# *Pushing the Envelope* of Sentiment Analysis *Beyond Words* and Polarities

**A thesis submitted in partial fulfilment**

**of the requirement for the degree of Doctor of Philosophy**

**Lowri A. Williams**

**2017**

**Cardiff University**
**School of Computer Science & Informatics**

## Declaration

This work has not been submitted in substance for any other degree or award at this or any other university or place of learning, nor is being submitted concurrently in candidature for any degree or other award.

Signed ..... (candidate) Date .....

## Statement 1

This thesis is being submitted in partial fulfilment of the requirements for the degree of PhD.

Signed ..... (candidate) Date .....

## Statement 2

This thesis is the result of my own independent work/investigation, except where otherwise stated, and the thesis has not been edited by a third party beyond what is permitted by Cardiff University's Policy on the Use of Third Party Editors by Research Degree Students. Other sources are acknowledged by explicit references. The views expressed are my own.

Signed ..... (candidate) Date .....

## Statement 3

I hereby give consent for my thesis, if accepted, to be available online in the University's Open Access repository and for inter-library loan, and for the title and summary to be made available to outside organisations.

Signed ..... (candidate) Date .....

# Acknowledgements

I would like to sincerely thank my academic supervisor, Professor Irena Spasić, for her support, guidance, and vision which created the opportunity for this research, in which she shared her expertise, provided constant clarification and perspectives for new research pathways, as well as challenging my capabilities and teaching me invaluable professional and research skills. My utmost gratitude is extended to her for her continual encouragement and patience, and for keeping me motivated and on track at all stages of my studies at Cardiff University, particularly during my PhD.

My time as a research student would not have been the same without the support from the School of Computer Science & Informatics. My colleagues provided feedback, discussions, recommendations, company, and friendship, which made researching my PhD both interesting and enjoyable. I'd like to particularly thank Liam, Max, Walter, Will, and Chris.

Thank you to Amy, who, ever critically, helped proof read this thesis and provided invaluable constructive feedback, as well as her personal support.

The continuous encouragement from my family and friends has been invaluable. I owe the biggest acknowledgement to my parents, my brother, Huw, and my sister in law, Annwyl, who have provided endless love and support. I'd also like to thank my friends, in particular Rachel, Hannah, Phyl, and other members of the 'Crewage', for their support throughout my time at university. Thank you to Irene, Amir, Wafi, and Catherine for both their academic support and friendship, and to Pete, Emily, Siwan,

and Kitty for keeping me motivated with their humour. Lastly, I'd like to thank my friends from Cardiff & Met Ladies Hockey Club for their support.

# Abstract

Idioms are multi-word expressions which hold a literal and figurative meaning which is conventionally understood by native speakers. Their overall meaning, often, cannot be deduced from the literal meaning of their constituent words. Sentiment analysis, also referred to as opinion mining, aims to automatically extract and classify sentiments, opinions, and emotions expressed in text. The research in this thesis is motivated by the fact that idioms, which often express an affective stance towards an entity or an event, are not featured systematically in sentiment analysis. To estimate the degree to which the inclusion of idioms as features may improve the results of traditional sentiment analysis, we compared our results to two state-of-the-art sentiment analysis approaches. Firstly, we collected a set of idioms that are relevant to sentiment analysis, i.e. those that can be mapped to an emotion. These mappings were obtained using a crowdsourcing approach. Secondly, to evaluate the results of sentiment analysis, we assembled a corpus of sentences in which idioms are used in context. Each sentence was annotated with an emotion, which formed the basis for the gold standard used for the comparison against the baseline methods. The classification performance was improved by almost 20 percentage points.

Given the positive findings from our initial experiments, the main limitation was the significant knowledge-engineering overhead involved in hand-crafting lexico-semantic resources used to support idiom-based features. To minimise the bottleneck associated with the acquisition of such resources, we scaled up our original approach by automating their engineering. Subsequently, these resources were used to replace the manually

engineered counterparts of such features in the originally proposed method. The fully automated approach outperformed the two baseline methods by 7 and 9 percentage points. These improvements, however, were poorer in comparison to those achieved in the initial study. Nevertheless, we have demonstrated, not only can idiom-based features be automatically engineered, but they too, improve sentiment classification results, when such features are present.

Taking a long-term view of the research in this thesis, we want to address the limitations of state-of-the-art sentiment analysis approaches by focusing on a full range of emotions, rather than sentiment polarity. However, there is no consensus among researchers on a standardised framework for classifying emotions. Proposing such a framework would be a major contribution to the field of sentiment analysis, as it would stimulate its evolution into fully-fledged emotion classification and allow for systematic comparison of independent studies. With this goal in mind, we investigated the utility of different classification frameworks for sentiment analysis. A comprehensive statistical analysis of our experimental results provided explicit evidence that, in relative terms, six basic emotions are best suited for sentiment analysis. However, we identified the major shortcoming of oversimplifying positive emotions.

# Contents

# List of Publications

Some of the work introduced in this thesis is based on the following publications:

[226] Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece & Irena Spasić. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 2015.

[192] Irena Spasić, Lowri Williams & Andreas Buerki. Scaling up the extraction of idiom-based features in sentiment analysis. Submitted to IEEE, Transactions on Affective Computing.

[225] Lowri Williams, Michael Arribas-Ayllon, Andreas Artemiou & Irena Spasić. Comparing the utility of different emotion classification schemes for emotive language analysis. Submitted to Springer, Journal of Classification.

# List of Figures

# List of Tables

# List of Acronyms

**BOW** Bag-of-Words

**BNC** British National Corpus

**EARL** Emotion Annotation and Representation Language

**FN** False Negative

**FP** False Positive

**IAA** Inter-Annotator Agreement

**MWE** Multi-Word Expressions

**NLP** Natural Language Processing

**POS** Part-of-Speech

**SVM** Support Vector Machines

**WNA** WordNet-Affect

**TN** True Negative

**TP** True Positive

# *Chapter 1*

# Introduction

*"Curiosity killed the cat."*

The proliferation of textual user-generated content (e.g. product reviews) on the Web 2.0 provides opportunities for a range of practical applications that require opinions (e.g. market research) as an alternative, or a supplement, to more traditional qualitative research methods, such as surveys, interviews, and focus groups. However, the sheer scale of text data acquired from the web poses challenges to qualitative analysis. Text mining has emerged as a potential solution to the problem of information overload associated with reading vast amounts of text originating from diverse sources. In particular, sentiment analysis, also referred to as opinion mining, aims to automatically extract and classify sentiments, opinions, and emotions expressed in text.

Most research in this domain considers sentiment analysis as a classification problem, in which sentiment bearing text segments (e.g. phrases, sentences or paragraphs) are classified in terms of their polarity (positive, negative or neutral). Some domains, however, require further differentiation between these classes, and associate text segments with specific emotions (e.g. happiness, sadness, anger, etc.).

The written expression of sentiment relies on words and the creative use of language which may infer or increase its salience. Strapparava & Mihalcea [195] note, that in discourse, each lexical unit, whether it be a word or a phrase, has the ability to contribute useful information regarding the sentiment that is being expressed. Features used

to support sentiment analysis include words, their Part-of-Speech (POS), syntactic dependencies, and negation. Most commonly, sentiment bearing words, or opinionated words that convey a subjective bias, are identified by using specialised lexicons (e.g. WordNet-Affect (WNA) [196]), and used to classify sentiment. An alternative approach to classifying sentiment, is the dynamic calculation of a word's semantic orientation based on its statistical association with a set of positive and negative paradigm words [211].

Other, more complex, features have included linguistic models based on lexical substitution, n-grams, and phrases. Using an n-gram graph based method to assign sentiment polarity to individual word senses, experiments have implied that figurative language (i.e. the language which digresses from literal meanings) not only conveys sentiment, but drives the polarity of a sentence [170].

## 1.1 Motivation

Although the value of phrase-level and figurative language features have been acknowledged in sentiment analysis, few approaches have extensively explored idioms as features of this kind.

Idioms (e.g. *happy as Larry, over the moon, raining cats and dogs*, etc.) are multiword expressions, or phrases, which hold a figurative or literal meaning, often both, which is conventionally understood by native speakers. Their overall meaning, often, cannot be deduced from the literal meaning of their constituent words.

The research in this thesis is motivated by two matters of principle. Firstly, idioms often imply an affective stance towards an entity or an event [24] (e.g. *have kittens* is used to refer to an extreme worry, or fear). This implies that an idiom itself may be sufficient in determining the underlying sentiment of its context. In fact, the error analysis of sentiment classification results often reveals that the largest percentage of

errors are neutral classifications when idioms are explicitly used to express sentiment [13], as idioms hold their meaning as a single semantic unit [20, 25, 56].

Whilst idioms have been extensively studied across many disciplines, such as linguistics and psychology, they are under-represented in sentiment analysis, primarily because there is no comprehensive knowledge base that systematically maps them to their sentiment. Moreover, to be included as features in sentiment analysis, idioms need to be recognised in text. Whereas frozen idioms can be recognised using a simple lexicon approach, other idioms are known to be syntactically flexible, often containing inflection and nominalization, which make their occurrences more difficult to automatically identify in text.

Taking a long-term view of the research in this thesis, we want to address the limitations of state-of-the-art sentiment analysis approaches by focusing on a full range of emotions, rather than sentiment polarity. However, there is no consensus among researchers on a standardised framework for measuring, conceptually organising, representing, and classifying emotions. Proposing such a framework would be a major contribution to the field of sentiment analysis, as it would stimulate its evolution into fully-fledged emotion classification and allow for systematic comparison of independent studies.

There is a large number of overlapping natural language expressions that can be used to describe, express, or refer to emotions [187, 27]. Therefore, emotion classification requires a standardised framework to represent emotion [117, 181]. Multiple classification frameworks have been proposed, with the main tension in the literature being whether emotions can be defined as discrete, universal categories of basic emotions (e.g. anger, sadness, happiness), whether they are characterised by one or more dimensions incorporating aspects of valence, arousal or intensity, or whether they are organised hierarchically. No study has investigated the utility of existing classification frameworks in the context of sentiment analysis. The absence of such studies hinders research in sentiment analysis from progressing into fine-grained approaches to emo-

tion classification not least due to an inability to establish measurable benchmarks for state-of-the-art.

## 1.2 Research Hypotheses and Contributions

In this Section, we outline the hypotheses and the main research contributions presented in this thesis. We believe that three significant contributions have been made, in which our experimental results serve to support our hypotheses.

The research in this thesis is motivated by the under-representation of idioms in sentiment analysis. We hypothesise that the inclusion of idiom-based features will reduce misclassification of sentiment when such features are present. Our primary contribution is supported by our experimental results, which **demonstrate the value of idioms as features of sentiment analysis, by showing that the presence of idiom-based features in traditional sentiment analysis approaches significantly improves sentiment classification results**.

In addition to the experimental results gained from this study, the assembled lexico-semantic resources can support further research into the subject. They include a comprehensive collection of 580 idioms manually annotated with sentiment polarity, which, to our knowledge, represents the largest lexico-semantic resource of this kind to be utilised in sentiment analysis. We implemented a set of local grammars that can be used to recognise occurrences of these idioms in text. Additionally, we assembled a corpus of 2,521 sentences with a wide range of idioms in context. Similarly to the idioms themselves, this corpus was also annotated with sentiment polarity, which can be used in systematic evaluations of sentiment analysis approaches that claim to use idioms as features.

The main limitation of the proposed approach is the significant knowledge-engineering overhead involved in hand-crafting local grammars for the recognition of idioms in text, as well as the manual effort associated with acquiring the sentiment polarity of

idioms. This means that, although our collection of idioms is comprehensive, it is still not representative of all relevant idioms. Having already provided evidence of the significance of idioms in sentiment analysis and the viability of their use as features, we wanted to generalise the original approach without resorting to further manual knowledge engineering. To minimise the bottleneck associated with the acquisition of lexico-semantic resources, we scaled up our original approach, by automating their engineering. In this case, we hypothesise that idiom-based features can be automatically engineered and used to support sentiment analysis. More specifically, we hypothesise that the canonical, or dictionary form of an idiom, can be used to automatically derive the variations of their occurrences in text. Additionally, we hypothesise that it is possible to automatically extract sentiment polarity of an idiom from its dictionary definition. This hypothesis is supported by experimental results, which **demonstrate that automatically engineered idiom-based features improve the results of sentiment analysis, when such features are present**. This approach allows existing idiom dictionaries aimed at English language learners, to be re-purposed for sentiment analysis.

One of our research interests is to further advance the field of sentiment analysis by expanding it beyond mere sentiment classification. However, we identified a lack of consensus among researchers on a standardised framework of emotion. Our research contribution here is **an investigation into the utility of emotion classification frameworks for sentiment analysis**. Our goal is to identify an appropriate classification framework in terms of completeness and complexity. We therefore investigated the utility of such frameworks, by exploring their ease of use by human annotators, as well as the performance of supervised machine learning algorithms when they were used to annotate training data.

To quantitatively measure their utility from a human annotator perspective, we measured Inter-Annotator Agreement (IAA). Assuming that classification frameworks with a better balance between completeness and complexity are easier to interpret and use,

we hypothesise that, when a correct class is available, unambiguous and readily identifiable, the likelihood of independent annotators selecting that particular class increases, thus leading to higher IAA. Higher IAA, however, does not necessarily mean that a framework has sufficient coverage of the emotion space. To complement our findings, and to illustrate the difference in the coverage of different frameworks, we performed correspondence analysis of annotations across each framework. Additionally, we conducted thematic analysis on semi-structured interviews, to gain a qualitative insight into how annotators interact with and interpret each framework.

To explore how well text classification algorithms can learn to differentiate between the classes within a given framework, we evaluated their performance when the corresponding annotations are used to train the classification model. Our hypothesis here is a general one; if the class distribution is imbalanced, and the degree of overlap among the classes is high, then classification performance is negatively affected [161]. We further explore the classification performance across each framework by analysing the confusion matrices, which show how the automatically predicted classes compare against the actual classes from the gold standard.

Our findings from this investigation demonstrate that six basic emotions [51] were ranked the highest in both criteria, and are therefore best suited for sentiment analysis. However, both quantitative and qualitative analysis highlighted the major shortcoming of oversimplifying positive emotions, which are all conflated into a single category, happiness. This work supports further investigation into ways of extending basic emotions to encompass a variety of positive ones. As with our primary contribution, the datasets from this investigation provide resources that can support further research into fine-grained approaches to emotion classification and sentiment analysis. This includes an assembled corpus of 500 emotionally charged text documents that have been annotated with emotions from six comprehensive classification frameworks.

## 1.3 Thesis Structure

The remaining Chapters are organised as follows:

- **Chapter 2** - *Background* - Introduces sentiment analysis, as well as its comprising topics, and the key concepts related to this research.
- **Chapter 3** - *Idioms as Features in Sentiment Analysis* - Investigates how the inclusion of idioms as features impact the results of traditional automated sentiment analysis approaches.
- **Chapter 4** - *Scaling Up the Extraction of Idiom-Based Features* - Scales up our approach in Chapter 3, to automate the engineering of lexico-semantic resources that support the use of idioms in sentiment analysis.
- **Chapter 5** - *Comparison of Emotion Classification Frameworks for Sentiment Analysis* - Investigates the utility of different classifications frameworks for sentiment analysis.
- **Chapter 6** - *Conclusion* - Concludes the thesis by summarising our contributions and findings, as well as highlighting proposals of future work.

# Review Example

We use the following headphone review to illustrate examples throughout Chapter 2. A number is associated with each sentence for reference.

> "(1) These headphones sound great with my iPod and my MacBook, and have superb clarity, treble and bass. (2) I recently bought the Sony MDR-V300 headphones, and these are much better! (3) The sound is much clearer, compared to the Sony which has too much bass in my opinion. (4) However, they are much more uncomfortable to wear for extended periods of time than the Sony." [5]

# *Chapter 2*

# Background

*"Scratching the surface."*

There exists substantial research on the subject of sentiment analysis, which, due to the proliferation of the Web 2.0 and its surrounding applications, has received considerable attention over the last decade. Sentiment analysis aims to automatically extract and classify sentiments, opinions, and emotions expressed in text [109, 134]. The field of sentiment analysis, which encompasses Natural Language Processing (NLP) and text mining - the computational task of analysing and processing natural language, has become a very active research area, as it coincides with the rapid growth and the availability of informal and opinionated textual user-generated content on the web.

This Chapter provides an overview of sentiment analysis. More specifically, it is divided into the following main sections: Section 2.1 defines the aims and objective of sentiment analysis, as well as providing context for its applications. Section 2.2 discusses how the related notions of sentiment and emotions are represented in this domain. Section 2.3 reviews the features used to identify sentiment in text, as well as the means of extracting these features. Specifically, we discuss idioms, and their role in sentiment analysis. Section 2.4 reviews both state-of-the-art machine learning approaches to sentiment analysis. Finally, Section 2.5 summarises the main topics discussed in this Chapter.

## 2.1 Sentiment Analysis

The proliferation of textual user-generated content on the web (e.g. product reviews, social media, blog posts, etc.) presents us with a wealth of information about people's opinions, which can influence some of our own decisions (e.g. "Should I buy this product if this review says that it broke soon after another buyer bought theirs?"). However, the sheer scale of text data available on the web impairs our ability to identify the most relevant information in real time.

Text mining has emerged as a potential solution to the problem of information overload that is associated with reading vast amounts of text originating from diverse sources. In particular, sentiment analysis, often referred to as opinion mining, aims to automatically identify, extract, summarise, and/or categorise opinions, evaluations, appraisals, attitudes, and emotions expressed within text [109, 134].

### 2.1.1 User-Generated Content

Opinions are central to almost all human activity, as they may influence the behaviour of others. If we need to make a decision, we are interested in other people's opinions. As individuals, for example, we want to know the opinions of existing users of a product before purchasing it, or others' opinions about political candidates before making a voting decision in an election. In business, companies and organisations acquire public opinions for marketing purposes, public relations, and political campaigning [110].

In the past, when an individual needed an opinion, they would ask friends and family members. When a company or organisation needed public or consumer opinions, it conducted surveys, opinion polls, and focus groups. With the proliferation of the Web 2.0, there came textual reviews, forum discussions, blogs, micro-blogs, comments, and postings on social networking sites, all generated by us, the users. These forms of media, collectively referred to as user-generated content, have produced an important shift

in the way in which people communicate and share textual knowledge and information, which influence social, political, and economic behaviour worldwide [129].

Sentiments, evaluations, and reviews are becoming very much evident with the growing interest in e-commerce, which is a prominent source of expressing and analysing opinions. Nowadays, customers on e-commerce sites, such as Amazon[1], mostly rely on reviews posted by other customers. For instance, in the Review Example (see page 9), we can be satisfied in deciding on purchasing the headphones, following a review describing a customer's positive experience in using the product.

In business, companies have increasingly realised that consumer voices influence and shape the opinions of other consumers and, ultimately, their brand loyalties and purchase decisions. In turn, customers' opinions are analysed for market research. For example, if a product seller is receiving negative feedback, in the interest of their reputation, the service provider might remove or suspend the seller. As well as product reviews, the web abounds with reviews for all kinds of services, ranging from restaurants (e.g. Yelp[2]), holiday destinations (e.g. Trip Advisor[3]), and films (e.g. IMDb [4] and Rotten Tomatoes[5]).

Online media and social networking sites (e.g. Facebook[6] and Twitter[7]) are also used for public self-disclosure. These open platforms provide and encourage users to interactively express their thoughts on global and personal issues across their social network, in real time. Most recently, users are expressing their opinions and sharing their experiences regarding the services they use and the products they buy on such social media sites. Companies have increasingly turned to social media to analyse and respond to consumer opinions.

[1] https://www.amazon.co.uk/
[2] https://www.yelp.com
[3] https://www.tripadvisor.co.uk/
[4] http://www.imdb.com/
[5] https://www.rottentomatoes.com/
[6] https://www.facebook.com/
[7] https://twitter.com/

Aside from short, mostly informal, social posts and product reviews, blogs have also emerged as a platform for expressing self-disclosure. Blogs are frequently updated series of archived posts, typically in reverse-chronological order. They may vary widely in nature and content, and have featured extensively in the popular media, entering political campaigns, news organisations, and businesses. As they have grown in popularity, they are often portrayed as online diaries or personal journals, often expressing sentiment towards more pensive subjects, such as health related topics. For example, Claire Greaves [71] openly blogs about her experiences with mental health, a subject which is surrounded by stigma.

Along with news media sites (e.g. BBC News[8]), such platforms offer a commenting feature, which prompts discussions within online social communities, where users are able to respond and contribute their experiences. This produces a stream of real time opinionated data, which can be used for applications which aim to, for example, predict election outcomes (e.g. [208]) and stock market direction (e.g. [3, 96]).

In its own right, user-generated content represents a rich source of information. The increasing popularity of publishing personal texts suggests that opinionated information is becoming an important aspect of the textual data on the web [207]. These forms of media serve as a platform to extract heterogeneous opinions that are publicly published by users from diverse societies. Due to the ever-growing volume of this type of information, the analysis of sentiments and opinions on a large scale is impractical without automatic classification and aggregation.

As a response, sentiment analysis methods and their applications have flourished in recent years. Such applications have spread across several domains, such as politics, finance, and health care [110].

The recent role of social media has produced several studies on the mining of online political speech [123]. Sentiment analysis has been used to understand voters' opinions (e.g. [103, 133]), to clarify what public figures support or oppose (e.g. [205]), tracking

[8]http://www.bbc.co.uk/news

discussion on political debates (e.g. [45]), predicting election outcomes (e.g. [208]), and determining whether news sources are biased in covering one political party more than another (e.g. [108]). Such applications help enhance the quality of the information that voters have access to during a political campaign, supporting their voting decisions.

Sentiment analysis has also been utilised in finance. Existing studies in this domain have focused on the relationship between the general mood of the public expressed on social media and the stock market direction (e.g. [22]), the sentiment expressed by expert investors on financial blog sources (e.g. [141]), and the opinions expressed on traditional news and finance media (e.g. [3, 96]). Such applications support financial and economic experts in identifying shifts in the markets. This provides a wider understanding about corporates, and a base to support economic knowledge in decision making, such as risk estimation before making an investment [73].

In health care, understanding a patient's experience is central to improving the quality of care they may receive in the future. Traditional measures for collating this information include surveys or structured questionnaires which ask specific and limited questions, and are expensive to administer. Patients have begun to report their health care experiences on the web in the form of blogs, social media postings, and health care rating websites. Existing studies in this domain vary, having focused on automatically analysing patient opinions and satisfaction ratings for services such as the UK's National Health Service (e.g. [4, 28]), estimating the sentiment of forum posts written by cancer survivors, as well as studying the changes in these forum members and their sentiments during their treatment [164], and the sentiment expressed in suicide notes (e.g. [191, 155]), which can be used to support online suicide prevention services.

