoo*
Pleasure	much fun, utter amazement, hunky, enjoying, this is so exciting, loving it, thrilling, enjoyed, hotties, fun, entertained, stunningly graceful*, endless respect*, best ever, thank you so much, smiley, she's amazing, entertainment, pleasurable, exciting
Love	hatred for none, obsession, love you, adorable, <3, 2cute, looking good, cuties, tenderness*, sexy, adore, in love*, luv*, sweet, grateful*, thank you so much*, attractive*, grace*, cute, adoration*
Awe	superhumans, how the heck, very impressed, question, awestruck, insanely talented, strength, impressed, spectacular, astonished, beyond me, speechless, admiring, astonishing, talented, oh my goodness, astonishment, dazzling, beautiful to watch, sooooo good
Relief	thank god, sigh of relief, phew, panic over, relief, about time*, relieved*, relaxed*, yayyyyyy*, incredulous*, satisfied*, accomplishment*, finally*, satisfaction*, content*, thank you*, glad*, defeat*, nuts*, oooh*
Surprise	that was crazy, surprising, wowsers, that was amazing, out of this world, shocked, shock, holy crap, holy fuck, feeling awe, surprise, holy cow, surprised, no way, amaze*, stunned*, oh my god, yikes*, incredulous*, woah*
Nostalgia	desire, conquer*, the bomb*, yes yes yes*,):*, grace*, oh my god*, wish*, buzzing*, crushed*, fine*, loved*, interest*, regret*, hold my breath*, :')*, goteamgb*, boring*, luv*, winning*
Pity	feel sorry for, poor ukraine, pitiful, feel so bad, what a shame, feel for, empathy, feel so sorry, sorrow, painful, sympathetic, heartbroken, aww, disapointed*, awww, too bad, feel bad*, pity, poor girl, poor
Sadness	:'(, heartbreaking, poor jordyn weiber, very sad, feel bad*, unjust, such a shame, heartbroken*, crushed, regret, a shame, empathy*, sympathy*, how sad*,):, sad, depressed*, sympathetic*, uncomfortable*, noo*
Worry	scared, frightened, fearful, nervousness, hold my breath, uh oh, anxiety, worried, worry, fear, exhausted, pity compassion, stressful, scare, how sad*, heartbreaking*, oh no, crushed*, uncomfortable*, tension
Shame	lazy, pathetic, remorse, shameful, booing*, facepalm*, awkward, uncomfortable, embarrassment*, dissapointed*, ashamed*, ouch*, oops*, shame*, rude*, wondering*, awful*, dumb*, a shame*, disgust*
Guilt	stink, eating disorder*, guilt*, sorry*, negative*, feel so sorry*, astonishing*, too bad*, exhausted*, blame*, regret*, horrible*, feel so bad*, yikes*, at least*, nbcfail*, guess*, addicted*, rude*, :(*
Regret	bummed out, this sucks, saddened, let down, not happy, sadly, oops, disappointed, not fair, bummer, what a pitty, nooooooo, poor quality, not gooood*, depressed, shame, what happened, dammit, unhappy, annoyance*
Envy	jealous, disapointed*, oh well*, eating disorder*, nooooooo*, enthralled*, embarrassment*, cheated*, annoying*, fml*, dammit*, neat*, hideous*, favorite*, wish*, lazy*, crushed*, satisfied*, ugh*, epic*
Disgust	squicks me out, gross, disgusting, ugly, offended, dislike, terrible, nasty, yuck*, hideous, loser, ashamed, odd*, prick*, lame, scorn, displeasure*, boring*, disgust*, horrible
Contempt	sobbing, get over it, get it right, asshole, not gooood*, dumb, facepalm, fuck you, nbcsucks*, what the hell*, loser, mean, screwed up*, scare*, dickheads*, disgrace, disgusted*, contempt, sucks*, frustration*
Anger	this is ridiculous, stupid pricks, fed up, frustrating, ignorant, furious, irritating, stupid japanese, still annoyed, fuming, pissed off, screw this, fuck dis, sort it out, piss me off, asswad, damn it, annoyed, stupid, irritated

Table 4.2: The top 20 terms from *OlympLex* for each emotion category. We report only the terms that were annotated at least 4 times as emotion indicators. The terms are ordered by decreasing the weight of the corresponding emotion category within their emotionalities. The terms marked with * do not have the highest weight for the stated emotion.

Based on the found emotionality $\vec{p}(d)$, we can extract two types of labels for the text document: polarity label and emotion multi-label. The output polarity label of our classifier is the dominant polarity within the emotionality, i.e. a polarity having the highest sum of the weights for corresponding emotions. The output emotion multi-label is defined as a set of dominant emotions in the emotionality. This set contains the emotions having the highest weights p_i.

4.5.4 Experimental Evaluation

We evaluate our lexicon as a classifier on the SREC corpus. As we build this lexicon essentially out of the same data, we apply ten-fold cross-validation to avoid possible overfitting. The existent universal lexicons are used for benchmarking. As they require no training, we test them directly over the full available data.

Polarity Classification

We consider the basic polarity classification task with 3 classes (*Positive*, *Negative* and *Neutral*). We use only $1,826$ tweets that have one dominant polarity based on the workers' answers. This dominant polarity is taken as a true polarity label of a tweet.

We compare our classifier (applied within polarity classification settings) with three existent sentiment lexicons and three emotion lexicons. As sentiment lexicons, we test GeneralInquirer [SDSO68], Bing Liu's lexicon [HL04], and OpinionFinder [WWH09]. All of them provide the lists of positive and neutral terms. As emotion lexicons, we test WordNet-Affect [SV04], NRC [MT13], and GALC [Sch05] (in the stemmed format). These lexicons enumerate the terms for each given emotion category. To apply them for polarity classification, we generate the lists of positive and negative terms from them: we assign a polarity to a term based on the polarity of an emotion associated with it. All of those lexicons are general-purpose lexicons, which are suitable to apply to any domain. To classify the text into the polarity with a given polarity lexicon, we apply a lexicon-based approach, similar to ours. We sum up the number of found lexicon terms in the tweet text for each polarity category and output the polarity having the highest sum value. If no polar terms are found, the *Neutral* polarity is assigned. Furthermore, if two polarities have the same sum weight, the output polarity is *Neutral* as well.

To evaluate the experiment results, we use the standard classification evaluation measures: accuracy, precision, recall and F1-score. We consider only non-neutral classes (*Positive* and *Negative*) for precision and recall. Table 4.3 shows the results of our classifier, compared with other presented lexicons. The proposed lexicon *OlympLex* outperforms every other lexicon, both in terms of accuracy and F1-score. This is because even though other studied lexicons have high precision, their recall is relatively low; whereas *OlympLex* has both high precision and recall. As *OlympLex* was the only lexicon fitted to the Olympic gymnastics data, its superiority reveals the advantage of domain-targeted lexicon construction and the ability to capture domain-specific emotional expressions.

Lexicon	P	R	F1	A
OlympLex***	**81.7**	**73.2**	**77.2**	**72.5**
BingLiu	80.4	52.9	63.8	53.6
OpinionFinder	66.0	46.6	54.6	46.6
GeneralInquirer	69.8	44.4	54.3	44.5
NRC*	60.6	39.7	48.0	40.4
WnAffect*	78.6	28.1	41.4	30.1
GALC*	81.6	25.6	39.0	27.9

Table 4.3: The results of polarity classification on the SREC data. We compare *OlympLex* with other sentiment and emotion lexicons. P=precision, R=recall, F1 = F1-score, A=accuracy. The sign * marks lexicons that employ several emotion categories instead of only polarities.

Emotion Classification

We evaluate emotion recognition results in the setting of a multi-label classification problem. The output is a set of labels instead of a standard single label answer. Our classifier returns a multi-label output O_C for a tweet extracted from the tweet emotionality $\vec{p}(d)$. The set of emotion labels given for this tweet by workers forms a true output – a set of true labels (O_T) of emotion classification.

As a baseline for multi-category emotion classification we consider the GALC lexicon [Sch05], described in section 3.3. In these experiments, we use its original stemmed format and detect any word in the text that starts from its stems.

Multi-label Evaluation We use the standard evaluation metrics adapted for multi-label output [TK07]. For each tweet, we compute the precision $P = \frac{|O_C \cap O_T|}{|O_C|}$, which shows how many of emotions outputted by the classifier are correct; the recall $R = \frac{|O_C \cap O_T|}{|O_T|}$, which shows how many of true labels are found by the classifier; and the accuracy $A = \frac{|O_C \cap O_T|}{|O_C \cup O_T|}$, which shows how close the sets of the classifier and true labels are. These values are averaged among all applicable tweets. For computing precision and recall, we use only the tweets with non-neutral answers in O_C and O_T correspondingly (meaning that *No emotion* label is not present in a set).

We show the comparative results of the two lexicons in Table 4.4. Compared to the GALC baseline, our classifier has both higher precision and recall. Higher recall is explained by the

Lexicon	P	R	F1	A
GALC	49.0	10.2	16.8	12.5
OlympLex	**53.5**	**24.9**	**34.0**	**25.4**

Table 4.4: The aggregated results of multi-label emotion classification on the SREC data. We compare *OlympLex* with the baseline GALC lexicon. P=precision, R=recall, F1 = F1-score, A=accuracy.

		GALC			*OlympLex*		
Polarity	Emotion	P	R	F1	P	R	F1
Positive	Involvement	52.4	2.4	4.6	49.4	17.6	**26**
	Amusement	51	11.6	18.9	55	24.6	**34**
	Pride	89.6	6.7	12.5	60.8	59.4	**60.1**
	Happiness	46.3	8.8	14.8	45.1	9.8	**16.1**
	Pleasure	44.8	5.9	10.4	48.8	17.9	**26.2**
	Love	38.1	27.4	**31.9**	48.0	8.2	14
	Awe	42.9	6.7	11.5	54.2	23.7	**33**
	Relief	100	17.1	**29.2**	50	4.9	8.9
	Surprise	38.3	9	**14.6**	33.3	6	10.2
	Nostalgia	20.5	14.5	**17**	28.6	3.2	5.8
Negative	Pity	75	2.5	4.9	57.8	31.4	**40.7**
	Sadness	52.5	19.6	**28.6**	41.7	9.3	15.3
	Worry	54.8	21.5	**30.9**	43.2	15	22.2
	Shame	18.5	9.8	**12.8**	25	3.9	6.8
	Guilt	25	5.6	**9.1**	0	0	-
	Regret	53.3	3.4	6.4	36.3	12.4	**18.5**
	Envy	100	11.1	20	55.6	13.9	**22.2**
	Disgust	50	1.4	2.8	39.4	9.4	**15.2**
	Contempt	-	0	-	42.1	4.7	**8.5**
	Anger	48.4	10.8	17.7	53.3	26	**35**

Table 4.5: The results of emotion classification on the SREC data at per-category level. We compare *OlympLex* with the baseline GALC lexicon. P=precision, R=recall, F1 = F1-score.

fact that our lexicon is larger and contains longer *n*-gram terms. In addition, it includes not only explicit emotion expressions (e.g. *sad* or *proud*), but also more implicit ones (e.g. *yes* or *mistakes*).

Per-Category Evaluation Another way to evaluate the output of multi-label classifier is to evaluate it for each emotion category separately. For each category we compute precision, recall and F1-score. The results of this evaluation in comparison with the benchmark GALC lexicon are presented in Table 4.5. Overall, our lexicon performs better on most of the categories (12 out of 20) in terms of F1-score. The highest F1-score is achieved for the Olympic-related emotion of *Pride*. The categories where GALC outperforms *OlympLex* are again mostly those that have less data in the annotations. For some of those categories, particularly for Relief, Nostalgia, and Guilt, the superiority of GALC can be explained by the lack of corresponding explicit terms in the built lexicon. Therefore, we hypothesize that adding the known explicit terms in the resultant lexicon can help further improve the quality of emotion recognition.

4.6 Discussion and Future Work

We designed an annotation task for simultaneous data labeling and indicator elicitation. In contrast with more common goal of having annotating data for building and validating the emotion recognition system, our task directly aims at building an emotion lexicon for fine-grained emotion recognition based on the human-provided features. In this section, we discuss the main findings revealed by this work and suggest directions for future work.

Emotion Granularity We show that, despite the relatively high granularity of chosen GEW emotions, every emotion category is present in the annotated dataset. Such a detailed emotion representation allows us to create a more accurate description of the sentiment evoked by the chosen event of the Olympic Games. For instance, we find that *Pride* is the dominant labeled emotion, followed by *Involvement*. If we have used less coarse-grained emotion models of commonly recognized basic emotions, we would not have been able to capture these details. This advocates for using more tailored, fine-grained emotion models.

The fine granularity of annotation results in the presence of several emotional states per tweet, in some cases due to co-experience of emotions as a mixed state and in other cases due to presence of several emotions towards different subjects (e.g. different athletes). This confirms the need to treat emotion recognition problem as multi-label classification, instead of multi-class classification.

The fine granularity of emotion categories also makes it more challenging to achieve high performance scores in emotion classification. An interesting future direction could be to evaluate the average human performance with the same metrics as for classifiers in order to establish the upper-bound on the classifiers' performance. It could be also beneficial to incorporate the notion of emotion similarity in performance evaluation, as it would allow us not to penalize classifiers on similar, but not-matching answers. More research is required on how to define similarity scores between the studied emotion categories.

Advantage of Human-Generated Domain-Specific Lexicons We applied the GALC lexicon to all the tweets related to the Olympics and found that its terms are found in 31% of data. This indicates that people do express their emotions explicitly with emotional terms. However, a list of currently available explicit emotional terms is not extensive. For instance, it does not cover slang terms. Moreover, people do not limit themselves to only explicit emotional terms. Our method highlights the possibility of employing the human common knowledge in the process of extracting novel emotion-bearing features.

Our lexicon, constructed based on the answers provided by non-expert humans, was built with a context-sensitive method and includes domain-specific expressions. This led to a significantly higher recall and accuracy on the target domain, compared to the general-purpose lexicons. We benchmarked the cross-validated version of created *OlympLex* lexicon with the existing universal-domain lexicons for both polarity and multi-emotion problems. In suggested settings, we showed that it can outperform general-purpose lexicons in the binary classification due to its domain specificity. We also obtained significant improvements over the baseline GALC lexicon, which was the only preexisting one compatible with the GEW.

However, high domain specificity of the created lexicon and restricted variety of data used in its construction implies possible limitations of its usage for other types of data. Its porting and generalization to other domains is one of the future directions.

Potential Tweet Selection Our analysis of the crowdsourced answers reveals several potential ways to improve the answers' usefulness by selecting more beneficial tweets to label. For instance, we observe the imbalance in the annotated emotions, as no constraints were placed to guarantee equal distribution of emotions or polarities. This results in extracting less textual indicators for negative emotions, such as *Sadness* and *Worry*. In order to increase the discriminating power of a built emotion recognition system, it can be better to provide equal amount of information on each emotion class, or at least on different polarities. Furthermore, we show that not all returned emotion indicators actually appear in other tweets, making them not applicable. Selecting tweets with higher number of applicable potential emotion indicators can help increase the coverage of an emotion recognition system built from the same amount of annotated tweets. Incorporating different criteria for selecting tweets to label in order to maximize the usefulness of the labels follows the principles of active learning [Set09], and thus can be beneficial to apply with annotation iterations [Set11].

Potential Task Redesign Our answers' analysis also shows the potential advantage of asking workers for more information about the tweet. For example, knowing whether the tweet's only purpose is to share news or facts (even if emotion-provoking) and whether the meaning of the tweet is non-understandable or contradictory would allow us to process such tweets separately from more explicit and clear emotional tweets.

We also find that the provided additional emotion indicators are frequently unique and not-repeated in the annotated data. Thus, to include such answers in the lexicon, we can either design a separate validation task or implement an iterative task flow where the new tweets to label will contain new expressions to validate.

In the future, it could also be interesting to explore how to extract similar emotion annotations using different gamification principles [DDKN11, VAD04, BMF14].

4.7 Chapter Summary

In this chapter, we present a human computation method for building a domain-specific fine-grained emotion lexicon. Our designed annotation task, unlike most previous approaches, involves both document-level emotion labeling and emotional features extraction. We show that non-expert human annotators, using their common sense, can successfully attach emotion labels to tweets, and also extract relevant emotional features. Using their answers, we carefully construct a linguistic resource for emotion classification. We show that such human-generated lexicon can be successfully used in fine-grained emotion classification, outperforming various existing lexicon-based methods. Compared to the previous domain-independent lexicons, the domain specificity of our lexicon, imposed by the presence of domain-specific emotional expressions, makes it more accurate for analyzing emotions within a given domain. The suggested method can be reused to construct additional lexicons for other domains.

An important aspect of our work is the fine granularity of the studied emotions. Recognizing fine-grained emotions enabled us to capture the subtleties of the emotional responses in the target domain—tweets regarding the 2012 summer Olympics in London. In this dataset, we found that the prevalent emotion is *Pride*, a detail which is unattainable using common coarse-grained emotion models. This research also shows that annotators can indeed distinguish the nuances of 20 fine-grained emotions from the Geneva Emotion Wheel (GEW). They agree on a specific dominant emotion label at least in one third of the annotated data. However, the frequent presence of mixed emotional states validates our approach to model emotion classification as the multi-label classification task. This also justifies the need to model weighted emotion associations in the emotion lexicon instead of storing hard links to emotions.

While such crowdsourcing approach to build emotion lexicons is shown to be successful, our analysis of the collected answers suggests that the task design itself should prevent errors in the annotation and stimulate workers to return truthful, acceptable answers. We investigate in the next chapter two different preemptive mechanisms aiming to improve the quality of answers in the crowdsourcing environment.

5 Preemptive Quality Control for Crowdsourcing

5.1 Introduction

In the previous chapter, we showed the possibility to use crowdsourcing for collecting emotion annotations of microblog posts. However, the quality of the answers returned by non-expert, non-experienced, remote workers presents an important concern for the usefulness of such crowdsourcing process. Due to anonymity and diversity of skills, each person's capability and motivation for performing the desired task is not known *a priori*. The problem is further pronounced for online labor markets, such as Amazon Mechanical Turk, or AMT (mturk.com). Unfortunately, the use of monetary rewards as motivation to perform tasks may attract untruthful or unskillful workers. Therefore, the quality control measures play a crucial role in making such human computation system viable.

The standard techniques for quality assurance, such as combining answers from different workers and posterior removal of non-satisfactory answers [IPW10], assume the existence of a correct answer. Our emotion annotation task requires subjective evaluation of given objects (text documents in our case), which is considered to be a judgment process. Thus, it is an example of judgment tasks. This class of tasks comprises many other tasks allowing to assess the content of objects, including evaluation of web pages' quality [KCS08], assessment of answers' relevance [Kaz11], and collection of opinions about products. The common characteristic of judgment tasks is that human raters are asked to judge objects (e.g. documents) with regard to their own opinions, beliefs, preferences, and feelings, as well as common sense. Therefore, judgment tasks are subjective in nature and, as such, lack unique correct answers. In the task of text emotion labeling, one worker can label the text "I have spent the full day shopping" as *Pleasure*, while another one can label it as *Regret*. Both answers should be considered as valid because both emotional experiences are possible in this situation. All this makes it difficult to establish the validity of a specific answer based only on the (dis-)agreement with the majority label, especially when only limited number of answers is available.

One solution is to detect repeated disagreement with peer answers, as adapted in previous section for the posterior quality control measure. But that requires having multiple answers from the same worker. Another solution could be to use “trap door” questions for validation of workers’ answers. Yet, we would need to use easier, less ambiguous texts to label in order to ensure a specific emotion to be correct, which could make these questions easier to recognize by malicious workers trying to answer only them correctly. Therefore, a special care has to be taken to preemptively ensure workers can and will put their effort to provide appropriate truthful answers to all questions.

In this chapter, we study the effects of two prominent mechanisms of such preemptive quality control: including obligatory tutorials and framing of financial incentives. Both studies employ our task of emotion annotation as a case study.

The first quality control mechanism we investigate is the inclusion of tutorials for inducing shared task understanding. It is directed to teach workers how to perform the task properly in order to increase their ability to produce acceptable answers. The main goal of tutorials is to explain to people the task they are going to perform in the form of simple understandable instructions. To achieve this goal, our designed tutorial includes detailed instructions for each task action as well as examples of potential answers. It additionally requires passing the comprehension quiz that tests for the correct understanding of the given instructions. Strict guidelines for offline annotation and inclusion of clear instructions in crowdsourcing are considered to be essential for collecting reliable human annotations [WHA12]. Using tutorials aims at establishing a common comprehension of the judgment process specifics and task requirements. With an increase of task comprehension, the quality of the result work is deemed to increase. Yet, the exact quantification of the tutorials’ impact remains understudied, especially for the class of judgment tasks. In the following study, we are interested in the general impact of the tutorial on the workers’ task performance and behavior. We investigate how inclusion of the tutorial that teaches task comprehension affects the quality of the workers’ answers in crowdsourcing, as well as their engagement. The findings suggest that including the tutorial for our emotion annotation tasks helps both to alleviate the common misunderstandings of the instructions and task questions and to make workers produce more content per one annotation. At the same time, the additional effort required to pass the tutorial stays within reasonable limits. Our analysis also shows the potential of using the same tutorial format as a qualification test based on the instructions’ comprehension quiz.

The second quality control mechanism that we study is inclusion of financial incentives in the form of qualitatively described bonuses. This is directed to motivate workers to put more effort into doing the task properly in order to obtain a bonus reward. Previous research showed the potential of motivating workers by introducing additional financial incentives, such as bonuses for good-quality or agreeing answers [HSSV15, Har11]. However, the conditions of such bonuses are usually formulated quantitatively in terms of mathematical formulas, whereas a lay person from MTurk is unlikely to fully understand the implications of such computations, especially in the case of non-mathematical judgment tasks (such as emotion labeling). We study in this work the effects of alternative, qualitative formulation of bonuses, which are described in plain English without any mathematical formulas. We refer to this approach as qualitative framing. We perform an online crowdsourcing experiment to investigate which bonus formulation is most effective. It compares several peer-oriented incentives (e.g. asking for agreement with other workers [FPTJ14]) with other social-based incentives (e.g. asking to comply with expert answers or rely on personal judgment of quality [SHC11]). We analyze their effects on the workers' answers quality, while varying the difficulty of the tasks. The results demonstrate better workers' performance under a well-formulated incentive framing inspired by game-theoretic bonus schema of Peer Truth Serum [FPTJ14]. This framing asks workers to provide answers that both agree with those of other workers and at the same time novel. The positive effect is observed only for categorical labeling and only when the difficulty of the task is high, while when the task is easy there is no difference of which incentive to use.

In short, we investigate the effects of two different quality control measures aiming to improve the quality and reliability of the annotations, as well as to motivate workers to put more efforts. We show that including a well-designed tutorial can help to avoid workers' misconceptions, while also motivate workers to return more information within each annotation. By varying the qualitative formulation of bonus incentives, we show that when the tweets are more difficult to annotate it can be beneficial to ask workers to think about other workers' answers while providing an original answer. This line of research helps finding the optimal presentation of the annotation task.

5.2 Related Work

Ensuring Quality of Crowdsourcing A variety of methods exist for improving the quality of the crowdsourced answers (for a more detailed review, read [QB11, ABI$^+$13]). One group of such techniques consolidates the posterior measures. They are applied after collecting the task outputs and imply statistical filtration of malicious answers or specification of answer aggregation methods to detect true labels [IPW10]. Such methods were designed for tasks with definite or verifiable answers. In case of judgment tasks, the variance can be higher both due to the differences in personal opinions and poor answers. The preemptive quality control measures should be a prior step allowing collecting the data of better quality before their posterior refinement. Such measures try to enforce the desired level of quality before acquisition of the results. Tutorials and motivational incentives belong to this category of methods, as well as worker's selection, task practicing, screening processes, and task design itself.

Selecting workers having the desired initial prerequisites is one mechanism of preemptive quality control. AMT allows to specify that only workers having specific qualifications are allowed to perform the task, e.g. those that have high approval rate in the previous tasks [Kaz11], or those from a certain location to target the demographics of the workers. Manually designed qualification tests are also possible, e.g. requesters can test the level of English comprehension for the language labeling tasks [HG14]. Another approach is to screen for non-attentive or malicious workers by including verifiable validation questions that test workers' attention and understanding of the task material [KCS08, MT13, DHSC10]. For example, Mohammad and Turney [MT13] included the validation question on the correct understanding of the meaning of the word being labeled. Simplifying the task flow for workers is another way of preemptive quality control, as the simplified task requires less cognitive effort and thus leads to less errors [CTIB15]. The simplicity of visual presentation [FKTC13], split of task units and payment schemas [MKC+13, KKMF13] also have effect on the result answers' quality. Optimization of parameters for monetary reward, such as the amount of work per pay rate, acceptance conditions, bonuses and penalizations policies, can be adopted to obtain the appropriate quality [HZP+10].

In prior research in human computation, several works mentioned the improvements in quality due to inclusion of the training or comprehension validation process, which can be considered as a tutorial process. For example, practicing on the task, especially when feedback is given, was shown to improve the quality of the results in objective tasks with specific answer [OSL+11, SDFF12]. Heer et al. [HB10] suggested to present a practice judgment task in the form of fixed multiple-choice questions with only one clearly correct answer. Giving clear instructions seems to be an indispensable step for achieving more reliable crowdsourced annotations [SBDS14].

The quality of the collected answers also depends on whether workers have motivation to perform a better job. Reward schemas and persuasive social messages were suggested as incentives to increase workers' motivation [SHC11].

We further review the studies of the two prominent ways of preemptive quality control that we will discuss in this chapter: tutorials and motivation incentives.

Tutorials Research Tutorials facilitate the learning of specific knowledge or skills. Being a method of educational instructions, computer-based tutorials were studied within the *instructional-design theories* [Rei13]. An instructional-design theory "offers explicit guidance on how to better help people learn and develop". It suggests the framework for formalization of the possible methods and design principles for instructions, along with the situations when (not) to apply them. However, such theories are rather conceptual and require further specification for concrete applications. Specific examples of tutorials were investigated thoroughly within *several domains*, including manuals for computer software [BBMC89, CSKFMR87, CR86, Har95, KP05], online library tutorials [Mes12], and tutorials in video games [AOL+12].

Research on tutorials generally involves investigation of different tutorial parameters. A number of works aimed to study the influence of various learning methods, such as learning by examples or by principles [BBMC89, CR86, EC11]. The different styles of instruction presentation were also investigated, from incorporation of visual illustrations, with or without animations and videos [Har95, Mes12], to varying levels of interface interactivity, visibility and freedom of actions [AOL+12, HT07, KP05]. Our work does not aim to study possible tutorial parameters, but rather only quantify the impact of the tutorial inclusion in the context of human computation.

The previous researchers on tutorials in other, non-crowdsourcing domains suggest to measure efficiency and effectiveness of tutorials. The design of our metrics for tutorials evaluation follows these best practices from the previous works in other domains. The efficiency is normally measured via the statistics of the learners' performance in tutorial, while the effectiveness of tutorials is evaluated via learners' performance on the real tasks requiring usage of the knowledge presented in the tutorial. In such evaluations, researchers record time spent on tutorial and real tasks, as well as the number of errors [BBMC89, HT07, KP05]. Some studies take into account the opinions of the participants [Har95, Mes12]. Andersen et al. [AOL$^+$12] investigated the impact of different tutorial settings on the users' *engagement*, using the online games as a case study. The engagement was measured from the behavioral data, such as the total time spent on the game, number of completed levels, and return rate. The authors have found that tutorials do not always have a positive influence, which indicates the importance of testing the tutorial implementations. Based on these works, we suggest to evaluate the tutorial impact both on the quality and engagement of the workers, as well as to investigate the workers' performance and effort in the tutorial.

Unfortunately, in crowdsourcing domain, the exact effects of tutorials are under-studied because inclusion of the tutorial or any sort of instructions is rather taken for granted by the task developers. We attempt to quantify the effects that a tutorial can bring to a judgment task, with the example of our designed human computation task for emotion annotation.

Incentives Research As in online labor market (e.g. MTurk) the primary incentive for doing tasks is a monetary reward, the effects of financial incentives are widely studied in the crowdsourcing research. However, the evidences of their positive effects on workers' performance are inconclusive. Several researchers showed that payment magnitude has no effect on the quality of the workers' answers, only on their willingness to perform more tasks [MW09, RKK$^+$11]. At the same time, placing additional performance-based bonuses (i.e. those that reward better workers' performance) was shown to have positive effect for crowdsourcing tasks [Har11, HSSV15, FPTJ14]. The experiments of Ho et al. [HSSV15] suggest that the effectiveness of financial bonuses can depend on the type of a studied task, on the amount of bonus and base payment, as well as on the criteria for obtaining a reward. There are additional evidences that such effects can depend on the difficulty of the task [MKC$^+$13]. In this work, we evaluate how much the effects of financial incentives depend on the framing of the bonus criteria, while varying the difficulty of the task.

The effects of extrinsic, financial incentives can be intertwined with different intrinsic incentives, such as fun, altruism, or glory. Rogstadius et al. [RKK$^+$11] showed that workers' accuracy can be improved significantly through intrinsic motivators (by framing the task as for charity), especially when the monetary reward is low. Shaw, Horton, and Chen [SHC11] extensively compared 14 different social, financial, and hybrid incentives for a content analysis task.

Only financial incentives that asked to prospectively think about the responses of their peers produced a more accurate output. Paying bonuses based on the answers matching other workers' answers was confirmed to improve accuracy in counting tasks [HF13, FPTJ14]. Our work aims at investigating alternative formulations of bonus criteria that are also based on peer agreement and social conformity, but framed in a qualitative simplified language without referencing to any mathematical formulas.

Another distinguishing property of our experiment is the use of a judgment task. The described studies of financial incentives mostly used syntactically generated tasks with directly verifiable answers, such as counting objects or finding spelling mistakes. That made it possible to set objective bonus criteria, e.g. percentage of correct answers or range of difference from the exact quantitative answer (e.g. ±3). Yet, in reality, not all crowdsourcing tasks have unique irrefutable ground-truth for validation and bonus computation. Punishing a non-conforming, but truthful answer in a judgment task would be undesirable. Our task of emotion annotation requires less strict bonus formulations.

Overall, using a well-designed judgment task of emotion annotation, we conduct the online experiment to discover whether any particular framing of financial incentive (formulated with the qualitative, non-mathematical description of bonus criteria) is more effective for obtaining better-quality answers. We also analyze the differences in such effects depending on the level of task difficulty, where we expect easy condition to have less or no effect of specific incentives compared to the difficult condition.

5.3 Impact from Including a Tutorial

We start our analysis of the preemptive quality control measures by analyzing the effects of the tutorial inclusion. To understand such effects, we compare the data collected previously from two batches for our emotion annotation process: one batch with workers without the tutorial, and another one with an obligatory tutorial. This comparison reveals the positive effect of the included tutorial on the quality of the workers' answers, both in terms of emotion labels and emotion indicators.

5.3.1 Study Data and Methodology

In this study, we analyze the observed differences of including the tutorial in the designed emotion annotation task. With this goal, we analyze the data from two last crowdsourcing batches described in section 4.4.3. Both of the batches used the same interface for labeling the tweets based on GEW emotions and indicators (section 4.4.1), while the first one (batch 4, or B_4) did not include any tutorial, and the second one (batch 5, or B_5) included the obligatory tutorial designed to help workers understand the task (section 4.4.2). Batch B_5 additionally included the manual review process for workers aiming to submit more than 50 answers, i.e. to label more tweets. The sizes of the batches were also different: the first batch B_4 aimed at

Actual batches	B_4	B_5
Tutorial	None	Obligatory
Limitation on # of answers	None	50; Authorization for > 50
# of tweets	200	1787
# of answers	800	7148
Transformed batches	B_{-T}	B_{+T}
Limitation on # of answers	50	50
Limitation of workers	None	No repeated
# of workers	57	654
# of answers	502	5667

Table 5.1: The summary of the batches used for our tutorial analysis. B_4 and B_5 represent the actual setup without changes; B_{-T} and B_{+T} represent a synthetic setup restricting differences to the tutorial inclusion.

a pilot annotation of small amount of data, and the second batch B_5, standing for the final annotation of the large number of the tweets. These differences between two batches are summarized in Table 5.1, under the section Actual batches.

To analyze the tutorial effect, we have to exclude the possible influence of the assignment number limitation. To achieve this goal, we use only the first 50 answers from every worker in both batches, resulting in the synthetic batches B_{-T} and B_{+T} respectively. This approach simulates a design where both batches have the same limitation on the maximal number of submitted assignments per worker without permission to continue. However, in our experiments several workers (15) participated in both batches: first in B_4 without tutorial, and later in B_5 with tutorial at the beginning. We exclude the answers of those workers from the batch B_{+T} as their task performance could also be contingent on the potential training effect appearing because of repeated practicing on the task.

As a result, we have the data from the two experiments differing in the tutorial's presence. Table 5.1, section Transformed batches, describes the statistics on these two resultant batches: B_{-T} without the tutorial and B_{+T} with the tutorial. Their comparison allows for the analysis of tutorial's effectiveness, while the batch with tutorial alone provides statistics of tutorial's efficiency.

5.3.2 Effectiveness: Quality and Engagement Analysis

We quantify the effect of the tutorial inclusion on the output quality and observed engagement of workers. All considered quality metrics are calculated first as the average metric for workers answers, and then averaged across the workers. Thus, we include for this analysis only

		B_{-T}	B_{+T}	
Statistics on the workers' engagement				
% of workers submitted only one answer		40.4	32.6	$p = 0.116$
# of HITs per worker	median	2	3	
	mean	8.81	8.67	$p = 0.470$
average time spent on one HIT (seconds)		87.5	94.2	$p = 0.204$
average total time spent on annotation (seconds)		544.7	611.9	$p = 0.304$
Statistics on the quality of workers' answers [a]				
average polarity conformity		87.2	90.2	$p = 0.129$
average emotion conformity		*46.4*	*55.8*	$p = 0.026$
average strength conformity		*70.1*	*76.5*	$p = 0.048$
average # of tweet indicators per answer		*1.41*	*1.77*	$p < 0.001$
average # of additional indicators per answer		*1.46*	*2.09*	$p < 0.001$
average % of accepted tweet indicators per answer		*78.3*	*93.5*	$p < 0.001$
average % of accepted additional indicators per answer		*72.1*	*95.9*	$p < 0.001$

Table 5.2: Comparison of workers' performance in the batches without the tutorial B_{-T} and with the tutorial B_{+T}. All the metrics are averaged on per-worker basis. The results indicate the gain in annotation quality after the tutorial was included in the task. Statistically significant changes (p-value < 0.05) are highlighted in italic.

[a] Only workers who submitted more than one HIT are considered in this case in order to have a more representative averaging.

workers who answered more than once. We apply the criterion of answer conformity[1] and compute *Polarity conformity* of a worker's answer with respect to the considered polarity classes (*Positive*, *Negative* and *Neutral*); *Emotion conformity* with respect to the available 21 emotion classes (20 emotion classes of GEW plus *No Emotion*); and *Strength conformity* with respect to four strength classes: *Low*, *Medium*, *High*, *No* (with *No* corresponding to *No emotion* answer in the wheel). We adapt again the idea of the answer acceptability: we define the acceptability criteria for both types of emotion indicators and compute the *percentage of accepted indicators* per answer. A tweet emotion indicator is accepted if it appears in the text of the judged tweet, while an additional emotion indicator is accepted if it does not coincide with one of emotion names used in GEW. As was discussed, the last was a common misconception workers had that we tried to avoid. Even though the submission of such non-accepted answers could be avoided by changing the task design, they were allowed in the current more straightforward implementation. In addition to the acceptability metrics, we include metrics evaluating the amount of worker-generated content. We record the *number of*

[1] To overcome introduced imbalance in numbers of answers per tweets in synthetic batches, we used all available answers for the tweet given in batches B_1 and B_2 for computing of answer conformity measures. However, as expected, only answers of B_{-T} and B_{+T} are taken into account in averaging the result statistics.

provided emotion indicators, both tweet and additional. We suppose that the more engaged the workers are, the more indicators they will return.

The results of batches comparison on these metrics are presented in Table 5.2 as Statistics on the quality of workers' answers. For all considered metrics the workers in the batch with tutorial B_{+T} have consistently better quality than in B_{-T} (i.e. without the tutorial). These gains are statistically significant for all metrics except polarity conformity. Higher emotion conformity signifies the importance of the tutorial in developing and establishing shared knowledge of the emotion judgment. The relatively similar level of polarity conformity can be due to two factors. First, workers can have the inherent higher consistency in use of polarity concept compared to use of emotion concepts. Second, no explicit instructions on the polarity distinction were given either in the task interface or in the tutorial. One can also observe the highly significant gain for all the metrics concerning the indicators, including their number and acceptability. This reveals ability of the tutorial to eliminate the common misconceptions in understanding of indicators concept.

We further evaluate the impact of the tutorial's presence on the workers' engagement. We compute general statistical metrics appropriate to assess implicitly the engagement level of workers for any repeatable human computation task. *Number of HIT answers submitted per worker*, both median and mean, reflects for how many times workers stayed engaged or motivated to continue performing our task. Similarly, we record the *percentage of workers submitted only one answer* without subsequent continuation (or non-return workers in other words). It shows the ability of the task to engage workers in the beginning. We also measure *time spent* per each tweet, as well as in total on annotation for each worker. These engagement statistics reveal the amount of effort workers input into tasks.

The statistics comparing the engagement level in two batches are presented in Table 5.2 as Statistics on the worker's engagement. One can observe a slightly higher retention rate in the tutorial-included batch, as shown by lower percentage of non-returning workers, higher median on the labeled tweets, and increased total time. However, none of the investigated engagement metrics showed a statistically significant change.

5.3.3 Efficiency: Tutorial Performance

To evaluate the efficiency of tutorial inclusion as an entrance barrier for non-serious workers, we captured the statistics on the tutorial requests and completions and measured the *drop-out rate*. We found out that only 78.6% of workers who were exposed to the tutorial finished it. This means that the other 21.4% of workers left the task after finding out about the tutorial existence or due to its difficulty. This number is not too high and we consider it to be acceptable in elimination of non-serious workers.

To understand better the behavior of workers during the tutorial, we save *time spent* on each tutorial and quiz steps, as well as the *number of errors* made in quiz questions for all workers

Tutorial steps	S1		S2	S3
Quiz questions	Q1	Q2	Q3	Q4
Average quiz time (s)	38.9		13.7	14.7
Average step time (s)	30.4		26.0	14.4
Average error number	0.36	0.31	0.14	0.38
% of workers made errors	23.8	17.6	11.7	32.4

Table 5.3: Quantitative statistics of workers' performance in the tutorial. We report average spent time and number of errors in each tutorial's step.

who submitted the tutorial.[2] As our quiz questions present multiple-choice options, we consider as one error any submission of an incorrect answer. The values of these statistics, clustered into steps and corresponding questions, are presented in Table 5.3. One can see that question Q4 (on the additional emotion indicators) caused the most problems for workers, as shown by the highest number of errors on this question. This might be due to the shortness of time spent on the related instructions, indicating a potential drop in attention, or because of higher internal ambiguity of the related task question. Nevertheless, workers afterwards spent more time reviewing the full example in the last step (24.5 seconds in average). From the overall statistics, we found that the average time spent on the tutorial is 2.8 minutes, which is less than one third of the average total time spent on the tweets' annotations (10.2 minutes, without counting the tutorial time). We also found that, out of all workers completed the tutorial, 49.1% managed to finish it errorlessly. The average number of errors throughout the full tutorial is less than the expected number of errors in case of random guessing (1.2 vs. 5.33).

5.3.4 Tutorial–Task Performance Relations

To show the efficiency of tutorial in dependence with the result task performance, we investigate the correlation between the number of errors made in tutorial and the quality of later answers. For task performance, we adopt the same answer quality metric, using again the answer-based averaging, without grouping them per worker, but employing only all the answers from the batch B_{+T}. The dependency results can be found in the Table 5.4. The total number of errors in the tutorial is categorized into four classes to reflect possible differences in workers' quality: no error (0) corresponds to ideal workers, one error indicates non-attentive workers as we presume it could be made due to lack of attention paid or misunderstanding of instructions, $2-5$ errors stands for problematic but sincere workers as it is yet less than the number of errors with random choice, whereas more than 5 errors implies workers answering randomly. The results reveal that with the increase of the number of errors in tutorial the quality of later answers is decreasing. By setting the limit on the accepted maximum number of errors we can also detect possible non-serious workers. As by the nature of the tutorial,

[2] Unfortunately, at the time of running this experiment, we did not save those statistics for workers who quited the tutorial.

# of errors in tutorial	0	1	2–5	6–9
# of HIT answers	2998	1259	1177	233
polarity conformity	90.5	90.1	**86.7**	**71.2**
emotion conformity	57.4	55.6	**46.2**	**37.3**
strength conformity	76.9	75.9	**72.6**	**60.9**
# of tweet indicators	1.77	1.73	**1.54**	**1.25**
# of additional indicators	2.04	2.16	**1.70**	**1.41**
% of accepted tweet indicators	97.4	**92.8**	**88.1**	**72.8**
% of accepted additional indicators	95.6	**92.9**	**93.2**	**81.6**

Table 5.4: Dependency between number of errors made in the tutorial and the quality of the annotations. We highlight in bold the scores that are statistically significantly different from those of the workers with no errors in the tutorial (p-value $< 10^{-3}$).

we aim to teach all workers, we suggest to develop additional tutorial steps for such workers, instead of eliminating such workers. The workers would have to continue training until they learn or leave the task.

5.3.5 Discussion

The results of this study confirmed most of the speculated impacts of the tutorial. Due to the enforced training process of task understanding, the tutorial inclusion increased the quality of workers' answers for all main considered quality metrics. For example, we achieved a 19% increase for the metric evaluating the quality of emotion labels (emotion conformity) after the tutorial inclusion. Moreover, in that case, with the tutorial we were able to collect on average 54% more of emotion indicators, and their accuracy was higher on 35.4%, all statistically significant. The evaluation of tutorial's efficiency showed that the efforts required by the tutorial are acceptable and that the performance in the tutorial is indeed representative of the future performance on the task.

Note that this analysis was purely observational, and thus there could be other compounding factors that affected the outcome. For example, we run the two batches under comparison at different times and the tweets used in the batches were different (even though similar in nature). Additionally, we could not control for the workers who left the task during the tutorial session. The last factor is the most difficult to ignore, as we cannot then separate whether the observed difference in quality is due to the effect of teaching during the tutorial or the workers' self-selection, leaving in the pool of contributing workers only those who are more engaged and motivated to perform this task a priori.

5.4 Qualitative Framing of Incentives[3]

As a preemptive measure of quality assurance, it is desirable to motivate workers beforehand to put more efforts and provide better-quality answers. Previous research showed the potential of motivating workers by introducing additional financial incentives, for example, by providing bonuses for good-quality or agreeing answers [HSSV15, Har11]. However, the conditions of such bonuses are usually formulated quantitatively in terms of mathematical formulas, whereas a lay person from MTurk is unlikely to fully understand the implications of such computations. In this study, we report the results of a crowdsourcing experiment comparing the workers' answers quality while qualitatively framing financial incentives based on different principles for motivating workers. We formulate qualitative framing of rewards as describing bonus conditions with plain English, without using mathematical formulas. We investigate the effects of different social and peer-oriented formulations of framing on obtaining better-quality answers in crowdsourcing judgment tasks. These effects are studied while using again our emotion annotation in text as a test task.

We also consider two different levels of task difficulty in order to analyze whether the effects differ. It is expected that, when the task is easy, the quality of the work should not vary much under different incentive conditions. However, when the task is more difficult and requires particular effort from the workers, one can expect that the incentive used has more impact on the quality of the workers' answers.

Our experimental results verify this for categorical labeling: for difficult tasks, it is beneficial to use one specific and well-formulated incentive, while there is no difference which one to use for easy tasks. More specifically, we find that the Peer Truth Serum, which asks workers to both agree with their peers and provide non-common answers, gives systematically better performance results (in terms of label agreement and correctness) compared to the other tested incentives in more difficult task condition.

5.4.1 Experiment Design

We again use the same emotion annotation task in our experiment. It asks workers to annotate emotions expressed in given tweets from the writer's perspective. The experiment design schema is depicted in Figure 5.1. In the preview of the HIT, all workers were shown the participatory call for annotating 10 tweets with emotion labels and indicators. They were also informed that they will be asked to follow a short obligatory tutorial before doing the task, and to answer optional demographic and task feedback questions. The base payment was set to \$0.5 USD, and there was information about a chance to obtain an additional bonus of \$0.1 USD (without specific details shown in the preview). Each worker could perform this HIT only once. Our experiment aims to distinguish the effects of different qualitative framing of financial incentives, while varying the level of the task difficulty. Thus, we follow

[3] The experiment described in this section is the result of joint work with Sephora Madjiheurem.

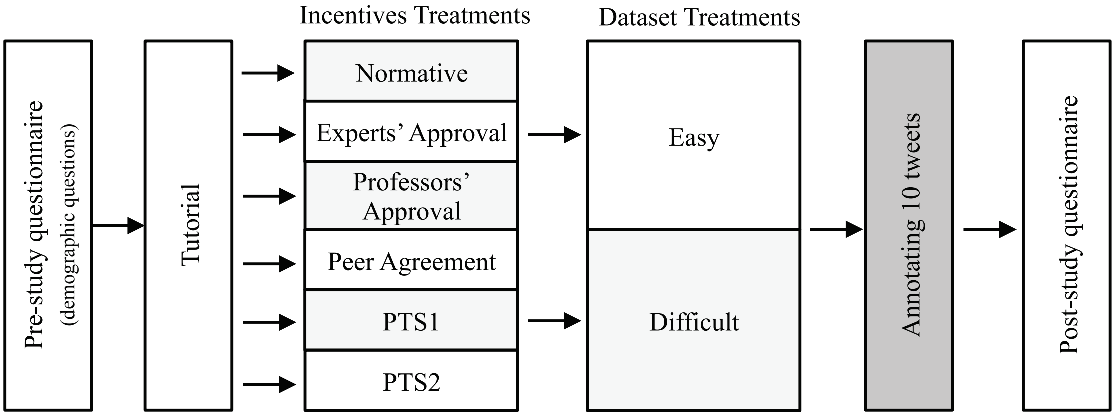


Figure 5.1: The schema of the experiment design for studying the effects of the different framing of financial incentives.

a between-subject factorial experiment design, with 2 levels of task difficulty and 6 bonus descriptions. After accepting the HIT, each worker was randomly assigned to one of the 2 × 6 treatment conditions, with the ceiling of 100 workers submitting their answers per condition.

We describe below which bonus descriptions we used to test different qualitative framing of financial reward in our experiment. We provide their exact text that was presented to the workers right before starting tweet annotation, and describe the underlying principle of each incentive in greater details.

Normative *“You can qualify for the additional* \$0.1 *bonus if your answers demonstrate an additional effort."*
Because it would be hard to imagine a way of quantifying “an additional effort" in such a task, this incentive only appeals to the worker's honesty. This also follows the norms of MTurk where obtaining the reward is implicitly contingent on the workers' performance, because a requester has an option to reject any work of insufficient quality.

Experts' Approval *“You can qualify for the additional* \$0.1 *bonus if your answers are extremely accurate according to our experts."*
This treatment condition is expected to work well on workers who respond to authority. Indeed, it implies that their work will be examined by experts, as if they were watched.

Professors' Approval *“You can qualify for the additional* \$0.1 *bonus if your answers are approved by our professors."*
By suggesting that the workers' answers will be reviewed by professors, this incentive appeals to the respect workers may have for accomplished and recognized academics. This also can lead to an increase of intrinsic motivation to help students or research projects.

Peer Agreement *"You can qualify for the additional* $0.1 *bonus if your answers agree with those of other workers."*
The goal of this incentive is to make the worker think about what answers other workers would give in order to match them.

Peer Truth Serum (PTS) We formulate two alternative descriptions of the same PTS principle: First formulation, **PTS1** *"You can qualify for the additional* $0.1 *bonus if your answers are more surprisingly common with other workers than collectively predicted."*
Second formulation, **PTS2** *"You can qualify for the additional* $0.1 *bonus if your answers both agree with those of other workers and at the same time novel."*
The previous incentive (Peer Agreement) assumes that the majority should be right and motivates workers to return answers that the majority of their peers would return. However, in case of judgment tasks, we want to motivate workers to provide their own believed answers even in case when they disagree with the majority. Peer Truth Serum, which combines elements of the Baysian Truth Serum [Pre04] with Peer Consistency [HF13] is a way to achieve this [FPTJ14]. We used here two different formulations of the Peer Truth Serum, both with the intent to make workers analyze whether the majority answer they expect is the truth and only report what they believe is the truth.

We do not include in our experiment any condition without bonuses (either with or without specific instructions) for several reasons. First of all, the focus of our study is to find which qualitative formulation of bonus conditions is more advantageous. The bonuses instructing for better quality answers were already shown to be more effective than no bonuses and no instructions [HSSV15, Har11, FPTJ14]. Using only socially framed instructions (e.g. Normative) without bonuses were shown not to be effective for crowdsourcing, while joining financial incentives with peer-oriented instructions was more successful [SHC11]. Moreover, providing instructions for better answers without bonuses could lead to the refusal of performing the task, because the conditions of answers' acceptance are not clear enough. Our goal is to advance the research on crowdsourcing incentives by comparing in more detail the less studied, but potentially effective incentives (i.e. with a bonus). The effects of bonus vs. no bonus are considered to be fixed and out of scope of the current experiment.

To discover whether the effects of financial incentives differ depending on the difficulty of the task, we prepare two non-overlapping datasets with different levels of difficulty: easy and difficult.

Easy dataset The first, "easy" dataset consisted of 70 tweets manually chosen such that they were emotional and not difficult to interpret, thus requiring less effort to annotate. They all are chosen from the tweets with emotional hashtags (e.g. #happy) or emojis, but those cues were removed from the text when presented to workers. The example tweet is "*You said it would be different but like usual nothing has changed.*" While selecting the tweets we tried to balance the presence of all 20 emotions based on the presumed labels.

Difficult dataset The second, "difficult" dataset comported 70 emotional tweets that were less obvious to interpret and required more attention to annotate. 35.7% were tweets that contained negated emotional terms, such as "*I do not regret a single thing being born as Dusun.*" We assumed that inaccurate or inattentive workers could not notice negation and interpret such tweets in a direct way. Also, assigning negated emotional expressions to specific categories could require more effort than assigning direct emotional expressions. We also included in this dataset 38.6% sarcastic tweets, for example *"Best part about rush hour is driving into it going Chicago!!! #not."* They are assumed to be more difficult to interpret because of their indirect sense. The remaining tweets were those that we discovered while exploring the tweets with emotional hashtags, e.g. those that we found confusing or those that expressed more than one emotion, such as "*Yesterday I had a pitty party for myself and today I'm feeling grrrreat and happy. #overit #backatit #strongminded #StayPositive.*"

The overall flow of the HIT was the following. First, we asked workers to optionally answer some demographic questions (pre-study questionnaire). Second, the workers followed the short tutorial containing detailed instructions on the expected answers and the comprehension quiz. After having completed the tutorial, workers were shown the description of conditions for obtaining a fixed bonus (the text was specific to each treatment condition). Then, they were asked to annotate 10 tweets, randomly selected from the corresponding treatment dataset (easy or difficult). In the end, workers could provide us with a feedback by answering a post-task questionnaire (this step was again optional). Notice that while the specific conditions for obtaining a bonus were described to workers, in reality, we gave the bonus (along with the base payment) to every worker who submitted the full HIT.

5.4.2 Experiment Run

The collection of data took place between May 13th and June 3rd 2015. During this time, $1,875$ workers accepted the HIT and $1,190$ of them (63.5%) completed the task fully (labeled all 10 tweets). Each worker was randomly assigned to one of the treatment condition. We aimed to have 100 workers per condition, but due to a technical issue with saving answers of some workers, this number was not always reached. To deal with the growing concern around cheating of workers [VdVE11, KCS08], unreliable labels were filtered out. More precisely, we removed the tweet labels that were generated in less than 3 seconds, as well as the answers from workers who provided only one or two labels for all of their tweets. In total, $5,647$ labels provided by 567 workers were retained for 70 easy tweets and $5,497$ labels collected from 568 workers for 70 difficult tweets. On average, each tweet was labeled by 79.6 different workers with a standard deviation of 2.9.

We also compute the drop-out rate under each treatment condition as the proportion of workers who did not finish labeling all 10 tweets among those who have completed the tutorial (thus, only the workers that gained access to the annotation task are taken into account). Overall, the drop-out rate is 27.3%. The Chi-Squared test of independence suggests that there is no significance difference of the drop-out rate between the different treatment conditions (on the easy dataset : $\chi^2 = 1.95$, DF = 5, p-value = 0.85, on the difficult dataset : $\chi^2 = 2.58$, DF = 5, p-value = 0.76).

Almost all workers (98.8%) answered at least one question from the demographic questionnaire. The collected answers showed that the workers were residing in at least 31 different countries, with a large majority of those who answered residing in the USA (86.6%) and India (10.3%). It also appears that most workers were between the ages of 18 to 25 or 26 to 35 (37.9% and 36% respectively). A large majority (70%) reported that they are regular users of social media, and most of the workers are native English speakers (79.9%) or have an advanced level of English (14.6%). A one-way ANOVA significance test of the distribution of demographic parameters among the different groups showed that the distributions are not significantly different across all six groups (p-value > 0.58).

We also observed that on average workers spent 13.2 seconds on the bonus instruction page, suggesting that the average worker indeed read instructions as opposed to skipping this step.

5.4.3 Performance Metrics

The worker's performance is evaluated both in terms of the provided emotion categories and emotion indicators. We compute the agreement with the peer answers from workers in the same condition. We also compute the correctness according to the gold annotations, which are extracted based on the workers' majority votes. The average of scores at the workers' level is reported, i.e. we start by computing each metric for all worker's answers and then compute average among workers. The computation of each metric for one worker's answer on one tweet is described in detail below.

Category Agreement The agreement of the worker's emotion category label is computed as the percentage of agreed peer labels for the same tweet.

Category Correctness To compute category correctness, we first obtain the ground-truth categories by aggregating the workers' labels from all incentive conditions. For each tweet, we extract the majority label (the one that is assigned most often for the tweet) and all the categories with a relatively high assignment number (we use the threshold of 0.5 minimum ratio from the majority label count). This means that for each tweet several emotion categories can be present in the ground-truth data. For easy dataset, each tweet has in average 1.51 ground-truth labels with a standard deviation of 0.85, and for difficult dataset — 2.00 labels with a standard deviation of 1.36. In average, majority labels on the easy dataset are returned by 42.5 workers, while the other agreed labels in the ground truth – by 18.5 workers. Similarly, on the difficult dataset, majority labels are returned by 34.6 workers in average, and other ground-truth labels – by 14.2 workers. These high coverage scores motivate our decision to use the workers' answers themselves to extract the ground-truth correct labels for the studied tweets. If many people specify that the tweet expresses a certain emotion, it is difficult to argue with that consensual association. Therefore, the worker's label is considered to be correct if it is within the set of extracted ground-truth labels for the tweet.

Tweet Indicator Agreement To compute indicator agreement, we split all provided indicators into tokens and generate a set of non-repeated tokens as the current tweet indicator answer. To compute agreement between two sets of tweet indicator tokens, we apply a standard Information Retrieval metric of Jaccard similarity. For each worker's answer the tweet indicator agreement is the average agreement of that token indicator set with the token sets from peer workers.

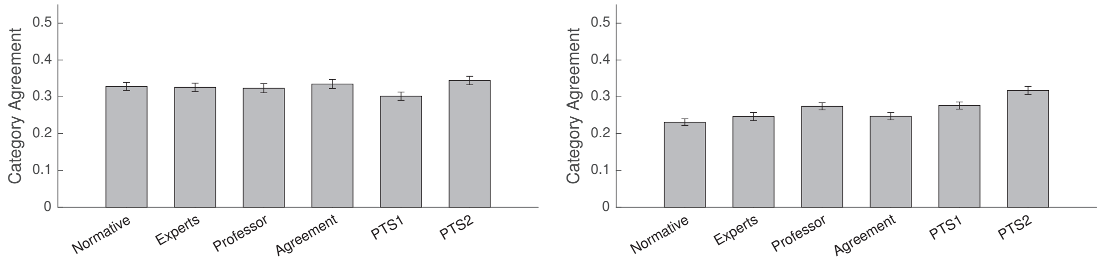
Tweet Indicator Correctness To construct the ground-truth set of tweet indicators, we detect the answers with correct category labels and aggregate them as follows. We consider as correct all the tokens that appear in the answers at least as often as the half of occurrence number of the most appearing token. The tweet indicator correctness is then computed as Jaccard similarity between the considered set of tweet indicator tokens and the extracted correct set.

Additional Indicator Number We also measure the engagement of a worker in producing new additional emotion indicators, as detected by the number of additional indicators.

In our analysis we used one-way ANOVA for testing the differences in mean agreement and correctness scores within each dataset. The multiple comparisons between incentives conditions were accounted for by correcting the p-values correspondingly, using the Tukey-Kramer correction.

5.4.4 Effects of Incentives on the Quality of Category Labels

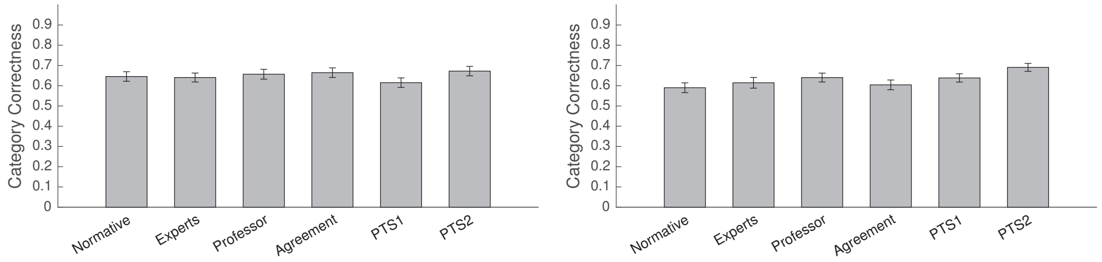
Figures 5.2 and 5.3 depict the comparisons across the incentives conditions on category agreement and category correctness respectively. Focusing on the easy dataset, we observe that for both metrics there is no significant effect of the incentives (p-value of one-way ANOVA is 0.183 for category agreement and 0.571 for category correctness). Nevertheless, the incentives significantly affect worker's performance on difficult dataset. Indeed, p-value of one-way ANOVA is $< 10^{-7}$ for category agreement and 0.033 for category correctness. Pairwise comparison of the incentives reveals the advantage of PTS2 incentive: it is significantly better than any other incentive in terms of category agreement (the highest p-value is 0.045 when comparing with PTS1), and it leads to the highest category correctness while significantly outperforming the Normative condition (with p-value 0.023). The Normative condition, which is the only condition that appeals to nothing else but workers' honesty, in its turn results in the lowest category agreement on the difficult dataset, significantly different from PTS1, PTS2, and Professor's Approval incentives (the highest p-value in those comparisons is 0.034). These observations can be summarized and explained as follows.



(a) Easy dataset

(b) Difficult dataset

Figure 5.2: The effect of incentives on average *category agreement* of workers. Error bars indicate a standard error.



(a) Easy dataset

(b) Difficult dataset

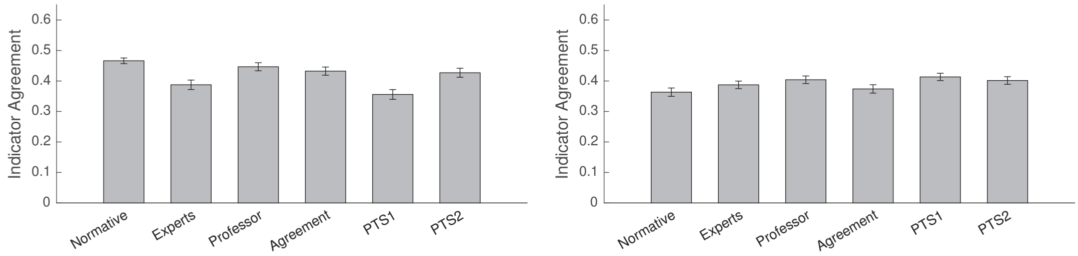
Figure 5.3: The effect of incentives on average *category correctness* of workers. Error bars indicate a standard error.

Finding 1. *No influence of incentives for category labels' quality on easy dataset.*

Finding 2. *Advantage of PTS2 and disadvantage of Normative for category labels' quality on difficult dataset.*

One potential explanation for the differences in findings between datasets could be that labeling easier tweets may require smaller effort than labeling more difficult tweets. Indeed, the category label of easy tweets is likely to be more obvious for workers, and thus they might return agreeing labels without thinking about the prospective validation of their answers. Thus, the specific framing of incentives is not important when the task is easy.

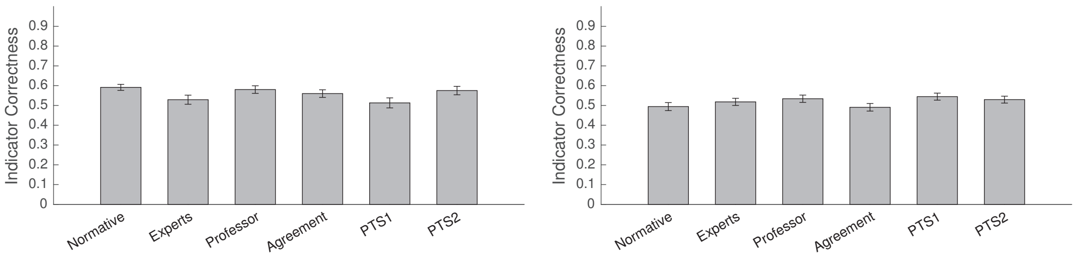
When the task is difficult, workers are required to put a larger effort to decide on a specific emotion. We hypothesize that this made the framing of incentives a more important factor, which led to the observed differences in the incentives' effects on the difficult dataset. The experiment results show higher category correctness and agreement of PTS2. Other researchers suggested that the peer-oriented incentive schemes (of which PTS is an example) can be effective because workers start to prospectively think about the answers of other workers [SHC11]. One hypothetical benefit of PTS-based incentive schema is overcoming workers' bias towards specific answers, e.g. those that they believe to be chosen by majority [FPTJ14].



(a) Easy dataset

(b) Difficult dataset

Figure 5.4: The effect of incentives on average *tweet indicator agreement* of workers. Error bars indicate a standard error.



(a) Easy dataset

(b) Difficult dataset

Figure 5.5: The effect of incentives on average *tweet indicator correctness* of workers. Error bars indicate a standard error.

The benefits of PTS-based incentive schema were shown before for counting tasks [FPTJ14]. This work confirms its advantage for a realistic judgment task (at least when more difficult data are being labeled). Additionally, in difference with the previous works, we also reveal that a more clear formulation of PTS2 (“agreeing and novel”) results in better category agreement than another commonly accepted formulation of PTS1 (“surprisingly common"). This shows the importance of designing more clear formulations of incentive instructions, at least when the task is difficult.

5.4.5 Effects of Incentives on the Quality of Indicators

Figures 5.4 and 5.5 show average tweet indicator agreement and correctness under the incentive treatments. The effects are different from the ones with category labels. First of all, the indicators output significantly differs among the incentives only for indicator agreement on easy dataset (with one-way ANOVA p-value $< 10^{-7}$). This is because of significantly lower agreement scores for PTS1 and Experts’ Approval schemes: they are both lower than either Professors’ Approval or Normative incentives (with p-value < 0.04), while the indicator agreement of PTS1 is also lower than of PTS2 and Peer Agreement. The Normative incentive, which was performed among the worst ones for category labels, achieves the highest indicator agreement on the easy dataset.

Second, no differences between incentives in terms of indicator agreement or correctness is detected on the difficult dataset. Yet, the results tend to suggest a reverse picture that on the easy dataset: Normative incentive has the lowest tweet indicator agreement with the borderline p-value of comparison with the highest agreement achieved by PTS1 (p-value = 0.061).

Additionally, we investigate the effects of incentives on the number of given additional indicators as the measure of workers' engagement in the task. The average number of additional indicators is 2.1 with standard deviation of 1.0. We again found no significant differences in those numbers between different incentive on either easy or difficult datasets.

These findings are summarized and explained below.

Finding 3. *Advantage of Normative and disadvantage of PTS1 and Experts' Approval incentives for tweet indicators' agreement on easy dataset.*

In search for the explanation of this effect, we investigated other properties of returned tweet indicators. We found that not all tweet indicator tokens actually appeared in the text of the tweets, and that tweet indicator agreement is highly correlated with the percentage of the tokens that do appear. The Pearson correlation coefficient between them is 0.87. Further investigation reveals that many of not-acceptable tweet indicators (36.2%) are names of emotion categories used in the labeling (e.g. "pride elation"). The differences in tweet indicator agreement among incentives could be then due to the subtle change in understanding of what is a tweet indicator. We hypothesize that financial incentives conditioned on answers of the peers or another validation process might bias workers to think about indicators as additional labels representation of tweets.

Finding 4. *No influence of incentives for tweet indicators on difficult dataset.*

We hypothesize that this can be due to the high difficulty of selecting specific indicators for difficult tweets. Within the easy tweets there are more explicit terms (e.g. "beyond excited") that are easy to identify as indicators of the chosen emotion. Annotating tweets in the difficult dataset requires selecting longer spans of implicit text, for which the expectations are less clear and output is harder to validate. This uncertainty might make workers to ignore the financial incentives.

Finding 5. *No influence of incentives for number of additional indicators.*

The number of additional indicators remains stable across different incentives conditions. This could be expected as none of the suggested incentives implied that bonus would depend on the amount of inputted data.

5.4.6 Perceived Evaluation

With the post-questionnaire, workers voluntarily reported how well they understood the task (task comprehension), how enjoyable it was to complete the task (task enjoyment), how easy they found the task to be (task easiness), and how much effort they invested in the task (cognitive effort). This information was retrieved using a 5-point Likert scale with the following options: (1) Strongly Disagree, (2) Disagree, (3) Neutral, (4) Agree, and (5) Strongly Agree.

While we did not find any significant difference in either task enjoyment or task comprehension between datasets or incentives conditions, we found a statistically significant difference in task easiness and cognitive effort between the aggregated groups of two datasets. Not surprisingly, workers who labeled tweets from the easy dataset found the task easier than those who labeled more difficult tweets: mean of *task easiness* is 3.86 (SD = 0.89) on the easy dataset vs. 3.57 (SD = 1.05) on the difficult dataset, p-value $< 10^{-}5$. Similarly, workers who labeled tweets from the easy dataset reported putting less cognitive effort for doing the task than workers from the difficult dataset condition: mean of *cognitive effort* is 4.64 (SD = 0.63) on the easy dataset vs. 4.53 (SD = 0.80) on the difficult dataset, p-value = 0.04.

Overall, the analysis of the post-questionnaire answers allows us to conclude that the incentives have no effect on the workers' perceived evaluation in general. The worker's answers also verify that the difficult dataset is indeed perceived as more difficult than the easy dataset and that it requires additional cognitive effort to label. The latter implicitly approves our methodology for selecting tweets for the difficult and easy datasets.

5.4.7 Discussion

In this study, we investigate how different qualitative framing of financial incentives can affect the workers' answers in a crowdsourcing judgment task. We conduct an online experiment on the crowdsourcing platform, while using emotion annotation in text as a main task. Six different qualitative incentives, based on different principles of motivation for better-quality answers, are investigated and tested with varied difficulty of the task and performance metrics.

The results of our experiment reveal the differences in the effects of the incentives depending on the difficulty of the task and on the performance metric of interest. When the task is easy, the incentives have no impact on the category labels' quality. However, when the task becomes more tedious, it can be beneficial to use a particular incentive: with a well-formulated Peer Truth Serum (PTS) bonus formulation workers tend to output more correct and agreeing labels. Moreover, only one of the two studied PTS-based incentives was proven to be advantageous, showing the importance of well-defined formulations of the incentive principles. On the other hand, only when the task condition is easy can we observe the effects of incentives on the provided emotion indicators, with the superiority of another incentive based on the trust to workers' own judgment of quality (Normative). We hypothesize that this effect might be related to the changes in effort split between our annotation subtasks caused by the different bonus formulations. Investigating this relation could be a prominent direction for the future work. Another direction could be introducing gamification principles, such as revealing to workers whether they agreed with their peers or produced a novel useful answer, or introducing game scores to further incentivize engaged participation of workers [BMF14].

In this work, we focus on investigating the differences of effects between specific qualitative bonus formulations. We did not experiment with different amount of bonus or no bonus placement because we considered reward-related effects to be fixed and non-varied across our incentives conditions. It is possible that changing the amount of the given bonus could result in more pronounced differences of incentives' framing or alternatively no effect. However, the fact that the qualitative framing of incentives affects the quality of answers at least for one fixed bonus amount encourages the future investigation, testing, and deployment of more advantageous qualitative bonus formulations.

5.5 Chapter Summary

Managing crowdsourcing of emotion annotations from non-expert online workers requires additional effort to ensure the quality of the obtained answers. Our experiments show that two types of preemptive quality control mechanisms can indeed affect the workers' performance.

First, including the tutorial to help workers understand the task's specifications leads to better annotation quality than not giving the separate instructions. We design the tutorial that not only teaches workers how to perform the task, but also validates their comprehension of the instructions. Its inclusion positively affects both the quality of emotion category labels and returned indicators. Particularly, it is effective in alleviating common misconceptions and in engaging workers to return more emotion indicators. As our tutorial also includes a knowledge quiz, it could be potentially integrated into a qualification task for filtering less attentive workers.

Second, the choice of language and principles for the formulation of bonus conditions can affect the quality of the annotations, at least when more difficult data are being labeled. We show that using a specific well-formulated incentive, asking workers to produce novel agreeing answers, results in better agreement and correctness of emotion category labels in more difficult task conditions. This indicates the importance of carefully designed qualitative framing of the performance-contingent bonuses for more effort-demanding tasks. However, the comparison of incentives' framing in the easier task conditions reveals no such impact. Thus, the choice of more effective incentives should be potentially adapted to the level of task difficulty.

Overall, this chapter proposes two mechanisms aiming to properly manage the online crowd for collecting relevant and truthful emotion annotations. The two studied preemptive quality control mechanisms are particularly suitable to apply in the context of judgment tasks in general. For such tasks, teaching and motivating workers is a preferable way to ensure the quality of their answers and to avoid excluding potentially correct answers by the posterior filtering methods. Therefore, our findings are of potential value not only for researchers and practitioners who conduct emotion annotation tasks, but also for those who conduct other judgment tasks.

6 Distant Supervision for Lexicon Construction

6.1 Introduction

One successful approach to textual emotion recognition is to formulate it as a text classification problem and to apply supervised machine-learning techniques in order to obtain emotion classifiers [AS07, SM08, RRJ$^+$12]. However, supervised methods require considerable quantities of annotated data (i.e. text documents with known emotion labels). In the previous chapters, we showed that crowdsourcing can be successfully employed to obtain manual emotion annotations of reasonable size (in the magnitude of thousands of labeled documents). However, crowsourcing data annotations still requires significant investment of human effort, being that to managing the crowd or to filtering and aggregating crowdsourced answers, not mentioning the time and cost of the human labelers. Moreover, while larger annotated corpora allow capturing greater variety of emotional linguistic expressions, it becomes excessively expensive to collect manual annotations in the magnitude of hundred thousands or millions of documents. Thus, it is desirable to explore the methods for building emotion recognition systems without relying on the costly and time-consuming manual annotations.

The research objective of this chapter is to develop a method to automatically build fine-grained emotion classifiers in the absence of manually annotated data. In this endeavor, we have resorted to distant learning (also known as distant or weak supervision)—a type of semi-supervised learning used by many researchers in text classification to generate textual classifiers without costly data annotation [GBH09, PB12, MBSJ09]. The main idea is to train the classifiers on the data with automatically assigned emotion labels (called pseudo-labels). The term 'distant learning' then refers to the process of supervised learning from some data that distantly approximate the ground-truth. In contrast to traditional semi-supervised learning, where classifiers are learned over partially annotated data (i.e. a mixture of annotated and unlabeled data) [Zhu05], the distant learning approach requires no manual annotation. Instead, annotated data are obtained automatically using some emotion labelers that are able to detect emotions of interest in the *subset* of available text documents.

In the domain of social media, researchers have successfully applied distant learning for topic-independent emotion recognition while using emoticons (e.g. ':)' or '>:(') and emotional hashtags (e.g. '*#happy*' or '*#angry*') as initial labels [YLC07, WCTS12, DCGC12, Moh12a, PB12, SI13]. They are considered to summarize emotions in the corresponding texts. However, such content cues are not always present in adequate amount within specialized topics of discussion (sports, politics, finance, or education) and are likely to be absent in text documents other than social media (reviews, news articles, or technical comments). Instead of relying on hashtags or emoticons for labeling, we aim to design and investigate a distant learning method that is more generally applicable.

To address this challenge, we suggest using terms from existing or easy-to-produce emotion lexicons as initial labelers. For instance, for any set of emotion categories, we can use a list of descriptive emotional terms (such as '*proud*' for *Pride*) and label texts according to the presence of these terms. Using such lexicon-based initial labelers ensures the generality of our methods: as they are not restricted to specific types of content cues, we can potentially detect emotional content within documents of any type or topic. A distant learning algorithm will then discover emotion associations of new terms based on their co-occurrences with given emotional terms. For example, it can recognize the phrase '*well done*' as an indicator of *Pride* emotion, if this phrase appears often enough together with known pride-related words, such as with the word '*proud*' in the text "So proud 2 be British! huge well done 2 all of Team GB! :D". At the same time, the more basic or smaller the given emotion lexicon is, the more there is a need to build emotion classifiers having higher recall and precision.

With this idea, we have developed Dystemo, a distant supervision method that generates fine-grained emotion classifiers from documents pseudo-labeled by some initial lexicon of limited coverage, accuracy, or both. We again focus on recognizing emotions in tweets,—short status updates from a popular social media website, Twitter. Twitter contains discussions on a variety of topics and events, and provides an easy opportunity to collect large datasets. Two main novelties lead to the success of the proposed method. First, we suggest a new Balanced Weighted Voting (BWV) algorithm that incorporates per-category rebalancing coefficients while learning. This overcomes the intrinsic imbalance of emotion distribution in initial pseudo-labeled dataset, which if left untreated can cause classifier's bias towards dominant emotions. Second, using social media as a source of textual data allowed us to include simple heuristics for detecting non-emotional (or neutral) tweets. These tweets turned out to be indispensable for training classifiers to discern neutral tweets from emotional ones. Both of these novelties significantly increase the accuracy of final emotion classifiers.

We validate the suggested method on tweets in the field of sports events using the fine-grained model containing 20 emotion categories. We show that with Dystemo we obtain the final classifiers of substantially better quality than the three tested initial emotion lexicons (the relative increase of micro-F1 score is between 41% and 236% on the large hashtag-based ground-truth data). In comparison with other distant learning algorithms, Dystemo achieves the best micro-F1 scores with two out of the three initial lexicons on the hashtag-based data, and shows competitive performance on small manually annotated data.

In summary, to the best of our knowledge, Dystemo is the first distant learning method for producing fine-grained emotion classifiers without the help of manually labeled text, nor of structured content features such as emoticons or hashtags. It relies on terms from emotion lexicons instead. Our carefully designed experiments confirm the viability of this approach, at least within the domain of tweets. We confirm that Dystemo is effective both in extending initial emotion lexicons of small coverage to find correctly more emotional tweets and in correcting emotion lexicons of low accuracy to perform more accurately.

6.2 Related Work

Distant learning (or distant supervision) is a specific type of semi-supervised learning, where initial partial knowledge is given in the form of the chosen pseudo-annotation heuristics (i.e. emotion lexicons in our case). Therefore, besides reviewing alternative distant learning approaches for emotion recognition, we also review other techniques of semi-supervised learning.