### 2.1.2 Sentiment

In order to explain the goal of sentiment analysis and different approaches towards achieving this goal, we first need to define the notion of a sentiment. Sentiment is often defined as a reflection of our attitudes, thoughts, opinions, or emotions towards an entity or an event [207, 134]. These terms, which are related to human subjectivity, are often understood similarly and used interchangeably in NLP [134]. Sentiment analysis is, therefore, often alternatively referred to as opinion mining, review mining, attitude analysis, appraisal analysis, subjectivity classification, and affective computing [129, 150].

There are notable differences between the meaning of emotions, opinions, and sentiments [207]. Munezero et al. [134] generally define emotion as our subjective, pre-conscious social expressions of feelings, influenced by culture. An opinion is a transitional concept based on objective and/or subjective probabilities of information, which reflect our attitude towards an entity or an event. On the other hand, sentiments are different from opinions in that they reflect our feelings or emotions, which may not always be directed towards an entity.

In Section 2.1.2.1 and Section 2.1.2.2, we focus on the concepts of opinion and subjectivity, and emotion and affect respectively. The discussions will be confined to a description of these notions, and studies related to recognising these notions in written text.

#### 2.1.2.1 Opinion and Subjectivity

A word that often expresses sentiment is ‘opinion’. Liu [109] and Kim & Hovy [92] define an opinion as a combination of four factors (entity, holder, claim, and sentiment), in which the opinion holder may believe a claim about an entity, and in many cases, associate a sentiment with that belief. For instance, in the Review Example, sentence

(2) demonstrates a holder's opinion, claiming an overall positive sentiment towards the headphones, i.e. the entity.

There are two types of opinions: regular opinions and comparative opinions [88]. A regular opinion expresses a sentiment towards a particular entity, or an aspect of the entity. For example, sentence (1) expresses a positive sentiment towards the aspect of the headphone's sound quality. Regular opinions can directly/explicitly (e.g. "The headphones are uncomfortable to wear."), or indirectly/implicitly (e.g. "After wearing the headphones, my ears started ringing."), express a sentiment towards an entity [110]. On the other hand, a comparative opinion compares multiple entities based on some of their shared aspects, e.g. sentence (4) compares the headphones with another brand based on their comfort, and expresses a preference towards the other headphones. Regular opinions have been the main focus of research, as their explicit nature is much easier to detect and classify [110].

A factor to consider in the construction of an opinion is its holder. We can look at an opinion from two perspectives: the author (opinion holder) who expresses the opinion, and the reader (opinion reader) who interprets the opinion. The opinion reader may stand with, or against the opinion of its holder. For example, the author of "The sound quality is great." is expressing a positive opinion, whereas another reader, who also has the same headphones, may disagree. Most of the research in this domain assumes that the opinion is that of the opinion holder [110].

Unlike factual and objective expressions, sentiments and opinions share an important characteristic: they are both subjective. Wiebe [220, 222] centred the idea of subjectivity around that of a private state, which is defined by Quirk [166] as "a state which encloses sentiment, opinions, emotions, evaluations, beliefs, and speculations, which is not open to objective observation or verification". For example, sentence (3) expresses both objective ("The sound is much clearer") and subjective ("compared to the Sony which has too much bass in my opinion") expressions. Similar to opinions, subjective statements may not always have an associated sentiment (e.g. "I think the

headphones are broken" is a subjective sentence with no explicit sentiment, yet a negative sentiment can be implied). Conversely, objective statements may not contain words which explicitly express a sentiment, but it may be implied (e.g. in sentence (3), "The sound is much clearer" states a desirable fact that implies a positive sentiment) [109].

Sentiment analysis may be viewed as a text classification problem, which can be divided into two sub-tasks: identifying whether the state is subjective or objective (subjectivity analysis), and classifying the orientation of a subjective text segment (opinion mining) [109, 207].

For some sentiment analysis applications, before determining the orientation of the sentiment, it is important to distinguish whether a text is subjective or not, or identify which portions of the text are subjective. Subjectivity analysis is the task of identifying when a private state is being expressed, and identifying related attributes of it, such as who is expressing it, about whom or what it is being expressed, the orientation of it, etc. [150, 14]. Early research solved subjectivity classification as a standalone problem, i.e. not for the purpose of sentiment classification. In more recent research, some studies treated subjectivity classification as the first step of sentiment classification, by using it to remove objective sentences which are assumed to express or imply no opinion [110].

A popular task which often denotes sentiment analysis, and a subsequent task to subjectivity analysis once a statement is classified as being subjective, is opinion mining. Opinion mining is a recent sub-discipline of information retrieval and computational linguistics [53]. As opposed to focusing on the topic being discussed, opinion mining is a task which focuses on the opinions that are expressed on a particular topic, and classifying the orientation of this opinion (e.g. whether the expressed opinions are positive or negative) [207]. Several applications (e.g. determining whether a product is received well by customers) require the study of opinions expressed by many people, as, due to their subjective nature, analysing the opinion from a single person is often insufficient where decision making is concerned [110].

#### 2.1.2.2 Emotion and Affect

Besides opinion, 'affect' and 'emotion' are often used to refer to sentiment. Affect is generally described as a predecessor to emotion, which is defined as our subjective, pre-conscious social expression of feelings, influenced by culture [134]. In this domain, these terms are used interchangeably in the literature.

A sub-task of sentiment analysis is the recognition and classification of emotion in text, otherwise referred to as emotion classification, analysis or recognition, or affect detection or identification. The concepts of emotions and sentiments have also been used interchangeably in this domain. Their concepts are similar, as emotions too, may not always have a target; e.g. one may feel content for no apparent reason [134]. The strength of a sentiment or opinion is typically linked to the intensity of our emotions [110].

Humans may interpret other people's emotions using cues such as facial expressions, physical reactions, gestures and postures, our speech, and most importantly, our writings [7, 109]. The written expression of sentiment relies on words and the creative use of language which may infer or increase its salience. The aim of emotion analysis is to understand how people express emotion through text, or how text may trigger different emotions [147, 21]. Emotion analysis can identify expressions of happiness, sadness, anger, etc. [116].

Although sentiment is easier to classify, using specific emotions provides a greater insight into an expressed feeling. Some domains require further differentiation to associate specific emotions with appropriate actions. For example, in monitoring counter-terrorism issues, sadness, fear, and anger may require a different targeted response, e.g. counselling, media communication, and anti-radicalisation. Most recently, emotion analysis has been applied to a range of texts from different domains. Studies have focused on emotions expressed in web logs (e.g. [62, 124, 137]), fairy tales (e.g. [6, 58]), short stories (e.g. [89, 167]), chat messages (e.g. [234, 115, 17]), e-mails (e.g.

[113]), reviews (e.g. [174]), Twitter posts (e.g. [208, 127, 16]), etc. (see Table 2.4).

One of the main challenges of emotion classification is the formal definition of emotion. In this respect, text classification problems use categories to represent an emotion. There is a large number of overlapping natural language expressions that can be used to describe, express, or refer to emotions [187, 29]. Additionally, the ambiguity of natural language does not allow us to describe mixed emotions in an unequivocal way, for example, love, anger, and fear, are emotional words with various meanings that are not clearly defined, and have different meanings to different people [29]. However, to be able to classify emotions, particularly in text, a standardised way of representing emotions, and their related states is needed [117, 181]. At present there is no consensus among researchers on how to conceptually organise emotions. We discuss the frameworks that have been used for emotion classification in Section 2.2.2.

### 2.1.3 Levels of Analysis

Sentiment can be expressed in texts of various lengths, from short, superficial, and informal texts, such as microblogging posts, to long, detailed, and more formal texts, such as blog posts. Words alone can incite sentiment, yet when combined with other words to form a larger lexical unit, such as phrases and sentences, different sentiments can be expressed [119]. Sentiment analysis can be applied to text of different levels of granularity: document, sentence, and word or phrase level. We summarise some examples of state-of-the-art sentiment analysis which classify sentiment at such levels in Table 2.9.

#### 2.1.3.1 Document Level

The task at this level is to classify whether a whole text document expresses a sentiment [151]. For instance, given the Review Example, the task is to determine whether sentences (1, 2, 3) and (4) collaboratively express an overall positive or negative sentiment

towards the headphones.

However, one of the drawbacks of analysing sentiment at this level is that it is assumed that the whole document expresses one opinion towards a single entity. In practise, an opinionated document may evaluate multiple entities, and express different sentiments towards those entities [110]. Consider, for example, sentence (1), which discusses the headphones' sound quality and expresses a positive sentiment, whereas sentence (4) discusses their comfort and expresses a negative one. In this case, it is impractical to assign one overall sentiment to the whole document, as it may overlook the occurrence of the different sentiments being expressed, and can lead to incorrect classification [151, 188].

#### 2.1.3.2 Sentence Level

Conversely, a finer level of analysis is to classify whether an individual sentence expresses sentiment [199]. For example, the task would be to determine the sentiment expressed in sentences (1, 2, 3) and (4) individually.

Much of the work which classifies sentiment at this level assumes that a sentence expresses a single sentiment from a single opinion holder. Needless to say, this is not always the case, as multiple sentiments may be expressed in compound and more complex sentences. Consider, for example: "The sound quality of these headphones is clear, but the material around the headband makes it uncomfortable to wear," which expresses a positive sentiment towards the headphones' sound quality, but a negative one towards the headphones' comfort. This too, can lead to incorrect classification.

#### 2.1.3.3 Word and Phrase Level

Sentiment can also reside in even smaller linguistic units, such as words (e.g. great, awful) and phrases (e.g. idioms such as *kick the bucket, raining cats and dogs,* etc.) [150]. Analysis at this level aims to classify the overall sentiment of a text segment

by distinguishing the sentiment of the individual unigrams and bigrams it contains [227, 200]. This task often involves using affective lexicons (e.g. WordNet-Affect [196]), which contain entries of words and phrases that convey a subjective bias, to support the extraction of sentiment words from contexts [198, 48, 114]. We discuss this approach in more detail in Section 2.4.1.

The main motivation behind extracting sentiment at such a level of granularity is that it provides an insight into the types of linguistic features that have an impact on the overall sentiment, as well as being able to capture a mixture of sentiments that may be expressed. However, an important factor to consider is that the sentiment of a word or phrase used in context, may differ from its sentiment when it is not used in context. We discuss unigram, bigram and n-gram features in more detail in Section 3.4.1.

#### 2.1.3.4 Concept Level

There are several methods of expressing sentiment in written text [118]. The main method is by using explicit words which convey a subjective bias (e.g. sentence (1) explicitly expresses a positive sentiment based on the presence of the word 'great'). Conversely, sentiment may be expressed in more indirect ways. In order to analyse this type of sentiment, one may consider concept level classification, which, unlike syntactic techniques, is able to detect sentiment that is expressed in a more subtle manner.

The task at this level focuses on the semantic analysis of text. Some approaches use web ontologies or semantic networks, which allow conceptual and affective information to be aggregated with natural language [27, 159]. For example, 'too quiet' can be identified as expressing a negative sentiment if it occurs in a headphone review. By relying on semantic knowledge bases, such approaches step away from the use of explicit sentiment words, and rely on natural language concepts which implicitly express sentiment [27].

#### 2.1.3.5 Aspect Level

Classifying texts at the document and sentence level is often insufficient for applications that require sentiments relating to a specific aspect or topic [110, 207]. Within a general topic (e.g. headphones), the author may discuss several specific aspects of an entity. For example, sentences (1) and (3) discuss the headphones' sound quality, whereas sentence (4) discusses the headphones' comfort. Thus, for a task which requires determining the sentiment of a specific topic, the classification is performed at the aspect level. In order to analyse sentiment at this granularity, the relevant aspects must be extracted. This introduces a new set of challenges that requires deeper NLP [110], which goes beyond the range of this thesis. However, upon extraction, sentiment classification can be performed using traditional machine learning approaches.

## 2.2 Representation of Sentiment

Sentiment analysis is considered to be a classification problem, in which, given a text segment, the task is to automatically classify its sentiment. In this respect, sentiment classification problems often use polarity categories (i.e. positive, negative or neutral) to represent and classify sentiment.

Emotion classification problems use categories of emotion (e.g. happiness, sadness, fear, etc.) to represent and classify emotion, and related states. However, there is a large number of overlapping natural language expressions that can be used to describe, express, or refer to emotions [187, 27]. Therefore, emotion classification requires a standardised framework to represent emotion [117, 181]. We discuss the range of frameworks which have been used for emotion classification, in Section 2.2.2.

### 2.2.1 Sentiment Polarity

Many sentiment analysis approaches consider the notion of sentiment to be sentiment polarity, which can be positive, negative or neutral. Sentiment polarity allows for the simple representation and management of sentiment [207]. Intuitively, positive sentiment is used to represent supportive opinions (e.g. “Brilliant night last night”). Conversely, negative sentiment is used to represent disagreement or discouraging opinions (e.g. “My throat is killing me”). The neutral category is used to represent the absence of any sentiment (e.g. “It is Monday today”).

### 2.2.2 Emotion Classification Frameworks

To be able to classify emotions, particularly in text, a standardised way of representing emotions and their related states is needed [117, 181]. At present, there is no consensus among researchers on how to conceptually organise emotions. Computational approaches to emotion classification have focused on various emotion frameworks, resulting in a large number of multi-modal, emotionally annotated data [7].

A range of frameworks exist, most of which were founded on a psychological theory. Those presented in the literature organise emotions as universal basic categories, dimensional characterisations, and hierarchical organisations of emotion.

#### 2.2.2.1 Categorical Representation of Emotion

Categorical representations of emotion are usually theory-driven accounts that suggest that basic emotions are the functional expression of underlying biological and evolutionary processes [37, 38, 105]. Researchers have investigated several aspects of human emotion in order to arrive at a set of emotion categories that are universally acceptable [157]. For instance, Ekman [51] defined basic emotions as those that have six universally accepted, distinctive facial expressions. The six basic emotions are

arguably the most popular within the field of computer science for emotion mining, recognition, and classification (e.g. [6, 8, 39, 127, 195, 113]) [15].

However, the emotions comprising each basic emotion set vary amongst theorists, as there is not always an agreement on which emotions are basic [146]. For example, can surprise be considered a basic emotion if it can take the form of a positive, negative or neutral valence? [15]. Thus, several sets of basic emotions exist (see Table 2.1 for examples).

| Theorists | Basic emotions |
|---|---|
| Ekman [51] | happiness, sadness, fear, anger, disgust, surprise |
| Tomkins [206] | joy, anguish, fear, anger, disgust, surprise, interest, shame |
| Izard [86] | enjoyment, sadness, fear, anger, disgust, surprise, interest, shame, shyness, guilt |
| Ortony et al. [145] | joy, sadness, fear, anger, disgust, surprise |
| Plutchik [158] | joy, trust, fear, surprise, sadness, disgust, anger, anticipation |

**Table 2.1: Basic emotion categories**

Ortony, Clore & Collins [145] challenge the notion of the universality of basic emotions, by questioning whether they can blend to form more complex, secondary emotions. Some theories suggest that secondary emotions can be created by fusing, blending, mixing or compounding basic emotions [146]. For example, Plutchik's wheel of emotion [158] (Figure 2.1) has eight basic emotions (joy, trust, fear, surprise, sadness, disgust, anger, and anticipation), where each has three levels of activation represented by colour boldness, e.g. annoyance is less intense, whereas rage is more intense, than anger. The emotion space is represented so that combinations of basic emotions derive secondary emotions (e.g. joy + trust = love, anger + anticipation = aggression, etc.). The wheel has been used to support an automated computational framework for finding emotions in Twitter conversations [93].

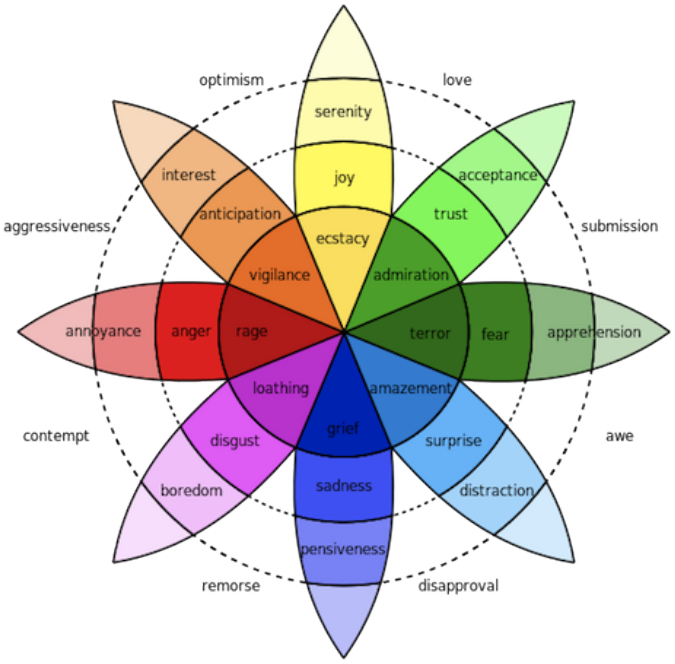


**Figure 2.1: Plutchik's wheel of emotion**

However, an important point to consider is that categorical representations of emotions are limited in size, often not covering the emotion space adequately. This may lead to forced-choice, where subjects are likely to discriminate among the available categories, rather than identifying the precise emotion itself. This can occasionally force the subjects to choose an appropriate category which might not necessarily reflect the true emotion being expressed. Nevertheless, due to their simplicity and familiarity, categorical models of emotion have been dominant in this field [116].

#### 2.2.2.2 Dimensional Characterisation of Emotion

Other researchers, such as Schlosberg [180], have referred to continuous dimensions of emotions. As opposed to distinct emotion categories, dimensional characterisations represent emotions as coordinates in a multi-dimensional space [29], capturing the positive and negative shifts in sentiment [207]. There are variations among these models,

many of which are formed of two or three dimensions [173]. These dimensions incorporate aspects of arousal and valence (e.g. [175, 97]), evaluation and activation (e.g. [217, 36]), positive and negative (e.g.[216]), tension and energy (e.g. [202]), pleasure, arousal, and dominance (e.g. [121]), dominance and co-operation (e.g [104]), etc. We summarise such examples in Table 2.3. Dimensional representations of emotion provide a way of describing emotional states that is more tractable than using distinct emotion categories. This is important, particularly when dealing with naturalistic data, where a range of emotions can be expressed at the same time [29, 116].

One of the very first empirical dimensional models is Russell's Circumplex Model of Affect [176] (Figure 2.2). This framework organises twenty eight discrete emotions around a circumplex, according to two dimensions: arousal and valence. The arousal dimension differentiates a direction of high and low arousal, whereas the valence dimension indicates a scale of positive and negative emotions. Subjects are able to choose a position located anywhere between two emotions. Numerical data are obtained from the relative position of the points in the two-dimensional bipolar space [116].

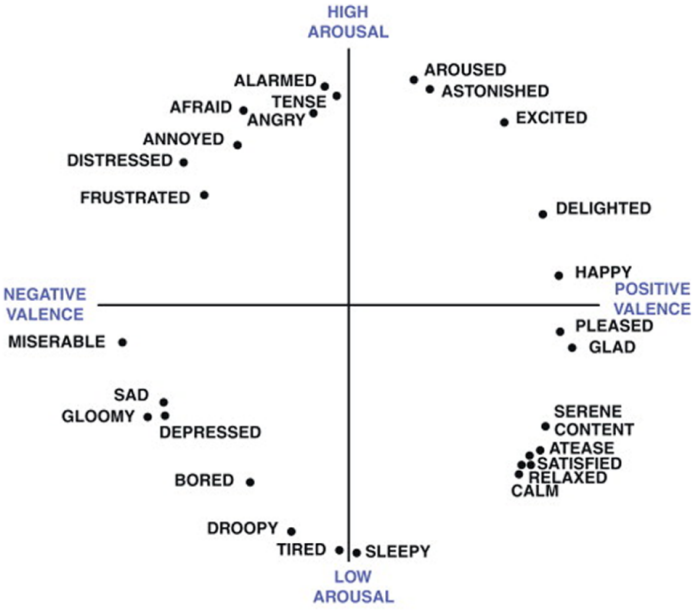


**Figure 2.2: Russell's Circumplex model of affect**

The Circumplex Theory of Affect [216] (Figure 2.3) identifies two main dimensions: positive and negative affect. There are four sub-dimensions: positive affect, engagement, negative affect, and pleasantness. Each dimension has two directions: high and low. Specific emotions are classified into one of eight categories on this scale. For example, excitement is classified as having high positive affect, calmness as having low negative affect, etc. The Circumplex has been used to support emotion classification [174, 167], and has been suggested as a useful model for quantifying and qualitatively describing emotions identified in text [174].

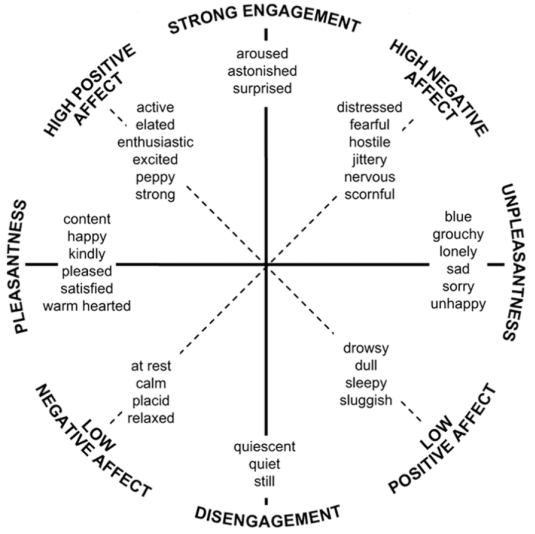


**Figure 2.3: Watson & Tellegen's Circumplex theory of affect**

Scherer's affect model [179] (Figure 2.4) organises one hundred and two discrete emotions around a circumplex, according to two dimensions: activity and evaluation. The activity scale differentiates a direction of passivity and activity. Similar to Russell's Circumplex Model of Affect, the evaluation dimension indicates a scale of positive and negative emotions. The proximity of two emotion categories in the circumplex represents conceptual similarity of the two categories [116]. Such a framework was used to support emotion classification of weblog texts [63].