Semi-Supervised Extension of Emotion Lexicons The given general-purpose emotion lexicons, such as GALC [Sch05], are unlikely to cover the full variety of emotional expressions used in language. Researchers have developed semi-supervised techniques to extend initial general (but limited) emotion lexicons, considered as seeds. These methods define several metrics of term similarity and then use them to cluster new terms into emotion categories based on their similarity to the seeds. The original WordNet-Affect lexicon [SV04] and one part of the Synesketch lexicon [KPJD13] were built in this way, starting from a small number of explicit emotional terms. Similarity metrics were defined using semantic relationships (such as synonymy). In the construction of the EmoSenticNet lexicon [PGC$^+$12, PGH$^+$13, PGC$^+$14], additional term similarities were derived from term co-occurrences in the database of emotional experiences by using Pointwise-Mutual Information (PMI) [TL03]. Other corpora used to construct emotion lexicons, using PMI-based scores and starting from a small number of seed emotional keywords or symbols, were web n-grams [PIMK13], sentences from weblogs [YLC07], and tweets [Moh12a]. Such lexicon-growing methods can, therefore, increase the coverage of used emotional expressions. Instead of focusing on term-level emotion associations, our method aims at building document-level emotion classifiers. Nevertheless, for comparison, we adapt PMI-based computation of term emotion scores [Moh12a] to be applied within our framework.

Distant Supervision in Emotion Recognition With Twitter, many researchers overcome the lack of annotated data by crawling tweets with emotional hashtags, such as *#happy* or *#angry* [Moh12a, WCTS12, DCGC12, SI13]. In accordance with the idea of distant supervision, such tweets serve as pseudo-labeled data and are used to train machine-learning classifiers in a supervised manner. Yet, only a small fraction of tweets is likely to contain such hashtags, making questionable the application of these restrictive heuristics throughout different datasets. In the present work, instead of using hashtags for the pseudo-labeling of tweets, we propose using more applicable initial labelers based on terms from a given emotion lexicon. The data labeled based on emotional hashtags are used only for automatically validating the constructed emotion classifiers.

Building emotion classifiers using a limited set of emotional terms and unlabeled data has been attempted before. One method is to represent the given text corpus in a reduced-dimensionality vector-space model and assign emotions based on similarities to computed emotion vectors [KVC10, DA08].

These methods were validated for a small set of emotion categories, whereas we design a methodology capable of dealing with a much more complete set of emotion categories. Moreover, those methods disregarded the treatment of neutral tweets (i.e. tweets without emotions); while we design and successfully apply heuristics to help classifiers recognize neutral tweets.

Semi-Supervised Learning for Other Tasks Many other algorithms have been designed for semi-supervised learning (Zhu gives an overview [Zhu05]). For multi-category text classification, a commonly applied method is Naïve Bayes with the Expectation-Maximization procedure [NMTM00]. It iteratively repeats two actions: first, it learns the parameters over the annotated data; second, it re-annotates the data using the learned parameters. In our experiments, we also applied a Naïve Bayes as one of the compared classifiers, but starting from the data that were pseudo-annotated by a given initial labeler.

We also review the advances in semi-supervised methods for polarity classification—a problem closely related to emotion recognition. Experiments show semi-supervised classifiers outperform supervised ones when few labeled data are available [WK09]. The idea of distant learning for building polarity classifiers has been successfully applied to Twitter data as well, where researchers use emoticons and hashtags as the sentiment pseudo-labels (positive or negative) and identify neutral tweets using objective hashtags or as tweets from the news websites [GBH09, PP10, KWM11]. Among other methods, an iterative self-training approach has been shown to be effective [QZHZ09]. To apply binary polarity classification methods to our multi-category emotion classification problem, we first split it into multiple independent binary classification problems, each distinguishing one emotion category from all the others. This setup allowed us testing machine-learning classifiers suitable for binary applications.

Overall, no related work has studied how to apply the distant supervision framework for multi-category emotion classification when neither manual labels nor labels from a content structure, e.g. hashtags, are accessible. This is the main problem tackled in this chapter.

6.3 Distant Supervision Method — Dystemo

We first (re)-introduce the definitions used to describe the problem and our suggested method. Our problem of emotion recognition was formulated in section 3.1. Shortly, we address it as a multi-label classification task with a given set of emotion categories E^0 as potential labels (including a neutral category e_0). The output of the classifier for a tweet d is a label set $Y_d = \{e_{i_k}\} \subseteq E^0$.

We also define the *emotionality* of the text $\vec{p} = (p_0, p_1, p_2, \ldots, p_{|E|})$ as the distribution of the emotion categories expressed in the text, with $\sum_{i=0}^{|E|} p_i = 1$ and $\forall i \; p_i \geq 0$, where p_i is the weight of the i th emotion. Emotionality can be transformed into a multi-label by applying a technique adapted from the alpha-cut for fuzzy sets [BB95]. We denote this operator as $\mathfrak{A}\colon (\vec{p}, \alpha) \to 2^{E^0}$, where α defines a threshold on the emotion weight for the emotion to be included in the multi-label. $\mathfrak{A}(\vec{p}, \alpha)$ returns all the labels e_i that have the weight $p_i \geq \alpha \cdot p^*$, where $p^* = \max_i p_i$ is the maximum emotion weight within the distribution $\vec{p}$. Thus, all the labels with a weight close enough to the maximum weight are output. If $\alpha = 1$, only the labels with the maximum weight are output. For example, for the emotionality ($p_2 = 0.2,\ p_3 = 0.3,\ p_4 = 0.5,\ \forall i \neq 2,3,4\ p_i = 0$) the multi-label $\{e_3, e_4\}$ would be found with $\alpha = 0.5$. In the opposite direction, a multi-label Y_d can be transformed into the emotionality by specifying the weight of each label in Y_d as $\frac{1}{|Y_d|}$.

6.3.1 Method Input

Figure 6.1 shows an overview of our distant learning method, Dystemo. It aims at building an emotion classifier for detecting emotions of the specified category set within a specific dataset of tweets, e.g. those on a certain topic. Correspondingly, as an input, it requires Twitter data collected for a desired application, denoted *unlabeled data* U, and emotion model specifying which category set E to recognize. The method also requires *emotion* and *neutral labelers*. The core of an emotion labeler is an emotion lexicon containing associations of linguistic expressions (terms) to the emotion categories of interest. Then, the *emotion labeler* is a simple initial emotion classifier assigning emotions to tweets based on the occurrence of terms from the given lexicon. The *neutral labeler* in its turn aims at identifying neutral tweets. It is essential to have neutral tweets in the training set. Otherwise, we risk obtaining classifiers that identify almost every tweet as emotional (as it will be shown in the experimentation section), which is unacceptable for a successful emotion recognition system. We suggest simple heuristics for detecting neutral tweets, namely based on the presence of URLs in the tweet and absence of potential emotional cues (these heuristics are described in detail in section 6.4.3).

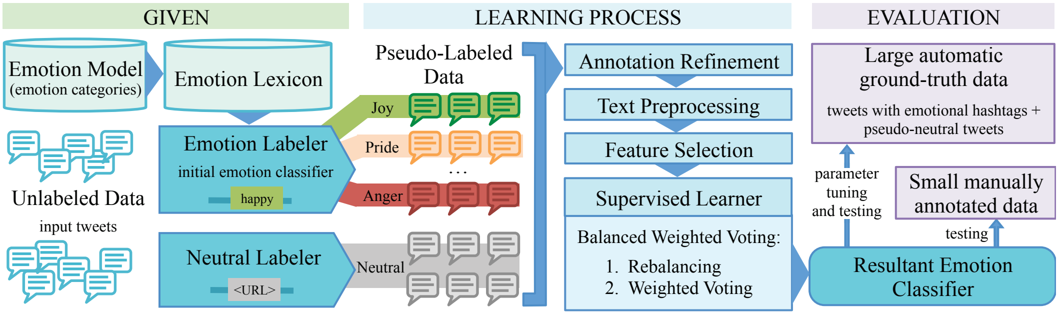


Figure 6.1: The framework for our distant learning method.

6.3.2 Initialization of Learning Process

The learning process starts with applying both *emotion* and *neutral labelers* to unlabeled data U to obtain the *pseudo-labeled data* L. We assume that the emotion labeler returns the emotionality $\vec{p}(d)$ for a given document $d \in U$, while the neutral labeler assigns a tweet to a neutral class e_0 by setting $p_0(d)$ to 1.0. Tweets detected by the neutral labeler are referred to as pseudo-neutral and are not considered to be labeled by the emotion labeler. Tweets from U where the emotion labeler found no emotion are not included in L, because they could be classified as neutral due to the lack of information about emotional expressions in the initial emotion lexicon. Overall, the pseudo-labeled data L comprise the set of tweets with mapped emotionalities, one part found by the emotion labeler, and another—by the neutral labeler.

The first step of actual learning process is *annotation refinement*. It is essential to apply it when emotion lexicons assign weights to terms, as we need to eliminate annotations of emotions with relatively low weights. The refinement is applied to each tweet individually. Given the parameter α_{ref}, it sets to zero the weights of those emotions that would not be included in the multi-label: $e_i \notin \mathfrak{A}(\vec{p}, \alpha_{ref})$, and then normalizes the distribution. Whether or not to apply this refinement is a parameter of the method.

The second step is to *preprocess the texts* of the tweets used in learning. This includes extraction of emoticons and punctuation marks as separate tokens, and lower-case transformation. We also replace usernames, numbers, and URLs with the corresponding placeholders. Furthermore, we normalize elongations (the multiple repetition of letters in a word) with their shortened form, thus “soooooo” is replaced with “soo$*$”.

The third step is to *extract* and *select features* over which the classifier will be learned. We use 1-, 2-, ..., n-grams as features. We exclude the n-grams containing only stop-words and mark n-grams as negated if a negation word is detected up to two words before them. For example, in the text “not extremely happy” we would extract the n-grams “not_happy”, “not_extremely”, and “not_extremely happy” as separate features. Also, we only retain the n-grams that appeared K or more times in the pseudo-labeled dataset L. From these, we select terms that are indicative of emotions by estimating their polarity. We compute a term’s semantic orientation using Pointwise-Mutual Information (PMI) [TL03]. First, the polarity label (l^+ or l^-) of each tweet $d \in L$ is identified as $\text{sign}\left(\sum_{i \in E^+} p_i(d) - \sum_{i \in E^-} p_i(d)\right)$, where $E^+ \subset E$ and $E^- \subset E$ are the corresponding sets of positive and negative emotions. Then, the semantic orientation $SO(t)$ of a term t is computed as

$$pmi(t, l^+) - pmi(t, l^-) = \log \frac{P(t, l^+)P(l^-)}{P(t, l^-)P(l^+)} = \log \left[\frac{1 + freq\left(t, l^+\right)}{1 + freq(t, l^-)} \cdot \frac{|V| + freq(l^-)}{|V| + freq(l^+)} \right] \tag{6.1}$$

where V is the set of extracted terms, $freq(l^\pm)$ is the number of positive (l^+) or negative (l^-) tweets, and $freq(t, l^\pm)$ is the number of tweets with the term t, which are either positive or negative. The formula uses smoothing: we add 1 to each term frequency computation, and $|V|$ to class frequency computations in order to compensate for the additions to term frequencies. The higher the absolute value of $SO(t)$, the more confident we are that the term t has strong polarity and is thus potentially emotional. We filter out the features that have an absolute score $|SO(t)|$ lower than a threshold τ. The remaining features are used for the feature representation of the tweets. As tweets are short, the terms' presence is used for features' values, instead of their frequency.

With the tweets represented as feature vectors and their associated emotionalities, the final resultant classifier can now be learned in a supervised manner. We apply Balanced Weighted Voting as the *supervised learner*. Its choice also defines how the *resultant classifier* will work.

6.3.3 Supervised Learner — Balanced Weighted Voting (BWV)

The BWV algorithm is a supervised learner that produces a lexicon of terms with the associated emotionalities based on their occurrences in pseudo-labeled data L. It takes as an input the list of terms (in our case, n-grams from the feature selection process), and for each term t computes its emotionality $\vec{w}(t) = \left(w_0(t), w_1(t), w_2(t), \ldots, w_{|E|}(t)\right)$, where $w_i(t)$ is the weight of the term t for the emotion i.

For learning, we know the emotionality of each tweet $d \in L$, $\vec{p}(d) = \left(p_0(d), p_1(d), \ldots, p_{|E|}(d)\right)$. In BWV, we first balance the distribution of emotions: we compute the rebalancing coefficient c_i for each emotion and multiply by it the corresponding emotion weight for each tweet. We then compute the weights of emotions for a term t as the normalized sum of rebalanced tweet emotion weights:

$$w_i(t) = \frac{\sum\limits_{d:\, t \in d} c_i \cdot p_i(d)}{\sum\limits_{j} \sum\limits_{d:\, t \in d} c_j \cdot p_j(d)} \tag{6.2}$$

We define the coefficient for the i-th emotion as $c_i = -\log \frac{\sum_{d \in L} p_i(d)}{|L|}$. Using a logarithm in the formula allows penalizing the emotion categories appearing more often without overestimating the weights of under-represented emotion categories.

This algorithm is inspired by the simple Weighted Voting (WV) approach used for the construction of the emotion lexicon in chapter 4. The original WV differs from BWV in that it lacks the rebalancing coefficients c_i. As a result, the lexicon created is biased towards dominant emotions: the more often an emotion appears in the labeled data, the greater its weight will be in the emotionalities of the terms. The BWV approach involves reweighting process of the emotional assignments of tweets, which is similar to the resampling approaches designed to cope with class imbalances for classification problems [Jap00].

The lexicon constructed via the BWV learner from pseudo-labeled data is the basis for the *resultant emotion classifier*. It is applied to the tweets as follows. To compute the emotionality of a tweet $\vec{p}(d)$, we search for the lexicon terms within its text, sum the emotionalities of the lexicon entries found, and normalize the vector. If no lexicon terms are found, the *Neutral* label is returned. When lexicon terms are found, the output is an emotion multi-label obtained from the computed emotionality with the operator $\mathfrak{A}\left(\vec{p}(d), \alpha_0\right)$, where α_0 is the parameter of the algorithm.

6.3.4 Parameter Tuning and Automatic Evaluation

Our distant learning method involves multiple parameters, e.g. the refinement parameter α_{ref} or the length n of n-gram features. To find its optimal parameters, we need to perform parameter tuning. For this, we suggest using an automatically generated set of ground-truth tweets labeled based on the presence of emotional hashtags. The used emotional hashtags are explicit descriptive words for the chosen emotions, such as *#happy* for *Happiness*. In a study of users' moods, De Choudhury et al. [DCCG12] found that an emotional hashtag at the end of a text corresponded to the author's mood in 83% of tweets. We considered this evaluation to be the indicator of the good enough quality for using such emotional hashtags as ground-truth labels for automatic evaluation and parameter tuning. As our emotion recognition system also should be able to recognize tweets without emotions, we additionally include pseudo-neutral tweets in these constructed data. Overall, having such large ground-truth data allows for an automated way to set the parameters of our method and to validate its performance.

6.4 Setup for Method Application

We present here a potential scenario of developing an emotion classifier for a new set of emotion categories to be detected within a specific topic of discussions. Consistent with our previous work, we again use the 20 emotion categories from the Geneva Emotion Wheel (GEW, v. 2.0) [Sch05], introduced in chapter 3.2. This section describes the data and initial labelers, providing the details of how to apply our distant learning method in the real application.

6.4.1 Data for Application: Olympic Tweets

We focus again on the domain of fans' Twitter reactions to sports events, which was studied in chapter 4. More particularly, we use the dataset OLYMP_DATASET, which contain tweets about the 2012 Olympic Games presented in section 3.4. This dataset consist of 33.2 million English tweets containing Olympic-related keywords.

Data Preparation We apply prior data filtering in order to select the tweets most useful for learning: we use only tweets containing at least 3 words (disregarding hashtags and usernames) to increase the probability of learning additional terms, and exclude retweets and tweets with duplicate text to avoid overfitting.[1]

Using sports-oriented Twitter discussions as an application, we will study the proposed method in the context of our motivation scenario to build domain-specific emotion classifiers. However, such domain specificity also requires us to extend the list of stop words with the dataset-specific words. Our stop-word list thus includes the frequently appearing words referring to the main topic of our Olympic-related tweets, e.g. *olympic*, *usa*, *team*. We additionally consider punctuation marks as stop-words.

6.4.2 Emotion Labelers

Three initial emotion lexicons are taken as emotion labelers for our distant learning method. Two are topic-independent: one lexicon of explicit emotional terms (GALC) and one weighted lexicon learned from general Twitter data (PMI-Hash). We also take one domain-specific lexicon that we built previously for analyzing reactions to sports events on Twitter (OlympLex).

GALC GALC is a domain-independent emotion lexicon of the unigram stems explicitly expressing an emotion, for example, 'happ∗' for *Happiness*. As described in section 3.3, we instantiated such stemmed words into the tokens, e.g. 'happy', and obtained the GALC-R lexicon. The terms are assigned to the GEW categories using hard links. To compute a document's emotionality using this lexicon, we sum the number of terms found for each emotion (excluding negated terms) and normalize the obtained vector. Using this lexicon of explicit terms as an initial emotion labeler will help us investigate how well the proposed distant learning method can build a new emotion lexicon using only the small number of seed emotional words.

OlympLex We focused on the same domain of reactions to sports events while building the emotion lexicon using human computation in chapter 4. The outcome was the small within-domain emotion lexicon, OlympLex. It contains the emotion indicators selected from

[1] Retweets were detected using the pattern "RT @username" (a better detection could be based on parsing retweet parameters in the tweet's JSON, but unfortunately they were not present in our current dataset). In order to detect duplicates, we compared tweets based on their full text with removed non-alphanumeric symbols and marked significantly overlapped tweets as duplicates.

the tweets by the annotators, as well as related user-entered emotional expressions. This emotion lexicon allocates a GEW-based emotionality for each of its terms (from unigrams to 5-grams). Furthermore, we removed 94 frequent terms related to a description of the Olympics rather than emotions, such as ‘event’. The average of the emotionalities of terms found in the tweet text (excluding negated terms) is the emotionality of the whole tweet. Using this lexicon as an initial emotion labeler will allow us to study how well the proposed distant learning method can improve the quality of a small within-domain lexicon.

PMI-Hash We also generate a topic-independent Twitter-specific emotion lexicon using the PMI-based learning method [Moh12a]. It computes emotion weights of terms using tweets with emotional hashtags. We use the dataset EMHASH_DATASET of tweets with GEW-associated emotional hashtags. It was presented in section 3.4. For computational reasons, instead of using the full dataset of $1,729,980$ tweets, we use only randomly selected $500,000$ tweets having at least 3 words. We applied the same preprocessing and n-gram extraction steps as in our method, and used unigrams and bigrams as terms for the lexicon. The weights of these terms are computed via *PMI-Based* learner. It computes the strength of association $SoA(t, e_i)$ of term t to the emotion e_i as the difference in PMI of term t towards the presence and absence of emotion e_i. The formula (6.1) is used again while considering the presence of emotion e_i^+ as a positive class and an absence of emotion e_i^- as a negative class. The positive values are saved as term emotion weights, i.e. $w_i(t) = \max(0, SoA(t, e_i))$. In total $85,530$ terms are extracted. When applying this lexicon to the text, we sum the weights of found lexicon terms and normalize the resultant vector to obtain an emotionality of the text. Using such automatically generated topic-independent lexicon as an initial emotion labeler will allow us to study whether the proposed distant learning method can improve its quality.

6.4.3 Neutral Labeler

The neutral labeler aims to find the tweets with a high probability of being neutral. To define the heuristics of such labeling, we assume that the presence of a URL indicates less emotional tweets, such as news or information sharing. We extracted such tweets and observed that to enforce tweet neutrality, we should avoid the presence of usernames and personal pronouns (which makes sharing more personal), emoticons, and other emotional cues. We exclude tweets that contain explicit emotional terms from the GALC lexicon, intensity shifters (exclamation marks, elongations, intensifier and diminisher words), and strong subjective terms (from MPQA Subjectivity Lexicon [WWH05]). The examples of such identified neutral tweets are “Sports Debates and Olympic Coverage <URL>” and “read more: history of the Olympic torch, flame, and relay <URL>.”

6.5 Evaluation Methodology

This section describes ground-truth data used for the evaluation of the obtained classifiers, as well as how we tune and evaluate the resultant emotion classifiers. Also, it introduces other classifiers used for comparison with BWV.

6.5.1 Ground-Truth Data

Large Automatically Labeled Data Following the idea introduced in section 6.3.4, we generate the pseudo-annotated tweets for evaluating the quality of built classifiers in an automatic way. We extracted from the full dataset of pre-filtered Olympic tweets (no short tweets, retweets, or duplicates) those that contain emotional hashtags at the end of a text (we used the same 167 hashtags that were used to collect the general emotional tweets from EMHASH_DATASET and presented in section 3.3). We additionally excluded tweets featuring several emotional hashtags. This procedure resulted in $52,218$ tweets labeled with emotional hashtags, i.e. only 0.16% of the full dataset of 33.2 million tweets. The distribution of emotion categories is given in Figure 6.2. Table 6.1 provides examples of such hashtagged tweets. As these data are intended for testing the algorithms' outputs, 'labeling' hashtags were removed from the texts.

The second half of the automatic ground-truth data consist of pseudo-neutral tweets, i.e. tweets detected by the introduced neutral labeler. We randomly selected the same number of such tweets ($52,218$) for inclusion in the evaluation set. The URLs were removed from their texts. We decided to use the same number of pseudo-neutral tweets as hashtagged tweets because the real proportion of emotional to non-emotional tweets is unknown and may vary between datasets or dataset subsets.

We split these automatic data into a validation set S_V to tune the algorithm's meta-parameters and test set S_T to evaluate the resultant classifiers in 1:2 proportion, that is $34,802$ tweets for S_V and $69,634$ tweets for S_T. This process preserved emotion distribution, meaning tweets for each emotion category were split proportionally, including pseudo-neutral tweets.

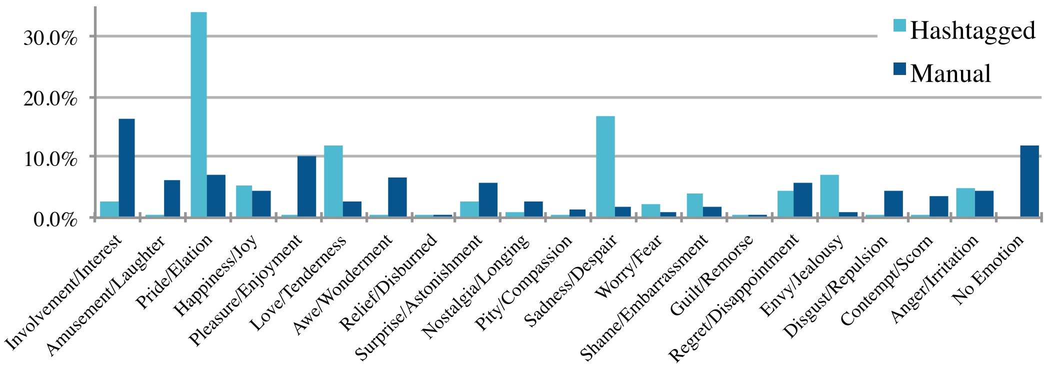


Figure 6.2: Distribution of emotion categories found in the hashtagged dataset and manual annotations within the Olympic-related tweets.

Emotion	Example hashtagged tweets from the Olympic data.
Involvement	Watching Olympic power walking.... #interesting
Amusement	Olympic triple jumping is really funny to watch... #laughing #lol
Pride	great day at the #Olympics for Canada ! #proud
Happiness	Pumped for the Olympics tonight! #happytweet
Pleasure	3 fleece blankets, coffee and watching the Olympics. #satisfied
Love	Destiny Hooker is my hero. #Olympics #USA #loveher
Awe	Just discovered the olympic decathlon #fascinated
Relief	are GB finally doing something in the Olympics #relieved
Surprise	15 years old and Olympic champion #amazed
Nostalgia	WOOHOO OLYMPICS ^__^ brings back memories :(#nostalgia
Pity	I feel sorry for the people who come last in the Olympics-#pity
Sadness	Working, then tutoring missing the olympics :(#sadtweet
Worry	This Olympic ceremony is just freaking me out... #worried
Shame	False start for Norway! #C1 #Olympics #Shame
Guilt	Skipped the gym today. Now I am watching the Olympics. #Guilt
Regret	Well.. No boze pravde in this olympics.. #disappointment
Envy	i wish i could be in the olympics when im 15. #jealous
Disgust	#Rogan that was just a bit bullshit #dislike #London2012
Contempt	They have ping pong in the olympics? #despicable
Anger	ESPN keeps ruining the results of the olympics for me #annoyed

Table 6.1: Examples of Olympic-related tweets automatically annotated based on the presence of an emotional hashtag.

Manually Annotated Data As we plan to evaluate the performance of the distant learning classifiers built starting from OlympLex, we refrained from using the same annotated corpus (SREC) as it was used for the construction of OlympLex in chapter 4. We generated new manually-annotated within-domain data. For that, we asked human annotators to annotate 600 Olympic tweets (again pre-filtered, without overlap with S_V). To ensure the presence of multiple emotions, we avoided using only random tweets. Instead, we selected three types of tweets for annotation: 200 random tweets, 200 tweets with emotional hashtags (10 per each emotion category, with removed emotional hashtags), and 200 pseudo-neutral tweets (with removed URLs). Every tweet was labeled by two annotators. They were asked to provide up to 3 emotion labels per tweet, with one marked as dominant. They could also choose to label *Other emotion* or *No emotion*. Additionally, we asked them to mark if a tweet's emotion is ambiguous or if the text is unclear. We excluded such tweets to have a dataset of higher quality, resulting in 492 tweets available for evaluation. The Fleiss Kappa [Fle71] of paired dominant labels is 0.31, showing a fair agreement. We also computed what proportion of the tweets have partial agreement: we counted that in 58.3% of tweets the dominant label from one annotator is within the full set of labels from another annotator. We found that disagreement comes frequently while discerning whether the tweets is emotional or not (19.3% of tweets). We asked for the third annotation of such tweets and excluded an annotation in disagreement with other two regarding whether the tweet is emotional or not.

In order to prepare the ground-truth dataset for testing, we assign to a tweet an emotion multi-label that includes two chosen dominant emotion categories and all other agreed categories from two annotators. The average number of labels per tweet is 1.71, showing the multiplicity of emotional experience and the need to treat this problem as multi-label classification. The distribution of outputted labels is shown in Figure 6.2. We name this evaluation set S_M. Examples of manually annotated tweets were given in Table 3.2.

6.5.2 Performance Metrics

We record the performance of the corresponding algorithm instances using multiple evaluation metrics suitable for multi-label classification [TK07, SL09]. We compute both macro- and micro-versions of precision, recall, and F1-score, as well as accuracy.

Let T_i be the set of tweets where the emotion e_i is present according to the ground truth, O_i be the set of tweets that a classifier outputs as belonging to emotion e_i, and $C_i = T_i \cap O_i$ be the set of tweets correctly classified as belonging to emotion e_i. Then, for emotion e_i, recall is $R_i = \frac{|C_i|}{|T_i|}$, precision is $P_i = \frac{|C_i|}{|O_i|}$, and F1-score is $F_i = \frac{2P_iR_i}{P_i+R_i}$.

To compute macro-recall ($\textit{macro-}R$), macro-precision ($\textit{macro-}P$), and macro-F1 score ($\textit{macro-}F1$), we average those values between emotion categories. Thus, $\textit{macro-}R = \frac{1}{|E|}\sum_{i=1}^{|E|} R_i$, $\textit{macro-}P = \frac{1}{|E|}\sum_{i=1}^{|E|} P_i$, $\textit{macro-}F1 = \frac{1}{|E|}\sum_{i=1}^{|E|} F_i$. It is noteworthy that the *Neutral* category e_0 is excluded from this averaging, as it is not the focus of the emotion recognition system. We also exclude the *Contempt* category, which was under-represented in the dataset. The benefit of using macro-scores is that they assign equal importance to each emotion category, regardless of their distribution. Hence, macro-scores penalize classifiers' mistakes on under-represented categories, which classifiers tend to ignore despite their potential importance.

We compute micro-recall ($\textit{micro-}R$), micro-precision ($\textit{micro-}P$), and micro-F1 score ($\textit{micro-}F1$) using the formulas for recall, precision, and F1-score with the total number of true labels, outputted labels, and correctly detected labels for all emotion categories. That is, $\textit{micro-}P = \frac{\sum_i |C_i|}{\sum_i |O_i|}$, $\textit{micro-}R = \frac{\sum_i |C_i|}{\sum_i |T_i|}$, and $\textit{micro-}F1 = \frac{2\,\textit{micro-}P \cdot \textit{micro-}R}{\textit{micro-}P\, +\, \textit{micro-}R}$. The labels for the *Neutral* category are again excluded. In contrast to macro-metrics, micro-metrics take into account the distribution of emotions in the dataset. Thus, they provide an estimation of how well an evaluated classifier can detect emotions while giving more weight to the most frequently appearing emotions.

We also evaluate the accuracy of classifiers. In the context of multi-label classification, the accuracy $A(d)$ for a tweet d is defined as the Jaccard measure between the set of its true labels $T(d)$ and the set of labels $O(d)$ that a classifier outputs for it, that is, $A(d) = \frac{|T(d)\cap O(d)|}{|T(d)\cup O(d)|}$. The overall accuracy A is the mean of $A(d)$ for all tweets in the dataset D: $A = \frac{1}{|D|}\sum_{d\in D} A(d)$. Accuracy evaluates how applicable the classifier is, in general, over the dataset, as it checks its performance at the per-document level, while also evaluating its ability to separate the neutral category from other emotions.

6.5.3 Comparison with Other Methods

Using the same distant learning framework, we compare the BWV classifier with the five other supervised classifiers used for emotion recognition and text classification. To apply them instead of BWV, we transform the format of the pseudo-labeled data from emotionalities into multi-labels: to each emotionality $\vec{p}$ we apply operator $\mathfrak{A}(\vec{p}, \alpha_{ref})$ with the parameter α_{ref} specified for the annotation refinement. The tweets are represented in the same feature space of filtered n-grams as defined for our method (described in section 6.3.2). We consider two ways to address such multi-label classification using standard machine-learning classifiers: Multi-Class (mcl) and One-vs.-Rest (1vR) transformations [TK07].

Multi-Class Transformation (mcl) This approach transforms the given multi-label classification problem into a multi-class problem: each document d with a multi-label $Y = \{e_{i_k}\} \subseteq E^0$ yields $|Y|$ documents in the new training set, one for each label $e_{i_k} \in Y$. We consider two classifiers: Multinomial Naïve Bayes (mcl-MNB) and Logistic Regression (mcl-LogReg), implemented using WEKA [HFH$^+$09] and LibLINEAR software [FCH$^+$08], respectively. Both of them return probabilistic output, which is treated as an emotionality of the text and is transformed back into the multi-label using the operator $\mathfrak{A}$ with the parameter α_0 again.

One-vs.-Rest Transformation (1vR) This approach transforms the given multi-label problem into $|E^0|$ independent binary classification tasks, one for each emotion category. A classifier for emotion e_i decides if it is present (class e_i^+) or not (e_i^-). We again evaluate Multinomial Naïve Bayes (1vR-MNB) and Logistic Regression (1vR-LogReg) classifiers in these settings (but for binary classification). As both classifiers support the probabilistic output, we specified that multi-label output for a text d should contain only those emotions e_i for which the probability of its presence is higher than some threshold r (a new parameter), i.e. when $P(e_i^+|d) > r$. We also applied an additional per-category feature selection with this transformation. To select features for emotion e_i, we used the term's strength of association to that emotion $SoA(t, e_i)$ computed in the same way as for PMI-based learner (described in section 6.4.2). The terms that have an absolute score $|SoA(t, e_i)|$ lower than a threshold θ are filtered out. For simplification, θ is fixed to the same value for all emotion categories.

PMI-Based Learner We additionally compare our method with the PMI-based learner used for generation of emotion lexicon from pseudo-labeled data by Mohammad [Moh12a] and described in detail in section 6.4.2, where it is used for generating PMI-Hash emotion lexicon. We only include here the threshold θ to filter out low values of $|SoA(t, e_i)|$, similar to per-category feature selection in 1vR transformation. The outputted emotionality of the tweets is transformed into multi-label output using the operator $\mathfrak{A}$ with the parameter α_0 again.

Random Baseline We also adapt a random baseline (Random) to estimate the problem's difficulty: it decides independently whether or not each emotion is present, with probability defined by the emotion distribution in the test dataset. Performance scores are averaged over $1,000$ runs.

6.5.4 Input Data Sampling, Parameter Tuning, and Testing

In our experiments, instead of applying emotion labelers to all the available tweets, we use only N_U random pre-filtered tweets (no retweets, duplicates, or short tweets) due to our limited computational resources. At the same time, as our Neutral Labeler applies more restrictive heuristics, we could apply it to all the available pre-filtered tweets, and use in the experiments the same amount N_U of pseudo-neutral tweets. Balancing the amount of pseudo-neutral and potentially emotional tweets in learning process allows us to give the same detection priority to both of these classes. All unlabeled data used in learning are disjoint from any considered ground-truth data.

To find the optimal parameters of each algorithm, we perform parameter tuning separately for each initial emotion labeler via exhaustive grid search of discrete parameters' values. The explored parameter space and found optimal parameters are described in the appendix, section A.5. The data for learning are obtained with $N_U = 100,000$, and the validation set S_V is used for evaluating the performance of the classifiers during the parameter tuning process. Among the obtained results, we find a set of parameters that yields the highest micro-F1 score on S_V. We chose to maximize the micro-F1 score because it was found to lead to a better balance between micro-precision and recall. The learning process for building final classifiers to test uses larger data obtained with $N_U = 500,000$. The resultant classifiers are then evaluated on automatic test set S_T and manual test set S_M.

6.6 Evaluation Results

This section presents the results of the tuned distant learning algorithms on the test datasets. We compare the performance of the resultant classifiers with the baseline performances of the initial emotion labelers applied without distant learning or the neutral labeler. They are reported as *Initial*. Further, we report how significantly each algorithm's performance metrics differ from those of the corresponding initial labeler, as estimated by randomization tests [Yeh00]. One asterisk * indicates a p-value ≤ 0.05; two asterisks ** indicate a p-value ≤ 0.01.

6.6.1 Improvement over the Initial Emotion Labelers

Table 6.2 presents the results on the test dataset S_T. They show that our proposed BWV method substantially improves the quality of initial emotion labelers on all of the main performance metrics: macro-F1, accuracy, and micro-F1 score. The only exception is a lower macro-F1 score when starting from OlympLex, but this result is insignificant (p-value = 0.065). The largest improvements are observed for micro-F1 scores: 41% when started from PMI-Hash, 53% from OlympLex, and 236% from GALC. The highest micro-F1 score is 40.6% with PMI-Hash as the input emotion labeler. The minimum relative increase in accuracy is 10.6% (with GALC). These findings confirm that BWV can build emotion classifiers that are far more accurate than the existent emotion labelers. The experiments also show that the other algorithms applied within the same distant learning framework can improve the performance of initial classifiers too.

EL	Algorithm	macro			A	micro			rank
		P	R	F1		P	R	F1	
-	Random	2.5	1.3	1.7	41.6	8.7	4.4	5.8	
GALC	Initial	20.6	3.6	4.8	52.2	23.6	5.1	8.4	
	mcl-MNB	**21.4**	12.2**	**10.3**** ↑	**62.0**** ↑	**30.6****	28.1**	**29.3**** ↑	1
	mcl-LogReg	7.5**	**23.9****	8.9** ↑	43.1** ↓	9.6**	30.4**	14.6** ↑	6
	1vR-MNB	11.8**	17.1**	9.7** ↑	57.0** ↑	16.9**	**34.6****	22.7** ↑	4
	1vR-LogReg	12.1**	8.8**	8.1** ↑	54.4** ↑	22.0**	20.9**	21.5** ↑	5
	PMI-based	12.7**	10.2**	9.3** ↑	53.1** ↑	28.0**	26.4**	27.2** ↑	3
	BWV	16.8**	11.5**	9.8** ↑	57.8** ↑	27.2*	29.1**	28.2** ↑	2
OlympLex	Initial	11.4	9.7	7.1	47.4	19.3	19.3	19.3	
	mcl-MNB	**19.7****	11.2**	6.8 ↓	58.5** ↑	26.3**	27.0**	26.7** ↑	3
	mcl-LogReg	9.1**	12.4**	7.6** ↑	42.9** ↓	16.1**	21.6**	18.4** ↓	6
	1vR-MNB	19.4**	12.3**	7.3 ↑	58.9** ↑	23.3**	28.3**	25.6** ↑	4
	1vR-LogReg	11.1*	**16.5****	**9.8**** ↑	51.3** ↑	17.1**	27.9**	21.2** ↑	5
	PMI-based	15.8**	9.6	7.3 ↑	58.8** ↑	28.3**	26.0**	27.1** ↑	2
	BWV	17.8**	9.4	6.7 ↓	**59.4**** ↑	**29.9****	**29.2****	**29.5**** ↑	1
PMI-Hash	Initial	12.1	17.0	11.5	23.7	21.8	42.0	28.7	
	mcl-MNB	22.8**	15.9**	13.1** ↑	64.4** ↑	37.6**	43.0**	40.1** ↑	3
	mcl-LogReg	14.4**	18.7**	14.8** ↑	52.7** ↑	30.9**	41.8	35.5** ↑	6
	1vR-MNB	19.9**	16.7	14.2** ↑	**64.6**** ↑	37.5**	43.3**	40.2** ↑	2
	1vR-LogReg	17.6**	**18.9****	**16.2**** ↑	60.6** ↑	35.4**	42.2	38.5** ↑	5
	PMI-based	22.3**	15.6**	14.4** ↑	63.8** ↑	**38.5****	41.3**	39.9** ↑	4
	BWV	**29.3****	15.5**	13.1** ↑	64.1** ↑	37.3**	**44.4****	**40.6**** ↑	1

Table 6.2: Evaluating distant learning algorithms on automatic test data S_T. All performance scores are percentages. The results of the learned classifiers are compared with those of the corresponding initial labelers. One asterisk * indicates a p-value ≤ 0.05; two asterisks ** indicate a p-value ≤ 0.01.

To our surprise, we observed greater macro- and micro-precision from the classifiers obtained through distant learning than from the initial classifiers. The best micro-precision of BWV is 37.3% starting from PMI-Hash. It is 71% better than that of the initial PMI-Hash lexicon. Based on our previous experiments [SMP14], we expected that a distant learning approach would only improve the classifiers' recall by finding more emotional expressions. However, this increase in precision indicates that, in many cases, the distant learning process corrects the terms' emotion distributions.