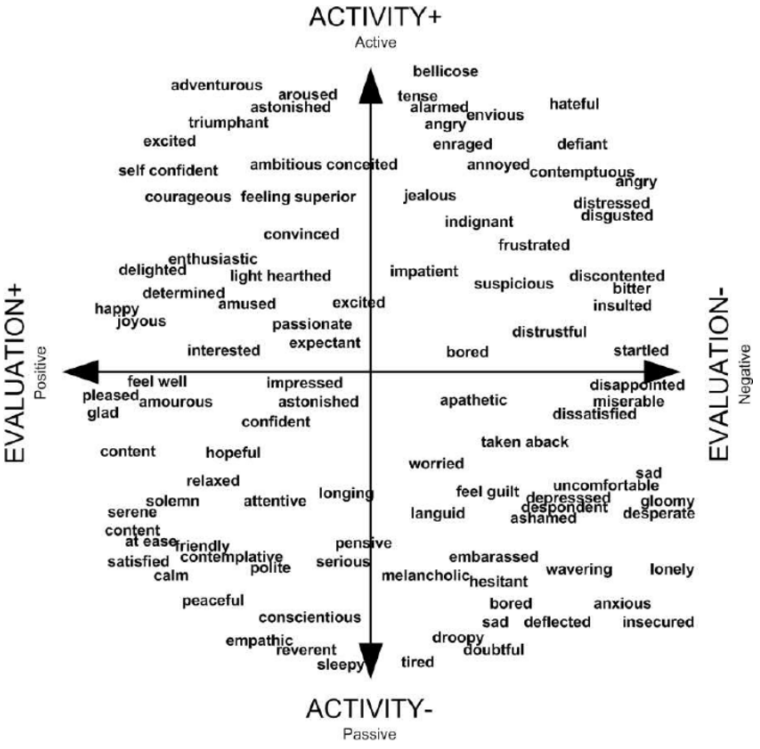


**Figure 2.4: Scherer's affect model**

Unlike most schemes discussed in this Section, which were founded on psychological theories and not devised with practical computational applications in mind, Emotion Annotation and Representation Language (EARL) Version 0.4.0 [57] is an XML-based language for representing emotions in a technological context, including corpus annotation and recognition. Similar to Watson & Tellgen's Circumplex theory of affect, EARL organises emotions into positive and negative categories, which are further refined based on intensity and attitude. There are five positive and five negative categories, where specific emotions are given as representative examples of each category. For example, agitation is exemplified by shock, stress, and tension (Table 2.2).

| **Positive category** | **Example** | **Negative category** | **Example** |
|---|---|---|---|
| Positive & lively | Delight, Elation. Joy | Negative & forceful | Anger, Contempt, Disgust |
| Caring | Affection, Empathy, Love | Negative & not in control | Fear, Worry, Anxiety |
| Positive thoughts | Hope, Pride, Trust | Negative thoughts | Doubt, Envy, Guilt |
| Quiet positive | Calm, Content, Relaxed | Negative & passive | Hurt, Sadness, Despair |
| Reactive | Interest, Politeness, Surprise | Agitation | Shock, Stress, Tension |

**Table 2.2: Emotion Annotation Representation Language**

| Framework | Dimensions | Sub-dimensions | Emotions |
|---|---|---|---|
| Russell's Circumplex Model of Affect [176] | Arousal & Valence | High & Low arousal, Positive & Negative valence | 28 |
| The Circumplex Theory of Affect [216] | Positive & Negative | See Figure 2.3 | 38 |
| Scherer's affect model [179] | Activity & Evaluation | Active, Passive, Positive, Negative | 102 |
| EARL [57] | Positive & Negative | See Table 2.2 | 48 |
| Kort's Affective Model of Interplay Between Emotions and Learning [97] | Arousal & Valence | Constructive & Nonconstructive learning, Positive & Negative affect | 11 |
| Thayer's two-dimensional model of emotion [202] | Tension & Energy | Energy, Stress | 10 |
| Whisell's Dictionary of Affect in Language [217] | Activity & Evaluation | Pleasantness, Activation | 348,000 |
| Feeltrace [36] | Activity & Evaluation | Very active, Very passive, Very positive, Very negative | 19 |
| Mehrabian's PAD [121] | Pleasure, Arousal, & Dominance | Pleasure, Arousal, Dominance | NA |
| Leary's rose [104] | Dominance & Co-operation | Dominant, Extrovert, Warm/Friendly, Agreeable/Trust, Submissive, Introvert, Cold/Unfriendly, Disagreeable | NA |
| Hevner's adjective circle [78] | Positive & Negative | Music features | 67 |
| Self Assessment Manikin (SAM) [102] | Pleasure, Arousal, & Dominance | Pleasant, Unpleasant, Relaxed, Energetic, Figure size | Rating score (1-9) |

**Table 2.3: Examples of dimensional frameworks of emotion**

#### 2.2.2.3 Hierarchical Organisation of Emotion

Unlike other frameworks discussed in this Section, which generally contain finite, but manageable sets of emotions, hierarchical models of emotion have been introduced to capture a much wider and richer set of emotions. The main focus of such frameworks has been on lexical aspects that can support text mining applications.

Affective hierarchies are structured so that specific emotions represent instances of more general, underlying emotions. It is suggested that emotions can be grouped into classes, with the most super-ordinate classes of most hierarchical models being positive and negative sentiment. The next level is considered as the more general, basic emotion level (e.g. happiness, sadness, love, anger, etc.). The lowest subordinate level consists of groups of individual emotions, that form a category named after the most typical emotion of that category (e.g. optimistic, miserable, passionate, frustrated, etc.) [177]. For example, see Figure 2.5 for an excerpt from Parrot's collection of emotions [152], which organises more than one hundred emotions across a hierarchical structure.

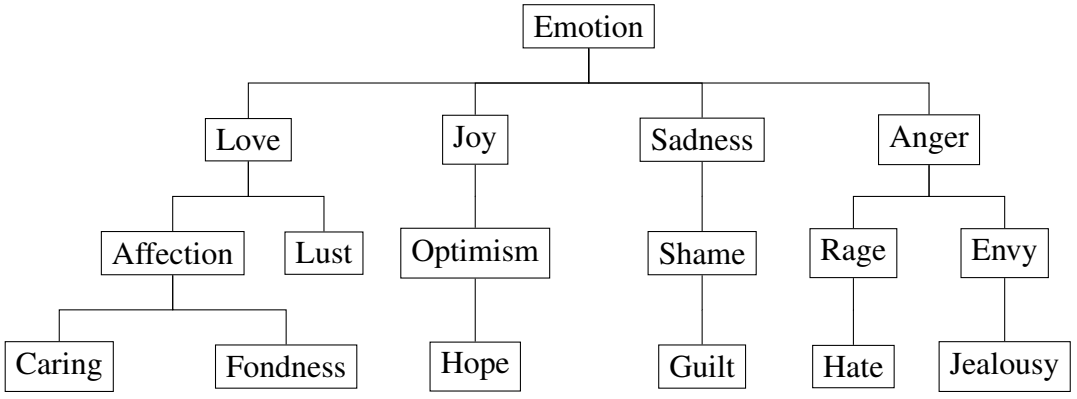


**Figure 2.5: An excerpt from Parrot's emotion hierarchy**

In summary, Table 2.4 provides examples of state-of-the-art emotion classification approaches, as well as the frameworks used to represent emotions. This brief overview demonstrates that there is a wide divergence in the frameworks used, with several studies classifying basic sets of emotions, and some using dimensional and hierarchical models. There is also variability in terms of the classification performance.

Naïve Bayes (NB), Support Vector Machines (SVM), Maximum Entropy (ME), Precision (P), Recall (R), F-measure (F), Accuracy (A)

| Reference | Framework type | Framework | Dataset | Size | Level | Features | Approach | Evaluation (%) |
|---|---|---|---|---|---|---|---|---|
| [80] | Categorical | 'neutral, anger, sad, afraid, disgusted, ironic, happy, surprise' | Dialogues | 1,201 | Sentence | Unigrams, Bigrams, Negation, Tense | k-nearest neighbour | A = 90 |
| [185] | Categorical | 'anger, fear, hope, sadness, happiness, love, thank, neutral' | News headlines | 1,250 | Sentence | Unigrams | Neutral Networks | A = 57.75 |
| [124] | Categorical | 'happy, sad' | Blog posts | 10,000 | Phrase | Unigrams | NB | A = 79.1 |
| [6] | Categorical | Ekman's six basic emotions | Fairy tales | 185 | Sentence | Unigrams, Emoticons, Punctuation | NB, SVM | A = 72.08, 73.89 |
| [8] | Categorical | Ekman's six basic emotions | Blog posts | 5,205 | Sentence | Unigrams, Emoticons, Punctuation | SVM, NB | A = 73.9, 72.1 |
| [190] | Categorical | 'sentiment, arguing, utterance arguing' | Dialogues | 6,504 | Sentence | BOW | SVM | A = 82.16 |
| [195] | Categorical | Ekman's six basic emotions | News headlines | 1,250 | Sentence | Unigrams | NB | P = 12.04, R = 18.01, F = 13.22 |
| [230] | Categorical | 'happiness, joy, sad, angry' | Blog posts | 5,410,933 | Sentence/ Docment | BOW | SVM | P = 74.02, R = 46.78, F = 57.33 |
| [62] | Dimensional | Scherer's affect model | Blog posts | 346,723 | Document | Unigrams, POS | SVM | A = 60.7 |
| [75] | Dimensional | Russell's Circumplex model of affect | Tweets | 134,100 | Sentence | Unigrams, Emoticons, Hashtags, Punctuation, Negation | SVM | A = 90 |
| [212] | Dimensional | Leary's Rose | Dutch sentences | 740 | Sentence | Unigrams, POS, Term frequency, Sentence length, Punctuation | SVM | F = 31 |
| [52] | Hierarchical | 3 levels: 'emotional, non-emotional, positive, negative,' Ekman's six basic emotions | Tweets | 2,809 | Sentence | Corpus based feature sets | SVM | F = 78 |
| [64] | Hierarchical | 3 levels: 'emotional, non-emotional, positive, negative,' Ekman's six basic emotions | Blog posts | 4.090 | Sentence | Unigrams | SVM | F = 62 |
| [91] | Hierarchical | 5 levels: 132 moods | Blog posts | 815,494 | Sentence | Unigrams, Emoticons | SVM | A = 24.73 |
| [126] | Hierarchical | 5 levels: 132 moods | Blog posts | 815,494 | Sentence | Term frequency, Sentence length, Unigrams, Punctuation, Emoticons | SVM | A = 58 |

**Table 2.4: State-of-the-art emotion classification**

## 2.3 Feature Space

While it is fairly easy for human readers to determine the sentiment expressed in text, it is quite a challenge to accomplish the same task automatically. In order to detect sentiment, machines often rely on the presence or frequency of textual features, which contribute useful information regarding the sentiment that is being expressed.

Features used to support sentiment analysis include: sentiment words (e.g. good, bad) which explicitly convey a subjective bias, negation (e.g. not, never) which can revert the polarity of sentiment words, textual conventions which have emerged from online texts (e.g. emoticons such as :( are used to express negative sentiment), the syntactic features of words (e.g. Part of Speech (POS)), and figurative language which are often used to enhance sentiment (e.g. irony, sarcasm, and idioms). We summarise some examples of state-of-the-art approaches to sentiment analysis which have included such features, in Table 2.9.

In Section 2.3.1.5.1, we pay particular attention to idioms - multi-word expressions holding a literal and figurative meaning which is conventionally understood by native speakers. We further reinforce the importance of idioms as features of sentiment analysis, by demonstrating that current state-of-the-art sentiment analysis tools, which do not consider idioms as features, often result in incorrect classifications when they are explicitly used to express sentiment.

Once the features have been selected, the next step is to extract them from a given text. We discuss feature extraction methods, in Section 2.3.2.

### 2.3.1 Features of Sentiment Analysis

The first step in classifying the overall sentiment of a text segment is to select the features which indicate sentiment. We discuss the textual features that are used to support sentiment analysis in the following Sections.

#### 2.3.1.1 Unigrams, Bigrams, and N-grams

Strapparava & Mihalcea [195] note that in discourse each lexical unit, whether it is a word or a phrase, has the ability to contribute useful information regarding the sentiment that is being expressed. In the literature, these features are referred to as sentiment words, opinion words, affective words or polar words [110]. Features take the form of single words (unigrams), short phrases (bigrams), and longer phrases (n-grams).

There are several methods of expressing sentiment in written text [118]. The first is the explicit use of individual words that convey a subjective bias. For example, in the Review Example, sentence (1) explicitly describes that the opinion holder's attitude towards the headphones is positive, based on the presence of the positive word, 'great'. Sentiment words are available from specialised lexicons (discussed in Section 2.4.1.1). These words are often adjectives (e.g. good, bad), adverbs (e.g. cheerfully, weirdly), nouns (e.g. blessing, rubbish), and verbs (e.g. love, hate) [110].

Sentiment may also be expressed by using comparative words (e.g. better, worse). Unlike explicit sentiment words, comparative words do not express a direct opinion towards an entity or event, but a comparative opinion [110]. Consider for example, "Samsung headphones are better than Sony headphones." The opinion holder does not explicitly express whether either of the branded headphones are a good or a bad product. Instead, based on the positive comparative word, 'better', they state a preference towards Samsung.

However, a text containing explicit or comparative sentiment words, may not always express a sentiment. This may occur in questions or conditional sentences, e.g. "Could you tell me which headphones are the best?" and "If the headphones are in the shop, I will buy them." Both examples implicitly imply a positive sentiment, without containing explicit sentiment words. Conversely, a text containing no sentiment words may implicitly express sentiment.

There are, however, drawbacks to using only unigrams as features. For example, nega-

tions, such as “not bad” or “not good,” will not be taken into account. In fact, unigram features can sometimes lead to misclassification. In this case, using frequent bigrams and n-grams as features, it is possible to capture some dependencies between words and the effects of individual phrases on the overall sentiment. N-grams have been included as features of several state-of-the-art sentiment analysis (e.g. [40, 227]). Nevertheless, the most successful features seem to be basic unigrams (see Table 2.9) [151, 178].

#### 2.3.1.2 Negation

Negation plays an important role in sentiment analysis as it can revert the polarity of sentiment words [122, 224]. Negation words such as ‘not’ or ‘never’ are often considered stop words. Stop words are commonly used words which have little meaning, e.g. ‘and’, ‘the’, ‘a’, etc., and as such, are removed from analysis together with their effect on the polarity of sentiment words. Consider for example how negation reverses the positive polarity expressed in “Turns out she was cool with it” to express a negative sentiment in “Turns out she was not cool with it.”

Negation may occur in two main forms: directly or longer-distance. Negative words (e.g. no, never, etc.), negative adverbs (e.g. hardly, barely, etc.), and negated verbs (e.g. doesn’t, isn’t, etc.) may directly effect the sentiment of the words closely surrounding their occurrence. Conversely, negation may also involve longer-distance dependencies. Consider for example “That doesn’t look very good.” The last word, ‘good’, is negated by ‘doesn’t’ at the beginning of the sentence.

Although negation has been a prominent feature of several state-of-the-art sentiment analysis studies (e.g. [150, 151, 224, 136, 135, 227] ), it must be handled with care. Not all occurrences of negative words changes the polarity of the sentiment [227]. Consider for example the negation in “Not only did the colour of the headphones attract me, but so did the sound quality,” which does not reverse the sentiment polarity, but enhances it.

#### 2.3.1.3 Other Textual Conventions

Zhang et al. [233] draw attention to the variety of challenges posed by the language used in informal text. Written online communication has led to the emergence of textual conventions [162] used to compensate for the absence of body language and intonation, which otherwise account for 93% of non-verbal communication [120]. These include the presence of emoticons - pictorial representations of facial expressions that are used to compensate for the lack of embodied communication (e.g. :) is commonly used to represent positive sentiment), slang (e.g. chuffed, do one's nut), abbreviations (e.g. complete waste of time - CWOT, great - GR8), onomatopoeic elements (e.g. gr, hm), as well as the use of upper case, punctuation (e.g. !!, ?!), and repetitions of letters (e.g. sweeeeet) for affective emphasis.

In particular, the use of emoticons is considered an effective way of conveying emotion in text [204, 42]. Several sentiment analysis studies which use data from the web, specifically Twitter data, have been based on the inclusion of emoticons as features, which demonstrated high accuracy when machine learning algorithms were trained with such data (e.g. [68, 168, 148, 98]). Other studies have included features such as capitalisation, punctuation, and emotional hashtags, which demonstrated an improved sentiment classification accuracy when such features are present (e.g. [98, 16, 2]).

#### 2.3.1.4 Part of Speech (POS)

It has been demonstrated that some adjectives (e.g. lovely, awful), nouns (e.g. concern, hope), verbs (e.g. love, hate), and adverbs (e.g. gently, harshly) are good indicators of sentiment [150]. Thus, several studies have exploited POS - the syntactic function of a word, as features of state-of-the-art sentiment analysis [150] (e.g. [76, 132, 219, 151, 172, 18, 135]).

#### 2.3.1.5 Figurative Language

An additional method of expressing sentiment in text is the use of figurative language [21]. In contrast to literal language which is related to the notion of true, exact or real meaning, figurative language covers a wide range of literary devices and techniques [69, 144]. These include figures of speech, metaphors, similes, personification, hyperboles, irony, sarcasm, symbolism, onomatopoeia, synecdoche, clichés, metonymy, and idioms. These allusions give the readers new insights into being more effective, persuasive, impactful, and colourful in their speech, to make a particular linguistic point [65].

Our pragmatic competence as human readers allows us to interpret and determine whether an expression is used literally or figuratively [69]. However, the use of figurative language which is pervasive in text and is often used with the intention for expressing opposite polarities and indirect meanings, has demonstrated to be challenging for sentiment analysis, with few approaches having been attempted. [65]. However, by using an n-gram graph based method to assign sentiment polarity to individual word senses, experiments implied that figurative language not only conveys sentiment, but actually drives the polarity of a sentence [170, 169, 156].

##### 2.3.1.5.1 Idioms

A prominent literary device of figurative language is the idiom. Traditionally, idioms (e.g. *Bob's your uncle, over the moon, kick the bucket*) have been defined as figurative statements whose overall meaning cannot be derived from the meanings of their individual compositional parts. Idioms are, therefore, considered to be fixed combinations of words referred to as Multi-Word Expressions (MWE). They are frequently used in everyday textual and verbal conversations, and are conventionally understood by native language speakers.

Most idioms can be distinguished from related linguistic categories such as formulae,

fixed phrases, collocations, clichés, sayings, proverbs, and allusions by considering the following properties [139, 70, 69]:

1. **Conventionality**: Generally, their overall meaning cannot be (entirely) predicted by considering the meaning of their constituent words independently. Idioms, therefore, hold their meaning as a single semantic unit [20, 25, 56] (e.g. *fish out of water* is used to refer to someone who feels uncomfortable in a particular situation).

2. **Inflexibility**: Their syntax is restricted, i.e. idioms do not vary much in the way they are composed (e.g. *raining cats and dogs* cannot be composed as *raining dogs and cats*).

3. **Figuration**: Idioms typically hold a figurative meaning, which stems from metaphors, hyperboles, and other types of figuration [138] (e.g. *on the rocks* is used to refer to a relationship experiencing difficulties and is likely to fail).

4. **Proverbiality**: They usually describe a recurrent social situation.

5. **Informality**: They are associated with less formal language such as colloquialisms.

6. **Affect**: Idioms typically imply an affective stance towards an entity or an event, rather than a neutral one [24]. Most idioms can be directly or indirectly classified into emotional categories [48], (e.g. *on cloud nine* is used to refer to being extremely happy, whilst *cry one's eyes out* is used to refer to crying bitterly, and at length).

The sixth property (affect) emphasises the importance of idioms in sentiment analysis, as it implies that they may be sufficient in determining the underlying sentiment of a text span. Whilst idioms have been extensively studied across many disciplines (e.g. linguistics, psychology, etc.), to the best of our knowledge, thus far, there is no comprehensive knowledge base that systematically maps English idioms to their sentiments.

This is the main reason why idioms are under-represented as features of sentiment analysis, with few exceptions (e.g. [229] describes a set of 8,160 Chinese idioms and [85] describes a set of 3,632 Arabic idioms).

To be included as features in sentiment analysis, idioms need to be recognised in text. The second property (inflexibility) makes this requirement feasible. Many idioms are frozen phrases, and can be recognised by simple string matching. Thus, idioms may be recognised using a lexicon-based approach, which can only identify idioms that are syntactically unproductive or frozen. For example, Shastri et al. [186] and Liu & Hwa [111] used a dictionary of idioms (e.g. *at a snail's pace*) to recognise them in text, and mapped them to their abstract meanings (e.g. slow), which were subsequently utilised to infer their sentiment. In another lexicon-based approach, the recognition of idioms in [94] was limited to 46 noun-noun compounds (e.g. *glass ceiling*). The use of idiom sentiment profiles was found to improve the performance of sentiment classification.

Conversely, some idioms are syntactically productive or flexible. The syntactic changes of idioms, such as the use of different grammar and verb tenses (active or passive voices), changes in word places, singular and plural forms, negation, and the inclusion of additional words, form one of the main criteria causing difficulties in recognising idioms in text [232]. For idioms that are flexible, lexico-syntactic patterns can be used to computationally model idioms and recognise their occurrences. For example, in Polish, a highly inflected language, idioms were recognised by using a cascade of regular expressions. Their effect on sentiment analysis results was evaluated on a corpus of product and service reviews, where idioms were found to occur rarely [23].

While recognising idioms in text, we must also consider their third property (figuration). In contrast to the literal meanings of some idioms, most hold a figurative meaning, proving a challenge for language learners, as they need to be taught to be understood [69]. Automatically identifying idioms in text using a lexicon-based or pattern matching approach alone will find all idiom occurrences. In order to recognise idioms, their sense, i.e. whether they are used literally or figuratively, needs to be disambigu-

ated in a given context. However, this is not a trivial task, with few approaches having attempted context-based idiom detection (e.g. [111, 193]).

To further reinforce the importance of idioms as features of sentiment analysis, we applied a selection of state-of-the-art sentiment analysis tools to contextual examples of idioms, shown in Table 2.5. Table 2.6 demonstrates these results, where only two tools correctly classified all four sentences. On average, the tools incorrectly classified half of the examples. In particular, S1, which expresses a positive sentiment, is correctly classified by ten out of sixteen tools, whereas S4, which expresses a negative one, is correctly classified by only two.

| ID | Sentence | Overall sentiment |
|---|---|---|
| S1 | "I was *over the moon* when I heard the news." | Positive |
| S2 | "Mr Jones was *grinning from ear to ear*." | Positive |
| S3 | "But I was *bored to tears*." | Negative |
| S4 | "I have a *bone to pick* with you." | Negative |

**Table 2.5: Contextual examples of idioms**

Nevertheless, idioms have not been completely ignored in sentiment analysis. We summarise some examples of state-of-the-art sentiment analysis approaches which have included idioms, amongst other features, as well as the language and size of their idiom dataset, in Table 2.7.

Very negative (- -), Negative (- or neg), Neutral (0), Positive (+ or pos), Very positive (++)

| Tool | Availability | S1 | S2 | S3 | S4 |
|---|---|---|---|---|---|
| SentiStrength | http://sentistrength.wlv.ac.uk/ | pos:1, neg:-1<br>Neutral | pos:2, neg:-1<br>Positive | pos:1, neg:-4<br>Negative | pos:1, neg:-1<br>Neutral |
| Stanford NLP | https://nlp.stanford.edu/sentiment/ | - -:2, -:13, 0:63, +:20, ++:2<br>Neutral | - -:4, -:24, 0:36, +:31, ++:5<br>Neutral | - -:24, -:61, 0:13, +:1, ++:1<br>Negative | - -:2, -:8, 0:41, +:44, ++:5<br>Positive |
| GATE | https://gate.ac.uk/sentiment/ | Neutral | Neutral | Neutral | Neutral |
| LingPipe | http://alias-i.com/lingpipe/demos/tutorial/sentiment/readme.html | pos:0.507<br>Positive | pos:0.668<br>Positive | neg:0.818<br>Negative | neg:0.842<br>Negative |
| Lexalytics | https://www.lexalytics.com/technology/sentiment | 0.000<br>Neutral | 0.000<br>Neutral | -0.593<br>Negative | 0.000<br>Neutral |
| Alchemy API | https://www.ibm.com/watson/alchemy-api.html | 0:0<br>Neutral | 0:0<br>Neutral | neg:-0.536676<br>Negative | neg:-0.559095<br>Negative |
| Google Prediction API | https://cloud.google.com/prediction/docs/sentiment_analysis | 0:0<br>Neutral | 0:0.1<br>Neutral | pos:05<br>Positive | 0:0.2<br>Neutral |
| NLTK | http://www.nltk.org/howto/sentiment.html | pos:0.244, 0:0.466, neg:0.756<br>Negative | pos:0.505, 0:0.801, neg:0.495<br>Neutral | pos:0.208, 0:0.052, neg:0.792<br>Negative | pos:0.583, 0:0.355, neg:0.417<br>Positive |
| Sentiment Analyser | http://www.danielsoper.com/sentimentanalysis/default.aspx | pos:100<br>Positive | pos:100<br>Positive | neg:-100<br>Negative | neg:-100<br>Negative |
| Text Analytics & Sentiment Analysis API | http://text2data.org/ | pos:0.090<br>Positive | pos:0.985<br>Positive | neg:-0.250<br>Negative | 0:-0.452<br>Neutral |
| TheySay Sentiment Analysis API | http://www.theysay.io/sentiment-analysis-api/ | 0:0.88<br>Neutral | pos:0.884<br>Positive | neg:0.859<br>Negative | 0:1<br>Neutral |
| ParallelDots Sentiment Analysis API | https://www.paralleldots.com/sentiment-analysis | 48%<br>Neutral | 97%<br>Positive | 7%<br>Negative | 92%<br>Positive |
| Repustate | https://www.repustate.com/ | pos:0.95<br>Positive | pos:0.95<br>Positive | neg:-0.95<br>Negative | 0:0<br>Neutral |
| Aylien Text Analysis API | http://aylien.com/ | pos:0.63<br>Positive | pos:0.59<br>Positive | neg:0.63<br>Negative | pos:0.74<br>Positive |
| Textgain | https://www.textgain.com/ | Positive | Positive | Negative | Positive |
| Sentiment Analyser | http://sentiment.vivekn.com/ | neg:82.1<br>Negative | 0:53:5<br>Neutral | neg:99.0<br>Negative | 0:61.8<br>Neutral |

**Table 2.6: The output produced for sentences S1-S4 from Table 2.5**

Naïve Bayes (NB), Support Vector Machines (SVM), Precision (P), Recall (R), F-measure (F), Accuracy (A).

| Reference | Language | Size | Level | Approach | Evaluation (%) |
|---|---|---|---|---|---|
| [229] | Chinese | 8,160 | Sentence | Unsupervised | A = 84 |
| [131] | English | 40 | Concept | Lexicon based/ Machine learning | A = 86.85 |
| [186] | English | NA | Document/ Sentence | Unsupervised/ Statistical parsing | P = 93, R = 85, F = 89 |
| [94] | English | 46 | Sentence | SVM, NB | A = 65, 62 |
| [48] | English | 1,000 | Sentence | Rule based lexicon approach | P = 92, R = 91, F = 91 |
| [12] | English | 89 | Sentence | SVM | A = 61.6 |
| [209] | English | 195 | Sentence | SVM | A = 80 |
| [143] | Spanish | NA | Sentence | Lexicon based/ Machine learning | A = 83.27 |
| [23] | Polish | NA | Sentence | Shallow parsing | A = 74.49 |
| [85] | Arabic | 3,632 | Sentence | Unsupervised | A = 98.62 |

**Table 2.7: State-of-the-art sentiment analysis which include idioms as features**

### 2.3.2 Feature Extraction

The second step in classifying the overall sentiment of a text segment is to extract the features which are described in Section 2.3.1. Feature extraction methods can be divided into two types: Bag-of-Words (BOW) and statistical methods. The BOW feature selection technique treats a text segment as a group of unrelated words, and identifies features that correspond to a specialised dictionary of known sentiment words and phrases. Statistical methods automatically extract textual features based on a numerical weighting. We discuss such methods in Section 2.3.2.1 and Section 2.3.2.2 respectively.

#### 2.3.2.1 Bag-of-Words

In sentiment analysis, a traditional method for extracting features is by representing text as an unordered group of words, often referred to as bag-of-words (BOW). To determine the sentiment attached to individual words, this method relies on specialised dictionaries (e.g. SentiWordNet [53]) in which words are pre-annotated with their polarity. For instance, the word 'beautiful' is considered to be positive, whereas 'horrible' is considered to be negative.

This feature selection technique ignores the internal structure or relationships between the linguistic units. Consider for example the individual words which construct the sentence "I feel happy today". The word 'happy' can be identified as a feature, based on its presence in both the text itself and in SentiWordNet [53], a lexicon of positive and negative sentiment words.

BOW features are traditionally used in lexicon-based approaches to classify sentiment in text. We discuss such classification, as well as examples of lexical resources used to support this feature extraction technique, in Section 2.4.1 and Section 2.4.1.1 respectively. Additionally, we further reinforce the importance of idioms as features of

sentiment analysis, and demonstrate that they pose a challenge for traditional lexicon-based sentiment analysis approaches.

#### 2.3.2.2 Feature Weighting

The second method of extracting features from text is to use a statistical approach. Statistical feature selection methods are used to automatically identify words of importance in text, based on a numerical weighting [123]. There exists several statistical measures for assigning weights to features in text:

- Feature frequency - the weight of the feature is the number of its occurrences in the text segment.
- Feature presence - the weight is binary, therefore taking the value of 0 or 1 based on the feature's absence or presence in the text segment.
- TF-IDF - a numerical statistic which reflects how important a feature is to document in a corpus
- Chi-square ($\chi^2$) - a measure for modelling the dependency between the features and the classes.
- Point-wise Mutual Information - a measure for modelling the mutual information between the features and the classes.
- Position information - a measure for determining the position of a feature within a text segment.

These feature selection methods have been shown to be effective for sentiment classification [110]. Term frequencies have been important in standard information retrieval. However, O'Keefe & Koprinska [142] compare the impact of different statistical measures for extracting features of sentiment analysis, and conclude that the feature presence method is the best among the others. That is, binary-valued feature vectors, in

which the entries merely indicate whether a word occurs (value 1) or not (value 0), formed a more effective basis for polarity classification, in comparison to feature vectors in which entry values increase with the occurrence of the corresponding word. This may indicate that for classification, the overall sentiment may not be highlighted through repeated use of the same terms [150].

This may also be the case with idioms. An obstacle in systematically investigating the role of idioms in sentiment analysis is their relative rarity. Thus, corpora commonly used for evaluation of sentiment analysis approaches are biased in their use of idioms, which prevents the findings on the role of idioms in sentiment analysis from being generalised. Nevertheless, it has been indicated that the presence of idioms, and not their frequencies, is enough to drive the sentiment of a text segment [170].

## 2.4 Classification of Sentiment

Sentiment analysis is considered to be a classification problem in which, given a text segment, the task is to automatically classify its sentiment as falling under one polarity category (i.e. positive, negative, or neutral). State-of-the-art sentiment classification often uses a lexicon based approach or machine learning techniques to accomplish this task. Lexicon-based approaches classify the overall sentiment of a text segment based on the orientations of their BOW features. A more sophisticated way of automatically classifying sentiment is by using machine learning. There are two state-of-the-art machine learning approaches to sentiment analysis: supervised and unsupervised. We summarise some examples of state-of-the-art sentiment classification in Table 2.9.

### 2.4.1 Lexicon-Based Approach

In Section 2.3.2.1, we discuss how text can be represented as a BOW. To perform sentiment classification on such representations, the lexicon-based approach uses an

algorithm to classify the overall sentiment of a text segment, by aggregating the sentiments of individual words as they were retrieved from the lexicon [210].

This method is a simple, yet naïve, approach to determine the overall sentiment of a text segment. One of the issues faced by using BOW or statistical methods for extracting features from text, is that unless they are explicitly encoded, features that hold their meaning as a single semantic unit may be overlooked. This is often the case when idioms, whose meaning is often figurative, are used to express sentiment in text.

To further reinforce this point, consider, for example, SentiStrength [203, 204], a state-of-the-art, rule based algorithm, which simultaneously extracts positive and negative sentiment from short texts, by using a lexicon of sentiment words with associated strength values. In Table 2.6, SentiStrength is demonstrated to incorrectly classify half of the examples (S1 and S4). If we explore these results in more detail, the reason behind why S1 and S4 are classified as neutral is intuitive under the BOW representation. Whereas S2 and S3 are classified as positive and negative, based on the presence of the words 'grinning' and 'bored' in the idioms *grinning from ear to ear* and *bored to tears* respectively, S1 and S4 do not contain sentiment words. For illustrative purposes, consider the following example, where the words which construct S1 are considered independently:

**Input:**

I was over the moon when I heard the news.

**Analysis:**

`I`[0] `was`[0] `over`[0] `the`[0] `moon`[0] `when`[0] `I`[0] `heard`[0] `the`[0] `news`[0]

**Output:**

result = 0, positive = 1, negative = -1

Frozen idioms and their associated sentiments may be explicitly encoded into lexicons. However, for idioms that are syntactically productive, i.e. they can be changed syntactically without losing their figurative meaning; this task becomes more challenging

as all possible syntactical changes need to be considered. Consider for example S2, where the verb 'grin' in the idiom *grinning from ear to ear*, is inflected. The idiom would not be recognised if its canonical form alone, i.e. *grin from ear to ear*, was explicitly encoded in the lexicon.

Furthermore, traditional lexicon based approaches are faced with the challenge of disambiguating the sense of an idiom in a given context. Consider for example,"Our relationship has been *on the rocks* recently," which expresses a negative sentiment, and "There are sea lions sleeping on the rocks." which expresses a neutral one. With no rule base for differentiating that the idiom *on the rocks* is used figuratively in the first example and literally in the second, the idiom may be incorrectly recognised as expressing a negative sentiment in both sentences.

#### 2.4.1.1 Lexical Resources

There exists several, well established, sentiment lexicons that have been used to support state-of-the-art sentiment analysis and emotion classification. Such resources are used to extract sentiment words from text segments, with the goal often being to represent text segments as BOW representations (discussed in Section 2.4.1).

A prominent resource used in NLP is WordNet [125], a lexical database of English nouns, verbs, adjectives, and adverbs, which are grouped together into sets of interlinked synonyms, referred to as synsets. WordNet has been used as a lexical resource for many text mining applications (e.g. [2, 55, 183]).

SentiWordNet [53] is an opinion lexicon derived from the WordNet database, where each term is associated with numerical scores indicating positive and negative sentiment information (e.g. 'happy' is scored as PosScore = 0.875 NegScore = 0.0 , whereas 'sad' is scored as PosScore = 0.125 NegScore = 0.75). It has been used as a lexical resource for extracting sentiment words from diverse texts in several sentiment analysis studies (e.g. [44, 31, 140, 198]).

One of the main lexical resources used to identify and extract emotions from text [160], is WordNet-Affect (WNA) [196]. Also derived from the WordNet database, WNA is a lexical model of affects, i.e. moods and situations eliciting emotions or emotional responses, providing direct (e.g. *joy, sad, happy,* etc.) and indirect (e.g. *pleasure, hurt, sorry*, etc.) affective terms. It was formed by selecting, assigning and linking a subset of WordNet synsets whose sense corresponds to an affect, and organising them into a hierarchy (see Figure 2.6 for an excerpt from this hierarchy). With a total of 1,903 words, WNA is a prominent lexical resource which has been used to support several sentiment analysis and emotion classification studies (e.g. [13, 195, 82]).

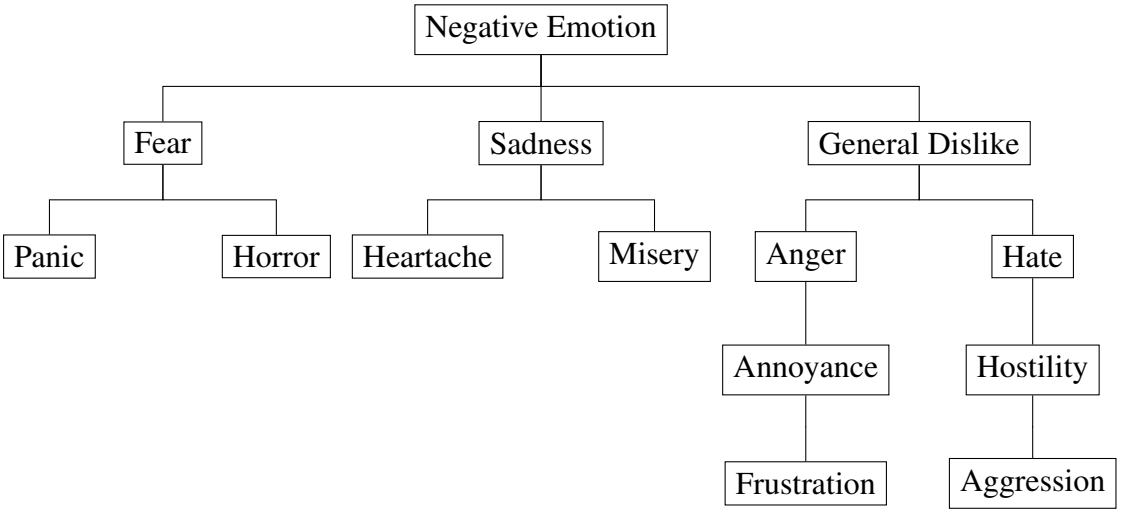


**Figure 2.6: An excerpt from the WNA hierarchy**

Several other lexical resources have been used to support the extraction of sentiment and emotion in state-of-the-art sentiment analysis and emotion classification. For example, EmoLex [128] and The General Inquirer [194] are lexicons compiled of more than 24,000 and 11,788 affective word senses respectively. Such resources were used to support sentiment analysis studies, such as [46, 76, 90, 231]. Other resources, such as Wiebe's et al. [223] subjectivity lexicon of 8,000 words, Whissell's [218] Dictionary of Affect of Language, and Linguistic Inquiry and Word Count (LIWC) [154], have been used to detect sentiment at both sentence and phrase level (e.g. [32, 95, 2, 208]).

### 2.4.2 Machine Learning Approaches

In machine learning, classification is the problem of identifying to which of a set of categories (i.e. sentiment polarity) an unseen instance belongs, based on a training dataset. There are two state-of-the-art machine learning approaches to sentiment analysis: supervised and unsupervised. Supervised machine learning relies on an annotated training dataset, where the category to which an instance belongs is known. The unsupervised approach is known as clustering. It involves grouping training instances, whose categories are unknown, into categories based on some measure of inherent similarity or distance between the features they contain. For both techniques, unseen instances, often referred to as testing data, are classified by comparing them to training instances using algorithms, referred to as classifiers.

#### 2.4.2.1 Supervised Machine Learning Approaches

In order to predict which category an unseen test instance belongs to, supervised learning methods depend on the existence of annotated training datasets. Training data usually consists of two parts: predictors and target values. Predictors are the text instances to be classified. They are often represented as feature vectors - an n-dimensional vector of numerical values that represent the presence of features in text. Target values, often referred to as class labels, are the known categories that an instance belongs to. There are several ways in which the class label of an instance can be determined.

##### 2.4.2.1.1 Annotating Data

Traditionally, supervised learning approaches often turn to humans to manually annotate corpora. In sentiment analysis, text segments are annotated with polarity categories to represent the sentiment that is being expressed.

A key component of any annotation task is the use of annotators [163]. Generally, recruiting a number of reliable and experienced annotators, who understand the task

and its difficulty, and are willing to contribute to such a time consuming task, is a challenge. With the proliferation of Web 2.0, the crowdsourcing of annotation tasks has become a popular method of obtaining gold standards to support supervised machine learning for a variety of text classification studies, including sentiment analysis (e.g. [162, 198]). Online crowdsourcing platforms, such as Amazon's Mechanical Turk[9] and CrowdFlower[10], are specific resources where people who require human intelligence for tasks such as annotating natural language texts, can be fulfilled by experienced annotators from across the world. Annotation tasks are generally formatted to be quick and relatively easy to perform, and workers are often paid for their contributions.

With no ground truth, several annotators are required to participate. As human judgements vary, the concern with manual annotations is their validity. Therefore, the methodology behind annotating data, is that to produce a reliable gold standard for classification, annotators need to demonstrate agreement. If different annotators produce consistently similar results, it can be inferred that they have internalised a similar understanding of the annotation guidelines, and that they have performed consistently under this understanding [10]. Thus, if agreement is high, the dataset is reliable for training. Annotator agreement can be measured by coefficients of agreement, referred to as Inter-Annotator Agreement (IAA), such as Cohen's Kappa [34] and Krippendorff's alpha coefficient [99].

Conversely, if various annotators disagree, the annotators need to be re-trained or replaced. With efficiency and cost-effectiveness, online recruitment of anonymous annotators has its disadvantages. Factors such as annotators' skills and focus, the clarity of the annotation guidelines, and a lack of specific training, may contribute to annotations being unreliable.

#### 2.4.2.1.2 Supervised Machine Learning Classifiers

[9] https://www.mturk.com

[10] https://www.crowdflower.com/

Once the annotated corpus has been evaluated, and IAA has demonstrated that the dataset is reliable, the ground truth for a given instance is adjudicated, often by choosing the majority annotation, forming a gold standard. A gold standard is often split into training and testing data. A supervised classification algorithm is then used to predict the class label of unseen test instances, based upon their correspondence with the training data. Several supervised classifiers exist. But the question of which classifier is the best performing for sentiment classification is left unanswered. The "no free lunch" theorem suggests that there is not a universally best learning algorithm [228]; in other words, the choice of an appropriate classification algorithm should be based on its performance for the particular problem at hand, and the properties of data that characterise that problem.

For sentiment analysis, Pang, Lee & Vaithyanathan [151], who conducted one of the very first empirical studies to classify opinions expressed in film reviews, evaluated three classifiers: Naïve Bayes, Maximum Entropy, and Support Vector Machines (SVM).

Naïve Bayes classifiers are simple probabilistic models based on Bayes' theorem. These models assume conditional independence, i.e. each individual feature, independent of other features, is assumed to be an indication of the assigned class. The goal is to analyse the relationships between features and the class, to estimate a conditional probability that correspond unseen instances to their sentiment.

Maximum Entropy classifiers are alternative probabilistic techniques which have proven effective in sentiment classification tasks (e.g. [151, 68]). Unlike Naïve Bayes, Maximum Entropy does not assume conditional independence amongst features. Instead, it is based on the Principle of Maximum Entropy. From all the models that fit the training data, it selects the model with the largest entropy.

SVM classifiers are known for their high performance, and have been widely used in sentiment classification problems (e.g. [90, 142, 151, 16]). They are non-probabilistic discriminative models which construct a hyperplane that optimally separates the train-

ing data into two classes.

As these supervised classifiers achieved high sentiment classification performances (see Table 2.9 where the average accuracy is 80%), they have had a prominent following in this domain, and have been used in several state-of-the-art sentiment classification studies. In order to gain a full understanding of these classifiers and how they are used in sentiment classification, see [150].

#### 2.4.2.1.3 Evaluation Measures

In text classification, evaluating the performance of a classifier concerns measuring its effectiveness, rather than its efficiency. As a result, the classifier's ability to correctly predict the category of an unseen instance is evaluated, and not its computational complexity [182].

Given a test dataset, sentiment classification is evaluated relative to the training dataset, producing four outputs: instances that are predicted as being positive, when they are indeed positive (True Positive (TP)), instances that are predicted as being negative, when they are indeed negative (True Negative (TN)), instances that are predicted as positive, when in fact, they are negative (False Positive (FP)), and instances that are predicted as negative, when in fact, they are positive (False Negative (FN)). These four counts constitute a confusion matrix shown in Table 2.8.

| | | **Predicted** | |
|---|---|---|---|
| | | **Positive** | **Negative** |
| **Actual** | **Positive** | True positive | False negative |
| | **Negative** | False positive | True negative |

**Table 2.8: Confusion matrix for binary classification**

There are several evaluation techniques which are used to evaluate how well a classifier performs on unseen test data. The most common measures used in text classification are precision, recall, F-measure, and accuracy. Often, the goal is to maximise

all measures, which range from 0 to 1. Therefore, higher values correspond to better classification performance.

Precision and recall are two metrics that are often simultaneously used to verify the performance of information retrieval, but are common statistics used in text classification. "Classic" precision and recall are derived from the ratios of relevant documents and non-relevant documents. They also consider the relevant documents that are not retrieved. More specifically, precision measures the number of retrieved documents that are relevant, whilst recall measures the number of all the relevant documents that are successfully retrieved. Both metrics can be calculated using the equations in (2.1).

$$Precision = \frac{TP}{TP + FP} \qquad Recall = \frac{TP}{TP + FN} \tag{2.1}$$

Precision and recall are not often taken into account alone. The two measures are often used together in the F-measure, which provides a single weighted metric to evaluate the overall performance. F-measure can be measured by calculating the harmonic mean of precision and recall (Equation (2.2)).

$$F - measure = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{2.2}$$

Others use accuracy as a metric to measure the performance of a classifier. Accuracy (Equation (2.3)) measures the number of instances that were correctly classified. However, the problem of using accuracy to measure the effectiveness of a classifier is that if the classifier always predicts one class, a strategy that defeats the purpose of building a classifier, it will achieve high accuracy. This is known as the accuracy paradox.

$$Accuracy = \frac{TP + TN}{TP + FN + FP + TN} \tag{2.3}$$

We summarise some examples of state-of-the-art sentiment classification, which use these measures to evaluate classification performance, in Table 2.9.

#### 2.4.2.2 Unsupervised Machine Learning Approaches

Unsupervised approaches to machine learning differ significantly from supervised approaches. As discussed in Section 2.4.2.1, supervised learning requires a large dataset of annotated texts to train classifiers. However, in reality, annotated training data are limited, and are not always available when building new classification models. Unsupervised learning techniques are applied in sentiment analysis to overcome this issue.

Unsupervised learning uses statistical inferences alone to learn structured patterns from unlabelled data. Without prior linguistic information regarding the sentiment being expressed, a number of unsupervised approaches take advantage of predefined lexicons (see Section 2.4.1.1 for examples) to extract sentiment features. In general, instances that exhibit a distinct similarity in the features they contain are grouped together, subsequently classifying similar texts with the same class. Early examples of such approaches include [76, 231, 48, 81, 198, 92, 210].