6.6.2 Comparison of BWV and Other Supervised Classifiers

To further compare the distant learning algorithms, we rank their performance using the micro-F1 score. We chose this metric because it aggregates the overall performance of classifiers while taking into account the significance of emotion categories and because it produces a more stable ranking for different emotion labelers than other metrics.

Mcl-LogReg performs worst, both in terms of micro-F1 and accuracy. This is due to its lower precision, possibly because it finds more tweets to be emotional (≥ 62% for all input emotion labelers) than other classifiers (≤ 54%), thus making more mistakes on neutral tweets.

1vR-LogReg is the next worst, with the moderate micro-F1 scores. Although 1vR-LogReg achieves the highest macro-F1 scores for OlympLex and PMI-Hash, these are accompanied by relatively low macro-precision (in comparison to BWV), which is undesirable for the real-world applications.

The third and forth ranks in the aggregated performance are shared between 1vR-MNB and PMI-based methods. 1vR-MNB performs best with PMI-Hash, achieving the top accuracy and high micro-F1 score; whereas PMI-based systematically increases accuracy, micro-precision and micro-F1 scores for all three emotion labelers.

The two classifiers with the highest ranks are mcl-MNB and BWV. When starting from GALC, mlc-MNB's performance is superior to BWV for all metrics except micro-recall. However, BWV produces the highest micro-F1 scores starting from the OlympLex and PMI-Hash lexicons. Moreover, its 40.6% micro-F1 score when starting from PMI-Hash is the highest score achieved in all our experiments, indicating that BWV was the most appropriate for real-world applications of emotion recognition in tweets.

6.6.3 Effects of Choosing Initial Emotion Labeler

The three evaluated initial emotion labelers differ not only in their basic performance but also in their results in conjunction with the distant learning method. Due to its explicit nature, the GALC lexicon has relatively high macro- and micro-precision, but low recall. Distant learning can improve its performance by increasing recall—it discovers new emotional terms that co-appear with the given terms in the unlabeled data. OlympLex's precision and recall are close to each other due to its higher coverage of emotion terms used in the sports domain. This lexicon's size is moderate, but our method can still discover new terms indicative of emotion and increase both micro- and macro-precision, probably because of a better adjustment of the distribution in emotion categories and better separation of the most frequent categories. Finally, PMI-Hash shows the highest macro- and micro-F1 scores of all the initial emotion labelers, yet its accuracy is the lowest and its recall is almost twice as large as precision. The PMI-Hash has this behavior because it was trained on data without neutral tweets, and thus it classifies most tweets (96%) as belonging to an emotion category and has low accuracy for neutral tweets. The distant learning approach successfully helps overcome this problem and increases PMI-Hash's precision up to the level of its recall.

Overall, this evaluation indicates that the described distant learning method is able to adjust all three initial emotion lexicons to an application domain. This is validated by a statistically significant increase in accuracy and micro-F1 score.

Algorithm	Parameters		macro			A	micro		
	Blc	Neut	P	R	F1		P	R	F1
Initial	-	-	12.1	**17.0**	11.5	23.7	21.8	42.0	28.7
BWV	Log	Incl	**29.3****	15.5**	**13.1**** ↑	**64.1**** ↑	37.3**	**44.4****	**40.6**** ↑
Alternative parameters									
WV	No		25.9**	11.2**	12.1** ↑	63.4** ↑	**45.3****	31.5**	37.1** ↑
BWV		No	16.0**	14.1**	11.1* ↓	34.0** ↑	25.9**	40.2**	31.5** ↑

Table 6.3: Improvement of the recognition quality in distant learning due to the inclusion of the rebalancing process and the incorporation of the pseudo-neutral tweets. The results are on automatic test set S_T with PMI-Hash as initial emotion labeler. All scores are given as %. The results of the learned classifiers are compared with those of the corresponding initial labeler.

6.6.4 Variations in Dystemo Configuration

Rebalancing Process The suggested BWV learning method originates from Weighted Voting (WV), which does not introduce rebalancing coefficients c_i (described in Section 6.3.3). Table 6.3 shows the benefits of having the rebalancing process. It compares BWV with WV on test set S_T while using PMI-hash as emotion labeler (the parameters of WV were tuned separately).

We observe that without rebalancing, WV is inferior to BWV for micro- and macro-F1 scores. Although WV showes the highest micro-precision, its recall is significantly lower than the initial labelers. With OlympLex and GALC as start points, WV's macro-F1 scores are even lower than those of the initial emotion labelers. This means that WV without rebalancing is unsuitable for distant learning, at least not within our method.

Using Neutral Tweets during Learning One part of pseudo-labeled data for learning comprises pseudo-neutral tweets. We investigate if adding them is helpful by learning additionally the BWV classifier without including the pseudo-neutral tweets in the learning process (with the parameters retuned accordingly). Its results on the test set S_T are indicated with parameter *Neut=No* in Table 6.3. It is noteworthy that, as S_T includes neutral tweets, the classifier's ability to recognize them is evaluated too.

We find that without neutral tweets BWV performs worse than with them in all metrics. This is because without exposure to neutral tweets during learning, resultant classifiers tend to classify most test tweets as emotional (up to 86%), even though it adapted higher feature selection threshold τ. This results in many errors on neutral tweets. Similar behavior is observed when using the other two initial labelers (OlympLex and GALC), but results are aggravated by a significant decrease in accuracy.

Emotion Labeler	Algorithm	A	micro			Coverage	#Labels
			P	R	F1		
GALC	Initial	25.6	**50.0**	5.4	9.8	14.2	1.14
	mcl-MNB	**30.0**** ↑	30.5** ↓	**12.2**** ↑	**17.4**** ↑	51.2	1.17
	BWV	28.9** ↑	27.8** ↓	11.1** ↑	15.9** ↑	51.8	1.16
OlympLex	Initial	32.7	**42.5**	16.4	**23.6**	54.5	1.06
	mcl-MNB	**34.9**** ↑	39.7 ↓	**16.6** ↑	23.5 ↓	63.0	1.00
	BWV	33.6** ↑	37.7 ↓	13.9* ↓	20.4* ↓	55.5	1.00
PMI-Hash	Initial	12.2	19.7	**12.9**	15.6	98.0	1.00
	mcl-MNB	28.4** ↑	**27.5**** ↑	**12.9** –	**17.5** ↑	65.9	1.06
	BWV	**28.5**** ↑	25.9** ↑	11.8 ↓	16.2 ↑	60.4	1.13

Table 6.4: Evaluating distant learning algorithms on manual test data S_M. All scores are percentages, except for the average number of emotion labels #Labels. The results of the learned classifiers are compared with those of the corresponding initial labelers.

6.6.5 Validation of Distant Learning on Manually Annotated Data

Testing algorithms on large ground-truth data S_T allowed us to automatically find the best parameters of the algorithms and cover more feature terms in evaluation. However, testing on manual data is essential to understand how the quality of classifier will be perceived in human's eye. Thus, we confirm the positive effects of distant learning on small manually annotated data, S_M, described in section 6.5.1. Table 6.4 presents the results of this test with two distant learning methods, BWV and mcl-MNB. We compare these two methods because they ranked high on automatic test data S_T. Notice that we do not report macro-scores because for many categories there are not enough tweets to obtain conclusive per-category metrics. However, we additionally report coverage of the methods which estimates how many tweets were detected as emotional, and the average number of emotion labels found in tweets classified as emotional.

When initial emotion labelers are GALC and PMI-Hash, the effects of distant learning algorithms remain similar to those discovered with automatic test data S_T: applying distant learning increases the accuracy and micro-F1 scores of initial labelers, due to recall increase for GALC (along with coverage increase) and precision increase for PMI-Hash. However, the improvements are smaller. This can be attributed to the fact that our manual annotation is less skewed towards dominant categories and requires from classifiers to perform better across more categories. Also, while comparing the performance of mcl-MNB and BWV algorithms starting from GALC and PMI-Hash, we can observe that mcl-MNB slightly outperforms BWV. However, we did not find a significant difference in their micro-F1 scores.

With OlympLex as a starting labeler, we obtain different effects. While both distant learning methods still increase the accuracy over the initial OlympLex labeler, neither mlc-MNB nor BWV improve the micro-F1 score despite previously observed significant increase. However, already on automatic data we have observed an insignificant decrease in their macro-F1 scores. This can signify the need to optimize for both macro- and micro-F1 scores in the parameter tuning process. Moreover, this evaluation shows that OlympLex, built using manual annotations of tweets, performs best on manually annotated data. This reveals the difficulty to improve emotion lexicons of better quality via distant learning and the need for more advanced methods in such cases.

We also observe all resultant classifiers have lower micro-recall scores on manual test set S_M compared to those scores on automatic set S_T. This can be due to the higher average number of emotion labels per tweet in the manual ground-truth (1.71 in S_M versus 1.0 in S_T). This means that, to achieve better recall scores, resultant classifiers have to find correctly more emotion labels per tweet. However, all the final classifiers return only up to 1.17 emotion labels per tweet. With OlympLex as an initial labeler, both mcl-MNB and BWV learn to return exactly one label per tweet. This leaves room for potentially better optimization of the classifiers' output parameter α_0.

Overall, our method, Dystemo applied with BWV as a learning algorithm is shown to be effective in extending initial emotion lexicons of small coverage to find more emotional tweets (coverage is 264% more and recall is 105% higher for GALC lexicon). Additionally, it can improve coarse emotion lexicons to perform more accurately (accuracy is 133% higher for PMI-Hash lexicon).

6.7 Discussion and Future Work

The present work showed that applying distant learning with emotion lexicons as initial labelers is a viable approach for building application-specific emotion classifiers. Experiments show that the resultant classifiers are able to achieve micro-F1 scores between 15.9% and 40.6% while recognizing twenty emotions. Previous work reported similar scores when fewer emotion categories were used: e.g. Mohammad [Moh12a] achieved a micro-F1 score of 49.9% for six basic emotion categories in cross-validation on hashtagged tweets and 43.7% on news headlines. Our classifiers deal with more emotion categories, and thus the performance baseline for guessing randomly is much lower (5.8% for twenty emotions versus 16.7% for six). This means the F1-scores of our method are more difficult to achieve given the challenging nature of the problem.

The suggested method was proven to be beneficial while using as an input three different kinds of initial lexicons. The performance of the resultant classifiers seems to vary depending on the amount of pseudo-labeled emotional data discovered by initial emotion lexicons. It would be interesting for future studies to examine what quantity of unlabeled data is required for the successful distant learning process. Moreover, we observe that the initial lexicons can have different best-detected categories. This can motivate future research in aggregating the classifiers obtained via distant learning from different initial lexicons in order to build the classifier having a better quality. It is also possible that repeating the learning process in iterations starting from the previously learned lexicons (by exploiting the ideas of self-training [QZHZ09, Zhu05]) could bring further improvements and requires more investigation.

We confirmed the contribution of the main components specific to our Dystemo method. The rebalancing, introduced in Balanced Weighted Voting (BWV) learner, leads to the relative increase of micro-F1 score by 9.2%. Techniques for balancing training data have never been tested for emotion recognition before. Applying other rebalancing techniques [BPM04] and testing how rebalancing processes help other learning algorithms for emotion recognition could be interesting avenues for future research as well. Another distinguishing property of our method is inclusion of novel heuristics to identify neutral tweets for learning. Our experiments show that this is essential to avoid constructing classifiers that find emotions in almost every tweet: when starting from PMI-Hash, accuracy grows from 34% to 64.1%. While a distant supervision over pseudo-neutral tweets was already proposed in the context of polarity classification [PP10, KWM11], for the problem of emotion classification, a *Neutral* category was only studied when training data were labeled manually, for example by Neviarouskaya et al. [NPI11a].

By comparing the suggested BWV learning method with other more advanced supervised classifiers, we show that even a simple lexicon-based classifier can achieve competitive performance. Yet, the additional advantage of BWV is that it produces an emotion lexicon, where each term (n-gram) is associated with an emotion distribution (called emotionality). This property opens a large perspective for potential future applications and improvements, such as extracting lexicon-based features for more advanced machine-learning classifiers [Moh12a, WCTS12].

While investigating the viability of distant learning starting from emotion lexicons, we used relatively simple features for classification, namely n-grams appearance, and simple feature aggregation techniques, that is averaging the distributions of appeared n-grams. We further review what mistakes our classifiers repeatedly made due to these simplifications (see example tweets with errors in the appendix section A.6). Many seem to appear in the tweets where emotional sense is captured within spans of texts longer than n-grams. Examples are “Why is <x> always on when I want to watch <y>?" and “<x>’s hopes for medal in <y> dashed". Our method would potentially benefit from incorporating more developed techniques of representing emotional meaning in text, such as parsing semantic concepts [PAG+14] or extracting main emotional parts [SEHHE14]. Similarly, modeling semantic compositionality could help to better aggregate detected lexicon features into tweet-level emotions. An example solution can involve treating emotions in composite phrases using hand-coded rules [NPI11a] or deep neural network representations [SPW+13, SM15]. Another source of mistakes is the lack of proper modeling of contextual modifiers that can change the emotional meaning of terms. In the future, we plan to include better treatment of such linguistic modifiers as negations (e.g. “lose interest") and downtoners (e.g. “least favorite") while applying both final and initial classifiers [CdAP13]. Finally, we observe tweets with the mixture of positive and negative emotions (e.g. “unlucky <x>, we are still proud of you"). Learning from them is likely to cause erroneous associations of terms to positive and negative emotions simultaneously. Future work should address how to limit the scope of corresponding emotion descriptions in the text, e.g. based on annotating parts of the texts with off-the-shelf polarity classifiers, such as SentiStrength [TBP12]. Another potential for achieving better recognition quality can be to adapt a hierarchical approach to the classification, where the classifiers should first decide whether a tweet is emotional or neutral, positive or negative, and only then classify it into specific emotions [GIS10].

The distant learning method developed and analyzed in this chapter is potentially valuable to many domains of textual emotion analysis lacking easily accessible labels. Further studies are required to determine whether these results can be generalized to those domains (e.g. reactions to other public events such as awards or elections, product reviews, or posts in support forums) with their corresponding sets of emotions.

6.8 Chapter Summary

This chapter presents an in-depth study of a distant learning method for multi-category emotion recognition in tweets, called Dystemo. The distant learning approach allows building new emotion classifiers using only automatically annotated data based on some heuristics. Instead of defining heuristics for detecting emotional tweets based on hashtags or emoticons, we argue for the use of existing or easy-to-produce emotion lexicons as a starting point. We describe a method that can either extend an initial lexicon to cover more emotional terms and expressions, or refine it to detect emotional tweets more correctly. Both improvements make the novel classifiers more suitable for the chosen application.

Using sports tweets as a dataset, we have shown a detailed validation process involving three different initial emotion lexicons for the classification of twenty emotion categories. The proposed distant learning method, applied with a novel supervised learner—Balanced Weighted Voting,—improves the micro F1-score in all three cases, with relative increases between 41% and 236%. Subsequent experiments suggest that rebalancing initially labeled data is an essential step in our method’s success. Among other contributions, we introduce heuristics to automatically find neutral tweets and show the importance of including them in the learning process.

Our research shows the viability of a distant learning method as an alternative way to build new emotion classifiers for a specific application domain and a specific set of emotion categories. In contrast with human computation method, it can built domain-specific emotion classifiers without costly manual labeling, while using only limited input resources in the form of an emotion lexicon. Yet, the same method can be applied along with human computation in order to further improve the quality of a lexicon obtained from the small manually annotated data.

Our method additionally avoids relying on special content cues such as hashtags, which makes it more general than other existing distant learning methods. Because of these properties, Dystemo can be used to build tailored emotion classifiers across multiple different domains of applications with minimal effort.

7 Modeling Effects of Modifiers on Emotional Statements

7.1 Introduction[1]

One of the major challenges in textual emotion recognition is to decipher the subtle ways humans express emotions. Even when the text contains an explicit emotional word, e.g. ‘happy’, this word can appear in the scope of a modifier that changes its emotional meaning, such as after the negation word ‘not’ in the phrase “I am not happy.” While emotion lexicons provide direct associations of lexical terms with emotions, the true feeling expressed in the text can change under a variety of modifiers. The question is how such modifiers impact the used emotional expressions? In order to improve the recognition quality of lexicon-based systems, we should be able to properly model the effects of different modifiers on emotions. Addressing this challenge is the subject of this chapter.

Multiple linguistic constructions can potentially affect the emotional meaning of words in their scope. Illustrative examples from Table 7.1 show how different modifiers can lead to different effects. Example (b) illustrates how negation (e.g. *not*, *never*) can shift the original emotion to another emotion. Examples (c) and (d) show how intensifiers (e.g. *so*, *very*) and diminishers (e.g. *a little*, *slightly*) can affect the emotion’s intensity. Example (e) shows how

[1] The analysis method described in this chapter was partially implemented by Margarita Bolívar Jiménez.

(a)	I am happy today	No modifier
(b)	I’m not ashamed to say it	Negation
(c)	I feel so relieved now	Intensifier
(d)	I feel a little sad tonight	Diminisher
(e)	I know I should be happy	Modality
(f)	I’ll be sad if you leave	Conditionality
(g)	Do you love her?	Interrogation
(h)	I was happy then	Past Tense

Table 7.1: Illustrative examples of the modifiers’ effects on emotional statements.

modality (e.g. *should*, *can*) can eliminate the presence of emotion. Example (f) reveals that conditional emotional statements (e.g. *if*-clause) can refer to a non-experienced emotion. Example (g) shows how interrogative mood (e.g. question mark *?*) can express uncertainty in emotional experience. Finally, past tense can convey emotion that is not experienced at the present time, as example (h) illustrates.

These linguistic modifiers are used ubiquitously in the language. Our analysis of tweets—the main application focus of this thesis— shows that 34% of explicit emotional terms appear in the scope of the mentioned modifiers. This large proportion of modified statements suggests we need to properly model the modifiers' effects.

Previous studies do not fully answer the questions of how modifiers affect specific emotions. They generally model only the effects of the most impactful modifiers—negation and intensification—in terms of polarity and intensity change of the words [HVIH$^+$11, HG14, Kou14]. The effects of other modifiers are either disregarded by ignoring the modifiers' presence or blocked by removing the modified statements [TBP$^+$10]. How different modifiers affect specific emotions is under-studied. In cases when effects are modeled at an emotion level [CdAP13], they are hand-coded, which makes their adaptation to other emotion categories more difficult.

This chapter suggests a unified framework for analyzing the effects of different modifier types on fine-grained emotion categories. We quantify the impact of six different linguistic modifiers on each specific emotion using a novel data analysis method. It is based on investigating emotion distributions of modified and non-modified occurrences of emotional terms in social media. The source of our data is Twitter, from which we collect tweets having an emotional hashtag, considered to be the author's self-revealed emotion. Having such labeled data, the emotion distributions in question are computed as the distributions of emotion labels in the corresponding tweets. In each of the emotion distribution comparisons, we will use Kullback-Leibler divergence [KL51, CT06] to determine the magnitude of difference between distributions. Instead of requiring manual hand-coded rules, this data-driven method derives the model of the modifiers' effects from the patterns of their usage in the large corpus of linguistic data automatically labeled with emotions. This makes this method easily adaptable to different emotion categories and modifiers to model.

The main focus of this chapter is the method to analyze whether the original emotion of an emotional statement is affected by the presence of a modifier affecting it and how it is affected. For each modifier type and for each considered emotion category, we answer three questions to understand and quantify the effect of modifiers:

1. *To what extent does the modified emotion differ from the original non-modified emotion?* This quantified difference will estimate to what extent the emotional meaning of emotional statements changes under each modifier. This helps to make informed decisions on how important it is to treat the effects of a specific modifier.

2. *Does the original emotion change under the modifier into another outcome emotion, or does it stay the same (shift or no shift)?* This will tell us whether the modifier changes the original emotion, and, if yes, towards which outcome emotion. This knowledge is necessary for properly modeling the modifier's effect within emotion classification.

3. *How confident are we that the discovered outcome emotion is actually expressed in the modified text?* Instead of an absolute value, we compute a confidence coefficient that estimates the confidence of the modified emotion relative to the confidence in the non-modified outcome emotion. Computing this coefficient pursues two goals. First, the effect of the modifier can be captured by an increase or decrease of confidence, even when no shift occurs. Second, knowing confidence coefficients of each modified case, we can eliminate less-confident modified statements in order to obtain more precise classifiers.

Overall, our method produces a fine-grained emotion-based model of modifiers' effects, describing how each emotion changes under each modifier. We discover that the effects of all studied modifiers are emotion-specific, and thus such a detailed model is essential for more precise treatment of modifiers. Our analysis sheds light on how to model negation relations between emotions and additionally reveals that not only negations can shift emotions, but also other less studied modifiers, such as modality and interrogation. Finally, we show the potential of the proposed modeling to find more precise emotional statements. All these findings lead to important implications for developing a modifier-aware emotion classification system.

7.2 Studied Modifiers Types

Different modifiers are presumed to influence the meaning of emotional statements. The modified statement might express another emotion, refer to a non-experienced emotion, have a different intensity of the same emotion, or even not express any emotion at all. In our study, we consider those modifiers that could influence the emotional meaning of terms based on their linguistic role, as inspired by the previous research in sentiment analysis.

Negation Negation is the most studied modifier of sentiment polarity. Researchers have proven that it can change both the polarity and intensity of the words within its scope [BCM$^+$12, HVIH$^+$11, JYM09]. Yet, its effect on specific emotion categories remains understudied. While the reversal of polarity is a straightforward operation (change of positive polarity to negative, and reverse), the reversal of emotion is not as simple because of complex relations between emotional concepts and between their linguistic expressions. For example, the phrase "I don't love you" implies rather *Sadness* than either original emotion of *Love* or the antonym emotion of *Hate*. We assume that negation of an emotional term may express the absence of any specific emotion, may refer to another emotion category, or may just change the intensity (or confidence) of the given emotion.

Intensification Intensification modifiers involve terms that change the intensity of emotional words. They are separated in two classes: *intensifiers* that increase the intensity, such as ‘very’ or ‘really’ (also called amplifiers), and *diminishers* that decrease the intensity, such as ‘less’ or ‘little’ (also called downtoners). While we hypothesize that neither intensifiers nor diminishers change the original emotion category, we assume that the confidence in the emotion presence would change according to the direction of intensity shift (an increase or decrease). For instance, in the sentence “I love you so much”, *Love* emotion is intensified, and we can be more certain that it is present.

Modality Modality is a linguistic construction used to distinguish non-factual situations (irrealis events) from events that happened or are happening (realis events). Modal operators can express a degree of uncertainty or possibility, and can also be used to express desires and needs [PZ06, CdAP13]. Consider, for instance, the phrases “I will regret it”, “I would be happy to have it”, and “I should be angry at you". In all of these examples the presence of modal verbs concealed whether the writer actually experienced the referenced emotion or not. Some modal expressions can directly imply the absence of the referenced emotion, as in “Would have loved to see you”. Most researchers that consider modality in sentiment analysis treat it as a polarity blocker, ignoring the occurrence of sentiment terms in its scope (e.g. [TBT+11]). Others suggest hand-coded coefficients for the change in certainty or confidence [NPI11b]. We suggest to further investigate the effects of modality on specific emotions using a data-driven modeling.

Conditionality Conditional sentences can also describe the irrealis event, i.e. those potential or hypothetical situations that are not yet known to happen. For instance, in the sentence “I’ll be sad if you leave”, the emotion *Sadness* is not experienced yet. Some studies already suggested that emotions in conditional sentences can be harder to classify [NLC09]. We hypothesize that emotional statements appearing in conditional sentences will have less confidence, because we can not be certain whether they were actually experienced.

Interrogation Interrogation represents the sentences where a question is asked. Similarly to conditionality, we cannot be certain whether the states or events mentioned in questions actually happened. Thus, interrogative sentences can change our confidence in detected emotion, as shown by the example question “Do you love me?” However, interrogation can even shift the original emotion to another one: e.g. the sentence “Are you angry at me?” implies rather *Worry* than *Anger*. As interrogative emotional statements are not common in the opinionated review texts, the effects of interrogation are traditionally neglected in sentiment analysis. Yet, they are frequent in personal communications and deserve further investigation.

Past Tense Past tense describes situations that happened in the past. Therefore, we can be more certain that the stated emotions were experienced (a confidence increase). Yet, expressing the emotion in the past can also mean that currently it is not experienced anymore (a confidence increase). Consider two phrases illustrating these two effects: “I was happy with you” and “I loved you so much before”. Thus, we investigate the effects of past tense to identify which case is more frequent.

7.3 Related Work

This paper studies the per-emotion effects of six linguistic modifiers, including negation, intensification, modality, interrogation, conditionality, and past tense. We review in this section previous related works on modeling and analyzing effects of the considered modifiers in the context of sentiment analysis.

The effect of negation has been widely studied over the recent years in the context of polarity classification. The most basic approach is to separate negative, positive, and neutral words, and to consider negation as a polarity reversal [PZ06]. Several other studies concluded that negation affects both polarity and intensity in the words that are within its scope [BCM$^+$12, HVIH$^+$11, JYM09, Kou14]. Other researchers found that negation effects depend on the prior polarity of words and the used negation expressions [ZGMK14]. In automatically learned systems for polarity and emotion classification, negations are often taken into account at the term level, by adding a separate feature for each negated term [PLV02, KZM14]. Taking this previous research into account, a more fine-grained negation treatment is needed. We suggest following an emotion-based approach, for which the effect of negation is modeled separately depending on the emotion of a modified term. With the antonym-based reversal of emotions under negation, this approach was proven to increase the accuracy of polarity and intensity classification [CdAP13].

To treat intensification, some methods add (subtract) points from the valence score of sentiment terms if they are preceded by an intensifier (a diminisher) [PZ06, KI06]. Other methods improve this idea and associate each intensifier and diminisher term with a multiplication coefficient representing its strength [TBT$^+$11]. The specific coefficient values are then hand-coded separately for each modifier term.

Regarding modality, most researchers suggest treating modal expressions as polarity blockers that remove the polarity of the terms they affect [PZ06, TBT$^+$11]. The linguistic study of Benamara et al. [BCM$^+$12] shows that modality has an effect on the strength and the degree of certainty in the opinion words that are within its scope. The effects of modality were also hand-coded with the per-expression certainty coefficients [NPI11b]. Carrillo-de-Albornoz and Plaza [CdAP13] manually developed a model of effects of modal verbs on the emotional expressions by investigating their use in the review texts. They suggested that some modal expressions can as well reverse the emotion.

Conditionality, interrogation, and past tense are treated by very few approaches. One example that studies sentiment expressions within conditional sentences is the work of Narayanan, Liu and Choudhary [NLC09]. They study the different types of conditional sentences and adapt a machine-learning algorithm to determine whether the modified opinion is positive, negative, or neutral. In addition to modality, Tabaoda et al. [TBT$^+$11] consider both conditionality and interrogation as polarity blockers (irrealis blocking), that is they ignore any sentiment terms within the scope of those modifiers. Past tense has not been deeply studied regarding its effects on polarity or emotion of terms, or their intensity. So we aim to conclude whether these understudied modifiers are relevant for consideration in emotion recognition.

The above-mentioned types of modifiers have not yet been studied altogether within the same analytical framework, while considering their effects on fine-grained emotions. We aim to study how these modifiers affect the emotions and how they should be treated, using a novel data analysis method that allows to quantitatively model the modifiers' effects. This method uses the Kullback-Leibler divergence—a measure of dissimilarity between two probability distributions—to investigate the differences between emotion distributions of modified and non-modified emotional statements. Furthermore, the same method can quantify the confidence level of the found outcome emotion association under a modifier, without requiring any confidence or intensity labels in the ground-truth data.

7.4 Our Analysis of Modifiers

Our analysis of modifiers aims to answer the following questions: Does a modifier change the original emotion of an emotional statement? How confident are we that a detected outcome emotion is associated with the modified emotional statement? Which modifier type has the highest impact on emotional expressions? Does the impact of modifiers differ between emotion categories? Answering those questions will allow building a computational model of the modifiers' effects.

This section describes in detail the method we apply to quantify and analyze the effects of modifiers on emotional expressions. We introduce below the model of emotion categories and the corresponding lexicon of explicit emotional terms that we use in our study. We describe the data labeled with those emotions and our approach to detect modifiers for the studied emotional terms. Finally, we explain how these components are employed altogether to quantify the impact of different modifiers types on the original emotion expressed by an emotional term.

7.4.1 Input Data

For the current modifiers' analysis, we again use the same fine-grained emotion model of 20 GEW emotion categories, described in section 3.2. We then use the GALC-R as an input affective lexicon of explicit emotional expressions associated with the studied emotions. This lexicon was presented in section 3.3. In this work, we study the effects of modifiers only on these explicit terms.

Twitter Data Labeled by Emotional Hashtags For the purposes of our analysis, we require a large dataset labeled with emotions. As manual annotation is not achievable at the desired scale, we resort to the pseudo-annotated dataset of tweets. We use the $1,729,980$ general tweets with emotional hashtag corresponding to one of the GEW emotions. This dataset EMHASH_DATASET was presented in section 3.4. All these tweets were converted to lowercase and preprocessed to correctly separate emoticons, usernames, and punctuation marks from other tokens. The emotion category associated with an emotional hashtag is considered to be an emotion label for the full tweet text, and referred to as *hashtag emotion*. We randomly sampled 1.5 million of the general hashtagged tweets to be used for studying the effects of the modifiers on emotional terms (analysis dataset D_A). The remaining $229,980$ tweets will be used later in the emotion classification experiments (test dataset D_T).

7.4.2 Data Preparation for Analysis

To prepare analysis data, we will first detect the tweets where given explicit lexicon terms (from GALC-R) appear in the tweet's content (while ignoring hashtags), and detect any modifiers applied to them. Essentially, this provides three data points for every tweet with a lexicon term: what is the associated emotion category of that term (*term emotion*), whether that term is modified and with which modifier (*non-overlapping modifier class*), and the tweet-level emotion label based on its emotional hashtag (*hashtag emotion*). For example, from the tweet "I don't love it #sad" we will extract *term emotion* category *Love* (based on term *love*), Negation as *modifier class* (based on negation term *don't*), and *hashtag emotion* label *Sadness* (based on hashtag *#sad*). Then, we will extract for analysis the emotion distributions of the *hashtag emotion* separately in each subset of the tweets with the fixed *term emotion* and *modifier class*. Note that multiple lexicon terms correspond to each emotion (in average 52.9 terms per emotion), and thus we will aggregate for each *term emotion* all the tweets with the corresponding emotional lexicon terms. The overview of this process is pictured in Figure 7.1.

Detection of Lexicon Terms

To start, we identify all the tweets that contain exactly one emotional term from the revised emotion lexicon GALC-R. We look for its terms in the content text of the tweet, i.e. while disregarding the emotional hashtags. For example, we will detect the emotional term "happy" in the tweet "Happy to see you #happyday". There are $255,467$ such tweets in the analysis dataset D_A.

Detection of Modifiers Affecting Lexicon Terms

The next step is to apply the modifiers detection module to discover which of those emotional terms are modified with which modifier type. We detect modifiers based on the presence of specific modifiers' words and multi-word expressions near the emotional term in question.

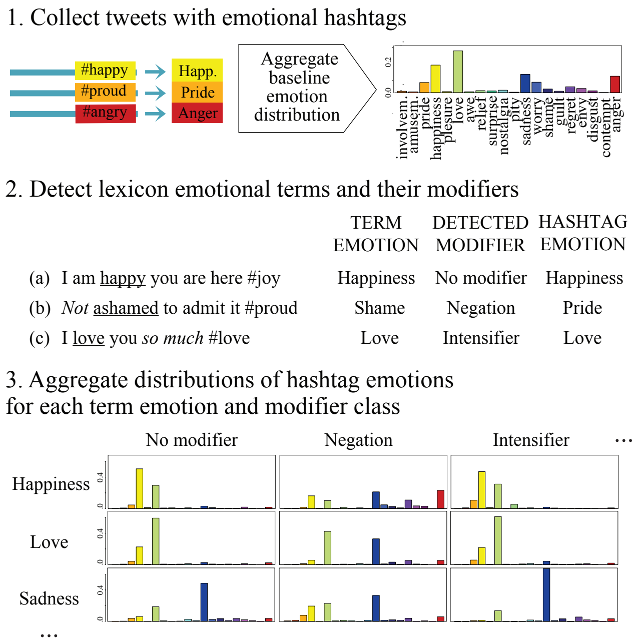


Figure 7.1: The process of data collection, preparation, and extraction of emotion distributions for the modifiers' analysis.

The scope of modifiers influence is defined per modifier type, but in most cases it is either three or four words after the modifier. Additional heuristics are applied to avoid errors in modifiers detection. One of them is not to allow for punctuation marks or emoticons in between the modifier and emotion terms. Another one is to exclude frequent false positive expressions with modifier words. The full list of the corresponding modifiers' terms can be found in the appendix, section A.7.

Negation The basis of our approach to detect negations is the list of negation expressions. It contains common negation words, such as *wasn't*, *not*, or *no*, and their misspelled variants, such as *werent* or *didnt*, taken from Carrillo-de-Albornoz and Plaza [CdAP13]. The list of negations also includes verbs such as *pretend*, *fail*, or *refuse* implying that a modified statement is not experienced or does not happen. Those verbs are taken from Lotan, Stern and Dagan [LSD13], where they are marked as having a negative signature (38 verbs in total). We additionally identify false negation expressions, i.e. multi-word expressions with negation words that actually do not negate the meaning of the statements, such as *nothing but*, *don't you just*, and *can't help*. Some of them are marked as intensifiers, e.g. *couldn't be more* and *never felt so*.

At the detection phase, we detect which negation expressions end up to three words before the emotional term in question. Among the found expressions, the longest one is selected and the term is marked as negated or not negated according to its label. We additionally ensure that no punctuation marks or emoticons appear in between the negation and emotion terms. And we deal with double negations, which are marked as not negated.

Intensification The method for detecting intensity change follows the same principles as for negation, detecting intensifiers and diminishers based on the corresponding lists of words and multi-word expressions. Those lists were collected from the related literature [CdAP13, HG14]. In order to refine their scope, we classify the phrases of intensifiers and diminishers according to their position in text, depending on whether they can appear before (e.g. *lots of*, *kind of*) or after (e.g. *very much*, *less*) the emotion term, or both. Several false positive phrases are also added: for example, when *kind of* appears within *that kind of* or *least* – within *at least*, they are not diminishers anymore. Additionally, we add frequent n-grams with intensity words that appear directly before or after the studied emotional terms and reveal a high confidence of changing the intensity accordingly, e.g. *how much i* and *you so much*. This allows us to limit the scope of these modifiers to one term directly before or after.

Modality Modality is also detected using the list of modal expressions, which have to appear before the emotional terms (we consider the scope of four terms in this case). The significant part of such expressions consists of modal verbs, such as *should*, *might*, or *can*. *Will*, *'ll*, *wont* are also in this list, i.e. the emotional terms in the future tense will be detected as modified with modality. Additionally, our list of modal expressions contains the expressions of desire (e.g. *wish*, *want*, and *hope*) and of uncertainty (e.g. *maybe*, *seems*, and *i doubt*). Note that we specifically avoided including modal expressions of certainty or 'trueness', such as *sure* or *indeed*, because they are assumed to have a different effect on emotions than other studied modal expressions [NPI11b].

Conditionality, Interrogation, and Past Tense The detection of the remaining modifier types follows the same principle of finding a modifier term and checking whether an emotional term is within its scope (which is different depending on the modifier type). Interrogation is detected by checking whether there is the interrogation sign '*?*' after the detected emotional term or the question-specific patterns, such as *am i* and *why does*, before the term. Conditionality detection is based on finding the word *if* before the emotional term in the same sentence (i.e. no punctuation marks for sentence separation or emoticons are in between). To detect past tense, we check for the part-of-speech tag specific for verbs in the past tense, using Stanford POS Tagger [TKMS03]. The emotional term is considered to be in past tense if this tag appears up to four tokens before it.

Separation of Non-Overlapping Modifier Classes

Several modifiers can modify the same term in the text, e.g. both negation and intensification are present in the phrase "not very interested". To exclude intertwining effects between modifiers from our analysis, we split the entries of modified terms into *non-overlapping modifier classes* using the following rules. We recognize Past tense modifier only if it does

Non-overlapping Modifier Class	Modifier Presence	Top 5 Modifier Entries	Modifier Class Composition
Negation	2.4%	not, n't, no, never, stop	Only, +Past
Mixed Negation	1.2%		Negation + Another
Intensifiers	14.9%	so, really, very, more, absolutely	Only, +Past
Diminishers	0.6%	kinda, only, a little, bit, kind of	Only, +Past
Past Tense	5.2%	was, had, loved, got, made	Only
Modality	4.3%	would, can, will, could, going to	Only, +Past
Conditionality	1.3%	if	Only, +Past, +Modality
Interrogation	2.9%	?, ??*, ?!, am i, do you	Only, +Past, +Modality
Mixed	2.2%		Other compositions
Any Modifiers	34.0%		

Table 7.2: Statistics on the presence of the modifiers' classes in the studied hashtagged tweets. We report the percentage of the modified terms from the GALC-R lexicon as a metric of the modifier's presence. We also enumerate the top modifier entries for each class and summarize our rules for deriving the non-overlapping modifiers' classes.

not overlap with any other modifier, otherwise we assign the overlapping modifier alone (i.e. Past Tense plus Negation will be assigned to the Negation class). The case of Modality and Conditionality is assigned to Conditionality only. The same is for Modality and Interrogation (assigned to Interrogation only). We also separate a class of Mixed Negation containing all the cases where other modifiers (except for Past Tense) overlap with Negation. All other overlapping cases of found modifiers are placed into the Mixed class and are not considered in the analysis of the modifiers' effects. Table 7.2 summarizes these rules and provides statistics on the presence of each studied non-overlapping modifier class within the analysis dataset D_A. We compute for each modifier class how often the terms from the GALC-R lexicon are modified by it.

Overall, 34% of emotional terms from GALC-R are modified by at least one modifier, with Intensifiers being the most common modifier (14.9% of entries), followed by Past Tense (5.2%). Negations in total modify 3.6% of terms, while 32% of them are Mixed Negation cases. Mixed class covers only 2.2% of emotional terms.

Extraction of Emotion Distributions

We will call *emotion distributions* the distributions of the *hashtag emotion* labels of the considered tweets. That is we compute for each hashtag emotion category what percentage of the studied hashtagged tweets have an emotional hashtag corresponding to that emotion. We compute the emotion distribution of all tweets in the analysis dataset D_A without any restriction and refer to it as the baseline distribution P_{BASE}.