There are various methods to estimate the similarity between features within text segments, such as cosine similarity. However, traditional unsupervised approaches (e.g. [107, 201]) use clustering algorithms, such as k-means and k-nearest neighbour, to find groups of related texts based on their similarities.

However, idioms still pose a challenge for sentiment classification using traditional unsupervised learning techniques. As discussed in Section 2.4.1, without a rule base for recognising idioms and disambiguating their figuration in text, relying on a dictionary of sentiment words is a naïve approach to including idioms as features of sentiment analysis.

## 2.5 Summary

In this Chapter, we have explored the subject of sentiment analysis. To gain a full understanding of the topic, we discussed the aims and objectives of sentiment analysis, as

well as providing context for its applications. Additionally, we outlined definitions of sentiment, as well as other related terms - opinion, subjectivity, and emotion. We explained the relationship between two main sub-tasks of sentiment analysis: subjectivity analysis and opinion mining. We also discussed emotion analysis, a sub-task of sentiment analysis which aims to recognised and classify emotions expressed in text.

Furthermore, we discussed how the related notions of sentiment and emotion are represented for the purposes of sentiment analysis. We paid particular attention to how emotion classification problems use frameworks of emotion (e.g. happiness, sadness, anger, etc.) to represent and classify emotions and related states. We also reviewed the textual features used to identify sentiment in text, as well as the means of extracting these features. We paid particular attention to idioms, their properties, and their role in sentiment analysis. Lastly, we reviewed both state-of-the-art machine learning approaches to sentiment analysis, as well as methods of evaluating their performances.

To summarise, the main highlights of this Chapter are as follows:

- The importance of idioms as features of sentiment analysis is emphasised by the fact that they typically imply an affective stance towards an entity or an event, as opposed to a neutral one.
- However, to the best of our knowledge, thus far, there is no comprehensive knowledge base that systematically maps English idioms to their sentiments. This is the main reason why idioms are under-represented as features of sentiment analysis.
- We reinforce the importance of idioms as features of sentiment analysis, by demonstrating that current state-of-the-art sentiment analysis tools, which do not consider idioms as features, often result in incorrect classifications when they are explicitly used to express sentiment.
- The lexicon-based approach often neglects idioms when they are explicitly used

to express sentiment. The reason behind this is because idioms hold their meaning, and thus, their overall sentiment, as a single semantic unit.

- In order to be included as features of sentiment analysis, idioms need to be recognised in text. Two important points must be considered:
  - As some idioms are syntactically flexible, their occurrences can be automatically recognised using lexico-syntactic patterns.
  - Idioms hold a literal and figurative meaning, therefore, their senses need to be disambiguated in a given context.
- The question of which framework of emotion is best suited for sentiment analysis is left unanswered. Several frameworks for recognising and classifying emotions, and their related states, have been used to support emotion classification. There are three main frameworks: categorical frameworks which represent distinct emotions (e.g. happiness, sadness, anger, etc.), dimensional frameworks which represent emotions in terms of dimensions (e.g. arousal and valence, positive and negative, etc.), and hierarchical frameworks which distribute richer sets of emotions across a hierarchical structure. We investigate the utility of each type of framework for sentiment analysis in this thesis.

We move forward with the knowledge gained in this Chapter, to choose suitable methods and resources for our experiments in the remainder of this thesis, which address our hypotheses discussed in Chapter 1.

Naïve Bayes (NB), Support Vector Machines (SVM), Maximum Entropy (ME), Precision (P), Recall (R), F-measure (F), Accuracy (A)

| Reference | Domain | Dataset | Size | Level | Features | Approach | Evaluation (%) |
|---|---|---|---|---|---|---|---|
| [151] | Film reviews | IMDb | 2,053 | Sentence | Unigrams, Bigrams, POS, Term position | NB, SVM, ME | A = 81.0, 82.9, 80.4 |
| [149] | Film reviews | Rotten Tomatoes | 2,000 | Document | Unigrams | NB, SVM | A = 86.4, 86.2 |
| [219] | Film reviews | IMDb | 2,000 | Document | BOW, Appraisal groups | SVM | A = 90.2 |
| [90] | Film reviews | Rotten Tomatoes | 2,000 | Document | Unigrams, Bigrams | SVM | A = 85.9 |
| [11] | Film reviews | Tweets | 2,890,000,000 | Sentence | URLs, Retweets | DynamicLMClassifier | A = 98 |
| [132] | Film reviews | IMDb | 1,380 | Document | Unigrams | SVM | A = 86 |
| [1] | Film reviews | IMDb | 2,000 | Document | Unigrams, Bigrams, N-grams | SVM | A = 95 |
| [81] | Product reviews | ClNet and Amazon | 500 | Sentence | POS, Term frequency | Unsupervised | P = 64.2, A = 84.2 |
| [48] | Reviews | Amazon | 445 | Sentence | Unigrams, N-grams, POS, Idioms | Rule based lexicon approach | P = 92, R = 91, F = 91 |
| [198] | Reviews | Epinions | 5,100 | Sentence | Unigrams, Negation, Intensifiers, POS | Unsupervised | A = 78.74 |
| [40] | Reviews | ClNet and Amazon | 37,494 | Sentence | Unigrams, N-grams | SVM, NB | A = 85.3 |
| [60] | Reviews | Global support services | 40,884 | Document | Bigrams, POS | SVM | A = 85.47 |
| [210] | Reviews | Epinions | 410 | Document | POS,,Unigrams, Sentence length | Unsupervised | A = 74 |
| [68] | Twitter data | Tweets | 1,600,000 | Sentence | Unigrams, Bigrams, POS | NB, ME, SVM | A = 82.7, 82.7, 81.6 |
| [148] | Twitter data | Tweets | 300,000 | Sentence | N-gram, Term presence, Term frequency | NB | A = 70 |
| [98] | Twitter data | Tweets | 607,966 | Sentence | Unigrams, Bigrams, POS, BOW, Emoticons, Abbreviations, Punctuation | AdaBoost.MH | A = 75 |
| [141] | Finance | Financial blogs | 232 sources | Sentence/ Word | BOW | SVM, NB | A = 66.06, 69.54 |
| [96] | Finance | Multex Significant Developments corpus | 12,000 | Sentence | Unigrams, Term presence | SVM | A = 65.9 |
| [133] | Politics | Politics.com | 77,854 | Word | Unigrams, Political features | NB | A = 60.37 |
| [4] | Health care | Rate my MD | 995 | Word | N-grams, Term presence | SVM, NB | F = 89, 94 |
| [214] | Public action | Tweets | 110,715 | Sentence | Emoticons, BOW, N-grams, Stop words | NB | F = 87.8 |
| [221] | News | Wall Street Journal Treebank | 1,004 | Sentence | Term presence, POS | NB | A = 72.17 |
| [231] | News | Wall Street Journal Treebank | 8,000 | Sentence/ Document | Unigrams, Bigrams, POS, BOW | NB | A = 91 |
| [222] | News | Wall Street Journal Treebank | 1,289,006 words | Sentence | Unigrams, Term frequency, POS | k-nearest neighbour | A = 94 |

**Table 2.9: State-of-the-art sentiment analysis**

# *Chapter 3*

# Idioms as Features of Sentiment Analysis

*"Raring to go."*

In this Chapter, we aim to estimate the degree to which the inclusion of idioms as features may improve the results of traditional sentiment analysis. More specifically, this Chapter is divided into the following main sections: Section 3.1.1 and Section 3.1.2 discuss the construction of a comprehensive dataset of emotionally charged idioms, as well as the construction of a comprehensive corpus of contextual examples of such idioms respectively. In Section 3.2, in order to support idioms as features in sentiment analysis, and to create a gold standard for classification experiments, individual idioms, as well as their contextual examples, were manually mapped to their sentiments. Additionally, in order to incorporate idioms as features, their occurrences need to be recognised in text. In this case, Section 3.3 discusses how idioms were modelled using regular expressions, whilst also taking into consideration negated and multiple idiom occurrences. In Section 3.4, we implement a sentiment analysis approach that incorporates idioms as features into two traditional sentiment analysis approaches, which we use as comparative baselines. Section 3.5 evaluates the classification results of our approach against the baselines. Finally, Section 3.6 summarises our findings.

## 3.1 Constructing an Idiom Corpus

Intuitively, our experiments in this Chapter require a comprehensive collection of idioms. In Section 3.1.1, we discuss the construction of a dataset of emotionally charged idioms. Additionally, in order to create a gold standard for classification experiments, in Section 3.1.2, we discuss the construction of a comprehensive corpus of idioms used in context.

### 3.1.1 A Selection of Emotionally Charged Idioms

Idioms and the properties they hold are discussed in Section 2.3.1.5.1. To reiterate, one of the main properties for distinguishing idioms from related linguistic categories is that most hold a figurative meaning conventionally understood by native speakers, which poses a challenge for English language learners. Failure to understand figurative idioms used in context can significantly effect one's understanding of language in a variety of personal and professional situations [138]. It is therefore unsurprising that most syllabi for English as a second language pay special attention to studying idioms, and as a result, there is an abundance of dedicated teaching material [112].

The idioms used in the experiments in this Chapter were collected from an online educational resource - Learn English Today[1], which organises individual idioms, along with their definitions, by themes, many of which can be directly (e.g. Happiness/Sadness) or indirectly (e.g. Success/Failure) mapped to an emotion. The focus here is specifically on emotion-related idioms, as it is anticipated that they have a substantial impact on sentiment analysis.

A total of 16 out of 60 available themes were selected, listed in Table 3.1, together with the number of associated idioms. A total of 580 idioms were collected.

[1] http://www.learn-english-today.com/idioms/idioms_proverbs.html

| Theme | Total | Theme | Total |
|---|---|---|---|
| Anger/Annoyance | 45 | Mistakes/Errors | 5 |
| Anxiety/Fear | 14 | Politeness | 8 |
| Arguments/Disagreements | 37 | Problems/Difficulties | 57 |
| Enthusiasm/ Motivation | 10 | Safety/Danger | 27 |
| Feelings/Emotions | 48 | Sleep/Tiredness | 11 |
| Fun/Enjoyment | 22 | Success/Failure | 84 |
| Happiness/Sadness | 21 | Surprise/Disbelief | 16 |
| Madness/Insanity | 11 | Violence | 6 |

**Table 3.1: Distribution of idioms across emotional themes**

### 3.1.2 Constructing a Corpus of Idioms Used in Context

One of the obstacles faced in systematically investigating the role of idioms in sentiment analysis is their relative rarity. Corpora commonly used for evaluation of sentiment analysis approaches are biased in their use of idioms, which prevents the findings on the role of idioms in sentiment analysis from being generalised.

The corpus of choice for the experiments in this Chapter is the British National Corpus (BNC) [19, 106], a large text corpus of both written and spoken English, compiled from a variety of sources. It has been the corpus of choice for several computational linguistic and NLP studies, including those focused on idioms (e.g. [70]). As such, the BNC was used to assemble a corpus of idioms used in context for reasons including:

- size, range, and representativeness - the BNC is a finite and balanced corpus made of 100 million words or informative and imaginative English written texts (90%), as well as conversational and task-oriented (e.g. lectures, TV broadcasting, commentaries, etc.) spoken English (10%).
- recency - most of the texts are from the period 1985-94.
- availability - the BNC is available online for research purposes.

- relevance - the focus is on British English idioms.

The BNC can be searched using its online search function for words or phrases, and can return up to 50 random sentences for each query. In order to find examples of idioms used in context, the BNC was searched for content words found in the idioms from the dataset described in Section 3.1.1. The results containing these expressions were manually matched to an idiom, resulting in a dataset of 2,521 sentences. A maximum of 10 sentences were selected for each idiom. Sentences were collected for a total of 423 idioms from the original dataset. Contextual examples of the remaining 157 idioms were not available in the BNC. The mean and median average number of sentences extracted for an idiom were both 6, with standard deviation of 3.39. Table 3.2 summarises the number of sentences collected for each theme associated with the idioms.

| **Theme** | **Total** | **Theme** | **Total** |
|---|---|---|---|
| Anger/Annoyance | 261 | Mistakes/Errors | 31 |
| Anxiety/Fear | 88 | Politeness | 42 |
| Arguments/Disagreements | 232 | Problems/Difficulties | 360 |
| Enthusiasm/ Motivation | 41 | Safety/Danger | 176 |
| Feelings/Emotions | 280 | Sleep/Tiredness | 64 |
| Fun/Enjoyment | 107 | Success/Failure | 519 |
| Happiness/Sadness | 128 | Surprise/Disbelief | 92 |
| Madness/Insanity | 47 | Violence | 50 |

**Table 3.2: Distribution of sentences across emotional themes associated with idioms.**

In most cases, expressions within the sentences have a figurative meaning associated with an idiom, whereas others convey a literal sense. Consider for example the following two sentences extracted for the expression *in the bag*, the figurative meaning of which is “to be virtually secured”:

"The Welsh farmer's son had the 1988 conditional jockeys' title already *in the bag*."

"I looked *in the bag*, it was full of fish."

In this case, some sentences may be FP. From a lexico-syntactic perspective, most idioms can be modelled with local grammars. However, it is more difficult to automate their recognition from a semantic perspective. It is necessary to include FPs in the corpus, in order to evaluate how incorrectly recognised idioms may affect the results of sentiment analysis.

## 3.2 Crowdsourcing of Sentiment Annotations

In order to incorporate idioms as features in sentiment analysis, their associated sentiments need to be explicitly encoded. Section 3.2.1 describes the choice of annotation frameworks used for annotating idioms, as well as their contextual examples, with their sentiments. In Section 3.2.3, we measure the reliability of our annotated datasets, by measuring IAA. Finally, in Section 3.2.4, we discuss how the annotated corpus of idioms used in context was used to form a gold standard for classification experiments.

### 3.2.1 Annotation Scheme

Taking a long-term view of the research in this thesis, we want to address the limitations of state-of-the-art sentiment analysis approaches by focusing on a full range of emotions, rather than merely sentiment polarity. In this case, we require a comprehensive, but practical, emotion classification framework. Having considered the literature discussed in Section 2.2.2, we based the annotation framework on the EARL [57], an XML-based language for representing emotions in technological contexts.

To reiterate, EARL organises 48 emotions into 5 positive and 5 negative categories (see Table 2.2). To facilitate the annotation task in this Chapter, we used the 10 top-level categories (e.g. *positive & lively, caring, negative & forceful, agitation,* etc.) as they provide a manageable number of choices for a human annotator, which are also evenly distributed between positive and negative polarities. For annotation purposes, as opposed to formal definitions, specific emotions were used as examples to explain each top-level category. For example, *caring* encompasses affection, empathy, friendliness, and love.

### 3.2.2 Annotation Process

Annotations were distributed by using a crowdsourcing approach. In this case, a bespoke web-based annotation platform was implemented. The simple interface was accessible via a web browser, which eliminated the installation overhead and minimised the need for special training. Online accessibility also provides the flexibility of choosing the physical location for annotation experiments.

The interface consisted of two panes. One pane contained randomly selected text to be annotated; in this case, either an idiom together with its definition, or a contextual example of an idiom. The second pane contained four annotation choices from EARL: *positive, negative, neutral,* and *ambiguous*. The selection of either *positive* or *negative* categories expanded the menu to provide additional choices (see Figure 3.1).

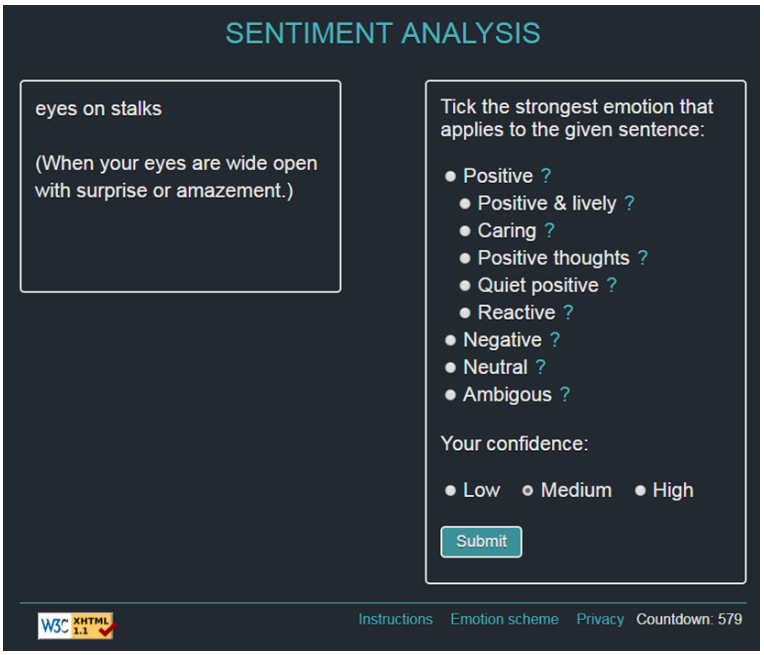


**Figure 3.1: Annotation platform interface**

We introduced a *neutral* category to allow the absence of a sentiment to be annotated explicitly. We also introduced an *ambiguous* category to annotate cases of sentiment that could not be determined as either *positive* or *negative* without additional information such as context, tone of voice or body language.

A help button was available next to each annotation choice to provide additional information about each category. The overall annotation framework could also be viewed in a separate window. Each annotation was scored on a three-point scale to account for the annotator's confidence in a particular choice: Low, Medium or High, with Medium being a default choice.

Group annotation sessions were conducted weekly, where new annotators were briefed about the study and their role as an annotator. All annotators were required to be of native, or native-like, English proficiency. All annotations were performed independently and no discussions about particular data items were allowed among the annotators during the task. The data were randomised individually for each annotator, so they were always annotated in a different order. Users were tracked by their IP addresses to avoid

duplication of annotations, not to identify individuals, and no other personal information was collected. All annotation results were stored securely in a relational database.

### 3.2.3 Annotation Results

A total of 18 annotators participated in this task. A total of 2,900 annotations were collected for all 580 idioms described in Section 3.1.1, with 5 annotations per idiom.

A total of 8,610 annotations were collected for all contextual examples of idioms described in Section 3.1.2, with at least 3 annotations per sentence. A total of 143 sentences had a maximum of 5 annotations. Overall, the mean and median average number of annotations per sentence were both 3, with a standard deviation of 0.60.

In order to compare this method to existing sentiment analysis approaches, which often classify text in terms of sentiment polarity, our experiments need to conform to the same classification framework. To create a gold standard that can be compared against a baseline, the specific categories in the EARL were projected onto *positive* and *negative* polarity, and *neutral* and *ambiguous* categories were merged into a single category called *other*. Nonetheless, we will be able to use the original annotations to re-train the machine learning method described in this Thesis to support emotion classification against the categories described in EARL, as part of the future work discussed in Chapter 6.

After projecting all annotations as sentiment polarity, we measure the reliability of the training dataset by measuring IAA using Krippendorff's alpha coefficient [99]. As a generalisation of known reliability indices, this measure was chosen as it applies to: (1) any number of annotators, not just two, (2) any number of categories, (3) incomplete or missing data, and (4) corrects for change expected agreement [100]. Krippendorff's alpha coefficient is calculated according to Equation (3.1):

$$\alpha = 1 - \left(\frac{D_o}{D_e}\right) \tag{3.1}$$

where $D_o$ is the observed disagreement, i.e. the proportion of items on which both annotators agree, and $D_e$ is the disagreement expected when annotations are given at random. Krippendorff suggests $\alpha$ = 0.667 as the lowest acceptable value to consider data as a reliable training set [100]. The values for Krippendorff's alpha coefficient were obtained using an online tool for calculating IAA [61]. The agreement on the idiom dataset was calculated as $D_e$ = 0.606, $D_o$ = 0.205, $\alpha$ = 0.662. The agreement on the corpus of idioms used in context was calculated as $D_e$ = 0.643, $D_o$ = 0.414, $\alpha$ = 0.355.

The relatively high agreement ($\alpha$ = 0.662) on idioms alone illustrates that they can be reliably mapped to their sentiment polarity. Significantly lower agreement ($\alpha$ = 0.355) on contextual examples of idioms, however, illustrates the complexity of sentiment interpretation, where a range of different emotions may often be conveyed in a single sentence. For example, "Brian, as ever, decided not to *sit on the fence"* received one *positive thoughts*, one *negative & forceful*, one *neutral*, and one *ambiguous* annotations.

### 3.2.4 Gold Standard

Annotated contextual examples of idioms were used to create a gold standard for sentiment analysis experiments. For each sentence, an annotation agreed by the relative majority of 50% of the annotators was adjudicated as the ground truth.

Prior to calculating the IAA discussed in Section 3.2.3, additional annotations from a new, independent annotator were sought to resolve any sentences with disagreeing annotations. A total of 282 additional annotations were collected. Table 3.3 shows the distribution of ground truth annotations across the three categories, together with an annotated example from each. A random subset of 500 sentences (20% of the dataset) was selected for testing, with the remaining 2,021 sentences used for training a classifier.

| Annotation | Total | % | Example |
|---|---|---|---|
| Positive | 677 | 26.9 | I shall *go the extra mile*. |
| Negative | 1,219 | 48.4 | All right, don't *jump down my throat*. |
| Other | 625 | 24.8 | Your mother used to *sleep like a log*. |

**Table 3.3: Distribution of annotations in the gold standard**

## 3.3 Recognising Idioms in Text

In order to incorporate idioms as features in sentiment analysis, their occurrences need to be recognised in text. In Section 3.3.1, we discuss modelling idioms using regular expressions. Section 3.3.2 discusses how negated idiom occurrences are also considered in this modelling.

### 3.3.1 Modelling Idioms with Regular Expressions

One of the main properties to consider when including idioms as features of sentiment analysis is their syntactic flexibility. When used in discourse, the syntax of many idioms may be restricted, i.e. they do not vary much in the way they are composed [139, 70]. Therefore, frozen idioms can be recognised by simple string matching, or encoded in sentiment analysis approaches that use lexicons. Less often, idioms are syntactically productive, i.e. they can be changed syntactically without losing their figurative meaning, e.g. "John *laid down the law*," can be passivized to "*the law was laid down* by John," while retaining the original figurative interpretation that John enforced the rules [67]. Idioms may also undergo more complex syntactic changes, such as nominalisation (e.g. "you *blew some steam off,"* vs. "you're *blowing off some steam*") [59]. Others may contain syntactic changes, such as inflection (e.g. verb tense change) [232].

Such linguistic phenomena can be modelled by regular expressions, e.g. *spill*[*s*|*t*|*ed*]

*the beans*. More complex idioms, which have variables for open argument places [87] (e.g. *put someone in one's place*), can still be modelled by means of lexico-syntactic patterns (e.g. *put* `NP` *in* `PRN's` *place*), and recognised in a linguistically pre-processed text.

In this Chapter, each idiom was computationally modelled by using lexico-syntactic patterns and local grammars [72]. For example, the following grammar:

```
<idiom> ::= <VB> <PRP$> heart on <PRP$> sleeve

<VB> ::= wear | wore | worn | wearing

<PRP$> ::= my, your, his, her, its, our, their
```

was used to successfully recognise the idiom *wear one's heart on one's sleeve* in the following sentence:

> "Rather than *wear your heart on your sleeve* you keep it under your hat."

Idiom recognition rules were implemented as expressions in Mixup (My Information eXtraction and Understanding Package), a simple pattern-matching language [35]. The pattern-matching rules were applied to the test dataset of 500 sentences, in which a single annotator marked up all idiom occurrences, disambiguating their figurative and literal sense, for example:

> "Phew, that was a `<idiom>` *close shave* `</idiom>`."
>
> "He has polished shoes, a `<nonidiom>` *close shave* `</nonidiom>`, and too much pride for a free drink."

A total of 471 sentences (94.2% of the dataset) were marked up to include figurative idiom occurrences, whereas the remaining 29 sentences (5.8% of the dataset) included

literally used idioms. An important point to consider here is that for some human readers, particularly for language learners or for those who are unfamiliar with the meaning of such language use, the skill of interpreting figurative language in text may be challenging [66]. In this sense, some marked up contextual occurrences of idioms may be FP. Nonetheless, the performance of recognising idioms was recorded as precision = 94.4%, recall = 100%, and F-measure = 97.1%. An idiom was considered to be correctly recognised if the suggested text span matched exactly the idiom marked up by the annotator.

### 3.3.2 Negation

As with other phrases, the polarity of idioms can be changed by negation. For example, the polarity of the idiom *jump for joy* is *positive*, but when negated (e.g. "I didn't exactly *jump for joy*"), the overall polarity can be reversed. It is, therefore, essential to consider negation when identifying idioms in context. In order to explicitly recognise negated idioms, pattern-matching rules were implemented based on clues such as negative words (e.g. no, never), negative adverbs (e.g. hardly, barely), and negated verbs (e.g. doesn't, isn't).

The performance of recognising negated idioms was recorded as precision = 86.2%, recall = 92.6%, and F-measure = 89.3%. Given a small number of negated idiom occurrences in the test dataset (25 sentences, i.e. 5% of the dataset), a larger corpus is required in order to better estimate the performance of this negation module.

## 3.4 Including Idioms as Features in Sentiment Analysis

In this Section, we implement a sentiment analysis approach which incorporates idioms as features of two state-of-the-art sentiment analysis approaches: SentiStrength [203,

204] and Stanford CoreNLP's sentiment annotator [189], which we use as baselines for evaluation.

### 3.4.1 Feature Selection

After projecting the original annotations onto sentiment polarity, the 5 annotations collected for each idiom (discussed in Section 3.2) were used to calculate their feature vectors. Each idiom was represented as a triple: (positive, negative, other), where each value represents the percentage of annotations in the corresponding category. For example, the idiom *wear one's heart on one's sleeve* received one *positive* annotation, zero *negative* annotations, and four *other* annotations. It was, therefore, represented as the following triple: (20, 0, 80).

We conducted two experiments in which idiom feature vectors were combined with the results of two baselines. SentiStrength [203, 204], is a state-of-the-art rule-based system that assigns sentiment polarity to a sentence by aggregating the polarity of individual words, and combines these values to predict the overall sentiment. In the first experiment, the output from SentiStrength was used as a feature, and combined with those based on idioms. For example, the sentence "The *party is over*," was analysed as follows:

**Analysis:** `The party[1] is over[-1].`

**Output:** `result = 0, positive = 1, negative = -1`

SentiStrength provides trinary classification outputs, which were converted into a three-dimensional vector to be used as features: (0, 1, -1). In our approach, the phrase *party is over* would be recognised as an idiom, which was annotated, and subsequently mapped to the following triple: (0, 100, 0), denoting that all annotators considered it to be Negative. The two vectors were appended to create a single feature vector for the given

sentence as follows:

$$\underbrace{(0, 1, -1,}_{\text{Sentiment Polarity}} \underbrace{0, 100, 0)}_{\text{Idiom Polarity}} \tag{3.2}$$

In the second experiment, we used a state-of-the-art sentiment annotator distributed as part of the Stanford CoreNLP [189], a suite of core NLP tools. This method uses a deep neural network approach to build up a sentiment representation of a sentence on top of its grammatical structure. In other words, the sentiment is predicted based on the way in which the words are combined into phrases, subsequently classifying sentiment on a 5 point scale: very negative, negative, neutral, positive, and very positive (see Figure 3.2, where "The *party is over*," is classified as *neutral*).

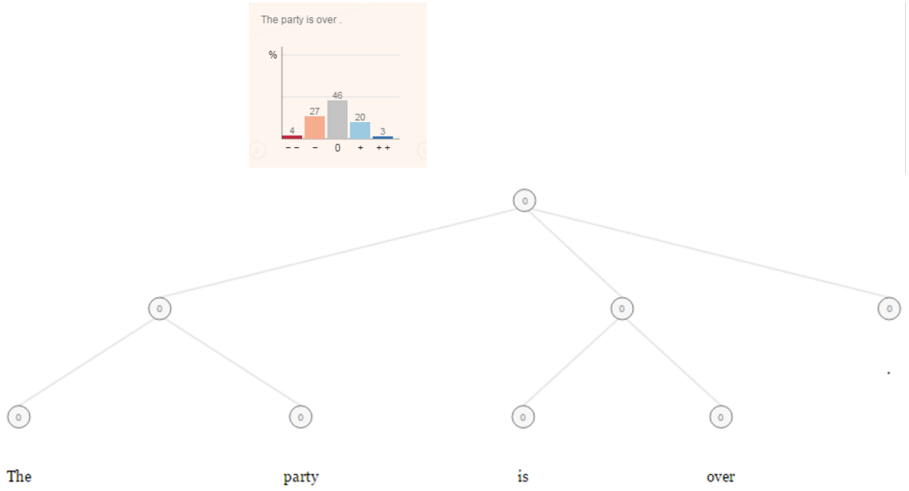


**Figure 3.2: Sentiment analysis results from Stanford CoreNLP**

Stanford CoreNLP's sentiment annotator distributes its probability across its classes. We converted these probabilities into a five dimensional vector, which was used as a feature in our approach. For example, "The *party is over*," was represented as the following vector: (4, 27, 46, 20, 3). As before, the idiom *party is over* would be recognised and mapped to its polarity triple (0, 100, 0), and appended to create a single

feature vector for the given sentence as follows:

$$\underbrace{(4, 27, 46, 20, 3,}_{\text{Sentiment Polarity}} \underbrace{0, 100, 0)}_{\text{Idiom Polarity}} \tag{3.3}$$

If an idiom was recognised to be negated, polarities in the idiom polarity vector were reversed, based on the assumption that negation converts *positive* to *negative* polarity, and vice versa. For instance, consider the sentence "The *party is not over* yet," in which the negatively charged idiom *party is over* is negated. In this particular example, the negation changes the overall polarity from *negative* to *positive*. Thus, the phenomenon associated with negation was modelled by reversing *positive* and *negative* polarities, i.e. the polarity triple for the idiom *party is over*, (0, 100, 0), would be converted to (100, 0, 0). A total of 121 sentences in the gold standard (4.8% of the dataset) were found to include negated idioms.

When multiple idioms occurred, the associated idiom polarity values were aggregated by summing up the polarity vectors, whilst taking into account the effects of negation. For instance, two idioms were recognised in the following sentence:

> "He <idiom>*stopped dead in his tracks*</idiom>, <idiom>*rooted to the spot*</idiom> with horror."

Their polarity triples, i.e. (0, 60, 40) and (0, 40, 60) respectively, were summed up to obtain (0, 100, 100). A total of 29 sentences in the gold standard (1.2% of the dataset) contained more than one idiom occurrence.

Finally, if no idiom was detected, the idiom polarity values were set to zero, i.e. (0, 0, 0). Idioms were not detected in a total of 34 sentences in the gold standard (1.3% of the dataset).

### 3.4.2 Sentiment Classification

Weka [74], a popular suite of machine learning software, was used to perform classification experiments. The choice of a machine learning method was based on the results of 10-fold cross-validation on the gold standard created in Section 3.2.4.

We performed cross-validation experiments by using a variety of classifiers distributed as part of Weka. Table 3.4 demonstrates the results following cross-validation, reporting classification models with the highest performance, including Naïve Bayes, logistic regression, SVM, Weka's implementation of J4.8 decision tree method [165] with no pruning, and a Decision Table majority classifier, each using the default parameters. The overall performance represents weighted-averaged results.

| | SentiStrength | | | Stanford CoreNLP | | |
|---|---|---|---|---|---|---|
| **Classifier** | **P** | **R** | **F** | **P** | **R** | **F** |
| Bayesian Network | 61.6 | 62.5 | 61.9 | 58.1 | 61.7 | 58.6 |
| Naïve Bayes | 60.8 | 63.0 | 61.2 | 58.2 | 61.1 | 58.8 |
| Logistic | 59.9 | 63.2 | 60.1 | 59.2 | 62.7 | 59.6 |
| Simple Logistic | 60.0 | 63.3 | 60.2 | 59.2 | 62.7 | 59.7 |
| SMO | 54.9 | 61.8 | 54.4 | 59.0 | 62.0 | 53.2 |
| LibSVM | 59.7 | 62.6 | 59.9 | 55.0 | 55.9 | 50.2 |
| J4.8 | 60.4 | 62.6 | 60.9 | 58.1 | 61.9 | 58.2 |
| Decision Table | 60.5 | 62.6 | 61.0 | 58.5 | 62.5 | 58.9 |

**Table 3.4: Weighted average results following cross-validation**

A Naïve Bayes classifier, more specifically a Bayesian Network classifier, outperformed other methods in terms of F-measure (61.9%), and provided a more balanced classification performance across the classes. These findings can be partially explained by the fact that a Naïve Bayes classifier does not necessarily require a lot of training data to perform well [49]. Consequently, the Bayesian Network classifier was selected for our classification experiments.

For both experiments, the feature vectors produced for each sentence (discussed in Section 3.4.1) were amended with their ground truth class label adjudicated in Section 3.2.4. For example, in the second experiment, the example "The *party is over"* is represented as a feature vector (3.4), where the numerical values correspond to the selected features, and *Negative* is the ground truth (class label).

$$\underbrace{(4, 27, 46, 20, 3,}_{\text{Sentiment Polarity}} \underbrace{0, 100, 0,}_{\text{Idiom Polarity}} \underbrace{Negative)}_{\text{Class Label}} \tag{3.4}$$

## 3.5 Evaluation

The classification performance was evaluated in terms of precision (P), recall (R), and F-measure (F), based on the numbers of TP, FP, and FN. Tables 3.5 and 3.6 provide the comparison of these values with the two baselines considered. The overall performance represents micro-averaged results across the three classes.

| Class | Method | TP | FP | FN | P | R | F |
|---|---|---|---|---|---|---|---|
| **Positive** | Baseline | 40 | 53 | 98 | 43.0 | 29.0 | 34.6 |
| | Our method | 102 | 63 | 36 | 61.8 | 73.9 | 67.3 |
| **Negative** | Baseline | 111 | 66 | 127 | 62.7 | 46.6 | 53.5 |
| | Our method | 170 | 54 | 68 | 75.9 | 71.4 | 73.6 |
| **Other** | Baseline | 72 | 158 | 52 | 31.3 | 58.1 | 40.7 |
| | Our method | 49 | 62 | 75 | 44.1 | 39.2 | 41.7 |
| **Overall** | Baseline | 223 | 277 | 277 | 44.6 | 44.6 | 44.6 |
| | Our method | 321 | 179 | 179 | 64.2 | 64.2 | 64.2 |

**Table 3.5: The evaluation results using SentiStrength as a baseline**

| Class | Method | TP | FP | FN | P | R | F |
|---|---|---|---|---|---|---|---|
| **Positive** | Baseline | 41 | 44 | 97 | 48.2 | 29.7 | 36.8 |
| | Our method | 104 | 81 | 34 | 56.2 | 75.4 | 64.4 |
| **Negative** | Baseline | 170 | 160 | 68 | 51.5 | 71.4 | 59.9 |
| | Our method | 181 | 82 | 57 | 68.8 | 76.1 | 72.3 |
| **Other** | Baseline | 19 | 66 | 105 | 22.4 | 15.3 | 18.2 |
| | Our method | 20 | 32 | 104 | 38.5 | 16.1 | 22.7 |
| **Overall** | Baseline | 230 | 270 | 270 | 46.0 | 46.0 | 46.0 |
| | Our method | 305 | 195 | 195 | 61.0 | 61.0 | 61.0 |

**Table 3.6: The evaluation results using Stanford CoreNLP sentiment annotator as a baseline.**

With the exception of recall for the *other* class in the first experiment (decreasing from 58.1% to 39.5%), our method demonstrates a considerable improvement across all three measures. The overall improvement in the first experiment is 19.6 percentage points (from 44.6% to 64.2% in Table 3.5), and 15.0 percentage points in the second experiment (from 46.0% to 61.0% in Table 3.6). In terms of F-measure, there is an improvement across all three classes, but most notably in *positive* classifications. This is mainly due to the considerable improvement of recall by 45 percentage points in both experiments without compromising precision.

Confusion matrices given in Table 3.7 and Table 3.8 show how classification outcomes are re-distributed across the three classes. Table 3.7 illustrates that SentiStrength as a baseline is conservative in making both *positive* and *negative* predictions, thus less often misclassifying instances of the *other* class. Its classification outcomes were improved in all other cases by the use of idiom-based features. Conversely, Stanford CoreNLP sentiment annotator proved to be more conservative in making *positive* predictions in comparison to making *negative* ones, thus making fewer misclassifications when making *positive* predictions. Nonetheless, its classification outcomes were improved in all other cases by the use of idiom-based features.

| Actual \ Predicted | P | N | O |
|---|---|---|---|
| **P** | 40 | 36 | 62 |
| **N** | 31 | 111 | 96 |
| **O** | 22 | 30 | 72 |

Baseline

| Actual \ Predicted | P | N | O |
|---|---|---|---|
| **P** | 102 | 18 | 18 |
| **N** | 24 | 170 | 44 |
| **O** | 39 | 36 | 49 |

Our method

**Table 3.7: Confusion matrices using SentiStrength as the baseline method**

| Actual \ Predicted | P | N | O |
|---|---|---|---|
| **P** | 41 | 74 | 23 |
| **N** | 25 | 170 | 43 |
| **O** | 19 | 86 | 19 |

Baseline

| Actual \ Predicted | P | N | O |
|---|---|---|---|
| **P** | 104 | 24 | 10 |
| **N** | 35 | 181 | 22 |
| **O** | 46 | 58 | 20 |

Our method

**Table 3.8: Confusion matrices using Stanford CoreNLP sentiment annotator as the baseline method.**

Finally, in order to determine the statistical significance of the improvement over the two baseline methods, we performed the analysis of paired observations. We compared the sentiment classification results for each sentence before and after taking idioms into consideration by using a continuity corrected version of McNemar's test [54] to check for statistically significant differences in error rates. Under the null hypothesis, the two methods compared should have the same error rate. McNemar's test is based on the $\chi^2$ test statistic and (approximately) distributed as $\chi^2$ with 1 degree of freedom. We used a variant of McNemar's test statistic that incorporates a correction for continuity to account for the fact that the statistic is discrete while the $\chi^2$ statistic is continuous. The choice of this particular test was based on the following two facts: (1) McNemar's test has been shown to have low type I error, in this case - the probability that it would

incorrectly detect a difference when no difference exists, and (2) its statistical power is improved when compared with the commonly used paired t-test [47]. The $\chi^2$ (1) and p-values recorded for the data produced in the first experiment, where SentiStrength was used as the baseline method, were $\chi^2$ (1) = 43.16 and p < 0.001. The values recorded for the second experiment were $\chi^2$ (1) = 29.28 and p < 0.001. Therefore, in both cases the results of McNemar's test confirmed that there was a statistically significant difference in error rates between the two methods.

## 3.6 Summary

In this Chapter, we have demonstrated the value of idioms as features of sentiment analysis, by showing that idiom-based features significantly improve sentiment classification results when such features are present. For this purpose, we assembled a collection of 580 emotionally charged idioms, as well as a corpus of 2,521 sentences containing examples of these idioms used in context. Both datasets were manually annotated with EARL categories, and were projected as sentiment polarity in order to be compared to the baselines. We used Krippendorff's alpha coefficient to measure the reliability of the training dataset. The agreement on the idiom dataset ($\alpha$ = 0.662) demonstrates that idioms can reliably be mapped to sentiment polarity. The significantly lower agreement ($\alpha$ = 0.355) on contextual examples of idioms, demonstrates the complexity of sentiment interpretation in such tasks.

In order to automatically recognise idiom occurrences in text, idioms were modelled using regular expressions. The performance of recognising idioms, as well as negated idioms, was recorded as F-measure = 97.1% and 89.3% respectively. Idioms were represented as vectors, and subsequently combined with the results of two state-of-the-art sentiment analysis approaches, following the classification of their contextual examples. We performed classification experiments using a Bayesian network classifier, and evaluated the performance against the baselines. The overall performance, in

terms of F-measure, was improved from 44.6% to 64.2% and from 46.0% to 61.0% in both experiments. These improvements were demonstrated to be statistically significant.

# *Chapter 4*

# Scaling Up the Extraction of Idiom-Based Features

*"Going the extra mile."*

Given the positive findings in Chapter 3, the main limitation is the significant knowledge-engineering overhead involved in hand-crafting lexico-semantic resources used to support idiom-based features. To minimise the bottleneck associated with the acquisition of the sentiment of idioms and their recognition in context, the aim of this Chapter is to scale up our original approach by automating their engineering. More specifically, this Chapter is divided into the following main sections: Section 4.1 and Section 4.3 discuss a systematic approach to automating the engineering of idiom-based features, by utilising the canonical form of an idiom to automatically derive the variations of their occurrences in text, and by automatically acquiring idiom polarity by extracting sentiment from their dictionary definitions respectively. In Section 4.4, the manually engineered counterparts of the idiom-based features from our initial approach are replaced with those that automatically engineered. We repeat the classification experiments and evaluate the performance against the original study. Finally, Section 4.5 summarises our findings.

## 4.1 Inducing Pattern-Matching Rules

Due to the extent of their possible variations, contextual idiom occurrences cannot be identified using exact or approximate string matching approaches. In Chapter 3, the extraction of idiom-based features was supported by a set of manually crafted, lexico-syntactic pattern-matching rules and local grammars [72].

The goal of this Chapter is to minimise the bottleneck associated with this task, by using the canonical form of an idiom, i.e. a particular fixed phrase which is recognised by a speaker of the language as the main form of an idiom, to automatically derive its variations in text. The difficulty associated with this task, however, is the fact that idioms are rather heterogeneous in terms of their transformational capacity [130]. Riehemann [171] thoroughly discussed the different types of variations involved in the use of idioms in context. We highlight and consider their knowledge in this Chapter, in order to systematically address these variations. For this purpose, we consider inflection, open slots, modification, passivisation, distribution over multiple clauses, and all other variations [171].

### 4.1.1 Inflection

The words which constitute an idiom may be subject to inflection. Thus, almost all verbs can be used in different tenses, and some nouns can be used in their singular or plural form. For example, the verb in the idiom *stir a hornet's nest* is used in the present perfect tense in the following sentence:

> "Forbes has *stirred up a hornet's nest*."

Similarly, the noun in the idiom *bone to pick* is used in its plural form in the following example:

> "He generously leaves us one or two *bones to pick*."

The problem of inflection can be addressed by lemmatising both the canonical form of an idiom and the text in which it occurs. For example, lemmatisation leaves both *stir up a hornet's nest* and *bone to pick* unchanged, but transforms the given sentences into forms in which the lemmatised idioms can be matched as strings:

"Forbes have *stir up a hornet's nest*."

"He generously leave us one or two *bone to pick*."

### 4.1.2 Open Slots

Other idioms may contain open slots, into which any noun phrase can be inserted. For example, in the idiom *send someone packing*, the open slot, which is indicated by the indefinite pronoun 'someone,' is replaced by a two-word noun phrase in the following example:

"New rule could *send some* insurers *packing*."

The problem of open slots in idioms can be addressed by using shallow parsing, or chunking, which is the linguistic process of grouping words into phrases. For example, the result of parsing the previous example is as follows:

```
[NP New rule] [VP could send] [NP some insurers]
[VP packing].
```

The elements of the imposed shallow structure can then be used to generalise the search for idioms with open slots using a pattern, e.g. *send* <NP> *packing*, or its lemmatised version, *send* <NP> *pack* [184], where indefinite pronouns within the idiom's canonical form is automatically replaced by <NP>, a non-terminal symbol that can be replaced by any noun phrase in the corresponding pattern matching rule. Although state-of-the-art noun phrase chunking methods perform at an F-measure of 94% [83], incorrectly

parsing noun phrases remains a potential problem in this approach. Alternatively, one may choose to ignore the syntactic structure altogether, and instead, search for a flexgram [213], a sequence of tokens with one or more open slots of variable length, e.g. `send * packing`.

### 4.1.3 Modification

The constituents of some idioms may be modifiable, e.g. by using adjectives to modify nouns, or by using adverbs to modify verbs. The following example of the idiom *grasp at straws* contains both types of modification:

> "You seem to want to *grasp* desperately *at* every single *straw*."

Some potentially modifiable noun and verb components can be identified using POS tagging (e.g. *grasp*/`VB` *at*/`IN` *straws*/`NN`). The results of lemmatisation and such tagging can be combined to automatically generate the corresponding flexgram, by inserting gaps before nouns, and after verbs. In the previous example, the automatically generated flexgram `grasp * at * straw` would match the modified idiom.

### 4.1.4 Passivisation

In addition to inflection, the occurrence of verbs in idioms may also vary in terms of their transitive forms. The passive form allows an object, of an otherwise active sentence to become the subject of a passive sentence. In this process, the order between the verb and its object is reversed, with the original idiom components becoming separated. For example, we may compare an active form of the idiom *bury the hatchet*:

> "Christmas looks to be a time for *burying the hatchet* or exhuming it for re-examination."

to a passive one:

> "From the look of things, *the hatchet* has been long *buried*."

To address the passivisation of idioms, automatically acquired POS information can be used to identify non-auxiliary verbs at the beginning of an idiom, and produce an additional flexgram for its passive form, in which the verb should appear at the end of the idiom structure, with space for an open slot inserted ahead of it. For example, the POS tagged version of the given idiom, *bury*/`VB` *the*/`DT` *hatchet*/`NN`, can be used to identify *bury* as the leading verb, and produce `the hatchet * bury` as the passive version of the matching flexgram. The flexgram can now recognise the idiom in the lemmatised passive sentence:

> "From the look of thing, *the hatchet* have be long *bury*."