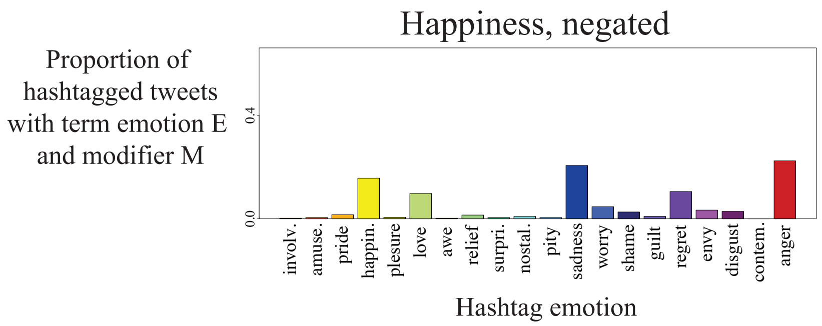


Figure 7.2: The annotated example of the modified emotion distribution $P_M(E)$ for term emotion E=Happiness and modifier M=Negation. Each bar of the distribution represents the percentage of tweets with the specific hashtag emotion (defined as a category of the tweet's hashtag) among the hashtagged tweets containing the lexicon terms that are associated with the studied term emotion and are modified by the studied modifier.

In the current format, each tweet with a found emotional term is represented as a triple of a *term emotion*, a *non-overlapping modifier class*, and a *hashtag emotion*. Then, for each *term emotion* E and for each *modifier class* M, we compute the *modified emotion distribution* $P_M(E)$ of the hashtag-based labels from the corresponding tweets. Figure 7.2 illustrates an example of such modified emotion distribution.

We also compute for each term emotion E the distribution of hashtag-based emotion labels of tweets with terms assigned to that emotion but without any modifiers in the text. That is we exclude the tweets where any modifiers' expressions appear, even when they are not applied to a lexicon term. This is to ensure that the non-modified emotion distributions are indeed non-modified, regardless of the potential mistakes of the modifiers' scope detection. We will refer to this distribution for term emotion E as the *non-modified emotion distribution* $P(E)$.

Our analysis of effects of each modifier class will be based on comparing the extracted emotion distributions to each other (e.g. non-modified vs. modified distribution for each term emotion).

7.4.3 Quantification of Modifiers Impact

For each modifier class, we study how it affects the emotion of terms based on the change in the emotion distributions of the tweets with modified and non-modified terms. We quantify the influence of the modifiers by comparing the corresponding distributions of hashtag emotion labels using the Kullback-Leibler (KL) divergence [KL51, CT06].

The KL divergence is an asymmetrical measure of the difference between two probability distributions S and Q. In our discrete case, it is computed with the following expression:

$$D(S||Q) = \sum_i s_i \log \frac{s_i}{q_i}$$

where s_i and q_i are the corresponding percentage of emotion E_i in the emotion distributions S and Q. The KL divergence measures how well the distribution S could be approximated by the distribution Q. The closer it is to zero, the better is the approximation. Thus, following the goals of our analysis to describe the modified distributions, we will always consider more restrictive modified distributions $P_M(E)$ as S in the formula, and take more general non-modified distributions as Q.

To obtain representative modifier emotion distributions, we include in the analysis of a modifier only those emotions for which at least 50 tweets contain such modified terms. Also, to avoid potential division by zero in the KL computation when emotion distributions are sparse, we add a smoothing constant of 0.05 to each emotion label count before normalizing the distributions to percentage values.

Our analysis of modifiers' effects aims to answer three questions regarding the effects of each modifier class M on a specific original term emotion E:

1. *To what extent does the modified emotion differ from the original non-modified emotion? (modifier divergence)*

We answer this question by comparing the emotion distributions of the modified cases with the ones without any modifier (e.g. distribution with negation vs. non-modified distribution for the original term emotion E). This means we compute the KL divergence $D\big(P_M(E)||P(E)\big)$. We refer to this metric as a *modifier divergence*.

2. *Does the original emotion change under the modifier into another outcome emotion, or does it stay the same? (shift or no shift)*

To detect which non-modified emotion approximates the best the extracted modified emotion, we compare the distribution of the modified emotion $P_M(E)$ with distributions of each non-modified emotion $P(E_i)$. The emotion E_i that provides the minimal KL divergence will be referred to as the *outcome emotion* E_{out} under that modifier, i.e.

$$E_{out} = \arg\min_{E_i} D\big(P_M\,(E)\,||P\,(E_i)\big)$$

We say that the modifier *shifts* the emotion E if the outcome emotion is different from the original emotion, i.e. if $E_{out} \neq E$. Otherwise, we say the emotion remains the same under the modifier ($E_{out} = E$).

3. *How confident are we that the discovered outcome emotion is actually expressed in the modified text? (confidence coefficient)*

Regardless of whether there was a shift of emotion or not, it is likely that the modified distribution $P_M(E)$ differs from the closest non-modified distribution of the outcome emotion $P(E_{out})$.

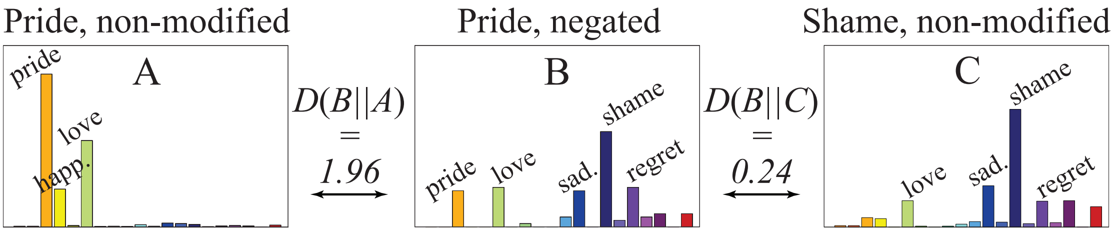


Figure 7.3: Examples of non-modified (A) and modified (B) emotion distributions for analyzing the effects of Negation on Pride. (C) visualizes the non-modified distribution of the outcome emotion Shame.

It can differ in two ways: the modified emotion distribution can have a more pronounced peak for the outcome emotion, or it can have a more random distribution, corresponding to a more mixed state of emotions or an absence of them. The first case intuitively increases our confidence that the outcome emotion is present, while the second one decreases this confidence.

To quantify these effects, we compute a *confidence coefficient* (CC) that measures a change of confidence in the presence of the outcome emotion in modified distribution relative to such confidence in the non-modified case. To compute it, we additionally compare both modified and non-modified distributions with the baseline emotion distribution of all analysis tweets P_{BASE}. We suggest computing the *confidence coefficient* (CC) as a ratio of two KL divergences: one between the modified and baseline distributions and one between non-modified distribution of the outcome emotion E_{out} and the baseline distribution, i.e.

$$CC = \frac{D\big(P_M(E)||P_{BASE}\big)}{D\big(P(E_{out})||P_{BASE}\big)}$$

The intuition behind this metric is the following. If the modified emotion distribution $P_M(E)$ is closer to the baseline than the non-modified distribution of E_{out}, it means that the modified emotion distribution $P_M(E)$ is more random than the non-modified $P(E_{out})$, leading to a smaller confidence in the outcome emotion's presence ($CC < 1$). Alternatively, if the modified emotion distribution $P_M(E)$ is further from the baseline distribution than the non-modified distribution of $P(E_{out})$, we can be more confident that the outcome emotion E_{out} is actually present ($CC > 1$).

To illustrate the suggested analysis method, we visualize the corresponding emotion distributions in the case of analyzing how Negation modifier affects original emotion of *Pride* (Figure 7.3). It can be observed that the distribution for negated Pride (B) is considerably different from the non-modified Pride distribution (A). More particularly, it has the peak on *Shame* instead of *Pride*, which leads to a high modifier divergence value of 1.96. Furthermore, this makes the non-modified Shame distribution (C) to be the closest to the negated Pride distribution. We thus infer that *Shame* is the outcome emotion of *Pride* under Negation. However, negated Pride distribution (B) has higher percentage of *Pride* and *Love* than non-modified Shame distribution (C), showing that it does not follow it exactly. In result, (B) is closer to the baseline distribution P_{BASE} than (C) (1.02 vs. 1.11), which in its turn results in a confidence coefficient less than one ($CC = 0.92$).

7.5 Discovered Effects of Modifiers

The presented method describes the effects of each modifier class M on each emotion E in terms of four characteristics: modifier divergence, the outcome emotion, whether there is a shift of emotion, and the confidence coefficient of the outcome emotion. In this section, we summarize the overall effects of each modifier and discuss the most interesting findings.

Table 7.3 presents the aggregated effects of modifiers. We report for each modifier the average of modifier divergences between emotion categories and their maximum (along with the names of the corresponding original and outcome emotions). We also summarize the shifting and confidence change behavior across the emotion categories: what proportion of the original emotions shifts into other outcome emotions with either increase or decrease of confidence, and what proportion of emotions remains the same under the modifier, also separated in the two groups for increase and decrease of confidence. Note that we use in our analysis (and thus in this aggregation) only the emotion categories for which enough modified entries are detected (≥ 50). The presented metrics show how each modifier affects the explicit emotional statements in general.

These results confirm the expected differences in the impact of different modifiers, with Intensifiers being the least influential modifiers and Negation—the most. At the same time, they show that the effects of each modifier differ depending on the emotion category it modifies. This is reflected in the fact that every modifier shifts at least one emotion to another, but none of modifiers has the same effect on all emotions. According to the overall shifting pattern, we separate three groups of effects: no shift, mixed, and shift.

Modifiers with No-Shift Effects The smallest value of average modifier divergence belongs to Intensifiers. This was expected given that Intensifiers should only increase the intensity of the emotion, but not change the category. This effect is confirmed for most emotions, with Intensifiers not shifting the original emotion, but increasing its confidence ($CC > 1$).

Modality and Past Tense modifiers also have a relatively low impact. Similar to each other, they mostly leave the emotion category non-shifted, but lead to its smaller confidence ($CC < 1$). This means that these two modifiers can introduce uncertainty in the expression of a specific emotion. Yet, some negative emotional terms expressed in the past tense are linked more confidently to the associated emotions, as in the cases of “It was disgusting” or “I was disappointed".

Modifier	Modifier Divergence			% of shifted		% of non-shifted	
Class	Mean	Max	(orig. → outcome)	$CC > 1$	$CC < 1$	$CC > 1$	$CC < 1$
Intensifiers	0.12	0.61	(Nostal. → Regret)	0%	11%	**78%**	11%
Past Tense	0.17	0.36	(Guilt → Guilt)	0%	6%	19%	**75%**
Modality	0.19	0.87	(Involv. → Worry)	6%	13%	6%	**75%**
Conditionality	0.26	0.52	(Involv. → Sadn.)	0%	27%	18%	**55%**
Diminishers	0.28	0.83	(Nostal. → Regret)	11%	11%	**56%**	22%
Interrogation	0.40	0.76	(Awe → Involv.)	29%	24%	**35%**	12%
Mixed Negation	0.52	1.30	(Pleasu. → Regret)	8%	**50%**	0%	42%
Negation	0.80	1.96	(Pride → Shame)	0%	**75%**	0%	25%

Table 7.3: Comparison of the different modifier classes using metrics aggregated across emotion categories. The modifier divergence for an emotion category is the KL divergence between modified and non-modified distributions. We count the percentage of shifted and non-shifted emotions under each modifier, aggregated by the confidence coefficient behavior (CC).

Conditionality and Diminishers have larger modifier divergence (between 0.26 and 0.28) and shift more emotions (27% and 22% correspondingly). Conditionality mostly introduces uncertainty in the emotion presence (confidence decrease) without shifting the original emotion. An interesting but counter-intuitive observation about Diminishers is an increase in confidence for many emotions. This is explained by the fact that when a person states he or she is "kinda/a little/only/a bit" "sad/in love/worry/disappointed" we can be more confident that the stated emotion is actually being experienced.

Even though the modifiers with the lowest modifier divergences mostly do not shift emotion, each of them shifts some positive emotions. For example, both Intensifiers and Diminishers have large effect on category *Nostalgia/Longing*, which is shifted towards negative emotion of *Regret*. This is because the expressions "really/kinda/kind of" "wish" are often used to express regrets. Another commonly affected category is *Involvement/Interest*. Modality shifts it into *Worry/Fear* with an increased confidence. This happens due to the widespread of the phrase "this will/shall/gonna/should be interesting" that expresses the author's worry about what is going to happen. Conditionality shifts it into *Sadness* because of the common pattern of desperate calls, for example "Looking for a roommate. If anyone interested let me know #desperate".

Modifiers with Mixed Effects Not every modifier has a clear overall shifting behavior. The effects of Interrogation and Mixed Negation depend largely on the emotions they modify.

We discover that Interrogation has its own pattern. It shifts many positive emotions into *Involvement*. This can be explained by the fact that asking other people questions about their positive emotions is an expression of *Involvement/Interest* by itself. Nevertheless, negative emotions mostly stay the same in interrogative sentences.

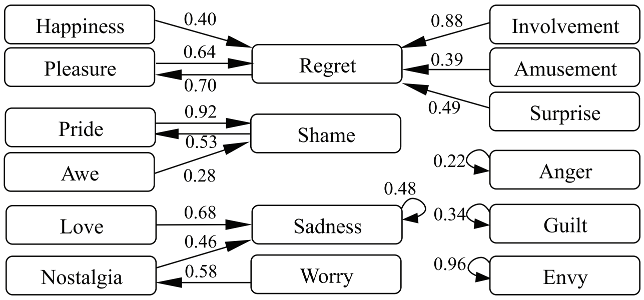


Figure 7.4: The extracted model of emotion shifts under negation. The arrows point to the outcome emotion of negating the original emotion. They are labeled with the confidence coefficients of the outcome emotions.

Mixed Negation has a mixed effect as well. For example, it shifts several positive emotions, including *Happiness* and *Pleasure*, into *Regret*, because of the dominance of Negation mixed with Modality (an example phrase is “I can’t be happy”). At the same time, many other emotions stay the same with the lower confidence.

Modifiers with Shift Effects The highest modifier divergence value corresponds, as expected, to Negation. In line with the previous findings in sentiment analysis, we observe that negation tends to shift emotions (it happens for 12 out of 16 analyzed emotions, i.e. for 75%) and decreases the confidence in the outcome for all emotions, even non-shifted (average *CC* is 0.56 < 1). However, our analytical method allows establishing which emotion the original emotion shifts to (i.e. the outcome emotion), and discovering emotions that do not shift. We investigate below in more details the impact of negation on each emotion category individually.

Figure 7.4 summarizes the shifting effects of negation. For each original emotion, the arrow points to the outcome emotion under negation and specifies the corresponding confidence coefficient (*CC*). Several clusters of negation effects can be observed.

1) Five of the positive emotions shift towards *Regret*, while *Regret* itself shifts back towards *Pleasure*. This cluster represents a standard notion of negation influence, where “not happy” and “not amused” are considered to have negative sentiment. It is noteworthy that *Happiness* does not shift to its direct antonym *Sadness*. Also, we do not have direct antonyms of *Amusement* and *Involvement* in the emotion model, thus under negation they shift into the most appropriate emotion category among the given ones.

2) We discover a reciprocal negation relationship along the antonym pair *Pride*-*Shame*. *Awe* also shifts towards *Shame*, which can be attributed to the frequently negated expression “no wonder”.

3) Negation of *Love* and *Nostalgia* becomes *Sadness*, as in the tweet “Nobody loves me enough to hang out with me”. At the same time, *Worry* shifts into *Nostalgia* based on the most frequent

phrase “don’t worry” that appears in many different contexts, both positive and negative. As the divergence between baseline and negated *Worry* distributions is small, it might rather represent a mixture of emotions than *Nostalgia*, even with a lower confidence.

4) There are four negative emotions, namely *Sadness*, *Anger*, *Envy*, and *Guilt*, for which there is no emotion shift under negation (i.e. they remain the same). This can be illustrated by the example texts “I’m not normally an angry ranty person” and “I’m trying not to get sad”. The confidence coefficients are small for all of these emotions, except *Envy*, for which it is close to one, meaning that “not envious” has almost the same meaning as “envious”.

Overall, all positive emotions are shifted to negative emotions, and several negative emotions shift towards positive ones. This confirms the expected power of negation to reverse polarity of emotions. Yet, we find no shift under negation for several negative emotions. This once again shows the importance of treating the effects of modifiers individually for each emotion.

7.6 Classification Quality of Modified Emotional Statements

We evaluate in this section how the classification quality of emotional statements depends on whether they are modified and how.

The aforementioned analysis constructed a quantitative model of modifiers’ effects: for each emotion category and a modifier, the model specifies what is the outcome emotion after modification and what is its confidence coefficient. For example, it specifies that a negated term of *Sadness* remains assigned to category *Sadness* with a confidence coefficient of 0.48.

As the basis of emotion classification, we use the GALC-R lexicon of explicit emotional terms. For each occurrence of a lexicon term, we detect which of the studied modifiers are present. We again use the non-overlapping classes of detected modifiers (as described in section 7.4.2). Based on the presence of modifiers and their discovered effects, we separate three cases of emotional terms’ occurrences:

- *Not Modified*: None of the studied modifiers is detected, we return the emotion associated with the emotional term;

- *No Shift*: Exactly one modifier is detected for the term, but it produces no shift of the emotion associated with the term. The term emotion is returned;

- *Shift*: Exactly one modifier is detected for the term, and it shifts the original term emotion into another outcome emotion. For this case, we also separate two scenarios for its treatment: whether we return the outcome emotion or the original emotion of the emotional term.

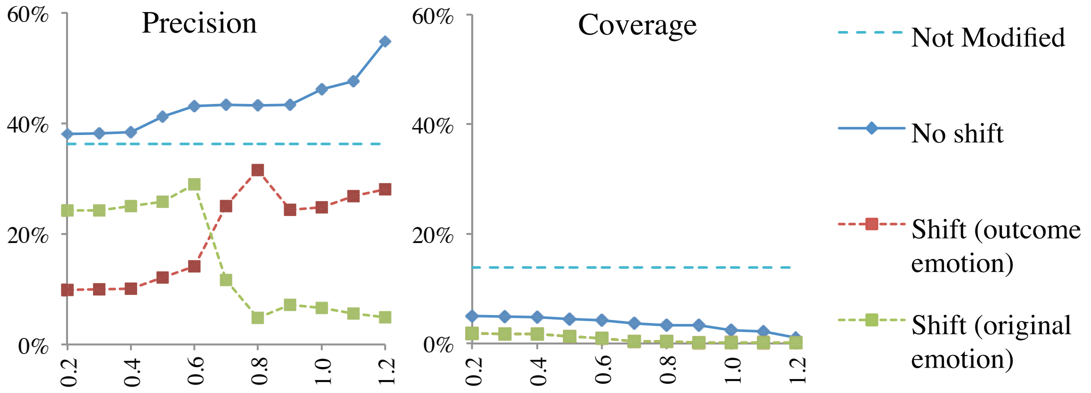


Figure 7.5: The precision and coverage of different modified cases of emotional terms depending on the confidence threshold τ. Only emotional terms with $CC \geq \tau$ are included in each case.

We exclude the mixed cases, where several non-overlapping modifiers are detected, and the cases where the modifier's effect is not modeled because not enough of such modified statements appeared in the analysis dataset D_A.

To compute the classification quality of each case of modified emotional occurrences, we use a test dataset of hashtagged tweets D_T, containing $229,980$ tweets with one of the emotional hashtags for 20 GEW emotion categories. This test set is separated from the analysis dataset D_A that was used to compute the model of modifiers' effects. We again consider the emotion category of the hashtag to be a ground-truth label, and remove the hashtags themselves from the tweet text. To evaluate the performance of each subset of modified emotional statements, we compute their precision and coverage on the test data. Precision is the ratio of correctly found hashtag labels among all labels returned based on the considered statements. Coverage is the percentage of test tweets in which the considered statements are found.[2]

We investigate the quality of classification depending on the filtering of modified cases: we ignore the modified statements with the confidence coefficient CC lower than a confidence threshold τ. Figure 7.5 shows the dependency of precision and coverage on this confidence threshold τ for the three considered modified cases. When $\tau = 0.2$ all corresponding modified statements are used without filtering. Notice that the non-modified case is independent of τ values.

The results show that no-shift modified cases have a higher precision than non-modified emotional entries for any values of τ. This means that when an emotional term appears in the scope of a no-shift modifier the precision of its association with the corresponding emotion is higher. Also, we can observe that the higher the confidence threshold τ is, the higher the precision of the no-shift modified cases is and the lower their coverage is. We can then use

[2] We report classification coverage instead of recall, because we are interested in analyzing the change of coverage separately from the effects of precision, whereas these effects are intertwined in the recall metric.

this mechanism for detecting more precise emotional statements by setting the appropriate τ value.

Considering shifted modified cases, we can observe that their precision is lower than of the non-modified case, regardless of what emotion (original or outcome) is returned. This means that we can exclude such shifted cases altogether in order to obtain more precise classification results. However, the plot also shows a clear change in the shifted modifiers' effects with an increase of τ value: when $\tau < 0.7$ the precision of returning the original emotion is higher than the precision of returning the shifted outcome emotion, whereas when $\tau \geq 0.7$ the reverse is true. This suggests how to potentially increase the overall precision without excluding shifted modified cases: we can return the original emotion for lower CC values, and the shifted output emotion for higher CC values.

In essence, knowledge about the shifting and non-shifting behavior of modified cases helps us to find more confident emotional statements. That is we can increase the precision of extracted emotional statements either by returning only those statements that appear in the scope of the non-shifting linguistic modifiers or by excluding the statements appearing in the shifting modifiers' scope. Therefore, we can construct higher-precision classifiers, which can be then used as starting points for distant learning algorithms. The opportunity to identify more confident emotional statements can also allow users of an application to ignore less confident classification examples.

7.7 Discussion and Future Work

We develop a data analysis method to model the effects of different linguistic modifiers on fine-grained emotional statements based on their usage in social media. Our analysis reveals multiple interesting patterns of modifiers' influence. First of all, the effects of modifiers are non-uniform across emotions, suggesting that for better modifiers' treatment we need to model the per-emotion effects. For example, we show that some under-studied modifiers can even shift emotion categories: conditionality and modality shift *Involvement* to negative emotions of *Sadness* and *Worry* correspondingly. Second, our data confirms that negations are the most notorious modifiers, shifting 75% of emotion categories. More interestingly, our model shows how the original emotions shift in the presence of negation and other modifiers. Third, we show the potential of incorporating the extracted modifiers' model along with its confidence coefficients to identify more precise emotional statements. All these discoveries present important opportunities for developing a modifier-aware emotion classification system.

Due to the specifics of our analysis, these findings are confined to the domain of Twitter environment (and potentially to other platforms of short-text social media). To make them more applicable to another domain, one could analyze tweets following the style of the target domain (e.g. having similar readability score) or apply directly the method to within-domain data, assuming that scalable bootstrapping techniques are attainable (e.g. using hashtags or emoticons). Also, our model of modifiers' effects depends on the available modifier detection modules and studied modifiers themselves, as well as on the considered emotional statements. Therefore, the resultant model aims at describing how to treat modifiers for classification purposes, not at universal modeling of emotion-modifiers relations.

This work takes a novel approach of using self-labeled Twitter data to model the effects of modifiers, without requiring any manual labeling. Thus, this method allows researchers to easily update the modifier model for new modifier types or emotion categories. This property opens a large window for future work. One application could be to investigate the effects of other linguistic or contextual modifiers, which would help testing their hypothesized impact. Furthermore, as there might be more than one modifier affecting an emotion term and their effects could amplify or cancel each other, a deeper analysis of such cumulative effects can help building a more precise modifiers' model. The method can be also adapted to automatically discover new modifying expressions. For example, finding more confident context of emotional expressions can guide more reliable and scalable initial data pseudo-annotation for distant learning process. Finally, more classification experiments are to be performed to investigate how modifiers' treatment affects the quality of emotion recognition.

7.8 Chapter Summary

This chapter studies the effects of six different linguistic modifiers, including negation, intensification, interrogation, past tense, conditionality, and modality, on the explicit emotional terms that are within their scope. We propose a novel data analysis method to computationally model their effects on emotions. It analyses how the modifiers change the associated emotion distributions, how they shift the original emotions of terms, and how they affect the confidence in the outcome emotions. As labeled data, we use a large number of tweets with author's self-revealed emotions, which are identified via emotional hashtags. This work, to the best of our knowledge, is the first systematic study of the effects of different modifiers at a fine-grained level of emotion categories. The obtained model is promising for treating modifiers more accurately in fine-grained emotion recognition tasks.

Understanding how to treat different modifiers is important for better quality of emotion recognition. When the goal is to use a novel less-studied emotion category set, understanding and coding the effects of modifiers in expert fashion can be challenging and time-consuming. The described method is a ready-to-apply solution in this case. Our data analysis approach allows modeling the different modifiers' effects within any data where bootstrapping enough emotional data of high quality is feasible, e.g. using hashtags or emoticons. This research also opens a window for multiple prospective research directions aiming at modeling and investigating the context of appeared emotional statements. One promising example is to reverse the proposed method and target it to find or validate the context indicating higher confidence in emotional expressions.

8 Conclusion

Our overarching goal is to advance fine-grained emotion recognition in short text. While many researchers investigate more separable, strong basic emotions, such as Happiness, Sadness, Anger, and Fear, we aim to enable computers to additionally distinguish more subtle, but still pronounced emotions, such as Pride, Pleasure, Amusement, Pity, Regret, and Shame. Our motivation scenario prescribes building emotion recognition systems using a novel fine-grained set of emotion categories to be recognized within a specific domain of interest. We build such emotion recognition systems for short text, particularly tweets, while focusing on the psychologically reduced model of 20 emotion categories from the Geneva Emotion Wheel [Sch05]. Our annotation procedure shows that human annotators can successfully separate such nuanced emotions. Yet, annotations frequently contain a mixture of emotions. This suggests casting emotion recognition as a multi-label classification problem. We show that fine-grained emotion recognition in this formulation is feasible to perform and allows detecting specific subtleties within a considered domain, such as detecting domain-specific dominant emotions.

This thesis identifies and investigates two main approaches to building such systems from scratch. The first is to manually collect emotional linguistic expressions via well-designed crowdsourcing processes. The second is to automatically construct emotion classifiers via a distant supervision process starting from pseudo-annotation of available data with lexicons of limited quality. Both are shown to outperform the available baseline models for the formulated fine-grained emotion classification when designed and evaluated for the data from a specific domain — Twitter-based reactions to Olympic sporting events.

Regarding the crowdsourcing process, we show that directly asking workers to provide emotion indicators is a viable way to generate emotion lexicons. Such aggregated crowdsourced answers are a valuable source of affective commonsense knowledge, which is crucial for accurate emotion recognition. The designed human computation task can be used both for annotating text documents with emotions and for collecting novel emotion indicators. This work reveals the potential of using crowdsourcing to build new emotion recognition systems and describes a method which can be adapted to any specific domain.

Managing annotations through crowdsourcing generally raises concerns about the quality of collected answers from online non-expert users. We investigate two preemptive quality control mechanisms aiming to teach and motivate workers to provide better-quality answers: tutorials and framed financial incentives. We show that both can affect the quality of the crowdsourced annotations and thus are important mechanisms to ensure the annotations' quality.

Regarding the distant supervision process, we develop a framework and a specific learning method that allow us to build new emotion classifiers out of the data pseudo-labeled with initial emotion lexicons of limited quality. Compared to more restrictive heuristics of using emotional hashtags, commonly applied for pseudo-labeling in distant supervision [Moh12a, DCGC12], our approach is likely to be more applicable to building emotion classifiers across different domains, especially when only limited within-domain datasets are available. We discover the most advantageous settings leading to building better-quality classifiers. These include adjusting the learning parameters on the pseudo-annotated validation set, introducing rebalancing coefficients for emotion distributions, and using additional heuristics to find pseudo-neutral tweets for the learning process.

Furthermore, we investigate how different modifiers affect the emotional meaning of emotional statements within their scope. This work compares six linguistic modifiers within the unified analysis framework and reveals the impact of less studied modifiers, such as modality and conditionality. More importantly, we provide an automatic way to build a computational model of the modifiers' effects for any detectable modifier type. This presents the potential for better treatment of modifiers in emotion classification. Computational modeling of their effects is the first step towards building more accurate modifier-aware emotion recognition systems.

Overall, our research addresses several main challenges in building novel emotion recognition systems working with fine-grained emotion category sets. We investigate how to crowdsource informative annotations for constructing an emotion lexicon, how to ensure the quality of such crowdsourced annotations, how to build emotion classifiers using limited resources by applying distant supervision, and how to model the effects of modifiers on emotional statements. Addressing these challenges and studying their multiple aspects is our contribution to the emotion recognition field. More accurate emotion classifiers are needed to foster the applications of this technology. While until now fine-grained emotion recognition remains a challenging task, we are hopeful that, with the current pace of technology development, fine-grained emotion recognition from different textual media and at different levels of emotion granularity will be available to scientists soon. This will help discover fascinating insights in human nature and build innovative affective applications, as well as help improve global emotional well-being.

8.1 Recommendations for Building Emotion Recognition Systems in Text

In the course of this work, we dealt with multiple issues appearing in the process of building textual emotion recognition systems. We summarize knowledge we obtained in a short list of direct recommendations that researchers and practitioners may find valuable.

For crowdsourcing informative emotion annotations:

- *Consider asking annotators to additionally provide emotion indicators besides a categorical label to text.* The elicited indicators have direct emotion annotations, and thus, with enough level of redundancy, can be used directly as features in emotion recognition. This thesis shows that an emotion lexicon generated in this way achieves reasonable recognition quality.

- *Consider ensuring that workers understand the task instructions and provide incentives to motivate them to perform better work.* Online crowdsourcing requires additional measures to prevent poor-quality answers. This thesis validates the importance of including tutorials for proper presentation and validation of task instructions. We also show the potential of using well-formulated bonus incentives to achieve better-quality annotations.

- *Consider adapting algorithmic selection of data for annotation.* Ensuring that the majority of annotated documents are emotional helps extract more of practical information about emotional linguistic expressions. Ensuring that all emotion categories of interest are present in the annotation in adequate amounts can help build emotion classifiers that are able to discriminate between all the studied emotions. Ensuring that the emotional expressions from the annotated text are repeatable within the investigated data can help build more applicable emotion recognition systems.

For building emotion recognition systems using distant supervision:

- *Consider choosing applicable heuristics for initial pseudo-labeling of the data.* Our work on distant supervision shows the viability of building emotion classifiers by pseudo-labeling the available data using basic emotion lexicons of limited quality. Such lexicons, in contrast with more restrictive heuristics, are more generally applicable, which results in more data available for training the classifiers.

- *Consider treating the emotion imbalance in the training data.* This thesis shows that the presence of skewed emotion distribution in the training data, if left ignored, can bias the classifier to return more dominant emotion categories. However, any distribution of emotion categories can theoretically be present in the data to which a classifier is applied. Incorporating per-emotion rebalancing coefficients can help avoid building biased classifiers.

- *Consider including heuristics for detecting neutral (or non-emotional) documents.* We show that adding pseudo-neutral tweets into the training data allows avoiding over-classification of the emotional content during the application of built classifiers. Having neutral tweets in the evaluation data allows validating the absence of such classifiers' behaviour.

For treating the effects of modifiers:

- *Consider modeling the effects of different modifiers on a per-emotion basis in the recognition model.* We show the possibility of automatically extracting a computational model of the effects of detectable modifiers. This describes how each emotion changes under each modifier in terms of shifting behaviour and confidence change. It can be beneficial to treat the influential modifiers on the basis of such model, instead of ignoring or blocking their effects.
- *Consider separating the level of confidence in emotional statements.* We show that the effects of some modifiers, such as modality, conditionality, and intensification, can be modeled by the resultant change of confidence in emotional statements. Ignoring less confident statements and up-weighting more confident ones can positively affect the quality of emotion recognition.

General recommendations:

- *Consider using a multi-label classification framework instead of multi-class.* Using a multi-class classification formulation implies one emotion category label per document. In reality, our annotation procedures show that a mixed emotional state is frequent, at least when more fine-grained emotion categorization is being employed.
- *Consider building a domain-specific emotion recognition system tailored to applications within a studied domain.* We study two operational approaches to build domain-specific emotion classifiers: via crowdsourcing and distant supervision. Both allow obtaining classifiers that perform better on within-domain data than general-purpose classifiers.

8.2 Perspectives for Future Research

Using our experience gained while developing fine-grained emotion recognition systems, we outline the following prominent directions of the future work in the area of text-based emotion recognition.[1]

Generalization of an Emotion Recognition System We motivated our research on building a new emotion recognition system by the need to apply it to a specific domain of collected data. Ideally, we would like to have a universal system that could handle different emotions and different domain-specific and general expressions. Such a system would automatically detect from which domain the text originates and apply an appropriate tailored recognition model. Building such a system would require developing generalization methods in order to ensure that the learned knowledge is transferable across the domains. We believe that future work in this direction could result in more universally applicable and accurate emotion recognition systems. Potentially, using the distant supervision framework across several representative domains could allow us to automatically discover more general emotional expressions, as well as to extend the emotion recognition model with domain-specific expressions suitable for reapplication within the same domain.

Detection of Causes of Emotional Experiences The focus of our work was on developing a system able to categorize emotions in tweets. However, the benefit of applying emotion recognition systems to social media data lies not only in quantifying the emotional experience, but also in understanding and summarizing the causes of the experience. Researchers are starting to address a problem of emotion cause detection [CLLH10, RCR$^+$11, NA13]. Its potential solution could be adapted from aspect-sentiment models, which aim to identify which sentiment is expressed towards which specific feature or aspect of a product [TM08, JO11, BE10]. Learning what caused a specific emotion is advantageous for two reasons. First, it can help discover deeper insights from the data. For example, we could not only compare the differences of emotional experiences across different populations, but also understand what causes such differences. Second, it would enable presenting the output of emotion recognition in a more interpretable way, where the detected emotions would be aligned with their causes.

Advanced Feature Representation of Affective Commonsense Knowledge Our research focused on building an emotion recognition system that works on lexical-level features, such as words and n-grams. However, emotions can be expressed and captured by the overall meaning (or semantics) of the text. How to represent that affective meaning by semantic-level features is an open research question. Researchers of language semantics develop different models to represent the text meaning, for example, by using the linguistic frames [BFL98]. In the areas of sentiment analysis and emotion recognition, researchers are starting to represent the concepts referred to in the text [GCHP11, PGH$^+$13] and model the descriptions of shared events and

[1] We present here only general perspectives for the future work. More specific suggestions for continuation of our research are enumerated at the end of each chapter.

situations using knowledge-based representations [BHM12]. The features derived using syntactic and semantic text parsing are viewed as providing more condensed, meaning-oriented representations of the text, which should result in more accurate emotion recognition.

Context Modeling of Emotional Expressions Our work analyzed the effects of main detectable classes of modifiers on emotional statements and showed how such effects can be extracted and modeled on a per-emotion basis. However, the modifying context of emotional words and expressions often remains undiscovered and thus untreated. Furthermore, the effect of a modifier can depend not only on the emotion category of an emotional expression, but also on the modifier's expression or emotional expression themselves. Automatic discovery of modifying expressions and modeling their effects on emotional statements can help to advance emotion recognition quality further. Future experiments could suggest how to better incorporate such detailed modifiers' models in the classification framework. Analogously, explicit modeling of the different contexts of emotional expressions, whether as specific syntactic relations [WWH05] or different word senses [MT13], can help distinguish the nuances of emotional statements.

Gamified Extraction of Affective Commonsense Knowledge This work showed the potential of using crowdsourcing to build emotion recognition systems. It aims at extracting affective commonsense knowledge from humans in the form of text emotion annotations. The more annotations the system would have access to, the more knowledgeable and accurate the system would become. An army of users now spend their time online and, with proper intrinsic motivation, they could be attracted to share their commonsense knowledge for free. Applying gamification principles, such as designing for entertainment, can make serious human computation tasks more attractive to a general audience [VAD08]. With proper game mechanics and incentives, even such tedious and attention-demanding task as emotion annotation and elicitation can be formulated as a game with a purpose [PS10]. This could allow obtaining annotations at a much larger scale than with paid crowdsourcing.

A Appendices

A.1 Refinement of the GALC Lexicon

We matched the GALC stems to instances appeared in the tweets and manually removed the incorrect or ambiguous matches (section 3.3). Table A.1 enumerates the top occurred removed instances. Tables A.2 and A.3 list the positive and negative terms from the resultant GALC-R.

Emotion	Top removed instances
Involvement	attention, alert, zealand, animation, animations
Amusement	play, playing, played, player, players, playlist, plays, smiles, playin, playoffs, function, playoff, funeral, fund, funk, amust, funky, playa, funding, funds, playstation
Happiness	happen, happened, happens, happening, happyhalloween, happybirthdaymichaelclifford, happybirthdaysrk, happybirthday, cheerleaders, happybirthdaymiley, joyce, cheerleader, happend, happyveteransday, happenin, happybirthdayjacob, happybirthdayartpop, happybirthdaysandarapark, cheerleading, happybirthdayjaxonbieber
Pleasure	please, pleasee, content, glow, thriller, pleas, please<num>, glowing, contents, pleasex<num>, pleasefollowme, pleass, contention, pleaser, glows
Love	friends, friend, friendship, friendships, lovemarriottrewards, friendzone, lovelies, lovemehardermusicvideo, tenders, friendsgiving, lovetheatre, friendzoned, lovehate, lovetanya, friendlies, loveless, raptors
Awe	raptors, adorbs, raptor, reverend, admiral, wonderland, wonderwall
Relief	relies
Surprise	wonderland, wonderwall, stunna
Nostalgia	long, longer, pink, pin, fantasy, longest, pineapple, wishy, pinterest, pint, pinky, ping, fantasyfootball, pine, longg, pins, pineapples, pinch, fantasysports, pints
Pity	pitch, pitbull, pit, pittsburgh, pitt, pitchers, pits, pitching, compass, pitcher, pitched, pitches, pita, pitbulls, pitchperfect<num>, pitvsten
Sadness	grier, sadie, saddle, sade, teared, sadistic, dole, sadies
Worry	alarm, terrorist, alarms, terrorists, terrorism, fearless, scarecrow, dreadlocks
Shame	crush, crushing, crushed, crushes, shameless, shamelessly, crusher
Guilt	blameitonthemistletoe
Regret	source, sour, sources, sourcing
Contempt	despite
Anger	made, cross, hat, haters, ireland, madison, crossed, madrid, hats, hater, madisonfollowme, crossing, raghavsinghania, hatin, maddie, madonna, crossoverweek, hath, crossover, irene

Table A.1: The top of removed GALC instances in the refinement process. We report up to 20 instances per emotion category and only the instances that appeared at least 50 times. There were no such instances for Pride, Envy, and Disgust.