### 4.1.5 Distribution Over Multiple Clauses

The components of some idioms may be distributed between a main and subordinate clause, as is the case in the following example:

> "You remember [`NP` *the hatchet*] [`SBAR` that we *buried* last year with such pomp and ceremony]?"

The issue associated with this phenomenon is that idiom components become separated with the introduction of a subordinate clause. Most of the examples of this type of variation are related to the use of the verb component of an idiom as the main verb of the subordinate clause [171]. They can be effectively resolved by the pattern-matching rule generated to address passivisation. For example, the same flexgram `the hatchet * bury` will also match the lemmatised version of the distributed idiom:

> "You remember *the hatchet* that we *bury* last year with such pomp and ceremony?"

### 4.1.6 Other Variations

Other types of idiom variations can be recognised using the pattern-matching rules generated to address passivisation (Section 4.1.4) [171].

## 4.2 Applying Automatically Generated Idiom Recognition Rules

In this Chapter, each idiom was computationally modelled based on its canonical form. Idiom recognition rules were implemented as expressions in Mixup (My Information eXtraction and Understanding Package), a simple pattern-matching language [35].

To conform to the methods used in the original experiment, the automatically generated rules were applied to the test dataset of 500 sentences, which were lemmatised, and in which a single annotator marked up all idiom occurrences, disambiguating their figurative and literal senses. An idiom was considered to be correctly recognised if the suggested text span matched exactly idiom the marked up by the annotator.

The performance of recognising idioms using automatically generated rules was recorded as precision = 92.7%, recall = 92.8%, and F-measure = 92.7%. We can observe that the precision of both handcrafted and automatically generated rules are comparable (94.4% and 92.7% respectively).

However, the automatically generated rules achieved lower recall (a decrease from 100% to 92.8%). A possible explanation for this decrease is that some lemmas were incorrectly matched, i.e. a word may be lemmatised differently in a sentence, where the context affects the POS of individual words, and consequently, its lemma. This is also the case for idioms, which were processed independently of context. Consider, for example, “I was riding high,” where the idiom *ride high* is inflected. Lemmatising the sentence achieves “I be rid high,” whereas lemmatising the idiom alone leaves it

unchanged. Consequently, the idiom occurrence is not recognised.

A potential solution to resolve this problem is to apply stemming. Stemming does not depend on the POS; thus, the previous example achieves "I was ride high," whereas stemming the idiom alone leaves it unchanged. The idiom can, therefore, be correctly recognised in context.

## 4.3 Extracting Sentiment from Idiom Definitions

An approach to automatically interpreting the figurative meaning of an idiom, is to instead interpret the literal meaning of its dictionary definition (e.g. [186]). For example, a dictionary definition for the idiom *live the life of Riley* is "a person who has a comfortable and enjoyable life, without having to make much effort." As most syllabi for English as a second language pay special attention to studying idioms, there is an abundance of teaching material, including dictionaries, dedicated specifically to their study [112]. These readily available pedagogical resources can be used to support the automated interpretation of the figurative meaning of an idiom, as well as the underlying sentiment of such phrases.

In our original experiment, we collected a set of 580 emotionally charged idioms, including their definitions, from an online educational resource - Learn English Today[1], which organises phrases by affective themes (see Table 3.1 for the distribution of idioms across 16 emotional categories). We used this dataset to support the functionality of the experiments discussed in this Chapter.

The sentiment polarities of the given dataset of idioms were originally obtained using a crowdsourcing approach (discussed in Section 3.2.2). One of the goals of this experiment is to instead automatically extract sentiment polarity from idiom definitions. We describe two approaches to this task; the first uses off-the-shelf sentiment analysis tools

[1] http://www.learn-english-today.com/idioms/idioms_proverbs.html

to classify the overall sentiment expressed within an idiom definition (Section 4.3.1), and the second is based on mapping idiom definitions to WordNet-Affect (WNA), a hierarchy which includes a subset of WordNet synsets suitable for representing affective concepts, such as moods and situations eliciting emotions or emotional responses [196] (Section 4.3.2).

### 4.3.1 Off-the-Shelf Sentiment Analysis

Off-the-shelf sentiment analysis tools, such as SentiStrength [203, 204] and Stanford CoreNLP [189] can be used to classify the sentiment expressed in short text segments. Such tools, however, struggle to identify the sentiment conveyed by the figurative meaning of idioms, as their meaning, including their overall associated sentiment, cannot be entirely predicted from their constituent words when they are considered independently [139]. For example, in the absence of any positive or negative words in the idiom *live the life of Riley*, SentiStrength classifies its sentiment as *neutral*. However, if we apply the same sentiment analysis approach to its definition, the tool correctly classifies its sentiment as *positive*, based on the presence of two positive bearing words, 'comfortable' and 'enjoyable.' Similarly, Stanford CoreNLP's sentiment annotator quantifies the positive, negative, and neutral sentiment of the idiom itself as 20, 3, and 77 respectively, also classifying the idiom as *neutral*. However, when applied to its definition, the annotator quantifies the sentiment values as 93, 2, and 5, thus agreeing with SentiStrength, and also correctly classifying the sentiment as *positive*.

To further reinforce this point, we applied such tools to all 580 individual idioms, along with their definitions. We subsequently compared the results to the originally crowdsourced sentiment polarity annotations by measuring IAA. The agreement was measured using three versions of Cohen's kappa coefficient [33]: simple unweighted, with linear weighting, and with quadratic weighting. The kappa coefficient is calcu-

lated according to the following formula:

$$\kappa = 1 - \left( \frac{1 - P_o}{1 - P_e} \right) \tag{4.1}$$

where $P_o$ is the observed agreement, i.e. the proportion of items on which both annotators agree, and $P_e$ is the expected chance agreement calculated under the assumption that: (1) both annotators act independently, and (2) random assignment of annotation categories to items is governed by the distribution of items across these categories. We report the values for the original kappa coefficient in order to interpret the agreement based on Landis & Koch's [101] agreement scale: 0-0.20 (poor), 0.21-0.40 (fair), 0.41-0.60 (moderate), 0.61-0.80 (good), 0.81-1.00 (very good).

Cohen's kappa coefficient treats all disagreements equally, which is not suitable when the annotation categories are ordered as they indeed are: negative < neutral < positive. In such case, it is preferable to use weighted kappa coefficient [34], which accounts for the degree of disagreement by assigning different weights $w_i$ to cases where annotations differ by $i$ categories. If there are $n$ categories, the weights can be calculated according to the following formulas for linear and quadratic weighting respectively:

$$w_i = 1 - \frac{i}{n-1} \qquad w_i = 1 - \frac{i^2}{(n-1)^2} \tag{4.2}$$

For example, for a total of 3 categories, linear weights would be set to 1, 0.5, and 0, when there is a difference of 0, 1, and 2 categories respectively. The quadratic weights would be set to 1, 0.75, and 0. The weights are then used to multiply the corresponding proportion of disagreements in the observed matrix, before calculating that kappa coefficient.

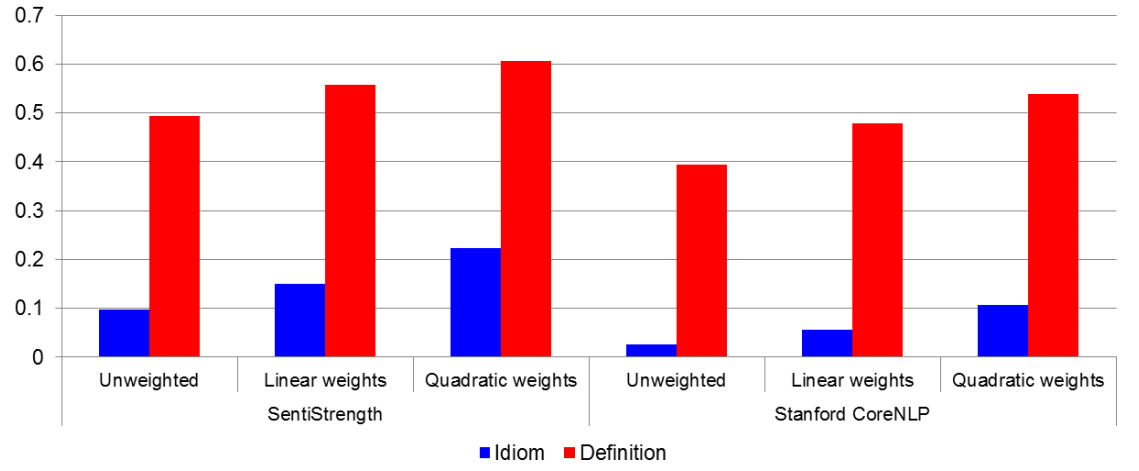


**Figure 4.1: Kappa agreement with the crowdsourced annotations**

The kappa values across each measure are shown in Figure 4.1. We can observe that the agreement with manual annotation increases across each measure, on average, by 0.4019, when sentiment analysis is applied to the definitions of the corresponding idioms. This improved the agreement from "very poor" to "moderate" on the agreement scale. In addition, we can observe that SentiStrength demonstrated a better performance on this particular dataset by, on average, 0.0824, in comparison to Stanford CoreNLP.

### 4.3.2 Identifying Affective Concepts

Another approach to consider when extracting sentiment from idiom definitions is to use a data driven method. In Chapter 2, we introduce WordNet, a lexical database of English nouns, verbs, adjectives, and adverbs, grouped together into sets of interlinked synonyms known as synsets [125]. An additional resource originating from WordNet, and one of the main lexical resources employed to detect sentiment in text, is WNA [196]. WNA was compiled specifically as a lexical model for classifying affects, such as moods, situational emotions, or emotional responses, providing direct (e.g. joy, sad, happy, etc.) and indirect (e.g. pleasure, hurt, sorry, etc.) affective terms. It was formed

by aggregating a subset of WordNet synsets into an affect hierarchy (see Figure 2.6). Our local version of the lexicon contains approximately 1,536 words, including all derivational forms of the word senses originally found in WNA.

WNA enables a more sophisticated interpretation of the sentiment(s) associated with an idiom. Each idiom can be represented by a vector, whose features correspond to nodes in the WNA hierarchy. Affective terms were extracted from idiom definitions by using simple string matching against affective WNA categories. A total of 278 affective nodes were identified across the 580 idiom definitions. Whilst 163 idiom definitions (28.1% of the dataset) were found to contain no affective words (e.g. *nothing doing* - "there is no way you would accept to do what is proposed"), 133 idiom definitions (22.9% of the dataset) were found to have more than one affective word occurrence (e.g. "*olive branch* - "to show that someone wants to end a disagreement and make peace.").

For each non-negated mention of an affective word found in an idiom definition, the corresponding feature is set to 1, together with all other features that correspond to its ancestors. This approach ensures that hierarchical relationships between affects are translated into a flat vector representation. For example, when interpreting the idiom *see red* using its definition "to suddenly become very angry or annoyed," two affective words are identified; 'angry' and 'annoyed'. As a result, the values corresponding to negative emotion, general dislike, anger, and annoyance (see Figure 2.6 for their hierarchical relationships) would be set to 1, whereas all other coordinates would remain as 0. Similarly, when interpreting the idiom *face like a wet weekend* using its definition "to look sad and miserable," two affective words are identified; 'sad' and 'miserable'. As a result, the values corresponding to negative emotion, sadness, and misery would be set to 1, whereas all other coordinates would remain as 0. Finally, when interpreting the idiom *bad blood* using its definition "intense hatred or hostility," two affective words are identified; 'hatred' and 'hostility'. As a result, the values corresponding to negative emotion, general dislike, hate, and hostility would be set to 1, whereas all

other coordinates would remain as 0. These values are summarised in Table 4.1.

| **Idiom** | Negative emotion | Sadness | Misery | General dislike | Anger | Annoyance | Hate | Hostility | ⋮ |
|---|---|---|---|---|---|---|---|---|---|
| *See red* | 1 | 0 | 0 | 1 | 1 | 1 | 0 | 0 | ... |
| *Face like a wet weekend* | 1 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | ... |
| *Bad blood* | 1 | 0 | 0 | 1 | 0 | 0 | 1 | 1 | ... |

**Table 4.1: Idioms represented as feature vectors**

Note that the vectors given in Table 4.1 are for illustrative purposes only and as such, focus only on a subspace of the WNA hierarchy. In practice, the length of the vector would be the total number of individual WNA hierarchical nodes that were identified across the idiom definitions, i.e. 278. This leads to feature vectors of high dimensionality, which may be associated with poor classification performance [84]. In Section 4.3.2.1 and Section 4.3.2.2, we discuss two potential approaches to reduce the dimensionality of such feature vectors.

#### 4.3.2.1 Feature Generalisation

In order to reduce the number of features, we can exploit the structure of the WNA hierarchy by simply projecting the original vectors onto a subspace that corresponds to the upper levels of the hierarchy, thereby, selecting more general features. For example, in Figure 2.6 we can focus on the two upper levels of the hierarchy, i.e. negative emotion, sadness, and general dislike, and simply remove the remaining features from the original vectors (see Table 4.2). A problem associated with this approach is that the structure of the WNA hierarchy is unbalanced (i.e. the negative sub-tree has 7 hierarchical levels, whereas the neutral sub-tree has 4); therefore, nodes that are at the

same level may not be of the same generality, which may introduce issues of biased representation.

| **Idiom** | Negative emotion | Sadness | General dislike | ... |
|---|---|---|---|---|
| *See red* | 1 | 0 | 1 | ... |
| *Face like a wet weekend* | 1 | 1 | 0 | ... |
| *Bad blood* | 1 | 0 | 1 | ... |

**Table 4.2: Idioms represented as generalised feature vectors**

#### 4.3.2.2 Clustering

The given vector representations (see Table 4.1 for examples) allow us to consider comparing idioms to one another in terms of their affective content, by, for example, using measures such as cosine similarity ($cos\theta$):

$$similarity(\vec{x}, \vec{y}) = cos(\theta) = \frac{\vec{x} \cdot \vec{y}}{\|\vec{x}\| \cdot \|\vec{y}\|} = \frac{\sum_{i=1}^{n} x_i y_i}{\sqrt{\sum_{i=1}^{n} x_i^2} \cdot \sqrt{\sum_{i=1}^{n} y_i^2}} \tag{4.3}$$

where $\vec{x}$ and $\vec{y}$ are two non-zero vectors of dimensionality, and $n$ and $\theta$ is the angle between them. In general, $cos\theta$ values range from -1 (which corresponds to 180°, thus indicating opposite direction) to 1 (which corresponds to 0°, thus indicating the same direction). The value of 0 indicates that the given vectors are orthogonal. In the case of the vector representation examples in Table 4.1, each coordinate is always a positive value, and thus, so are the corresponding $cos\theta$ values. Therefore, in this special case, the $cos\theta$ will range from 0 to 1, with higher values indicating higher similarity. In this representation, positive and negative affects will be orthogonal to one another. For example, in Table 4.1, we can establish that *see red* is more similar to *bad blood* ($cos\theta$

= 0.50), in comparison to *face like a wet weekend* ($cos\theta$ = 0.29), as they share two features (negative emotion and general dislike) as a direct consequence of encoding hierarchical relationships from WNA in a flat vector representation.

To visualise the similarity of affect between idioms, we applied multidimensional scaling to a distance matrix based on $cos\theta$. Figure 4.2 shows a clear separation between idioms in terms of their affect. The first direction (along the x-axis) separates idioms of positive affect on the left side (e.g. *place in the sun, happy as Larry, in the good books*), from those that are negative on the right side (e.g. *get worked up, dicey situation, haul over coals*). The second direction (along the y-axis) separates idioms that are associated with themes of anger at the bottom (e.g. *caught in the crossfire, get on someone's nerves, fight like cat and dog*), from those that are associated with themes of anxiety at the top (e.g. *back to the wall, have kittens, on tenterhooks*).

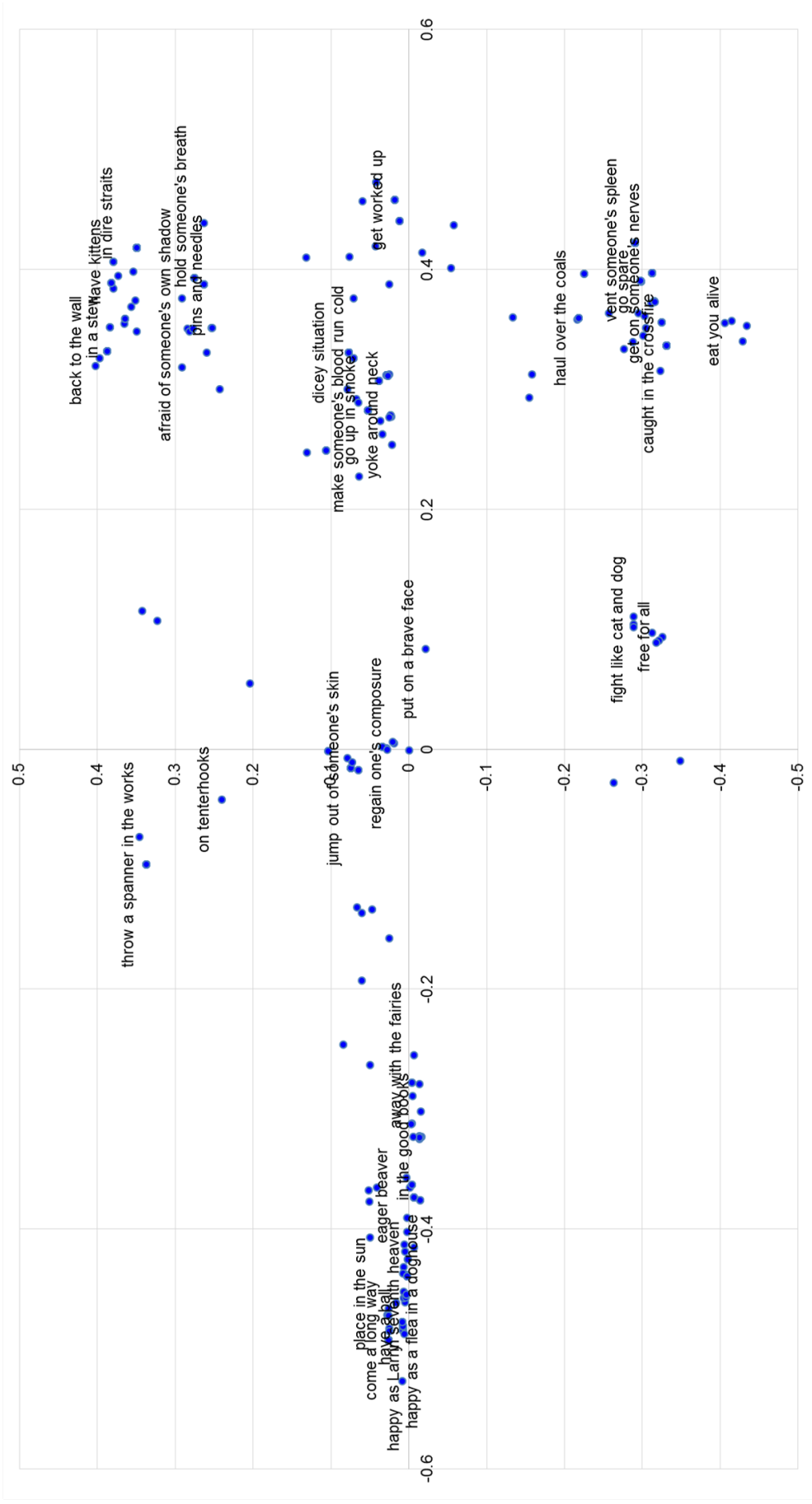


**Figure 4.2: Multidimensional scaling results**

A clustering algorithm can be used to identify clusters of related idioms. Table 4.3 illustrates the results of applying k-means clustering (k = 10), where, in principle, clusters can be mapped to affects. Nonetheless, uncategorised clusters can still be used as affective features to support sentiment analysis. The dimensionality of the problem can be controlled by reducing the number of clusters.

| Cluster | Interpretation | Members | Size |
|---|---|---|---|
| 1 | Surprise | *bolt from the blue; drop a bombshell; jaw drop; mixed feelings; knock down with a feather* | 28 |
| 2 | Frustration | *get someone's knickers in a twist; fish out of water; groan inwardly; put foot in mouth; slip through fingers* | 81 |
| 3 | Relief | *take a load off someone's mind; break the back of the beast; come up roses; save the day; weather the storm* | 11 |
| 4 | Affection | *eat, sleep and breathe something; have a soft spot; on cloud nine; knock someone's socks off; in the good books* | 30 |
| 5 | Anxiety | *break out in a cold sweat; cat on hot bricks; shake like a leaf; alarm bells ringing; cloud on the horizon* | 89 |
| 6 | Happiness | *lick someone's lips; pleased as punch; live the life of Riley; grin like a Cheshire cat; walking on air* | 24 |
| 7 | Excitement | *have a ball; have a whale of a time; paint the town red; over the moon; in seventh heaven* | 22 |
| 8 | Anger | *come down like a ton of bricks; fly off the handle; go through the roof; hot under the collar; see red* | 100 |
| 9 | Satisfaction | *bear fruit; reach first base; foot in the door; place in the sun; have the world by its tail* | 42 |
| 10 | Contempt | *fight like cat and dog; good riddance; steamed up; fit of pique; free for all* | 10 |

**Table 4.3: Clustering results**

#### 4.3.2.3 Mapping Affects to Sentiment Polarities

Both the generalisation and clustering approaches can be used to support the extraction of affective aspects of idiom-based features. However, to support the compatibility with the original study so that its results can be used as the baseline, we must map affects to sentiment polarities.

In this case, we used the generalisation approach. A total of 421 idiom definitions were successfully mapped onto affects and subsequently generalised as sentiment polarities. In Figure 4.3, we compared these results to the originally crowdsourced annotations by measuring IAA. In comparison to the agreement achieved by off-the-shelf sentiment analysis tools (see Figure 4.1), we can observe that the generalisation approach achieved higher agreement by, on average, 0.1572.

However, 159 out of 580 idiom definitions (i.e. 27% of the dataset) remained unclassified, as they did not contain non-negated mentions of affective words listed in WNA. In this case, and to take full advantage of both off-the-shelf sentiment analysis tools and our generalisation approach to sentiment polarity classification, their results were combined. Following our generalisation approach, we applied SentiStrength, which outperformed Stanford CoreNLP's sentiment annotator on these data (see Figure 4.1), to the remaining 159 unclassified idiom definitions. In Figure 4.3 we compare the overall results of the combined approach to the crowdsourced annotations by measuring IAA. We can observe that in comparison to the agreement achieved by off-the-shelf sentiment analysis tools in Figure 4.1, the combined approach also achieved higher agreement by, on average, 0.0639.