Involvement: absorbed, absorb, absorbing, animated, animatedly, ardor, attentive, attentively, attentions, attentionseeker, curious, curiosity, curiousity, curiously, curios, eager, eagerly, eagerness, engrossed, engrossing, enthusiasm, enthusiastic, enthusiast, enthusiasts, enthusiastically, fervent, fervently, fervor, interesting, interested, interest, interests, interestingly, interestin, interesante, interestingfact, zeal, zealous, zealousness, fervid, involvement

Amusement: amusing, amused, amuse, amusement, amuses, amusingly, fun, funny, funniest, funnier, funn, funfact, funnest, funy, funnyy, funnies, funtimes, funnysms, funner, funnily, funfacts, funhouse, funfunfun, funfactfriday, funsies, funnyshit, funtime, funnight, funtastic, funfun, funstuff, funyons, funnybone, humor, humorous, humors, humored, laugh, laughing, laughed, laughs, laughter, laughin, laughable, laughtrip, laught, laughign, laugher, laughh, laughably, laughn, playfully, playingg, playgirl, playinh, smile, smiling, smiled, smilee, smilling, smily, playful

Pride: pride, prideful, proud, proudly, prouder, proudest, prouddad, proudmom, proudd, proudalumni, proudof<int_num>sos, proudmoment, proudsister, proudalum, proudmama, proudfan, proudofyou, proudcoach, prouddirectioner, proudofmyself, proudtobeanamerican, proudness, proudlysa, proudtobe, proudmomma, proudindian, proudlysouthafrican, proudtobebritish, proudcousin, prouddaughter, proudamerican, prouda, proudpapa, proudmum, proudtobeadirectioner, proudofit, proudtobepinoy, pridefulness

Happiness: bliss, blissful, blissfully, blissfulness, cheer, cheers, cheering, cheered, cheerful, cheery, cheerfully, cheerfulness, cheerin, cheerdance, cheerss, cheerr, delectable, delectation, elated, elation, enchanted, enchanting, enchant, enchantment, euphoria, euphoric, exalted, exalt, exaltation, exhilarating, exhilaration, exhilarated, felicity, felicitats, felicitaciones, flush, flushed, flushes, glee, gleefully, gleeful, happy, happiness, happiest, happier, happily, happyy, happyfriday, happymonday, happy<int_num>thbirthdaymichael, happysunday, happines, happytopday, happyhour, happydays, happy<int_num>, happygirl, happyness, happyending, happyland, happyday, happysaturday, happythursday, happie, happyhappyhappy, happytweet, happydance, happytuesday, happyholidays, happyplace, happyanniversary, happyme, happycamper, happykid, happyhappy, happnd, happytimes, happeing, happnin, happytears, happin, happinessoverload, happythoughts, happykids, happieness, happymeal, joy, joys, joyful, joyous, joyfully, joyless, joyed, jubilant, jubilation, merry, merrier, merrily, merriment, merrit, merriweather, overjoyed, ravishing, rejoice, rejoicing, rejoices, rejoiced, exultant, cheerless, ecstatic, delighted, delight, delightful, delights, delightfully, enjoy, enjoying, enjoyed, enjoyable, enjoys, enjoyment, enjoyin, enjoyy, enjoylife, enjoyit, enjoyably, enjoyng, enjoyful

Pleasure: comfortable, comfortably, contented, contentment, contentious, glowed, glowy, pleasure, pleased, pleasant, pleasing, pleassee, pleases, pleasse, pleasantly, pleasurable, pleasingly, satisfied, satisfaction, satisfying, satisfy, satisfactory, satisfies, thrill, thrilled, thrilling, thrills, zest, zesty, delighted, delight, delightful, delights, delightfully, enjoy, enjoying, enjoyed, enjoyable, enjoys, enjoyment, enjoyin, enjoyy, enjoylife, enjoyit, enjoyably, enjoyng, enjoyful

Love: affection, affectionate, affections, affectionately, fond, fonder, fondly, fondness, fondest, fondling, friendly, friendss, friendliest, friendliness, friendlier, friending, love, loves, loved, lovely, lovetv<int_num>, lovee, lover, lovers, loveyou, lovelyz, lovesanny<int_num>, loveit, loveshoe<year>, love<int_num>, loveteam, lovelife, loveher, lovehim, loveu, loveliest, lovethem, lovemyjob, loveyouu, lovess, loveya, loveable, lovemeharder, lovelovelove, loveforlife, lovestory, lovebug, lovelyy, lovelove, lovejoy, lovelys, loverboy, lovecraft, lovelive, lovees, loveme, lovedd, lovebirds, lovedit, love<username>, loveyousomuch, loveofmylife, lovemylife, loveliness, lovethis, lovestruck, lovelondon, loveyouguys, loveyall, loven, lovegood, loveyoutoo, loveyoumore, lovemyteam, lovethissong, loveyouall, loveing, lovemusic, lovemyfamily, lovethatsong, lovemaking, lovethegame, lovexx, lovethat, loveem, lovesit, lovemyfriends, loveeyouu, lovesundays, lovelace, loveandhiphop, lovehersomuch, loveyourwork, lovethiskid, lovethemsomuch, lovemymomma, lovebeinghome, lovemydad, lovetwitter, lovesport, lovemyparents, tender, tenderness, tenderly, loveatfirstbite

Awe: admire, admired, admiration, admiring, admirable, admirer, admires, admirably, adores, adored, adoration, adoring, adorablee, adorably, adorableness, adorbz, adorkable, adorables, adoreyou, awesome, awe, awesomeness, awee, awesomee, awesomely, awesomest, awesomesauce, awestruck, awesomer, awesoome, awesone, awesme, awes, awesomenesstv, awesomeday, awesom, awesomeexposure, awesomenight, awesomes, awesomness, awesomme, dazed, dazzling, dazzle, dazzled, enthralling, enthralled, enthrall, fascinating, fascinated, fascination, fascinates, fascinate, fascinator, fascinatingly, marveling, rapt, rapturous, raptorss, spellbound, worship, worshipping, worshiping, worshipped, worshippers, worships, worshiped, adorable, awed, wonder, wonderful, wondering, wonders, wondered, wonderfully, wonderfull, wonderin, wonderous, wonderment, wonderfulness, wonderkid, wonderboy

Relief: relief, relieved, relieve, reliever, relieves, relieving, relief<int_num>liberia, disburned

Surprise: amazed, amaze, amazes, amazement, amazeballz, astonishing, astonished, astonish, astonishment, astonishingly, astonishes, dumbfounded, startled, startling, startle, startlingly, stunning, stunned, stunningly, stunnin, stunners, stunn, stunnas, surprise, surprised, surprises, surprising, surprisingly, surprisesurprise, thunderstruck, wonder, wonderful, wondering, wonders, wondered, wonderfully, wonderfull, wonderin, wonderous, wonderment, wonderfulness, wonderkid, wonderboy

Nostalgia: craving, crave, cravings, craves, craved, cravin, cravingg, craviing, cravingss, cravee, daydream, daydreaming, daydreams, daydreamin, daydreamer, daydreamed, daydreamers, desire, desires, desirable, desiree, desirous, fantastic, fantasies, fantasia, fantasize, fantasizing, fantastically, fantasylife, fantastical, fantastico, fantasmic, fantasic, fantasising, hankering, hark, harkness, harkless, homesick, homesickness, longing, nostalgia, nostalgic, nostalgiachat, wish, wishing, wished, wishin, wishful, wishfulthinking, wishh, wishyouwerehere, wishiwasthere, wishs, wishfullthinking, wistful, wistfully, yearn, yearning, yearns, yearned

Table A.2: The list of revised GALC instances for positive emotions.

Pity: commiserations, commiserating, commiserate, compassion, compassionate, empathy, empathetic, empathic, pity, pitiful, pitty, pitied, pitying, pitifully, piteous

Sadness: chagrin, dejected, desolation, desolate, despair, despairing, desperate, desperately, desperation, desperatetimes, despondent, gloomy, gloom, glum, grief, grieve, grieving, grievances, grievance, grievous, grieved, hopeless, hopelessly, inconsolable, melancholy, melancholic, melancholia, mourn, mourning, mourns, mournful, sad, sadness, sadly, saddest, sadder, saddens, saddened, sadd, saddening, sadface, sadtweet, sadderday, sadbuttrue, sadday, sadlife, sadtimes, sadden, sadtruth, sadc, sadstory, sadcase, saddays, sadpanda, sorrow, sorrows, sorrowful, tears, tear, tearing, teary, teardrops, teardrop, tearful, tearss, tearjerker, tearfully, weep, weeping, weeps, weepy, disconsolate, grief-stricken

Worry: afraid, afraidd, aghast, alarming, alarmed, alarmingly, anguish, anxiety, anxious, anxiously, anxieties, anxietyproblems, apprehensive, apprehension, apprehensions, dreads, dreading, dread, dreadful, dreaded, dreadfully, dreadlock, fear, fears, feared, fearing, fearful, fearfully, fearthebeard, fearthefin, fright, frightening, frightened, frightful, frighten, frightens, frightfully, frighteningly, frightfest, horrible, horror, horrific, horrendous, horribly, horrid, horrifying, horrors, horrified, horrifically, horrendously, jittery, jitters, jitterbug, nervous, nervously, nervousness, panic, panicking, panicked, panics, panicky, panick, panicatthedisco, scared, scare, scares, scaredd, scaredy, scarey, scarecrows, scaremongering, scaredshitless, terror, terrors, terrorize, terrorized, terrorizing, terrorising, wary, worried, worry, worrying, worryin, worryingly, anxiousness, diffident

Shame: abashed, ashamed, ashame, crushin, crushedit, crushers, crushingly, disgrace, disgraceful, disgraced, disgracefully, embarrassing, embarrassed, embarrassment, embarrass, embarrassingly, embarrasing, embarrasses, embarrased, embarrassments, embarrassin, embarrasment, humility, humiliation, humiliated, humiliate, humiliating, shame, shameful, shamed, shames, shameonlumsvc, shameonpatwarkhana, shamefully, shamee, shameonyou, shameonme, shamefull, shamefaced, abash

Guilt: blame, blamed, blames, blamee, blamegame, guilty, guilt, guiltypleasure, guiltyascharged, guiltypleasures, guilted, remorse, remorseful, repent, repentance, repented, blameworthy

Regret: regret, regrets, regretting, regretted, regretful, regrettable, regretfully, regrettably, regreted, comedown, disappointed, disappointment, disappointing, disappoint, disappointments, disappoints, disappointingly, discontent, discontented, disgruntled, disillusioned, frustrated, frustrating, frustration, frustrates, frustrations, frustrate, frustratingly, jilted, letdown, resign, resigned, resignation, resigns, resigning, sours, sourz, sourness, thwarted, thwart, thwarting

Envy: envious, envy, envying, jealous, jealousy, jealously, jealouss, jealouslines, jealousmuch

Disgust: abhor, abhorrent, abhors, abhorrence, averse, aversion, detest, detests, disgusting, disgusted, disgust, disgusts, disgustingly, dislike, dislikes, disliked, disliking, distasteful, distaste, loathing, loathe, loathsome, loath, loathed, nauseous, nausea, nauseating, nauseas, nauseated, nauseum, nause, queasy, repugnant, repulsive, repulsed, repulse, repulsion, revolt, revolting, revolts, revolted, sickening, sickens, sicken, sickened

Contempt: contempt, denigrate, deprecation, depreciation, deprecating, derision, despise, despicable, despised, despises, despising, disdain, disdainful, scorn, scorned, scornful, despiteful

Anger: acrimonious, acrimony, anger, angry, angrily, angrier, angriest, angryy, angrytweet, angryface, annoying, annoyed, annoy, annoys, annoyance, annoyin, annoyingly, annoyingg, annoyingteachers, annoyedd, annoyances, annoyingaf, annoyinh, annoyn, annoyig, annoyinng, annoyingness, annoyings, enraged, enraging, enrages, exasperated, furious, fury, grumpy, grumps, grump, grumpiest, grumpiness, grumpier, hate, hates, hating, hated, hatred, hateful, hatee, hateit, haterade, hateyou, hatemondays, hateing, hatethis, hatemylife, haterr, hatersgonhate, hateonit, hatered, hatinglife, incensed, indignation, indignant, infuriating, infuriates, infuriate, infuriated, irate, irritated, irritating, irritate, irritates, irritable, irritation, irritatin, irritability, mad, madness, madden, madly, maddest, madz, madding, rage, raging, raged, ragequit, ragedy, resentment, resent, resentful, resenting, sullen, temper, vex, vexed, vexing, wrath, wrought, wrathful

Table A.3: The list of revised GALC instances for negative emotions.

A.2 A List of Emotional Hashtags

Table A.4 enumerates 167 hashtags compiled from the GALC lexicon to identify the 20 GEW emotions. The process of its compilation was described in section 3.3. These hashtags are used for collecting the general emotional tweets (EMHASH_DATASET, section 3.4) and for identifying the pseudo-labeled emotional tweets among the Olympic-related tweets (section 6.5.1).

Emotion Category	Corresponding Emotional Hashtags
Involvement	#involvement, #interest, #curious, #eager, #interested, #enthusiastic, #enthusiasm
Amusement	#amusement, #laughter, #amused, #laughing, #smiling, #humorous
Pride/Elation	#pride, #elation, #pridefulness, #proud, #prideful, #elated, #soproud
Happiness	#happiness, #joy, #cheerful, #enjoyful, #happy, #joyful, #gleeful, #joyous, #happytweet, #sohappy, #happydays
Pleasure	#pleasure, #enjoyment, #contented, #delightful, #delighted, #satisfied, #delight, #pleased
Love/Tenderness	#love, #tenderness, #tender, #fondness, #loveyou, #lovethem, #lovehim, #loveher
Awe/Wonderment	#awe, #wonderment, #admired, #admiration, #adored, #worship, #fascinated, #awed, #dazed
Relief/Disburned	#relief, #disburned, #relieved
Surprise	#surprise, #astonishment, #amazed, #astonished, #surprised, #startled, #amazement
Nostalgia/Longing	#nostalgia, #longing, #wishful, #daydreaming, #nostalgic, #desirous
Pity/Compassion	#pity, #compassion, #empathy, #compassionate, #empathic, #pitiful
Sadness/Despair	#sadness, #despair, #grievous, #grief, #sad, #gloomy, #grieving, #desperate, #sorrow, #sorrowful, #sadden, #weepy, #sadtweet, #sosad
Worry/Fear	#worry, #fear, #scared, #worried, #panic, #anxious, #dreadful, #alarmed, #wary, #nervous, #afraid, #horrified, #worrying
Shame/ Embarrassment	#shame, #embarrassment, #embarrased, #disgraceful, #shameful, #humiliated, #shamefaced, #shamed, #ashamed, #abash, #abashed
Guilt/Remorse	#guilt, #remorse, #blameworthy, #guilty, #remorseful, #blamed
Regret	#regret, #disappointment, #comedown, #sour, #souring, #discontent, #disappointed, #regretful, #regrets
Envy/Jealousy	#envy, #jealousy, #jealous, #envious
Disgust/Repulsion	#disgust, #repulsion, #loathsome, #dislike, #loathe, #disgusting, #repulse
Contempt/Scorn	#contempt, #scorn, #disdainful, #despising, #despise, #disdain, #scornful, #despiteful, #despicable
Anger/Irritation	#anger, #irritation, #mad, #hatred, #hateful, #madden, #fury, #annoyed, #irritated, #angry, #rage, #wrathful, #hate, #angrytweet, #hateit, #irritating

Table A.4: The list of emotional hashtags associated with 20 GEW emotion categories.

A.3 Descriptions of Launched Crowdsourcing Tasks

Generating the SREC Dataset and OlympLex Lexicon Table A.5 describes the parameters of the annotation task that we conducted for generating the SREC dataset and OlympLex lexicon (chapter 4).

Title	Annotate emotions within tweets about sport (takes about 1 min)
Description	You will read the tweets related to the Olympic Games 2012; For each tweet you will choose category of the emotion which the tweet author felt; You will also show us based on which text expressions you made your decision
Keywords	sentiment, tweet, tweets, twitter, emotion, categorization, social, English, fast, language, research, opinion, classify, creative
Reward	0.04 $
Time Allotted	15 minutes
Qualifications	U.S. location
Task Size	1 tweet

Table A.5: The details of the HIT presentation on Amazon Mechanical Turk for the collection of the SREC dataset.

Experiment on Bonus Incentives Table A.6 describes the parameters of the annotation task that we used in the experiment on bonus incentives' framing (chapter 5).

Title	Annotate tweets with emotions
Description	You will help us find emotional expressions in 10 tweets.
Keywords	annotate, tweets, emotions, research, experiment
Reward	0.5 $
Time Allotted	2 hours
Task Size	10 tweets

Table A.6: The details of the HIT presentation on Amazon Mechanical Turk for the experiment on bonus incentives.

A.4 Instructions and Tutorials for the Annotation Task

Detailed Instructions

The following instructions were shown when the users hovered a mouse over the question marks in the task interface. The same instructions were also part of the tutorial. No further detailed instructions were given to users.

Instructions for Action 2: Selecting Emotion Category and Strength

1. Choose one emotion category that was your dominant feeling at the moment of writing this tweet

- ! “No emotion” and “Other emotion” in the center are also emotion categories

2. Decide how strong your emotion was and click on one circle of the corresponding size
   - ! You have three options: weak, medium, strong. The bigger the circle is, the stronger the emotion is.
3. If you chose “Other emotion”, input the most appropriate description for it in the textbox

Instructions for Action 3: Selecting Tweet Emotion Indicators

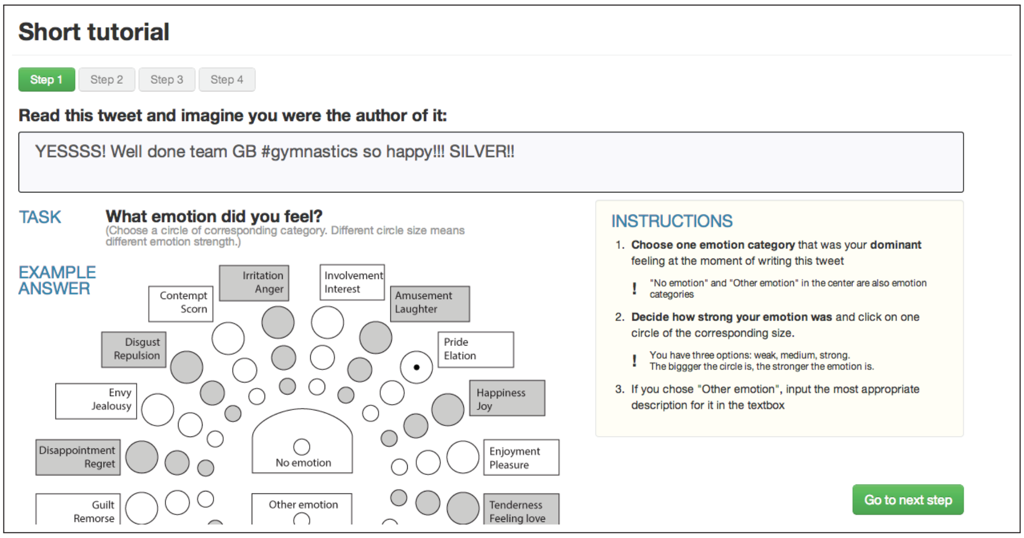
1. Detect all parts of the tweet revealing your emotion if present
   - ! Detect not only the separate emotional words (e.g. "happy"), but also the emotional expressions with several words (e.g. "well done") and emoticons (e.g. ":)")
   - ! Select the shortest expressions, but do not omit words for emotion strength (e.g. write the full expression "so excited")
2. Copy-paste found indicators from the tweet text into a textbox
   - ! You can select the text and drap&drop it into textbox
   - ! Each distinct expression should begin a new line

Instructions for Action 4: Providing Additional Emotion Indicators

1. Input words or phrases which you would use to express the emotion from this tweet
   - ! You should think about them as of emotion indicators possible to appear in a real tweet
   - ! Each distinct expression should begin a new line

Tutorial Steps

Figures A.1, A.2, A.3, A.4 presents the steps of our tutorial for labeling Olympic Tweets in order of their appearance (the design of the tutorial was presented in section 4.4.2).



Short tutorial

Read this tweet and imagine you were the author of it:

YESSSS! Well done team GB #gymnastics so happy!!! SILVER!!

TASK

What emotion did you feel?
(Choose a circle of corresponding category. Different circle size means different emotion strength.)

EXAMPLE ANSWER

INSTRUCTIONS

1. **Choose one emotion category** that was your **dominant** feeling at the moment of writing this tweet

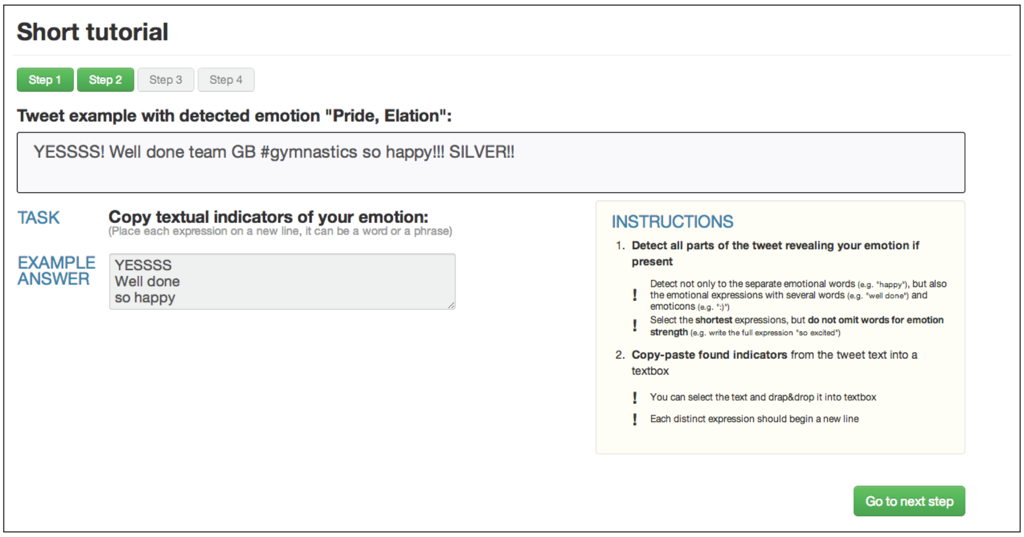
   ! "No emotion" and "Other emotion" in the center are also emotion categories

2. **Decide how strong your emotion was** and click on one circle of the corresponding size.

   ! You have three options: weak, medium, strong.
   The bigger the circle is, the stronger the emotion is.

3. If you chose "Other emotion", input the most appropriate description for it in the textbox

Figure A.1: Tutorial, Step 1.



Short tutorial

Tweet example with detected emotion "Pride, Elation":

YESSSS! Well done team GB #gymnastics so happy!!! SILVER!!

TASK

Copy textual indicators of your emotion:
(Place each expression on a new line, it can be a word or a phrase)

EXAMPLE ANSWER

YESSSS
Well done
so happy

INSTRUCTIONS

1. **Detect all parts of the tweet revealing your emotion if present**

   ! Detect not only to the separate emotional words (e.g. "happy"), but also the emotional expressions with several words (e.g. "well done") and emoticons (e.g. ":)")

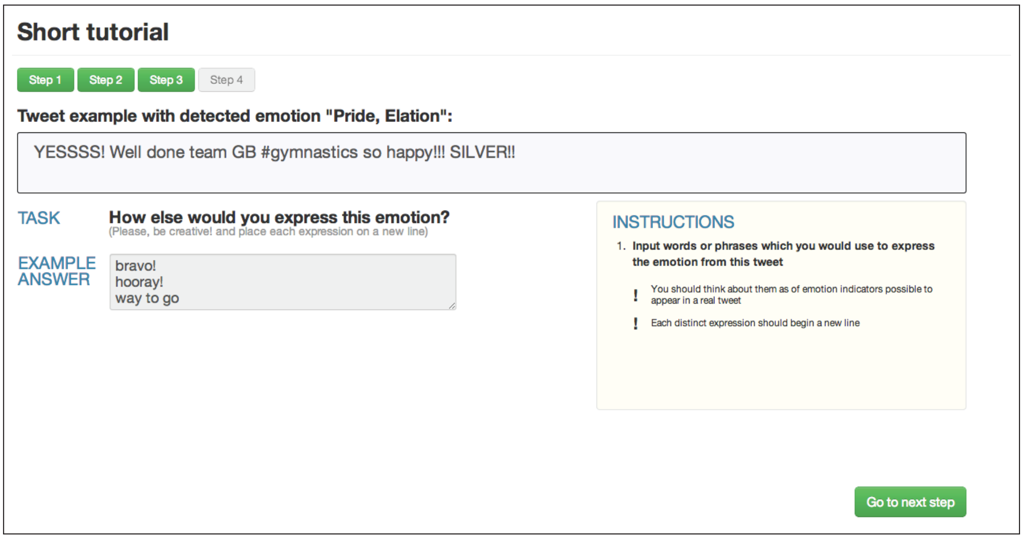
   ! Select the **shortest** expressions, but **do not omit words for emotion strength** (e.g. write the full expression "so excited")

2. **Copy-paste found indicators** from the tweet text into a textbox

   ! You can select the text and drap&drop it into textbox

   ! Each distinct expression should begin a new line

Figure A.2: Tutorial, Step 2.



Short tutorial

Tweet example with detected emotion "Pride, Elation":

YESSSS! Well done team GB #gymnastics so happy!!! SILVER!!

TASK **How else would you express this emotion?**
(Please, be creative! and place each expression on a new line)

EXAMPLE ANSWER

bravo!
hooray!
way to go

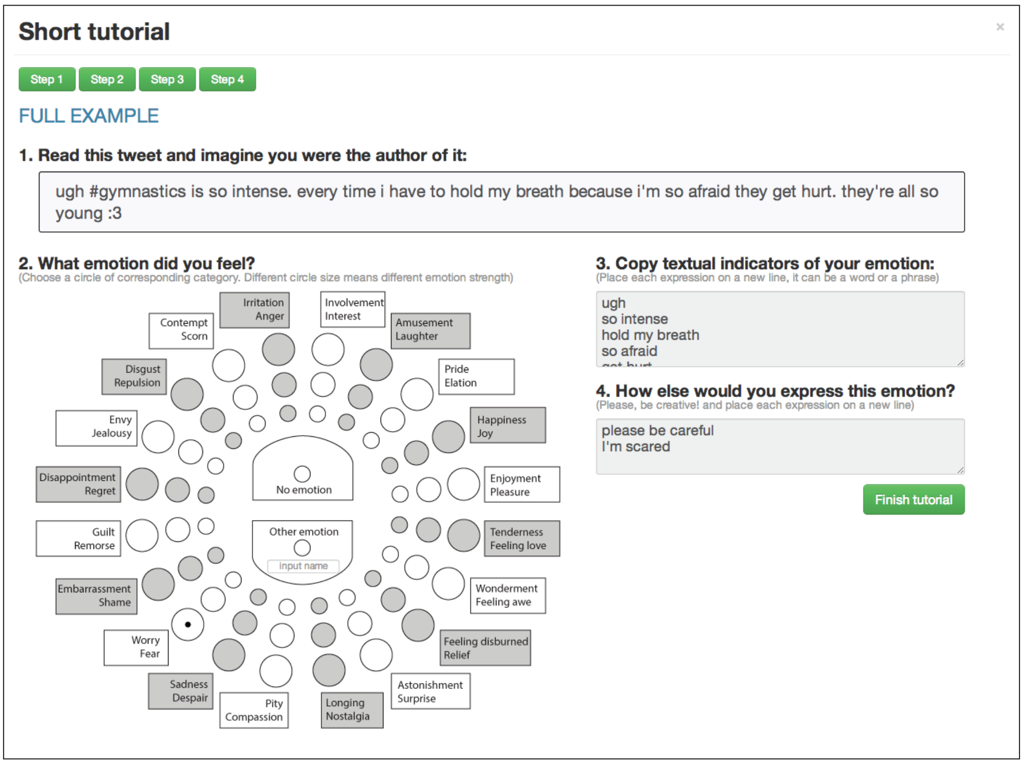
INSTRUCTIONS

1. **Input words or phrases which you would use to express the emotion from this tweet**

   ! You should think about them as of emotion indicators possible to appear in a real tweet

   ! Each distinct expression should begin a new line

Figure A.3: Tutorial, Step 3.



Short tutorial

FULL EXAMPLE

1. Read this tweet and imagine you were the author of it:

ugh #gymnastics is so intense. every time i have to hold my breath because i'm so afraid they get hurt. they're all so young :3

2. What emotion did you feel?
(Choose a circle of corresponding category. Different circle size means different emotion strength)

3. Copy textual indicators of your emotion:
(Place each expression on a new line, it can be a word or a phrase)

ugh
so intense
hold my breath
so afraid
get hurt

4. How else would you express this emotion?
(Please, be creative! and place each expression on a new line)

please be careful
I'm scared

Figure A.4: Tutorial, Step 4.

Tutorial Quiz Questions

1. What does it mean “to evaluate the emotion you felt as if you were the author of the tweet”?
   (a) to evaluate your emotion as a reaction to the tweet you’ve read
   (b) to evaluate your emotion at this moment
   (c) to evaluate your emotion in a situation in which you would write this tweet
2. What does it mean “to choose the dominant emotion at the moment of writing this tweet”?
   (a) to choose your dominant personality trait from the given wheel
   (b) to choose one emotion label that would best describe what you felt
   (c) to choose the first emotion label that could apply
3. What does it mean “to copy textual indicators of your emotion”?
   (a) to copy-paste all the words or short phrases indicative of the emotion in the tweet
   (b) to copy-paste the whole tweet text
   (c) to input a few phrases of your own representing the chosen emotion
4. What does it mean “to provide other ways to express this emotion in a tweet”?
   (a) to input additional emotion labels appropriate to describe what you felt
   (b) to copy-paste all the words or short phrases indicative of the emotion in the tweet
   (c) to input your own words and phrases to describe this emotion

Correct answers for the tutorial quiz: 1 (c), 2 (b), 3 (a), 4 (c)

A.5 Parameter Tuning in Distant Learning

This section gives more details about the selected algorithms’ optimal parameters. In the process of parameter tuning, we varied the following parameters:

- the length n of n-grams features: from 1 to 2;
- the minimum occurrence of n-grams, K: it was fixed to be 5;
- the threshold τ of feature selection: 0 (no selection), 0.1, 0.3, 0.5, 0.7, and 0.9;
- α_{ref} used in the annotation refinement (not applicable with GALC as an emotion labeler): 0 (no refinement), 0.7, 0.9, and 1.0;
- α_0 of the multi-label selection for output (where applicable): 0.7, 0.9, and 1.0;
- for two algorithms in One-vs.-Rest setting, the probability threshold for output r: 0.5, 0.7, and 0.9;
- the threshold θ for per-category feature selection (where applicable): 0 (no selection), 0.1, 0.3, 0.5, 0.7, and 0.9.

Emotion Labeler	Algorithm	n	τ	α_{ref}	α_0	Other
GALC	mcl-MNB	2	0.1	-	0.7	
	mcl-LogReg	2	0	-	0.9	
	1vR-MNB	2	0.1	-	-	r=0.5, θ=0
	1vR-LogReg	1	0.7	-	-	r=0.7, θ=0.9
	PMI-based	1	0.7	-	1	θ=0.7
	BWV	2	0.1	-	0.9	
	WV	2	0.9	-	1	
	BWV-NoNeut	1	0.7	-	1	
OlympLex	mcl-MNB	2	0	1	1	
	mcl-LogReg	2	0	0.9	1	
	1vR-MNB	2	0	1	-	r=0.7, θ=0
	1vR-LogReg	2	0.7	0.7	-	r=0.9, θ=0.9
	PMI-based	2	0.1	0.7	1	θ=0.1
	BWV	2	0	0.7	1	
	WV	2	0.9	1	1	
	BWV-NoNeut	1	0.7	1	1	
PMI-Hash	mcl-MNB	2	0.1	1	0.7	
	mcl-LogReg	2	0	0.9	1	
	1vR-MNB	2	0.3	0.9	-	r=0.5, θ=0.7
	1vR-LogReg	2	0.7	0.7	-	r=0.9, θ=0.1
	PMI-based	2	0.1	0.9	0.7	θ=0.7
	BWV	2	0	1	0.9	
	WV	2	0.9	1	1	
	BWV-NoNeut	1	0.9	0.9	1	

Table A.7: The optimal parameters of the distant learning algorithms chosen by the validation process.

Table A.7 presents which parameters were found as optimal for each distant learning algorithm, separately for each initial emotion labeler. The optimization was based on the maximization of micro-F1 score on the validation set S_V. We summarize below the selected optimal parameters to provide insights about which parameters are preferred by each algorithm in the distant learning process.

First of all, most of the algorithms obtain better results while using both unigrams and bigrams as features ($n = 2$). Using only unigrams ($n = 1$) results in consistently better micro-F1 scores for BWV-NoNeut—the BWV algorithm without using pseudo-neutral tweets in the learning process,— and for 1vR-LogReg and PMI-based method when started with GALC lexicon.

Second, we can separate two classes of algorithms in respect to their feature selection preferences: those that prefer no feature selection or use low τ values up to 0.3 (mcl-MNB, mcl-LogReg, 1vR-MNB, and BWV), and those that prefer higher τ values of 0.7 or 0.9 (1vR-LogReg, WV, BWV-NoNeut). One can notice that 1vR-LogReg additionally performs per-category feature selection ($\theta \geq 0.1$) and sets a higher threshold for outputting emotions ($r \geq 0.7$). PMI-based method always applies feature selection, but adapts its thresholds τ and θ depending on the initial emotion labeler.

Third, all the algorithms prefer to apply annotation refinement ($\alpha_{ref} > 0$) for the initial emotion lexicons incorporating term weights (OlympLex and PMI-Hash). The threshold is mostly high ($\alpha_{ref} \geq 0.9$), i.e. only the emotions with the highest weights are remaining in the pseudo-labeled emotionalities.

Forth, most of the classifiers that make their final decisions based on the computed emotionality (i.e. those that specify α_0 parameter) require output multi-label selection operator with high $\alpha_0 \geq 0.9$. This means that they learn to output only those emotions that are the closest to the one having the maximal weight in the emotionality.

A.6 Examples of Classified Tweets via Dystemo with BWV Algorithm

This section provides the example tweets in our experiments and how they were classified by the classifiers produced by Balanced Weighted Voting algorithm with the Dystemo's distant learning framework (chapter 6). Table A.8 lists for each emotion labeler three example tweets for which BWV found a correct emotion, whereas the corresponding initial emotion labeler did not. Those tweets are from manually annotated set S_M. In many cases, the BWV classifier is superior because it can find correct emotions in the tweets where initial labeler found no emotion (examples R.a-R.c, R.e-R.f). At the same time, other examples confirm the ability of the BWV classifier to correct the initial lexicons to perform better. Consider example R.i: while the expression "not_care about" is associated with *Sadness* by PMI-Hash, BWV correctly reassigns it to both *Sadness* and *Anger* and uses additional cues from the text to classify the text correctly as *Anger*.

Table A.9 gives the examples of tweets where all the resultant BWV classifiers make classification errors, regardless of an initial emotion labeler. It is noteworthy that many of these mistakes are due to the inability of the considered classifiers to capture longer emotional expressions. This can be observed in examples W.a-W.f. Furthermore, neglecting the modifying effect of the word "miss" leads to an error in example W.g. The errors in examples W.h and W.i are due to the lack of proper sentence structure modeling. It results in the classifiers' inability to identify the emotional phrases conveying the main emotional sense, as the phrase "thank goodness" does in example W.h. Finally, examples W.b and W.j reveal the usefulness of expressions formed only from the stop-words, such as "it's on" and "how can i ever", and the need for their inclusion in classification.

ref	Example tweet	Given labels	EL labels	BWV labels	EL
R.a	#TeamDunford Jason Dunford has qualified for nxt round! #KenyaTwaweza #TeamKenya	Pride	No Emotion	Pride	GALC
R.b	No offence but Someones last name in the Olympics is Hooker! Phahahaha!ah ok back to normal	Amusement	No Emotion	Amusement	
R.c	I am SO SICK of seeing #Olympics spoilers on Twitter and Facebook.	Disgust + Anger	No Emotion	Anger	
R.d	I’m so jealous of Olympic athletes who are also good looking. I am so inferior	Envy	Love	Envy	OlympLex
R.e	Thankyou Saina and the humble Gopichand. We shall break the chinese supremacy very soon. #badminton #proudmoment	Pride	No Emotion	Pride	
R.f	NBC ruined the Olympics for me	Regr. + Disg. + Anger	No Emotion	Anger	
R.g	Video: Serena Williams Gets Walk On At Olympics	No Emotion	Anger	No Emotion	PMI-Hash
R.h	Can’t believe the Olympics is over tomorrow already	Nostalgia + Sadness	Worry	Sadness	
R.i	I don’t care about the Olympics okay	Contempt + Anger	Sadness	Anger	

Table A.8: Examples of the tweets correctly classified by BWV for all three initial emotion labelers. EL=Emotion Labeler.

ref	Example tweet	Emotions
W.a	Watching the Olympics makes me fully unaware of how unsuccessful my life is #whydidIstoptraining	Guilt + Regret
W.b	Olympic level skill... How can I ever reach that level?!	Awe + Envy
W.c	Olympic girls look like men...	Disgust
W.d	Watchin olympic instead of doin project	Involvement + Guilt
W.e	only 6 days until @onedirection plays at the 2012 Olympic final ceremony’D	Involvement
W.f	Watching volleyball on the #Olympics makes me miss it oh so much.	Nostalgia
W.g	Work be crampin my style with these Olympic games! I always miss the good stuff!	Regret
W.h	Thank goodness all 204 flags are coming in at once !!! #closingceremony	Relief
W.i	Easily amused British public. Makes me sick. Crack on Olympics, let’s get it over with.	Contempt
W.j	It’s on. Men’s 800m final. Olympics!	Involvement

Table A.9: Examples of the tweets *in*correctly classified by BWV for all three initial emotion labelers. The emotion labels are those from the manual annotations.

A.7 Lists of Modifiers' Terms

We enumerate below the terms used for detecting each modifier class. *Normal* terms correspond to the terms indicative of the modifier's presence. *False positive* terms are phrases that contain normal modifier's terms, but do not imply the modifier's effect. In some cases, we also add conditions on when to detect such terms, e.g. whether the modifier's term should appear before or after the emotional term, or if it should be detected only when the emotional term is in comparative form. The details of the modifiers' detection process are given in section 7.4.2.

Negation

Normal: doesnt mean, does n't mean, do n't have to, do not have to, dont have to, does n't have to, does not have to, doesnt have to, not like, n't like, i do n't, i do not, i dont, you do n't, you do not, you dont, she does n't, she does not, she doesnt, he does n't, he does not, he doesnt, we do n't, we do not, we dont, they do n't, they do not, they dont, ai n't even, ca n't stay, ca n't be, not even, i do n't get, i did n't, was n't, was not, could never, wo n't, wont, wo n't be, wont be, will not, will not be, will never, will never be, 'll never, 'll never be, i would n't be, with no, get no, this ai n't, show no, gets no, got no, do n't, do not, dont, do n't be, do not be, dont be, do n't you, do not you, dont you, do n't get, do not get, dont get, never be, not going to, not gonna, no, non, none, nor, not, nothing, neither, nobody, nowhere, n't, don't, doesn't, doesnt, dont, cant, can't, cannot, couldn't, couldnt, shouldn't, shouldnt, shan't, shant, never, aren't, arent, isn't, isnt, didn't, didnt, havent, haven't, hasn't, hasnt, hadn't, hadnt, weren't, werent, wasn't, wasnt, wouldn't, wouldnt, needn't, neednt, daren't, oughtn't, oughtnt, mightn't, mightnt, mustn't, mustnt, aint, unlikely, to avoid, lack of, pretend to be, pretend it is, pretend to, refuse to be, annul, avert, avoid, betray, bury, carelessness, counteract, decay, decline, deflect, deny, descent, diminution, disregard, disuse, fail, false, fake, ignore, invalid, invalidate, lack, neglect, nullify, obviate, pretend, quash, refrain, refuse, sabotage, scraps, subvert, unable, undermine, void, wane, weaken, annuls, averts, avoids, betrays, buries, counteracts, decays, declines, deflects, denies, descents, disregards, disuses, fails, falses, fakes, ignores, invalids, invalidates, lacks, neglects, nullifies, obviates, pretends, quashes, refrains, refuses, sabotages, subverts, undermines, voids, wanes, weakens, annulled, averted, avoided, betrayed, buried, counteracted, decayed, declined, deflected, denied, disregarded, disused, failed, falsed, faked, ignored, invalided, invalidated, lacked, neglected, nullified, obviated, pretended, quashed, refrained, refused, sabotaged, scraped, subverted, unabled, undermined, voided, waned, weakened, stop, stops, stopped, stopped

False positive: ca n't stop, cant stop, cannot stop, can not stop, should n't be, shouldnt be, should n't have, shouldnt have, do n't you just, dont you just, if i did n't, wish i did n't, ca n't describe the, ca n't describe my, can not describe the, can not describe my, never fail to, never forget, cant forget, ca n't forget, can not forget, cannot forget, n't forget, not to forget, n't just forget, hope to forget, wish to forget, dont ignore, do n't ignore, n't ignore, have no idea how, ca n't help but, ca n't help, cant help but, cannot help but, can not help but, nothing but, who does n't, nothing is more, ca n't wait to, cant wait to, can not wait to, cannot wait to, do n't get how, do n't u, is no greater, you do n't know what, you do n't know how, have n't stopped, havent stopped, hard not to, ca n't believe how, cant believe how, can not believe how, cannot believe how, nothing like, can not express how, cannot express how, ca n't express how, can not express my, cannot express my, ca n't express my, never fails, do n't u just, cant explain how, cannot explain how, ca n't explain how, can not explain how, dont u just, never cease to, never ceases to, nobody knows how, ai n't it, never realized how, never seizes to, did n't realize, never knew, n't wipe, forget how, forgot how, forget why, forgot why, n't <term> this much, not <term> this much, not <term> so much, n't <term> so much, n't <term> this hard, not <term> this hard, n't <term> like this, not <term> like this, never lose, could n't hide my, ca n't hide my, nothing beats, not only, did n't know, you do n't know the, you do n't know my, you dont know the, you dont know my, know nothing about, have n't seen <term> until, no i, no you, no u, no we, no they, no it, no he, no she, ca n't deal, can not deal, cannot deal, cant deal, no one <term> but, no wonder, not to mention, no doubt, no doubts

Intensifiers

Normal, a modifier term should be *before* the emotional term: more, very, absolutely, so, so much, so fucking, fucking, fuckin, really, how much i, the most, lots of, my biggest, as much as i, too much, pretty, so in, really do, lot of, super, how much you, such a, how much, extremely, so incredibly, so unbelievably, big, in total, strongly, how much they, <username> much, quite, the biggest, biggest, soo, soo*, absolute, amazingly, awfully, bigger, certainly, complete, completely, considerably, decidedly, deeply, definitely, effing, enormously, entirely, especially, exceedingly, exceptionally, extra, extraordinarily, fabulously, fairly, far, flipping, flippin, fricking, frickin, frigging, friggin, fully, great, greatly, hella, high, higher, highest, highly, huge, hugely, immensely, incredible, incredibly, intensely, lot, lots, major, majorly, massively, much, obviously, particularly, perfectly, positively, purely, rather, real, remarkably, significantly, simply, some, substantially, such, terribly, thoroughly, total, totally, tremendous, tremendously, uber, unbelievably, unusually, utter, utterly, vastly, well

Normal, a modifier term should be *after* the emotional term: strongly, more, it so, you more, u more, so much, you much, me some, you so, u so, u soo much, you soo much, you soo* much, u soo* much, you very much, you a lot, u a lot, you lots, u lots, him so much, her so much, me so much, them so much, you guys so, u guys so, <username> so much, this so much, the most, you all so, too much, him more, her more, <username> more, you too much, myself so much, school so much, ya so much, u so fucking much, u very, you very, much, you both so much, this song so much, you soso much, someone so much, you so very much, ya lots, you most

False positive: as well

Intensifiers with negation words (they are considered as false positive for negation modifier): never been so, never been soo, never been soo*, never been more, never been this, never been as, nothing i <term> more, ca n't be more, ca n't be anymore, cant be more, cant be anymore, cannot be more, cannot be anymore, can not be more, can not be anymore, have n't been this, have n't been so, have n't been soo, have n't been soo*, havent been this, havent been so, havent been soo, havent been soo*, ca n't describe how, cant describe how, cannot describe how, can not describe how, nothing makes me more, nothing makes me <term> more, could not be any, could not be more, could not be anymore, could n't be any, could n't be more, could n't be anymore, couldnt be more, couldnt be any, couldnt be anymore, nothing more <term> than, nothing more i <term> than, nothing more <term> then, nothing more, nothing 's more, nothing is more, never felt so, never felt soo, never felt soo*, never felt this, never felt more, have n't felt so, have n't felt soo, have n't felt soo*, have n't felt this, have n't felt more, never felt <term> like this, never felt <term> so much, never felt <term> this much, nothing short of, never <term> myself more than, never <term> you more than, never <term> me more than, nothing <term> me more, nothing <term> more, nothing <term> as much, nothing <term> so much, nothing that <term> me more

Intensifiers with negation words, with condition that the emotional term should be in *comparative* form: never been, ca n't be, cant be, cannot be, can not be, have n't been, havent been, nothing makes me, could not be, could n't be, couldnt be, never felt, have n't felt, nothing <term> than, nothing <term> then, nothing is <term> than, nothing is <term> then

Diminishers

Normal, a modifier term should be *before* the emotional term: a little, kinda feel, kinda, bit, 'm kind of, i kind of, 's kind of, 're kind of, it kind of, is kind of, are kind of, actually kind of, kind of, little bit, slightly, little bit of, little bit in, a bit of, 's less, is less, be less, had little, almost, barely, difficult, few, fewer, fewest, hardly, just enough, kindof, kind-of, least, less, little, low, lower, lowest, marginally, minor, occasionally, partly, relatively, ridiculously, scarcely, small, somewhat, sort of, sorta, sortof, sort-of, not really, unlikely, unlikely to, unlikely to be, only

Normal, a modifier term should be *after* the emotional term: less, you less

False positive: that kinda, that kind of, this kind of, what kind of, the kind of, any kind of, a kind of, at least

Modality

Normal: 's meant to, is meant to, 're meant to, are meant to, 'm meant to, am meant to, meant to, should have, should 've, should <any> have, should <any> 've, was going to, was gonna, should, would <any> have, would <any> 've, would have, would 've, might <any> have, could <any> have, might have, could have, might <any> 've, could <any> 've, might 've, could 've, could, need, 'd better, you better, i better, they better, we better, it better, she better, he better, would rather, 'd rather, supposed to, suppose to, i suppose, supposedly, reportedly, presumably, presumed to, assume, reputedly, seemingly, purportedly, arguably, allegedly, if only, i doubt, i highly doubt, doubtful, doubtfully, conceivably, would like, would <any> like, would love to, would <any> love to, 'd like, 'd <any> like, 'd love to, 'd <any> love to, would, 'd, may have, may, maybe, probably, possible, possibly, probable, potentially, perhaps, most likely, more likely, very likely, arguably, likely, unlikely, hopefully, hope, wish, want, wanna, not sure, not certain, shall, will, 'll, wo, wont, wouldnt, shouldnt, couldnt, is going to, 's going to, going to, gonna, ought, oughta, must <any> have, must have, must <any> 've, must 've, must, mustnt, can, ca, cant, cannot, able, unable, have to, has to, had to, have got to, has got to, had got to, gotta, necessarily, seem to, seems to, seemed to, seem like, seems like, seemed like, seem, seems, seemed, i think, obliged to, about to, appears to, it appears, this appears, got to, bound to, i guess, i suspect, i believe, i 'm guessing, i expect, i presume, i assume, i may assume, it seems

False positive, no condition: want to say, ca n't stop, cant stop, cannot stop, can not stop, cant forget, ca n't forget, can not forget, cannot forget, hope to forget, wish to forget, ca n't help, cant help, cannot help, can not help, ca n't describe, cant describe, cannot describe, can not describe, ca n't express, cant express, cannot express, can not express, ca n't explain, cant explain, cannot explain, can not explain, ca n't wipe, ca n't be more, ca n't be anymore, cant be more, cant be anymore, could not be any, could not be more, could not be anymore, could n't be any, could n't be more, could n't be anymore, couldnt be more, couldnt be any, couldnt be anymore

False positive, with condition that the emotional term should be in *comparative* form: ca n't be, ca n't be more, ca n't be anymore, cant be, cant be more, cant be anymore, could not be, could not be any, could not be more, could not be anymore, could n't be, could n't be any, could n't be more, could n't be anymore, couldnt be

Interrogation

Normal, a modifier term should be *before* the emotional term: will you, will u, will i, will they, will we, will he, will she, will it, do you, do u, do i, does it, does he, does she, do we, do they, am i, 'm i, 's it, is it, are you, are u, is he, 's he, is she, 's she, are we, are they, what will, what can, what cant, what ca n't, what do, what should, what could, what may, what might, what about, what is, what has, what have, can you, can u, can i, can we, can they, can it, can he, can she, why is, why can, why cant, why ca n't, why do, why does, why do n't, why does n't, why will, why wont, why wo n't, why has, why have, why should, why could, why did, why did n't, did i, did you, did u, did it, did he, did she, did we, did they, have i, have you, have u, has it, has he, has she, have we, have they, have n't i, have n't you, have n't u, has n't it, has n't he, has n't she, have n't we, have n't they, where is, where can, where cant, where ca n't, where do, where does, where dont, where do n't, where does n't, where doesnt, where will, where has, where have, where should, where could, where did, where did n't, when is, when can, when cant, when ca n't, when do, when does, when dont, when do n't, when doesnt, when does n't, when will, when has, when have, when should, when could, when did, when did n't, how is, how can, how cant, how ca n't, how do, how dont, how do n't, how does, how doesnt, how does n't, how will, how has, how have, how should, how could, how did, how did n't, do n't i, do n't you, do n't u, do n't we, does n't it, does n't he, does n't she, do n't they, did n't i, did n't you, did n't u, did n't we, did n't it, did n't he, did n't she, did n't they

Normal, a modifier term should be *after* the emotional term: ?, ?*, ?!, ?!*, ?!?, ?!?*, ?*!, ?*!*, !?, !?*, !*?*

Conditionality

Normal, a modifier term should be *before* the emotional term: if

A.8 Source Code and Resultant Classifiers

We provide to the research community the emotion lexicon generated in our human computation task (OlympLex, chapter 4), the source code of the distant learning framework (Dystemo) and the main classifiers obtained through it (chapter 6), as well as the code for detecting and analyzing the linguistic modifiers (chapter 7). These supplementary material can be found at http://hci.epfl.ch/valentina/thesis-materials.

Bibliography

[ABI$^+$13] Mohammad Allahbakhsh, Boualem Benatallah, Aleksandar Ignjatovic, Hamid Reza Motahari-Nezhad, Elisa Bertino, and Schahram Dustdar. Quality control in crowdsourcing systems: Issues and directions. *IEEE Internet Computing*, 17(2):76–81, 2013.

[ABY11] Omar Alonso and Ricardo Baeza-Yates. Design and implementation of relevance assessments using crowdsourcing. In *Advances in Information Retrieval, Proceedings of the European Conference on Information Retrieval (ECIR), Volume 6611 of the series Lecture Notes in Computer Science*, pages 153–164. Springer, 2011.

[Ado02] Ralph Adolphs. Neural systems for recognizing emotion. *Current Opinion in Neurobiology*, 12(2):169–177, 2002.

[AMPI07] Shaikh Mostafa Al Masum, Helmut Prendinger, and Mitsuru Ishizuka. Emotion sensitive news agent: An approach towards user centric emotion sensing from the news. In *Proceedings of IEEE/WIC/ACM International Conference on Web Intelligence (WI)*, pages 614–620. IEEE, 2007.

[AOL$^+$12] Erik Andersen, Eleanor O'Rourke, Yun-En Liu, Rich Snider, Jeff Lowdermilk, David Truong, Seth Cooper, and Zoran Popovic. The impact of tutorials on games of varying complexity. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 59–68. ACM, 2012.

[AP08] Ron Artstein and Massimo Poesio. Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596, 2008.

[AR10] Shazia Afzal and Peter Robinson. Modelling affect in learning environments—Motivation and methods. In *Proceedings of the 10th IEEE International Conference on Advanced Learning Technologies (ICALT)*, pages 438–442. IEEE, 2010.

[AR11] Andra Adams and Peter Robinson. An android head for social-emotional intervention for children with autism spectrum conditions. In *Affective Computing and Intelligent Interaction, Conference Proceedings (ACII), Volume 6975 of the series Lecture Notes in Computer Science*, pages 183–190. Springer, 2011.

[ARS05] Cecilia Ovesdotter Alm, Dan Roth, and Richard Sproat. 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 (HLT)*, pages 579–586. ACL, 2005.

[AS07] Saima Aman and Stan Szpakowicz. Identifying expressions of emotion in text. In *Text, Speech and Dialogue, Conference Proceedings (TSD), Volume 4629 of the series Lecture Notes in Computer Science*, pages 196–205. Springer, 2007.

[Bar06] Lisa Feldman Barrett. Solving the emotion paradox: Categorization and the experience of emotion. *Personality and Social Psychology Review*, 10(1):20–46, 2006.

[BB95] George Bojadziev and Maria Bojadziev. *Fuzzy sets, fuzzy logic, applications*, volume 5. World Scientific, 1995.

[BB13] Eugene Y Bann and Joanna J Bryson. The conceptualisation of emotion qualia: Semantic clustering of emotional tweets. In *Proceedings of the 13th Neural Computation and Psychology Workshop on Computational Models of Cognitive Processes*, pages 249–263. World Scientific, 2013.

Bibliography

[BBMC89] John B Black, J Scott Bechtold, Marco Mitrani, and John M Carroll. On-line tutorials: What kind of inference leads to the most effective learning? *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI), ACM SIGCHI Bulletin*, 20(SI):81–83, 1989.

[BCM$^+$12] Farah Benamara, Baptiste Chardon, Yannick Mathieu, Vladimir Popescu, and Nicholas Asher. How do negation and modality impact on opinions? In *Proceedings of the Workshop on Extra-Propositional Aspects of Meaning in Computational Linguistics*, pages 10–18. ACL, 2012.

[BDY$^+$04] Carlos Busso, Zhigang Deng, Serdar Yildirim, Murtaza Bulut, Chul Min Lee, Abe Kazemzadeh, Sungbok Lee, Ulrich Neumann, and Shrikanth Narayanan. Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th International Conference on Multimodal Interfaces (ICMI)*, pages 205–211. ACM, 2004.

[BE10] Samuel Brody and Noemie Elhadad. An unsupervised aspect-sentiment model for online reviews. In *Proceedings of Human Language Technologies (HLT): The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 804–812. ACL, 2010.

[BFL98] Collin F Baker, Charles J Fillmore, and John B Lowe. The Berkeley FrameNet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL) and 17th International Conference on Computational Linguistics (COLING)*, volume 1, pages 86–90. ACL, 1998.

[BGRM11] Johan Bollen, Bruno Gonçalves, Guangchen Ruan, and Huina Mao. Happiness is assortative in online social networks. *Artificial life*, 17(3):237–251, 2011.

[BHM12] Alexandra Balahur, Jesus M Hermida, and Andres Montoyo. 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, 2012.

[BJJW14] Joel Brynielsson, Fredrik Johansson, Carl Jonsson, and Anders Westling. Emotion classification of social media posts for estimating people's reactions to communicated alert messages during crises. *Security Informatics*, 3(1):1–11, 2014.

[BL99] Margaret M Bradley and Peter J Lang. Affective norms for English words (ANEW): Instruction manual and affective ratings. Technical report C-1, The Center for Research in Psychophysiology, University of Florida, 1999.

[BMF14] Marina Boia, Claudiu Cristian Musat, and Boi Faltings. Acquiring commonsense knowledge for sentiment analysis through human computation. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, pages 901–907, 2014.

[BMMP13] Felipe Bravo-Marquez, Marcelo Mendoza, and Barbara Poblete. Combining strengths, emotions and polarities for boosting Twitter sentiment analysis. In *Proceedings of the Second International Workshop on Issues of Sentiment Discovery and Opinion Mining (WISDOM)*. ACM, 2013.

[BMP11] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. In *Proceedings of Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 450–453, 2011.

[BMZ11] Johan Bollen, Huina Mao, and Xiaojun Zeng. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8, 2011.

[BP12] Haji Binali and Vidyasagar Potdar. Emotion detection state of the art. In *Proceedings of the CUBE International Information Technology Conference*, pages 501–507. ACM, 2012.

[BPM04] Gustavo E A P A Batista, Ronaldo C Prati, and Maria Carolina Monard. A study of the behavior of several methods for balancing machine learning training data. *ACM SIGKDD Explorations Newsletter*, 6(1):20–29, 2004.

[BT14] Alexandra Balahur and Marco Turchi. Comparative experiments using supervised learning and machine translation for multilingual sentiment analysis. *Computer Speech & Language*, 28(1):56–75, 2014.

[CD10] Rafael A Calvo and Sidney D'Mello. Affect detection: An interdisciplinary review of models, methods, and their applications. *IEEE Transactions on Affective Computing*, 1(1):18–37, 2010.

[CdAP13] Jorge Carrillo-de Albornoz and Laura Plaza. An emotion-based model of negation, intensifiers, and modality for polarity and intensity classification. *Journal of the American Society for Information Science and Technology*, 64(8):1618–1633, 2013.

[CDCT$^+$01] Roddy Cowie, Ellen Douglas-Cowie, Nicolas Tsapatsoulis, George Votsis, Stefanos Kollias, Winfried Fellenz, and John G Taylor. Emotion recognition in human-computer interaction. *IEEE Signal Processing Magazine*, 18(1):32–80, 2001.

[CDH14] Glen Coppersmith, Mark Dredze, and Craig Harman. Quantifying mental health signals in Twitter. In *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality*, pages 51–60. ACL, 2014.

[CGSG04] Scotty Craig, Arthur Graesser, Jeremiah Sullins, and Barry Gholson. Affect and learning: An exploratory look into the role of affect in learning with AutoTutor. *Journal of Educational Media*, 29(3):241–250, 2004.

[Che15] Yu Chen. *Social Interface and Interaction Design for Group Recommender Systems*. Doctoral dissertation, École Polytechnique Fédérale de Lausanne, 2015.

[CLH12] Erik Cambria, Andrew Livingstone, and Amir Hussain. The hourglass of emotions. In *Cognitive Behavioral Systems*, pages 144–157. Springer, 2012.

[CLLH10] Ying Chen, Sophia Yat Mei Lee, Shoushan Li, and Chu-Ren Huang. Emotion cause detection with linguistic constructions. In *Proceedings of the 23rd International Conference on Computational Linguistics (COLING)*, pages 179–187. ACL, 2010.

[CR86] Davida H Charney and Lynne M Reder. Designing interactive tutorials for computer users. *Human-Computer Interaction*, 2(4):297–317, 1986.

[CS14] Ru Shan Chen and Yuta Sakamoto. Feelings and perspective matter: Sharing of crisis information in social media. In *Proceedings of the 47th Hawaii International Conference on System Sciences (HICSS)*, pages 1958–1967. IEEE, 2014.

[CSK$^+$14] Lorenzo Coviello, Yunkyu Sohn, Adam DI Kramer, Cameron Marlow, Massimo Franceschetti, Nicholas A Christakis, and James H Fowler. Detecting emotional contagion in massive social networks. *PLoS ONE*, 9(3):e90315, 2014.

[CSKFMR87] John M Carroll, Penny L Smith-Kerker, James R Ford, and Sandra A Mazur-Rimetz. The minimal manual. *Human-Computer Interaction*, 3(2):123–153, 1987.

[CT06] Thomas M Cover and Joy A Thomas. *Elements of information theory*. John Wiley & Sons, 2006.

[CTIB15] Justin Cheng, Jaime Teevan, Shamsi T Iqbal, and Michael S Bernstein. Break it down: A comparison of macro- and microtasks. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI)*, pages 4061–4064. ACM, 2015.

[CW04] Ze-Jing Chuang and Chung-Hsien Wu. Multi-modal emotion recognition from speech and text. *Computational Linguistics and Chinese Language Processing*, 9(2):45–62, 2004.

[DA08] Taner Danisman and Adil Alpkocak. Feeler: Emotion classification of text using vector space model. In *Proceedings of the AISB 2008 Convention Communication, Interaction and Social Intelligence, Volume 2: Proceedings of the AISB 2008 Symposium on Affective Language in Human and Machine*, pages 53–59, 2008.

[Dam94] Antonio R Damasio. *Descartes' error: Emotion, rationality and the human brain*. Putnam (Grosset Books), New York, NY, 1994.

[DCCG12] Munmun De Choudhury, Scott Counts, and Michael Gamon. Not all moods are created equal! Exploring human emotional states in social media. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 66–73, 2012.

[DCCH13a] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 3267–3276. ACM, 2013.

[DCCH13b] Munmun De Choudhury, Scott Counts, and Eric Horvitz. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th Annual ACM Web Science Conference*, pages 47–56. ACM, 2013.

[DCGC12] Munmun De Choudhury, Michael Gamon, and Scott Counts. Happy, nervous or surprised? Classification of human affective states in social media. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 435–438, 2012.

[DD10] Peter Sheridan Dodds and Christopher M Danforth. Measuring the happiness of large-scale written expression: Songs, blogs, and presidents. *Journal of Happiness Studies*, 11(4):441–456, 2010.

[DDKN11] Sebastian Deterding, Dan Dixon, Rilla Khaled, and Lennart Nacke. From game design elements to gamefulness: Defining gamification. In *Proceedings of the 15th International Academic MindTrek Conference: Envisioning Future Media Environments*, pages 9–15. ACM, 2011.

[Den08] Kerstin Denecke. Using SentiWordNet for multilingual sentiment analysis. In *Proceedings of the IEEE 24th International Conference on Data Engineering Workshops (ICDEW), Workshop on Data Engineering for Blogs, Social Media, and Web 2.0*, pages 507–512. IEEE, 2008.

[Des12] Pieter M A Desmet. Faces of product pleasure: 25 positive emotions in human-product interactions. *International Journal of Design*, 6(2), 2012.

[DH13] Bart Desmet and Véronique Hoste. Emotion detection in suicide notes. *Expert Systems with Applications*, 40(16):6351–6358, 2013.

[DHSC10] Julie S Downs, Mandy B Holbrook, Steve Sheng, and Lorrie Faith Cranor. Are your participants gaming the system?: Screening Mechanical Turk workers. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 2399–2402. ACM, 2010.

[DNKS10] Nicholas Diakopoulos, Mor Naaman, and Funda Kivran-Swaine. Diamonds in the rough: Social media visual analytics for journalistic inquiry. In *Proceedings of the IEEE Symposium on Visual Analytics Science and Technology (VAST)*, pages 115–122. IEEE, 2010.

[EAKK11] Moataz El Ayadi, Mohamed S Kamel, and Fakhri Karray. Survey on speech emotion recognition: Features, classification schemes, and databases. *Pattern Recognition*, 44(3):572–587, 2011.

[EC11] Elsa Eiriksdottir and Richard Catrambone. Procedural instructions, principles, and examples how to structure instructions for procedural tasks to enhance performance, learning, and transfer. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 53(6):749–770, 2011.

[Ekm92] Paul Ekman. An argument for basic emotions. *Cognition & Emotion*, 6(3-4):169–200, 1992.

[Ell92] Clark Davidson Elliott. *The affective reasoner: A process model of emotions in a multi-agent system*. Doctoral dissertation, Northwestern University, 1992.

[FCH$^+$08] Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874, 2008.

[FG13] Virginia Francisco and Pablo Gervás. EMOTAG: An approach to automated markup of emotions in texts. *Computational Intelligence*, 29(4):680–721, 2013.

[FKTC13] Ailbhe Finnerty, Pavel Kucherbaev, Stefano Tranquillini, and Gregorio Convertino. Keep it simple: Reward and task design in crowdsourcing. In *Proceedings of the Biannual Conference of the Italian Chapter of SIGCHI (CHIItaly)*. ACM, 2013.

[FL03] Beat Fasel and Juergen Luettin. Automatic facial expression analysis: A survey. *Pattern Recognition*, 36(1):259–275, 2003.

[Fle71] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[FMK+10] Tim Finin, Will Murnane, Anand Karandikar, Nicholas Keller, Justin Martineau, and Mark Dredze. Annotating named entities in Twitter data with crowdsourcing. In *Proceedings of the NAACL-HLT Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, pages 80–88. ACL, 2010.

[FPTJ14] Boi Faltings, Pearl Pu, Bao Duy Tran, and Radu Jurca. Incentives to counter bias in human computation. In *Proceedings of the Second AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, pages 59–66. AAAI, 2014.

[Fri87] Nico H Frijda. Emotion, cognitive structure, and action tendency. *Cognition and emotion*, 1(2):115–143, 1987.

[FSRE07] Johnny R J Fontaine, Klaus R Scherer, Etienne B Roesch, and Phoebe C Ellsworth. The world of emotions is not two-dimensional. *Psychological science*, 18(12):1050–1057, 2007.

[GAAES14] Benjamin Guthier, Rajwa Alharthi, Rana Abaalkhail, and Abdulmotaleb El Saddik. Detection and visualization of emotions in an affect-aware city. In *Proceedings of the 1st International Workshop on Emerging Multimedia Applications and Services for Smart Cities (EMASC)*, pages 23–28. ACM, 2014.

[GBH09] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. Technical report, Stanford, 2009.

[GCHP11] Marco Grassi, Erik Cambria, Amir Hussain, and Francesco Piazza. Sentic web: A new paradigm for managing social media affective information. *Cognitive Computation*, 3(3):480–489, 2011.

[Geb05] Patrick Gebhard. ALMA: A layered model of affect. In *Proceedings of the Fourth International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 29–36. ACM, 2005.

[GIS10] Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Hierarchical versus flat classification of emotions in text. In *Proceedings of the NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 140–146. ACL, 2010.

[GLHG16] Luciano Gallegos, Kristina Lerman, Arhur Huang, and David Garcia. Geography of emotion: Where in a city are people happier? In *Proceedings of the Companion Publication of the 25th International Conference on World Wide Web (WWW Companion), the 7th International Workshop on Modeling Social Media - Behavioral Analytics in Social Media, Big Data and the Web (MSM)*, pages 569–574. IW3C2, 2016.

[GMC+14] Huiji Gao, Jalal Mahmud, Jilin Chen, Jeffrey Nichols, and Michelle X Zhou. Modeling user attitude toward controversial topics in online social media. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 121–130, 2014.

[Gra09] Marco Grassi. Developing HEO human emotions ontology. In *Biometric ID Management and Multimodal Communication*, pages 244–251. Springer, 2009.

[GYT14] Hua Gao, Anil Yüce, and Jean-Philippe Thiran. Detecting emotional stress from facial expressions for driving safety. In *2014 IEEE International Conference on Image Processing (ICIP)*, pages 5961–5965. IEEE, 2014.

[Har95] Susan M Harrison. A comparison of still, animated, or nonillustrated on-line help with written or spoken instructions in a graphical user interface. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 82–89. ACM Press, 1995.

[Har11] Christopher Harris. You're hired! An examination of crowdsourcing incentive models in human resource tasks. In *Proceedings of the Workshop on Crowdsourcing for Search and Data Mining (CSDM) at the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 15–18, 2011.

[Har13] Marcus Hartner. The lingering after-effects in the reader's mind – An investigation into the affective dimension of literary reading. *Journal of Literary Theory Online*, 2013. (Review of: Michael Burke. *Literary Reading, Cognition and Emotion. An Exploration of the Oceanic Mind*. Routledge, New York/London, 2011).

[HB10] Jeffrey Heer and Michael Bostock. Crowdsourcing graphical perception: Using Mechanical Turk to assess visualization design. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 203–212. ACM, 2010.

[HB12] Amaç Herdağdelen and Marco Baroni. Bootstrapping a game with a purpose for commonsense collection. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 3(4):Article No. 59, 2012.

[HCSM11] Janna Hastings, Werner Ceusters, Barry Smith, and Kevin Mulligan. The emotion ontology: Enabling interdisciplinary research in the affective sciences. In *Modeling and Using Context*, pages 119–123. Springer, 2011.

[HF13] Shih-Wen Huang and Wai-Tat Fu. Enhancing reliability using peer consistency evaluation in human computation. In *Proceedings of the 2013 Conference on Computer Supported Cooperative Work (CSCW)*, pages 639–648. ACM, 2013.

[HFH$^+$09] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H Witten. The WEKA data mining software: An update. *ACM SIGKDD Explorations Newsletter*, 11(1):10–18, 2009.

[HG14] Clayton J Hutto and Eric Gilbert. VADER: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 216–225, 2014.

[HL04] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *Proceedings of the Tenth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 168–177. ACM, 2004.

[HLH11] Sudheendra Hangal, Monica S Lam, and Jeffrey Heer. MUSE: Reviving memories using email archives. In *Proceedings of the 24th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 75–84. ACM, 2011.

[HMnd] Hyisung Hwang and David Matsumoto. Functions of emotions. http://nobaproject.com/modules/functions-of-emotions, n.d. [Online; accessed 19-April-2014].

[HSSV15] Chien-Ju Ho, Aleksandrs Slivkins, Siddharth Suri, and Jennifer Wortman Vaughan. Incentivizing high quality crowdwork. In *Proceedings of the 24th International Conference on World Wide Web (WWW)*, pages 419–429. IW3C2, 2015.

[HT07] Jeff Huang and Michael B Twidale. Graphstract: Minimal graphical help for computers. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology (UIST)*, pages 203–212. ACM, 2007.

[HUM06] HUMAINE Emotion Annotation and Representation Language (EARL). http://emotion-research.net/projects/humaine/earl/proposal, 2006. [Online; accessed 15-June-2016].

[HVIH$^+$11] Alexander Hogenboom, Paul Van Iterson, Bas Heerschop, Flavius Frasincar, and Uzay Kaymak. Determining negation scope and strength in sentiment analysis. In *Proceedings of the IEEE International Conference on Systems, Man, and Cybernetics (SMC)*, pages 2589–2594. IEEE, 2011.

[HX05] Alan Hanjalic and Li-Qun Xu. Affective video content representation and modeling. *IEEE Transactions on Multimedia*, 7(1):143–154, 2005.

[HZP$^+$10] Eric Huang, Haoqi Zhang, David C Parkes, Krzysztof Z Gajos, and Yiling Chen. Toward automatic task design: A progress report. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP)*, pages 77–85. ACM, 2010.

[ILC$^+$14] Daniela Iosub, David Laniado, Carlos Castillo, Mayo Fuster Morell, and Andreas Kaltenbrunner. Emotions under discussion: Gender, status and communication in online collaboration. *PLoS ONE*, 9(8):e104880, 2014.

[IPW10] Panagiotis G Ipeirotis, Foster Provost, and Jing Wang. Quality management on Amazon Mechanical Turk. In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP)*, pages 64–67. ACM, 2010.

[Iza71] Carroll E Izard. *The face of emotion.* Appleton-Century-Crofts, 1971.

[Iza13] Carroll E Izard. *Human emotions.* Springer Science & Business Media, 2013.

[Jap00] Nathalie Japkowicz. The class imbalance problem: Significance and strategies. In *Proceedings of the International Conference on Artificial Intelligence (ICAI)*, volume 1, pages 111–117, 2000.

[JO11] Yohan Jo and Alice H Oh. Aspect and sentiment unification model for online review analysis. In *Proceedings of the Fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 815–824. ACM, 2011.

[Joa98] Thorsten Joachims. Text categorization with Support Vector Machines: Learning with many relevant features. In *European Conference on Machine Learning*, pages 137–142. Springer, 1998.

[JYM09] Lifeng Jia, Clement Yu, and Weiyi Meng. The effect of negation on sentiment analysis and retrieval effectiveness. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, pages 1827–1830. ACM, 2009.

[Kaz11] Gabriella Kazai. In search of quality in crowdsourcing for search engine evaluation. In *Advances in Information Retrieval, Proceedings of the European Conference on Information Retrieval (ECIR), Volume 6611 of the series Lecture Notes in Computer Science*, pages 165–176. Springer, 2011.

[KBO12] Suin Kim, JinYeong Bak, and Alice Haeyun Oh. Do you feel what I feel? Social aspects of emotions in Twitter conversations. In *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 495–498, 2012.

[KCS08] Aniket Kittur, Ed H Chi, and Bongwon Suh. Crowdsourcing user studies with Mechanical Turk. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 453–456. ACM, 2008.

[KH99] Dacher Keltner and Jonathan Haidt. Social functions of emotions at four levels of analysis. *Cognition & Emotion*, 13(5):505–521, 1999.

[KH05] Soo-Min Kim and Eduard Hovy. Identifying opinion holders for question answering in opinion texts. In *Proceedings of AAAI Workshop on Question Answering in Restricted Domains*, pages 1367–1373, 2005.

[KH11] Sepandar D Kamvar and Jonathan Harris. We feel fine and searching the emotional web. In *Proceedings of the fourth ACM International Conference on Web Search and Data Mining (WSDM)*, pages 117–126. ACM, 2011.

[KI06] Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125, 2006.

[KJM07] Youngho Kim, Yuchul Jung, and Sung-Hyon Myaeng. Identifying opinion holders in opinion text from online newspapers. In *Proceedings of the International Conference on Granular Computing (GrC)*, pages 699–702, 2007.

[KKMF13] Gabriella Kazai, Jaap Kamps, and Natasa Milic-Frayling. An analysis of human factors and label accuracy in crowdsourcing relevance judgments. *Information Retrieval*, 16(2):138–178, 2013.

[KL51] Solomon Kullback and Richard A Leibler. On information and sufficiency. *The annals of mathematical statistics*, 22(1):79–86, 1951.

[Kle11] Jochen Kleres. Emotions and narrative analysis: A methodological approach. *Journal for the Theory of Social Behaviour*, 41(2):182–202, 2011.

[KLY$^+$09] Edward Chao-Chun Kao, Chun-Chieh Liu, Ting-Hao Yang, Chang-Tai Hsieh, and Von-Wun Soo. Towards text-based emotion detection: A survey and possible improvements. In *Proceedings of International Conference on Information Management and Engineering (ICIME)*, pages 70–74. IEEE, 2009.

[KMP99] Jonathan Klein, Youngme Moon, and Rosalind W Picard. This computer responds to user frustration. In *Proceedings of the CHI Extended Abstracts on Human Factors in Computing Systems*, pages 242–243. ACM, 1999.

[KNS+11] Kathrin Knautz, Diane Rasmussen Neal, Stefanie Schmidt, Tobias Siebenlist, and Wolfgang G Stock. Finding emotional-laden resources on the world wide web. *Information*, 2(1):217–246, 2011.

[Koł13] Agata Kołakowska. A review of emotion recognition methods based on keystroke dynamics and mouse movements. In *Proceedings of the 6th International Conference on Human System Interactions (HSI)*, pages 548–555. IEEE, 2013.

[Kou14] Xin Kou. The effect of modifiers for sentiment analysis. In *Chinese Lexical Semantics*, pages 240–250. Springer, 2014.

[KP05] Caitlin Kelleher and Randy Pausch. Stencils-based tutorials: Design and evaluation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 541–550. ACM, 2005.

[KPJD13] Uros Krcadinac, Philippe Pasquier, Jelena Jovanovic, and Vladan Devedzic. Synesketch: An open source library for sentence-based emotion recognition. *IEEE Transactions on Affective Computing*, 4(3):312–325, 2013.

[KPV+14] Andreas Kanavos, Isidoros Perikos, Pantelis Vikatos, Ioannis Hatzilygeroudis, Christos Makris, and Athanasios Tsakalidis. Conversation emotional modeling in social networks. In *Proceedings of the IEEE 26th International Conference on Tools with Artificial Intelligence*, pages 478–484. IEEE, 2014.

[KR12] Shashidhar G Koolagudi and K Sreenivasa Rao. Emotion recognition from speech: A review. *International Journal of Speech Technology*, 15(2):99–117, 2012.

[KSM+10] Youngmoo E Kim, Erik M Schmidt, Raymond Migneco, Brandon G Morton, Patrick Richardson, Jeffrey Scott, Jacquelin A Speck, and Douglas Turnbull. Music emotion recognition: A state of the art review. In *Proceedings of the 11th International Society for Music Information Retrieval Conference (ISMIR)*, pages 255–266, 2010.

[KSMP14] Renato Kempter, Valentina Sintsova, Claudiu Musat, and Pearl Pu. EmotionWatch: Visualizing fine-grained emotions in event-related tweets. In *Proceedings of the Eighth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 236–245, 2014.

[KVC10] Sunghwan Mac Kim, Alessandro Valitutti, and Rafael A Calvo. Evaluation of unsupervised emotion models to textual affect recognition. In *Proceedings of the NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 62–70. ACL, 2010.

[KWM11] Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. Twitter sentiment analysis: The good the bad and the OMG! *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 538–541, 2011.

[KZM14] Svetlana Kiritchenko, Xiaodan Zhu, and Saif M Mohammad. Sentiment analysis of short informal texts. *Journal of Artificial Intelligence Research*, pages 723–762, 2014.

[Laz91] Richard S Lazarus. *Emotion and adaptation*. Oxford University Press, 1991.

[LGG+08] Juan Miguel López, Rosa Gil, Roberto García, Idoia Cearreta, and Nestor Garay. Towards an ontology for describing emotions. In *Emerging Technologies and Information Systems for the Knowledge Society*, pages 96–104. Springer, 2008.

[Liu12] Bing Liu. Sentiment analysis and opinion mining. *Synthesis Lectures on Human Language Technologies*, 5(1):1–167, 2012.

[LK77] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *Biometrics*, pages 159–174, 1977.

[LK06] Gilly Leshed and Joseph 'Jofish' Kaye. Understanding how bloggers feel: Recognizing affect in blog posts. In *Proceedings of the CHI Extended Abstracts on Human Factors in Computing Systems*, pages 1019–1024. ACM, 2006.

[LLS03] Hugo Liu, Henry Lieberman, and Ted Selker. A model of textual affect sensing using real-world knowledge. In *Proceedings of the 8th International Conference on Intelligent User Interfaces (IUI)*, pages 125–132. ACM, 2003.

[LSD13] Amnon Lotan, Asher Stern, and Ido Dagan. TruthTeller: Annotating predicate truth. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 752–757, 2013.

[LSH+06] Madelene Lindström, Anna Ståhl, Kristina Höök, Petra Sundström, Jarmo Laaksolathi, Marco Combetto, Alex Taylor, and Roberto Bresin. Affective diary: Designing for bodily expressiveness and self-reflection. In *Proceedings of CHI Extended Abstracts on Human Factors in Computing Systems*, pages 1037–1042. ACM, 2006.

[LSZ12] Jerry Lin, Marc Spraragen, and Michael Zyda. Computational models of emotion and cognition. *Advances in Cognitive Systems*, 2:59–76, 2012.

[LYC08] Kevin Hsin-Yih Lin, Changhua Yang, and Hsin-Hsi Chen. Emotion classification of online news articles from the reader's perspective. In *Proceedings of the 2008 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology*, volume 1, pages 220–226. IEEE Computer Society, 2008.

[LYTL14] Anthony J T Lee, Fu-Chen Yang, Hsin-Chieh Tsai, and Yi-Yu Lai. Discovering content-based behavioral roles in social networks. *Decision Support Systems*, 59:250–261, 2014.

[Man15] Jason Mander. Internet users have average of 5.54 social media accounts. http://www.globalwebindex.net/blog/internet-users-have-average-of-5-social-media-accounts, January 2015. [Online; accessed 15-June-2016].

[MBSJ09] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 2-Volume 2*, pages 1003–1011. ACL, 2009.

[MdR06] Gilad Mishne and Maarten de Rijke. MoodViews: Tools for blog mood analysis. In *Proceedings of the AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 153–154, 2006.

[Meh96] Albert Mehrabian. Pleasure-Arousal-Dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[Mes12] Lori S Mestre. Student preference for tutorial design: A usability study. *Reference Services Review*, 40(2):258–276, 2012.

[MFL+05] Joseph A Mikels, Barbara L Fredrickson, Gregory R Larkin, Casey M Lindberg, Sam J Maglio, and Patricia A Reuter-Lorenz. Emotional category data on images from the International Affective Picture System. *Behavior Research Methods*, 37(4):626–630, 2005.

[MG09] Stacy C Marsella and Jonathan Gratch. EMA: A process model of appraisal dynamics. *Cognitive Systems Research*, 10(1):70–90, 2009.

[MGP10] Stacy Marsella, Jonathan Gratch, and Paolo Petta. Computational models of emotion. *A Blueprint for Affective Computing: A Sourcebook and Manual*, pages 21–46, 2010.

[Mis05] Gilad Mishne. Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*, volume 19, pages 321–327, 2005.

[MK15] Saif M Mohammad and Svetlana Kiritchenko. Using hashtags to capture fine emotion categories from tweets. *Computational Intelligence*, 31(2):301–326, 2015.

[MKC+13] Andrew Mao, Ece Kamar, Yiling Chen, Eric Horvitz, Megan E Schwamb, Chris J Lintott, and Arfon M Smith. Volunteering versus work for pay: Incentives and tradeoffs in crowdsourcing. In *Proceedings of the First AAAI Conference on Human Computation and Crowdsourcing (HCOMP)*, 2013.

[MKK+12] Daniel McDuff, Amy Karlson, Ashish Kapoor, Asta Roseway, and Mary Czerwinski. AffectAura: An intelligent system for emotional memory. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 849–858. ACM, 2012.

[MKSR11] Kazuyuki Matsumoto, Yusuke Konishi, Hidemichi Sayama, and Fuji Ren. Analysis of Wakamono Kotoba emotion corpus and its application in emotion estimation. *International Journal of Advanced Intelligence*, 3(1):1–24, 2011.

[MM14] Robert R Morris and Daniel McDuff. Crowdsourcing techniques for affective computing. In Rafael A Calvo, Sidney D'Mello, Jonathan Gratch, and Arvid Kappas, editors, *The Oxford Handbook of Affective Computing*, pages 384–394. Oxford University Press, 2014.

[Moh12a] Saif M Mohammad. #Emotional tweets. In *Proceedings of the 1st Joint Conference on Lexical and Computational Semantics (*SEM)*, pages 246–255. ACL, 2012.

[Moh12b] Saif M Mohammad. From once upon a time to happily ever after: Tracking emotions in mail and books. *Decision Support Systems*, 53(4):730–741, 2012.

[Moh16] Saif M Mohammad. Sentiment analysis: Detecting valence, emotions, and other affectual states from text. In Herbert L Meiselman, editor, *Emotion Measurement*, pages 201–237. Elsevier, 2016.

[MP14] Lionel Martin and Pearl Pu. Prediction of helpful reviews using emotions extraction. In *Proceedings of the Twenty-Eighth AAAI Conference on Artificial Intelligence (AAAI)*, pages 1551–1557, 2014.

[MPI05] Chunling Ma, Helmut Prendinger, and Mitsuru Ishizuka. Emotion estimation and reasoning based on affective textual interaction. In *International Conference on Affective Computing and Intelligent Interaction*, pages 622–628. Springer, 2005.

[MSC$^+$13] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*, pages 3111–3119, 2013.

[MT13] Saif M Mohammad and Peter D Turney. Crowdsourcing a word–emotion association lexicon. *Computational Intelligence*, 29(3):436—-465, 2013.

[MTGF12] Claudiu-Cristian Musat Thisone, Alireza Ghasemi, and Boi Faltings. Sentiment analysis using a novel human computation game. In *Proceedings of the 3rd Workshop on the People's Web Meets NLP: Collaboratively Constructed Semantic Resources and their Applications to NLP*, pages 1–9. ACL, 2012.

[MW09] Winter Mason and Duncan J. Watts. Financial incentives and the "Performance of crowds". In *Proceedings of the ACM SIGKDD Workshop on Human Computation (HCOMP)*, pages 77–85. ACM, 2009.

[MZKM15] Saif M Mohammad, Xiaodan Zhu, Svetlana Kiritchenko, and Joel Martin. Sentiment, emotion, purpose, and style in electoral tweets. *Information Processing & Management*, 51(4):480–499, 2015.

[NA13] Alena Neviarouskaya and Masaki Aono. Extracting causes of emotions from text. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pages 932–936, 2013.

[NALF04] Fatma Nasoz, Kaye Alvarez, Christine L Lisetti, and Neal Finkelstein. Emotion recognition from physiological signals using wireless sensors for presence technologies. *Cognition, Technology & Work*, 6(1):4–14, 2004.

[NKR$^+$13] Preslav Nakov, Zornitsa Kozareva, Alan Ritter, Sara Rosenthal, Veselin Stoyanov, and Theresa Wilson. SemEval-2013 Task 2: Sentiment analysis in Twitter. In *Proceedings of Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 312–320. ACL, 2013.

[NLC09] Ramanathan Narayanan, Bing Liu, and Alok Choudhary. Sentiment analysis of conditional sentences. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 1, pages 180–189. ACL, 2009.

[NMTM00] Kamal Nigam, Andrew Kachites McCallum, Sebastian Thrun, and Tom Mitchell. Text classification from labeled and unlabeled documents using EM. *Machine learning*, 39(2-3):103–134, 2000.

[NPI07] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Textual affect sensing for sociable and expressive online communication. In *Affective Computing and Intelligent Interaction, Conference Proceedings (ACII), Volume 4738 of the series Lecture Notes in Computer Science*, pages 218–229. Springer, 2007.

[NPI10a] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. EmoHeart: Conveying emotions in second life based on affect sensing from text. *Advances in Human-Computer Interaction*, 2010 (Special Issue on Emotion-Aware Natural Interaction):Article No. 1, 2010.

[NPI10b] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. User study on AffectIM, an avatar-based instant messaging system employing rule-based affect sensing from text. *International Journal of Human-Computer Studies*, 68(7):432–450, 2010.

[NPI11a] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Affect Analysis Model: Novel rule-based approach to affect sensing from text. *Natural Language Engineering*, 17(1):95–135, 2011.

[NPI11b] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. SentiFul: A lexicon for sentiment analysis. *IEEE Transactions on Affective Computing*, 2(1):22–36, 2011.

[OCC88] Andrew Ortony, Gerald L Clore, and Allan Collins. *The cognitive structure of emotions*. Cambridge University Press, New York, NY, 1988.

[OCF87] Andrew Ortony, Gerald L Clore, and Mark A Foss. The referential structure of the affective lexicon. *Cognitive science*, 11(3):341–364, 1987.

[OMM75] Charles Egerton Osgood, William H May, and Murray S Miron. *Cross-cultural universals of affective meaning*. University of Illinois Press, 1975.

[OSL$^+$11] David Oleson, Alexander Sorokin, Greg Laughlin, Vaughn Hester, John Le, and Lukas Biewald. Programmatic gold: Targeted and scalable quality assurance in crowdsourcing. In *Proceedings of the AAAI Workhop on Human Computation (HCOMP)*, pages 43–48, 2011.

[PAG$^+$14] Soujanya Poria, Basant Agarwal, Alexander Gelbukh, Amir Hussain, and Newton Howard. Dependency-based semantic parsing for concept-level text analysis. In *Computational Linguistics and Intelligent Text Processing, Conference Proceedings (CICLing), Volume 8403 of the series Lecture Notes in Computer Science*, pages 113–127. Springer, 2014.

[Pan82] Jaak Panksepp. Toward a general psychobiological theory of emotions. *Behavioral and Brain sciences*, 5(03):407–422, 1982.

[Pan98] Jaak Panksepp. *Affective neuroscience: The foundations of human and animal emotions*. Oxford university press, 1998.

[PB12] Matthew Purver and Stuart Battersby. Experimenting with distant supervision for emotion classification. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pages 482–491. ACL, 2012.

[Pep14] Meet Pepper, Aldebaran's new personal robot with an "Emotion engine". http://spectrum.ieee.org/automaton/robotics/home-robots/pepper-aldebaran-softbank-personal-robot, 2014. [Online; accessed 15-June-2016].

[PGC$^+$12] Soujanya Poria, Alexander Gelbukh, Erik Cambria, Dipankar Das, and Sivaji Bandyopadhyay. Enriching SenticNet polarity scores through semi-supervised fuzzy clustering. In *Proceedings of the 12th International Conference on Data Mining Workshops (ICDMW), Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE)*, pages 709–716. IEEE, 2012.

[PGC$^+$14] Soujanya Poria, Alexander Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. EmoSenticSpace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems*, 69 (Special Issue on Big Social Data Analysis):108–123, 2014.

[PGH$^+$13] Soujanya Poria, Alexander Gelbukh, Amir Hussain, Dipankar Das, and Sivaji Bandyopadhyay. Enhanced SenticNet with affective labels for concept-based opinion mining. *IEEE Intelligent Systems*, 28(2):31–38, 2013.

[PGS12] René Pfitzner, Antonios Garas, and Frank Schweitzer. Emotional divergence influences information spreading in Twitter. *Proceedings of the Sixth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 543–546, 2012.

[Pic95] Rosalind W Picard. Affective computing. M.I.T Media Laboratory Perceptual Computing Section Technical Report 321, Massachusetts Institute of Technology, 1995.

[PIMK13] Jessica Perrie, Aminul Islam, Evangelos Milios, and Vlado Keselj. Using Google n-grams to expand word-emotion association lexicon. In *Computational Linguistics and Intelligent Text Processing, Conference Proceedings (CICLing), Volume 7817 of the series Lecture Notes in Computer Science*, pages 137–148. Springer, 2013.

[PL08] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[Plu80] Robert Plutchik. A general psychoevolutionary theory of emotion. *Emotion: Theory, research, and experience*, 1(3):3–33, 1980.

[Plu01] Robert Plutchik. The nature of emotions. *American Scientist*, 89(4):344–350, 2001.

[PLV02] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up?: Sentiment classification using machine learning techniques. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, volume 10, pages 79–86. ACL, 2002.

[PP10] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, pages 1320–1326, 2010.

[PPB13] Nikolaos Pappas and Andrei Popescu-Belis. Sentiment analysis of user comments for one-class collaborative filtering over ted talks. In *Proceedings of the 36th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 773–776. ACM, 2013.

[PPB14] Nikolaos Pappas and Andrei Popescu-Belis. Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 455–466. ACL, 2014.

[PPSP$^+$16] Daniel Preotiuc-Pietro, H Andrew Schwartz, Gregory Park, Johannes C Eichstaedt, Margaret Kern, Lyle Ungar, and Elizabeth P Shulman. Modelling valence and arousal in facebook posts. In *Proceedings of Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, 2016.

[Pre04] Dražen Prelec. A Bayesian Truth Serum for Subjective Data. *Science*, 306(5695):462–466, 2004.

[PS10] Lisa Pearl and Mark Steyvers. Identifying emotions, intentions, and attitudes in text using a game with a purpose. In *Proceedings of the NAACL-HLT Workshop on Computational Approaches to Analysis and Generation of Emotion in Text*, pages 71–79. ACL, 2010.

[PZ06] Livia Polanyi and Annie Zaenen. Contextual valence shifters. In James G Shanahan, Yan Qu, and Janyce Wiebe, editors, *Computing attitude and affect in text: Theory and applications*, pages 1–10. Springer, 2006.

[QB11] Alexander J Quinn and Benjamin B Bederson. Human computation: A survey and taxonomy of a growing field. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 1403–1412. ACM, 2011.

[QOC14] Daniele Quercia, Neil Keith O'Hare, and Henriette Cramer. Aesthetic capital: What makes London look beautiful, quiet, and happy? In *Proceedings of the 17th ACM conference on Computer Supported Cooperative Work & Social Computing (CSCW)*, pages 945–955. ACM, 2014.

[QR10] Changqin Quan and Fuji Ren. A blog emotion corpus for emotional expression analysis in Chinese. *Computer Speech & Language*, 24(4):726–749, 2010.

[QZHZ09] Likun Qiu, Weishi Zhang, Changjian Hu, and Kai Zhao. SELC: A self-supervised model for sentiment classification. In *Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM)*, pages 929–936. ACM, 2009.

[RCR$^+$11] Irene Russo, Tommaso Caselli, Francesco Rubino, Ester Boldrini, and Patricio Martínez-Barco. EMOCause: An easy-adaptable approach to emotion cause contexts. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA)*, pages 153–160. ACL, 2011.

[Rei13] Charles M Reigeluth. What is instructional-design theory and how is it changing? In Charles M Reigeluth, editor, *Instructional-design Theories and Models: A New Paradigm of Instructional Theory*, volume 2, pages 5–29. Routledge, 2013.

[RKK+11] Jakob Rogstadius, Vassilis Kostakos, Aniket Kittur, Boris Smus, Jim Laredo, and Maja Vukovic. An assessment of intrinsic and extrinsic motivation on task performance in crowdsourcing markets. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 321–328, 2011.

[RN96] Byron Reeves and Clifford Nass. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places.* Cambridge University Press, 1996.

[RQ12] Fuji Ren and Changqin Quan. Linguistic-based emotion analysis and recognition for measuring consumer satisfaction: An application of affective computing. *Information Technology and Management*, 13(4):321–332, 2012.

[RRJ+12] Kirk Roberts, Michael A Roach, Joseph Johnson, Josh Guthrie, and Sanda M Harabagiu. EmpaTweet: Annotating and detecting emotions on Twitter. In *Proceedings of International Conference on Language Resources and Evaluation (LREC)*, pages 3806–3813, 2012.

[Rum11] Anna Rumshisky. Crowdsourcing word sense definition. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 74–81. ACL, 2011.

[Rus03] James A Russell. Core affect and the psychological construction of emotion. *Psychological review*, 110(1):145–172, 2003.

[SBB+11] Marc Schröder, Paolo Baggia, Felix Burkhardt, Catherine Pelachaud, Christian Peter, and Enrico Zovato. EmotionML–An upcoming standard for representing emotions and related states. In *Affective Computing and Intelligent Interaction, Conference Proceedings (ACII), Volume 6974 of the series Lecture Notes in Computer Science*, pages 316–325. Springer, 2011.

[SBDS14] Marta Sabou, Kalina Bontcheva, Leon Derczynski, and Arno Scharl. Corpus annotation through crowdsourcing: Towards best practice guidelines. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, pages 859–866, 2014.

[Sch84] Klaus R Scherer. On the nature and function of emotion: A component process approach. In Klaus R Scherer and Paul Ekman, editors, *Approaches to Emotion*, pages 293–317. Erlbaum, Hillsdale, NJ, 1984.

[Sch00] Klaus R Scherer. Psychological models of emotion. *The neuropsychology of emotion*, 137(3):137–162, 2000.

[Sch01] Klaus R Scherer. Appraisal considered as a process of multilevel sequential checking. In Klaus R Scherer, Angela Schorr, and Tom Johnstone, editors, *Appraisal processes in emotion: Theory, methods, research*, pages 92–120. Oxford University Press, 2001.

[Sch05] Klaus R Scherer. What are emotions? And how can they be measured? *Social science information*, 44(4):695–729, 2005.

[SCS+13] Mohammad Soleymani, Micheal N Caro, Erik M Schmidt, Cheng-Ya Sha, and Yi-Hsuan Yang. 1000 songs for emotional analysis of music. In *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, pages 1–6. ACM, 2013.

[SDB+15] Chad A Steed, Margaret Drouhard, Justin Beaver, Joshua Pyle, and Paul L Bogen. Matisse: A visual analytics system for exploring emotion trends in social media text streams. In *Proceedings of IEEE International Conference on Big Data*, pages 807–814. IEEE, 2015.

[SDFF12] Hao Su, Jia Deng, and Li Fei-Fei. Crowdsourcing annotations for visual object detection. In *Proceedings of the AAAI Workhop on Human Computation (HCOMP)*, pages 40–46, 2012.

[SDSO68] Philip J Stone, Dexter C Dunphy, Marshall S Smith, and Daniel M Ogilvie. The General Inquirer: A computer approach to content analysis. *Journal of Regional Science*, 8(1):113–116, 1968.

[SEHHE14] Shadi Shaheen, Wassim El-Hajj, Hazem Hajj, and Shady Elbassuoni. Emotion recognition from text based on automatically generated rules. In *Proceedings of the 14th International Conference on Data Mining Workshops (ICDMW), Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE)*, pages 383–392. IEEE, 2014.

[SEK$^+$13] Hansen Andrew Schwartz, Johannes C Eichstaedt, Margaret L Kern, Lukasz Dziurzynski, Richard E Lucas, Megha Agrawal, Gregory J Park, Shrinidhi K Lakshmikanth, Sneha Jha, Martin EP Seligman, and Lyle Ungar. Characterizing geographic variation in well-being using tweets. In *Proceedings of the Seventh International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 583–591, 2013.

[Set09] Burr Settles. Active learning literature survey. Computer Sciences Technical Report 1648, University of Wisconsin–Madison, 2009.

[Set11] Burr Settles. Closing the loop: Fast, interactive semi-supervised annotation with queries on features and instances. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1467–1478. ACL, 2011.

[SG14] Jacopo Staiano and Marco Guerini. DepecheMood: A lexicon for emotion analysis from crowd-annotated news. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 427–433, 2014.

[SH01] Pero Subasic and Alison Huettner. Affect analysis of text using fuzzy semantic typing. *IEEE Transactions on Fuzzy Systems*, 9(4):483–496, 2001.

[SHC11] Aaron D Shaw, John J Horton, and Daniel L Chen. Designing incentives for inexpert human raters. In *Proceedings of the ACM 2011 Conference on Computer Supported Cooperative Work (CSCW)*, pages 275–284. ACM, 2011.

[SI13] Jared Suttles and Nancy Ide. Distant supervision for emotion classification with discrete binary values. In *Computational Linguistics and Intelligent Text Processing, Conference Proceedings (CICLing), Volume 7817 of the series Lecture Notes in Computer Science*, pages 121–136. Springer, 2013.

[SKCL09] Man-Kwan Shan, Fang-Fei Kuo, Meng-Fen Chiang, and Suh-Yin Lee. Emotion-based music recommendation by affinity discovery from film music. *Expert systems with applications*, 36(4):7666–7674, 2009.

[SL09] Marina Sokolova and Guy Lapalme. A systematic analysis of performance measures for classification tasks. *Information Processing & Management*, 45(4):427–437, 2009.

[SM08] Carlo Strapparava and Rada Mihalcea. Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied Computing*, pages 1556–1560. ACM Press, 2008.

[SM15] Aliaksei Severyn and Alessandro Moschitti. Twitter sentiment analysis with deep convolutional neural networks. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR)*, pages 959–962. ACM, 2015.

[SMP14] Valentina Sintsova, Claudiu Musat, and Pearl Pu. Semi-supervised method for multi-category emotion recognition in tweets. In *Proceedings of the 14th International Conference on Data Mining Workshops (ICDMW), Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction (SENTIRE)*, pages 393–402. IEEE, 2014.

[SOJN08] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y Ng. Cheap and fast—but is it good?: Evaluating non-expert annotations for natural language tasks. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 254–263. ACL, 2008.

[SPH$^+$11] Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 151–161. ACL, 2011.

[SPI09] Mostafa Al Masum Shaikh, Helmut Prendinger, and Mitsuru Ishizuka. A linguistic interpretation of the OCC emotion model for affect sensing from text. In Jianhua Tao and Tieniu Tan, editors, *Affective Information Processing*, pages 45–73. Springer, 2009.

[SPL06] Marc Schröder, Hannes Pirker, and Myriam Lamolle. First suggestions for an Emotion Annotation and Representation Language. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, volume 6, pages 88–92, 2006.

[SPP12] Mohammad Soleymani, Maja Pantic, and Thierry Pun. Multimodal emotion recognition in response to videos. *IEEE Transactions on Affective Computing*, 3(2):211–223, 2012.

[SPW$^+$13] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1631–1642. ACL, 2013.

[SRI16] J Fernando Sánchez-Rada and Carlos A Iglesias. Onyx: A linked data approach to emotion representation. *Information Processing & Management*, 52(1):99–114, 2016.

[SSKO87] Phillip Shaver, Judith Schwartz, Donald Kirson, and Cary O'connor. Emotion knowledge: Further exploration of a prototype approach. *Journal of Personality and Social Psychology*, 52(6):1061–1086, 1987.

[SSS12] Vera Sacharin, Katja Schlegel, and Klaus R Scherer. Geneva Emotion Wheel rating study. Unpublished report, University of Geneva, Swiss Center for Affective Sciences, 2012.

[STCD12] John Snel, Alexey Tarasov, Charlie Cullen, and Sarah Jane Delany. A crowdsourcing approach to labelling a mood induced speech corpus. In *Proceedings of the 4th International Workshop on Corpora for Research on Emotion Sentiment and Social Signals (ES3), Satellite of LREC*. ELRA, 2012.

[SV04] Carlo Strapparava and Alessandro Valitutti. WordNet-Affect: An affective extension of WordNet. In *Proceedings of Language Resources and Evaluation Conference (LREC)*, volume 4, pages 1083–1086, 2004.

[SW94] Klaus R Scherer and Harald G Wallbott. Evidence for universality and cultural variation of differential emotion response patterning. *Journal of Personality and Social Psychology*, 66(2):310–328, 1994.

[TBK10] Marko Tkalčič, Urban Burnik, and Andrej Košir. Using affective parameters in a content-based recommender system for images. *User Modeling and User-Adapted Interaction*, 20(4):279–311, 2010.

[TBP$^+$10] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *Journal of the American Society for Information Science and Technology*, 61(12):2544–2558, 2010.

[TBP11] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment in Twitter events. *Journal of the American Society for Information Science and Technology*, 62(2):406–418, 2011.

[TBP12] Mike Thelwall, Kevan Buckley, and Georgios Paltoglou. Sentiment strength detection for the social web. *Journal of the American Society for Information Science and Technology*, 63(1):163–173, 2012.

[TBT$^+$11] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[TIM08] Ryoko Tokuhisa, Kentaro Inui, and Yuji Matsumoto. Emotion classification using massive examples extracted from the web. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)*, volume 1, pages 881–888. ACL, 2008.

[TK07] Grigorios Tsoumakas and Ioannis Katakis. Multi-label classification: An overview. *International Journal of Data Warehousing and Mining (IJDWM)*, 3(3):1–13, 2007.

[TKMS03] Kristina Toutanova, Dan Klein, Christopher D Manning, and Yoram Singer. Feature-rich part-of-speech tagging with a cyclic dependency network. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology–Volume 1 (NAACL-HLT)*, pages 173–180. ACL, 2003.

[TL03] Peter D Turney and Michael L Littman. Measuring praise and criticism: Inference of semantic orientation from association. *ACM Transactions on Information Systems (TOIS)*, 21(4):315–346, 2003.

[TM08] Ivan Titov and Ryan T McDonald. A joint model of text and aspect ratings for sentiment summarization. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics (ACL) : Human Language Technologies (HLT)*, pages 308–316. ACL, 2008.

[TMM13] Gonçalo Tavares, André Mourão, and João Magalhaes. Crowdsourcing for affective-interaction in computer games. In *Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia*, pages 7–12. ACM, 2013.

[TP10] Yla R Tausczik and James W Pennebaker. The psychological meaning of words: LIWC and computerized text analysis methods. *Journal of Language and Social Psychology*, 29(1):24–54, 2010.

[TSSW10] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media (ICWSM)*, pages 178–185, 2010.

[Tur02] Peter D Turney. Thumbs up or thumbs down?: Semantic orientation applied to unsupervised classification of reviews. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics (ACL)*, pages 417–424. ACL, 2002.

[UH15] Orizu Udochukwu and Yulan He. A rule-based approach to implicit emotion detection in text. In *Natural Language Processing and Information Systems, Conference Proceedings (NLDB), Volume 9103 of the series Lecture Notes in Computer Science*, pages 197–203. Springer, 2015.

[VAD04] Luis Von Ahn and Laura Dabbish. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 319–326. ACM, 2004.

[VAD08] Luis Von Ahn and Laura Dabbish. Designing games with a purpose. *Communications of the ACM*, 51(8):58–67, 2008.

[VAKB06] Luis Von Ahn, Mihir Kedia, and Manuel Blum. Verbosity: A game for collecting common-sense facts. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 75–78. ACM, 2006.

[VD11] Frederik Vaassen and Walter Daelemans. Automatic emotion classification for interpersonal communication. In *Proceedings of the 2nd Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis (WASSA)*, pages 104–110. ACL, 2011.

[VdVE11] Jeroen Vuurens, Arjen P de Vries, and Carsten Eickhoff. How much spam can you take? An analysis of crowdsourcing results to increase accuracy. In *Proceedings of the ACM SIGIR Workshop on Crowdsourcing for Information Retrieval (CIR)*, pages 21–26, 2011.

[Wal98] Harald G Wallbott. Bodily expression of emotion. *European Journal of Social Psychology*, 28(6):879–896, 1998.

[WC99] David Watson and Lee Anna Clark. The PANAS-X: Manual for the positive and negative affect schedule – Expanded form. Unpublished manuscript, University of Iowa, Department of Psychology, 1999.

[WCL06] Chung-Hsien Wu, Ze-Jing Chuang, and Yu-Chung Lin. Emotion recognition from text using semantic labels and separable mixture models. *ACM transactions on Asian language information processing (TALIP)*, 5(2):165–183, 2006.

[WCTS12] Wenbo Wang, Lu Chen, Krishnaprasad Thirunarayan, and Amit P Sheth. Harnessing Twitter "Big data" for automatic emotion identification. In *Proceedings of the 2012 International Conference on Privacy, Security, Risk and Trust (PASSAT) and 2012 International Confernece on Social Computing (SocialCom)*, pages 587–592. IEEE, 2012.

[WHA12] Wesley Willett, Jeffrey Heer, and Maneesh Agrawala. Strategies for crowdsourcing social data analysis. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI)*, pages 227–236. ACM, 2012.

[WK09] Michael Wiegand and Dietrich Klakow. Predictive features in semi-supervised learning for polarity classification and the role of adjectives. In *Proceedings of the Nordic Conference on Computational Linguistics (NoDaLiDa)*, pages 198–205, 2009.

[WKB13] Amy Beth Warriner, Victor Kuperman, and Marc Brysbaert. Norms of valence, arousal, and dominance for 13,915 English lemmas. *Behavior research methods*, 45(4):1191–1207, 2013.

[WP15] Yichen Wang and Aditya Pal. Detecting emotions in social media: A constrained optimization approach. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*, pages 996–1002. AAAI Press, 2015.

[WR05] Janyce Wiebe and Ellen Riloff. Creating subjective and objective sentence classifiers from unannotated texts. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 486–497. Springer, 2005.

[WWB$^+$04] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Computational linguistics*, 30(3):277–308, 2004.

[WWH05] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing (HLT)*, pages 347–354. ACL, 2005.

[WWH09] Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. Recognizing contextual polarity: An exploration of features for phrase-level sentiment analysis. *Computational linguistics*, 35(3):399–433, 2009.

[Yeh00] Alexander Yeh. More accurate tests for the statistical significance of result differences. In *Proceedings of the 18th Conference on Computational Linguistics (COLING)*, volume 2, pages 947–953. ACL, 2000.

[YLC07] Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. Building emotion lexicon from weblog corpora. In *Proceedings of the 45th Annual Meeting of the ACL on Interactive Poster and Demonstration Sessions*, pages 133–136. ACL, 2007.

[ZDWX12] Jichang Zhao, Li Dong, Junjie Wu, and Ke Xu. MoodLens: An emoticon-based sentiment analysis system for Chinese tweets. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1528–1531. ACM, 2012.

[ZGMK14] Xiaodan Zhu, Hongyu Guo, Saif Mohammad, and Svetlana Kiritchenko. An empirical study on the effect of negation words on sentiment. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 304–313. ACL, 2014.

[Zhu05] Xiaojin Zhu. Semi-supervised learning literature survey. Computer Sciences Technical Report 1530, University of Wisconsin–Madison, 2005.

VALENTINA SINTSOVA

46, av. Victor-Ruffy, Lausanne
1012, Switzerland

Phone: +41 (786) 720320
E-mail: valentinasintsova@gmail.com

EDUCATION

Ph.D. in Computer Science **Sep 2011 – Oct 2016**
École Polytechnique Fédérale de Lausanne (EPFL), Switzerland
Advisors: Dr. Pearl Pu, the head of HCI research group at EPFL,
and Prof. Boi Faltings, the head of LIA lab at EPFL
Thesis: Advancing Fine-Grained Emotion Recognition in Short Text

M.S. in Applied Mathematics and Physics **Sep 2009 – Jun 2011**
Moscow Institute of Physics and Technology (MIPT), Russia
Department of Control and Applied Mathematics
Specialization in Mathematics and Information Technologies
GPA: 5.0 out of 5.0
Advisor: Rudakov K.V., the head of Intelligent Systems chair in MIPT
Thesis: The Description of the Function Sets with Finite Relations on Domain and Range

B.S. in Applied Mathematics and Physics, summa cum laude **Sep 2005 – Jun 2009**
MIPT, Department of Control and Applied Mathematics
GPA: 5.0 out of 5.0

WORK EXPERIENCE

EPFL, Lausanne, Switzerland **Sep 2011 – Aug 2016**
Position: Ph.D. candidate, research assistant
Worked on developing fine-grained emotion recognition systems in tweets. Designed human computation task for building emotion lexicons, developed and tested distant learning framework for building emotion classifiers, developed a method for analyzing the effects of modifiers on emotional statements.

GOOGLE Switzerland GmbH., Zurich, Switzerland **July 2015 – Oct 2015**
Position: Software Engineer Intern
Analyzed the performance of new features for predicting how well advertiser keywords match user queries. Wrote large-scale data processing pipelines for feature extraction and statistical analysis. Mostly used C++.

DOMASHNIE DENGI (micro-credit company), Moscow, Russia **Feb 2011 – Jun 2011**
Position: Risk management specialist
Constructed and analyzed the financial weekly and monthly reports; participated in the development of the algorithm for forecasting the amount of financial reserves for the full company for the year.

FORECSYS (forecasting and recognition systems), Moscow, Russia **Aug 2009 – Sep 2010**
Position: Junior developer
Was the front-end developer for the “Poligon” project (poligon.machinelearning.ru) – the site for testing machine-learning algorithms on different classification problems. Mostly used C# and ASP.NET.

EXTRACURRICULAR COURSES

Yandex School of Data Analysis, Moscow, Russia **Sep 2010 – May 2011**
(taken 1 year out of 2 due to moving to Switzerland)
Courses: Algorithms and data structures; Stochastic models; Recovery of functional laws from empirical evidence

PUBLICATIONS

Sintsova V., Pu P. **“Dystemo: Distant Supervision Method for Multi-Category Emotion Recognition in Tweets.”** ACM Transactions on Intelligent Systems and Technology (TIST), 8(1):Article No.13, 2016

Sintsova V., Musat C., and Pu P. **“Semi-Supervised Method for Multi-Category Emotion Recognition in Tweets.”** In Proceedings of the 4th Workshop on Sentiment Elicitation from Natural Text for Information Retrieval and Extraction, 2014

Martin L., Sintsova V., and Pu P. **“Are Influential Writers More Objective? An Analysis of Emotionality in Review Comments.”** In the 5th International Workshop on Social Recommender Systems, in conjunction with 23rd International World Wide Web Conference, 2014

Kempter R., Sintsova V., Musat C., and Pu P. **“EmotionWatch: Visualizing Fine-Grained Emotions in Event-Related Tweets.”** In Proceedings of the 8th International AAAI Conference on Weblogs and Social Media, 2014

Sintsova V., Musat C., and Pu P. **“Fine-Grained Emotion Recognition in Olympic Tweets Based on Human Computation.”** In Proceedings of the 4th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, 2013

Sintsova V., Pu P. **“Importance of Emotion Awareness for Emotional Well-Being and for Improvement of Social Communications”** Position paper for the ARV Workshop on Tools and Technologies for Emotion Awareness in Computer-Mediated Collaboration and Learning, 2013

Lisitsa A. V., Vorontsov K. V., Ivahnenko A. A., Inyakin A. S., Sintsova V. V. **“Online services for testing machine learning algorithms: MLComp, TunedIt, and Poligon.”** In Proceedings of the International conference “Intellectualization of information processing”, 2010 (in Russian)

GRANTS AND HONOURS

Best application award in AI Video Competition, IJCAI 2013:
Kempter R., Sintsova V., Musat C., and Pu P. “Discover Emotions in Tweets & Weibos during the 2012 Olympic Games,” 2013

Medal for Outstanding Achievements in studies, 2009
Scholarship of Abramov and Frolov Charitable Foundation, 2007-2009
“Gazprombank” scholarship for excellent results in studies, 2006

EPFL PROJECTS

Emotions in the tweets about Olympic Games **Sep 2012 – Sep 2013**

(research project)

Developed the system for the fine-grained emotion recognition from the tweets. To acquire the annotation of words with emotions, designed and launched a human computation task on Amazon Mechanical Turk. It involved the design of the tutorial and the posterior quality control measure. Also participated in designing the visualization system for detected emotions and in evaluating it with users. Mostly used Java and PHP.

Porting the Sentiment Analysis into another language **Dec 2012**
(course work, group project of 2 PhD students)

Analyzed different methods to transfer the sentiment analysis systems existing in English into Russian language, including machine translation of full texts and the lexicons. Mostly written in Java; adapted data crawler for movie reviews in Python.

Collective Classification in Sentiment Analysis **Dec 2011**
(course work, individual project)

The application of Iterative Classification Algorithm to the settings of sentiment analysis in the reviews, where the graph structure was extracted using different text similarities. Written in C#.

RESEARCH AND WORK INTERESTS

Text mining, Emotion Recognition, Social Media Analysis, Sentiment Analysis, Machine Learning, Data Mining, Human-Computer Interaction, Crowdsourcing

LANGUAGE SKILLS

Russian (native), English (advanced), French (intermediate)

OUTSIDE INTERESTS AND ACTIVITIES

Swimming, dances, board games, musicals, snowboarding, travelling