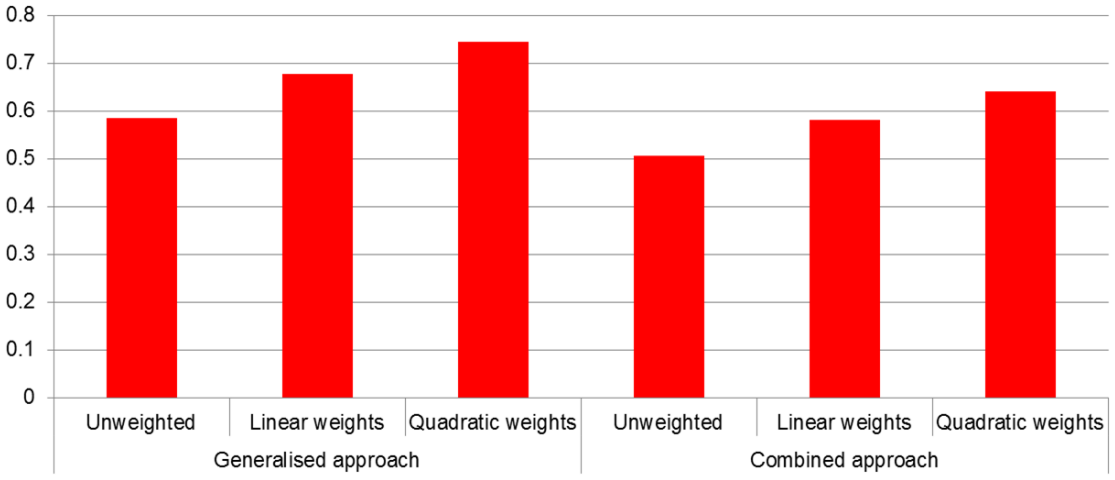


**Figure 4.3: Generalised and combined kappa agreement with the crowdsourced annotations.**

Table 4.4 demonstrates the differences in the distribution of the three categories across the 580 idioms when they are both manually and automatically mapped to sentiment polarity.

| | Manual | | Automated | |
|---|---|---|---|---|
| **Polarity** | **Total** | **%** | **Total** | **%** |
| Positive | 166 | 28.6 | 137 | 23.6 |
| Negative | 312 | 53.8 | 322 | 55.5 |
| Other | 102 | 17.6 | 121 | 20.9 |

**Table 4.4: Distribution of polarities across the idiom dataset**

Subsequently, the resulting sentiment polarity values from our combined approach were used to form idiom triples, and replaced the crowdsourced sentiment polarity annotations in the original sentiment analysis experiments discussed in Chapter 3. For example, in 4.4 we demonstrate how the sentiment of the idiom *battle of wills*, was

manually annotated as being predominantly *other*, i.e. (0, 40, 60):

$$\underbrace{(15, 58, 22, 3, 2,}_{\text{Sentiment Polarity}} \underbrace{0, 40, 60)}_{\text{Idiom Polarity}} \tag{4.4}$$

However, in 4.5, when such a feature is automatically generated using our combined approach, which extracts its sentiment polarity from its definition "when two parties are determined to win a conflict, argument, or struggle," the idiom is replaced by a *negative* triple, i.e. (0, 100, 0):

$$\underbrace{(15, 58, 22, 3, 2,}_{\text{Sentiment Polarity}} \underbrace{0, 100, 0)}_{\text{Idiom Polarity}} \tag{4.5}$$

In correspondence to our original approach, the phenomenon associated with negated idioms was modelled by reversing polarities in the idiom polarity vector, based on the assumption that negation converts *positive* to *negative* polarity, and vice versa. Additionally, when multiple idioms occurred, the associated idiom polarity values were aggregated by summing up the polarity vectors, whilst taking into account the effects of negation. Finally, if no idiom was detected, the idiom polarity values were set to zero, i.e. (0, 0, 0).

Table 4.5 demonstrates the differences in the distribution of idiom sentiment polarities across the gold standards for both our original and our automated method.

| | Original method | | Automated method | |
|---|---|---|---|---|
| **Polarity** | **Total** | **%** | **Total** | **%** |
| Positive | 793 | 31.5 | 596 | 23.6 |
| Negative | 1,193 | 47.3 | 1,286 | 51.0 |
| Other | 535 | 21.2 | 639 | 25.3 |

**Table 4.5: Distribution of idiom features across polarities**

## 4.4 Evaluation

In order to evaluate our experiments in this Chapter, and consequently, to compare the results with the those achieved in Chapter 3, we re-used the gold standard from the original study. To reiterate, the dataset consists of 2,521 sentences collected from the BNC [19, 106], which contain both figurative and literal occurrences of idioms used in context. Subsequently, this dataset was manually annotated with sentiment polarity, by crowdsourcing annotations using a web-based annotation platform (discussed in Section 3.2.2).

In Section 3.2.4, we created a gold standard for sentiment analysis experiments, where for each sentence, an annotation agreed by the relative majority of 50% of the annotators was adjudicated as the ground truth. We utilised the original testing and training data, which contained a random subject of 500 (20% of the dataset) and 2,021 sentences respectively. To conform to the methods used in the original experiment, a Bayesian Network classifier was used for classification, using Weka [74].

The main goal of this experiment is to investigate whether the results of sentiment analysis enriched with idiom-based features are comparable when manually engineered lexico-semantic resources are replaced by those that are automatically generated. In expectation that a fully automated approach may underperform in comparison to manually crafted features, we also compare whether the idiom-based approach would still outperform the original baseline methods, i.e. SentiStrength, and a sentiment annotator distributed as part of the Stanford CoreNLP, which do not incorporate idioms as features.

The classification performance was evaluated in terms of precision (P), recall (R), and F-measure (F), based on the numbers of TP, FP, and FN. Table 4.6 and Table 4.7 provide the comparison of these values for the two baselines considered, as well as the results achieved in the original study. The overall performance represents micro-averaged results across the three classes.

| Class | Method | TP | FP | FN | P | R | F |
|---|---|---|---|---|---|---|---|
| **Positive** | Baseline | 40 | 53 | 98 | 43.0 | 29.0 | 34.6 |
| | Original method | 102 | 63 | 36 | 61.8 | 73.9 | 67.3 |
| | Automated method | 75 | 73 | 63 | 50.7 | 54.3 | 52.4 |
| **Negative** | Baseline | 111 | 66 | 127 | 62.7 | 46.6 | 53.5 |
| | Original method | 170 | 54 | 68 | 75.9 | 71.4 | 73.6 |
| | Automated method | 168 | 102 | 70 | 62.2 | 70.6 | 66.1 |
| **Other** | Baseline | 72 | 158 | 52 | 31.3 | 58.1 | 40.7 |
| | Original method | 49 | 62 | 75 | 44.1 | 39.2 | 41.7 |
| | Automated method | 26 | 56 | 98 | 31.7 | 21.0 | 25.2 |
| **Overall** | Baseline | 223 | 277 | 277 | 44.6 | 44.6 | 44.6 |
| | Original method | 321 | 179 | 179 | 64.2 | 64.2 | 64.2 |
| | Automated method | 269 | 231 | 231 | 53.8 | 53.8 | 53.8 |

**Table 4.6: The evaluation results using SentiStrength as a baseline**

| Class | Method | TP | FP | FN | P | R | F |
|---|---|---|---|---|---|---|---|
| **Positive** | Baseline | 41 | 44 | 97 | 48.2 | 29.7 | 36.8 |
| | Original method | 104 | 81 | 34 | 56.2 | 75.4 | 64.4 |
| | Automated method | 64 | 57 | 74 | 52.9 | 46.4 | 49.4 |
| **Negative** | Baseline | 170 | 160 | 68 | 51.5 | 71.4 | 59.9 |
| | Original method | 181 | 82 | 57 | 68.8 | 76.1 | 72.3 |
| | Automated method | 190 | 148 | 48 | 56.2 | 79.8 | 66.0 |
| **Other** | Baseline | 19 | 66 | 105 | 22.4 | 15.3 | 18.2 |
| | Original method | 20 | 32 | 104 | 38.5 | 16.1 | 22.7 |
| | Automated method | 13 | 28 | 111 | 31.7 | 10.5 | 15.8 |
| **Overall** | Baseline | 230 | 270 | 270 | 46.0 | 46.0 | 46.0 |
| | Original method | 305 | 195 | 195 | 61.0 | 61.0 | 61.0 |
| | Automated method | 267 | 233 | 233 | 53.4 | 53.4 | 53.4 |

**Table 4.7: The evaluation results using Stanford CoreNLP sentiment annotator as a baseline.**

Confusion matrices given in Table 4.8 and Table 4.9 show how classification outcomes are re-distributed across the three classes. As expected, when manually crafted lexico-semantic resources were replaced by those that were automatically engineered, the performance, in terms of F-measure, in comparison to our original method, is poorer by 10.6 percentage points (from 64.2% to 53.8% in Table 4.6) and 7.6 percentage points (from 61.0% to 53.4% in Table 4.7) respectively. However, the use of automatically generated lexico-semantic resources still improves the performance of the baseline sentiment analysis methods by 9.0 percentage points (from 44.6% to 53.5%) and 7.4 percentage points (from 46.0% to 53.4%) respectively.

Baseline

| Actual \ Predicted | P | N | O |
|---|---|---|---|
| **P** | 40 | 36 | 62 |
| **N** | 31 | 111 | 96 |
| **O** | 22 | 30 | 72 |

Original method

| Actual \ Predicted | P | N | O |
|---|---|---|---|
| **P** | 102 | 18 | 18 |
| **N** | 24 | 170 | 44 |
| **O** | 39 | 36 | 49 |

Automated method

| Actual \ Predicted | P | N | O |
|---|---|---|---|
| **P** | 75 | 45 | 18 |
| **N** | 32 | 168 | 38 |
| **O** | 41 | 57 | 26 |

**Table 4.8: Confusion matrices using SentiStrength as the baseline method**

Baseline

| Actual \ Predicted | P | N | O |
|---|---|---|---|
| **P** | 41 | 74 | 23 |
| **N** | 25 | 170 | 43 |
| **O** | 19 | 86 | 19 |

Original method

| Actual \ Predicted | P | N | O |
|---|---|---|---|
| **P** | 104 | 24 | 10 |
| **N** | 35 | 181 | 22 |
| **O** | 46 | 58 | 20 |

Automated method

| Actual \ Predicted | P | N | O |
|---|---|---|---|
| **P** | 64 | 61 | 13 |
| **N** | 33 | 190 | 15 |
| **O** | 24 | 87 | 13 |

**Table 4.9: Confusion matrices using Stanford CoreNLP sentiment annotator as the baseline method.**

## 4.5 Summary

The goal of this Chapter was to address the main limitation of our original approach in Chapter 3 - the knowledge-engineering overhead involved in hand-crafting lexico-semantic patterns for the recognition of idioms in text, as well as the manual effort associated with acquiring the sentiment polarity of idioms. For this purpose, we re-used our collection of emotionally charged idioms, as well as their dictionary definitions.

To minimise the bottleneck associated with automatically recognising idioms in discourse, we used the canonical form of each idiom to derive rules to recognise their lexico-syntactic variations. The performance of recognising idioms was recorded as F-measure = 92.7%. We observed that although the precision of both handcrafted and automatically generated rules were comparable (94.4% and 92.7% respectively), the rules generated in this Chapter achieved a lower recall (a decrease from 100% to 92.8%). We suspect that the reason for this decrease is because some lemmas were incorrectly matched. We propose that stemming may resolve this issue. Nevertheless, we have demonstrated that it is possible to automatically recognise idioms in discourse, by using the canonical forms of each idiom to derive rules that are robust enough for this task.

In order to automate the acquisition of idioms' sentiment polarity, we automatically extracted sentiment from their definitions. We investigated four possible approaches for this task. In order to support the compatibility with the original study so that its results could be used as the baseline, an approach using a combination of WNA hierarchy and SentiStrength was demonstrated to be the most effective. We compared this approach to the results of the originally crowdsourced annotations by measuring IAA. In comparison to applying off-the-shelf sentiment analysis tools alone, the combined approach achieved a higher agreement by, on average, $\kappa = 0.0639$.

To evaluate the feasibility of this approach, we replaced the manually engineered lexico-semantic resources with their automatically generated counterparts and repeated

the same classification experiments as described in Chapter 3. As before, the overall classification performance of sentiment analysis when such idiom-based features are present, was improved from 44.6% to 53.8% and from 46.0% to 53.4% in both experiments. These results, however, are still poorer in comparison to those that were achieved in Chapter 3, where the inclusion of manually engineered features improved sentiment analysis results from 44.6% to 64.2% and from 46.0% to 61.0% in both experiments. Nevertheless, as we have demonstrated, not only can idiom-based features be automatically engineered, but they also improve sentiment classification results when such features are present.

# *Chapter 5*

# Comparison of Emotion Classification Frameworks for Sentiment Analysis

*"Reading between the lines."*

One of our research interests is to further advance the field of sentiment analysis by expanding it beyond mere sentiment classification. However, we identified a lack of consensus among researchers on a standardised framework of emotion. The goal of this Chapter is to investigate the utility of emotion classification frameworks for sentiment analysis. Our goal is to identify an appropriate classification framework in terms of completeness and complexity. We therefore investigate the utility of such frameworks by exploring their ease of use by human annotators, as well as the performance of supervised machine learning algorithms when they are used to annotate training data.

More specifically, this Chapter is divided into the following main sections: Section 5.1 introduces the selected frameworks under investigation. Section 5.2.1 discusses the construction of a corpus of emotionally charged text documents. In Section 5.2.2, we discuss how a crowdsourcing approach was used to manually annotate these text documents within each framework. To measure the utility of each framework from a human perspective, in Section 5.3.1, we measure and interpret IAA using Krippendorff's alpha coefficient. In Section 5.3.2, we discuss how we established a ground truth, in which we use as a gold standard for our supervised machine learning experiments. We also

utilise the ground truth to address the differences in the coverage of each framework using correspondence analysis. To complement the results of our quantitative analysis, in Section 5.3.4, we discuss the semi-structured interviews that were conducted to gain a qualitative insight into how annotators interact with and interpret each framework. We also discuss the results of the thematic analysis of the interview transcripts in this Section.

To measure the utility of each framework from a machine's perspective, in Section 5.4.1, the gold standard for each framework was used in cross-validation experiments to evaluate classification performance, where we specifically investigate the classification confusions. Finally, Section 5.5 summarises our findings.

## 5.1 Emotion Classification Frameworks

Intuitively, our investigation in this Chapter requires a selection of emotion classification frameworks. Our review in Chapter 2 demonstrates that the main tensions in the literature are whether emotions can be defined as discrete, universal basic categories, whether they are characterised by one or more dimensions, or whether they are organised hierarchically. Having considered these types, we selected five classification frameworks for this investigation, which vary size and in the way they represent emotion.

Categorical models of emotions appear to be dominant in this domain. Therefore, due to their popularity, simplicity, and familiarity within emotion classification, we selected Ekman's [51] six basic emotions (*happiness, sadness, disgust, fear, surprise, anger*). Additionally, we selected Plutchik's [158] wheel of emotion (Figure 2.1), a categorical model which represents the emotion space so that combinations of eight basic emotions (*joy, trust, fear, surprise, sadness, disgust, anger, anticipation*) derive secondary emotions (e.g. *joy* + *trust* = *love, anger* + *anticipation* = *aggression*, etc.), which we use as classes in this investigation.

Dimensional approaches represent emotions as coordinates in a multi-dimensional space [29]. We selected Watson and Tellegen's Circumplex theory of affect (Figure 2.3), due to its recommendation for describing emotions expressed in text [174]. Additionally, we investigate EARL [57] (Table 2.2), an XML-based language for representing emotions in technological contexts. This framework was designed to support practical computational applications, and for this reason, was selected for investigation. For both dimensional frameworks, similar to our experiments in Chapter 3, specific emotions were used as examples to explain each top-level category. For example, *caring* encompasses affection, *strong engagement* encompasses surprise, etc.

Hierarchical frameworks are used to capture a richer set of emotions, with the main focus being on lexical aspects that can support text mining applications. WordNet-Affect (WNA) [196] (Figure 2.6) is a lexical model of affects, such as moods, situational emotions, or emotional responses, providing directly (e.g. *joy, sad, happy*) and indirectly (e.g. *pleasure, hurt, sorry*) affective terms. It was formed by selecting, assigning, and linking a subset of WordNet synsets (see Section 2.4.1.1) whose sense corresponds to an affect, and organising them into a hierarchy. WNA has been used to support several sentiment analysis and emotion analysis studies, and thus was selected for this study.

In addition to those discussed in the literature, we included free text classification, where the choice of emotion was unrestricted. We specifically wanted to investigate whether a folksonomy would naturally emerge from annotators' free text choices, and could give rise to a suitable emotion classification framework.

Table 5.1 summarises the six frameworks used in this investigation.

| Type | Framework | Size |
|---|---|---|
| Categorical | Six basic emotions | 6 classes |
| | Plutchik's wheel of emotion | 16 classes |
| Dimensional | Circumplex | 4 dimensions, 8 classes |
| | EARL | 2 dimensions, 10 classes |
| Hierarchical | WNA | 7 levels, 1,536 classes |
| Unrestricted | Free text | $\infty$ |

**Table 5.1: Selected emotion classification frameworks for investigation**

## 5.2 Data Collection

### 5.2.1 Constructing an Emotionally Charged Corpus

State-of-the-art emotion classification has been applied to a range of texts from different domains (see Table 2.4). Twitter is a social networking service that enables users to send and read short, 140 character texts, referred to as tweets. Twitter provides an open platform for users from diverse demographic groups. An estimated 500 million tweets are posted each day [77]. The information content of tweets varies from daily life updates, sharing content (e.g. news, music, articles, etc.), expressing opinions, etc. The use of Twitter as a means of self-disclosure makes it a valuable source for emotionally-charged texts. Thus, it has become a recent popular source for textual data in this domain (e.g. [68, 148, 98]). For these reasons, Twitter was selected as a source of data in this Chapter.

We assembled a corpus of 500 self-contained tweets, i.e. those that did not appear to be a part of a conversation. More specifically, we excluded re-tweets and replies, as well as tweets that contained URLs or mentioned other users, to maximise the likelihood of an emotion expressed in a tweet to refer to the tweet itself and not an external source

(e.g. content corresponding to a URL). We used four criteria to identify emotionally-charged tweets: idioms, emoticons, hashtags, as well as automatically calculated sentiment polarity. The remainder of this section provides more detail on the selection criteria.

In Chapter 3, our study demonstrated that idioms are pertinent features in sentiment analysis, which significantly improve sentiment classification results. Using the set of emotionally-charged idioms described in Table 3.1, we collected 100 tweets containing references to such idioms (e.g. "If I see a mouse in this house I will *go ballistic"*).

Written online communication has led to the emergence of informal, sometimes ungrammatical, textual conventions [162] used to compensate for the absence of body language and intonation, which otherwise account for 93% of non-verbal communication [120]. Emoticons are pictorial representations of facial expressions that seem to compensate for the lack of embodied communication. For example, the smiley face :) is commonly used to represent positive emotions. We collected 100 tweets containing emoticons. Table 5.2 summarises the distribution of emoticons across these tweets.

| **Emoticon** | **Example** | **Total** |
|---|---|---|
| :( | Need to stop having nightmares :( | 56 |
| :) | Early finish in work for a change :) | 35 |
| :D | So proud of myself right now :D | 10 |
| :P | OK... so I have a huge crush! There!! :P | 2 |
| <3 | I <3 you | 2 |

**Table 5.2: Distribution of emoticons across 100 tweets**

One of Twitter's main features is the use of the hashtag (#), which adds contexts and metadata to the main content of a tweet, which in turn makes it easier for other users to find messages on a specific topic [30]. Hashtags are often used to imply or identify the users' emotional state [215] (e.g. "Sometimes I just wonder...I don't know what to think *#pensive*"). To systematically search Twitter for emotive hashtags, we used

WNA as a comprehensive lexicon of emotive words. Our local version of the lexicon consists of 1,484 words, including all derivational and inflectional forms of the word senses originally found in WNA. We searched Twitter using these surface forms as hashtags to collect 100 tweets. The hashtags were subsequently removed from the original tweets for the following two reasons. First, we wanted the annotators to infer the emotion themselves from the main content. Second, we did not want to skew the IAA in favour of the WNA as a classification framework.

The fourth strategy for identifying emotionally-charged tweets involved automatically calculated sentiment polarity. We collected 116,903 tweets at random, and processed them with a sentiment annotator distributed as part of the Stanford CoreNLP [189]. To reiterate, this method uses recursive neutral networks to perform sentiment analysis at all levels of compositionality across a parse tree by classifying a sub-tree on a 5-point scale: very negative, negative, neutral, positive, and very positive. We used the sentiment analysis results to select a random subset of 50 very positive and 50 very negative tweets. See Figure 5.1, where "Why do I even waste my time?" is classified as very negative.

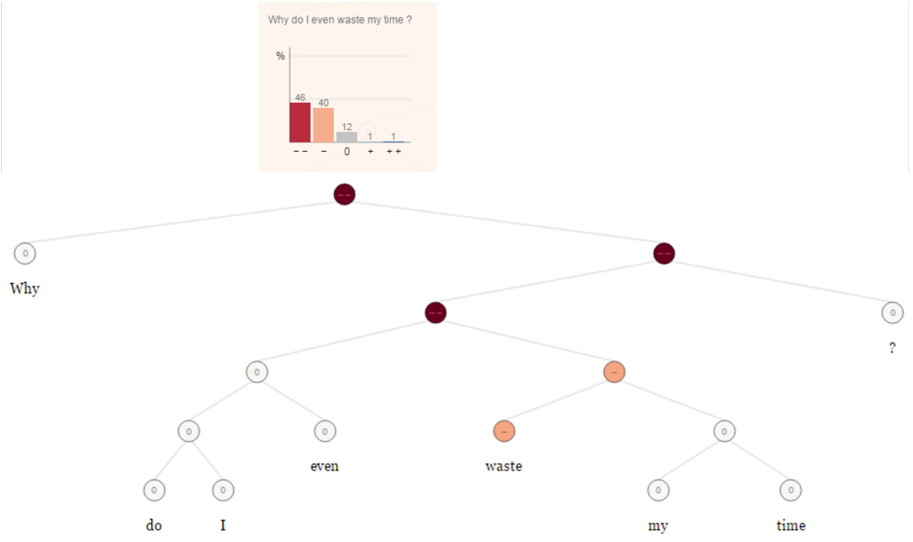


**Figure 5.1: Sentiment analysis results from Stanford CoreNLP**

Finally, we collected 100 additional tweets at random to include emotionally neutral or ambiguous tweets, while correcting for bias towards certain emotions based on the choice of idioms, emoticons, and hashtags. Table 5.3 summarises the corpus selection criteria and distribution of the corresponding tweets selected for inclusion in the corpus.

| Criterion | Example | Total |
|---|---|---|
| Idiom | If I see a mouse in this house I will *go ballistic*. | 100 |
| Emoticon | What a day!!!! *:(* | 100 |
| Hashtag | Why are people so mean? *#frustrated* | 100 |
| Sentiment polarity | Why do I even waste my time? | 100 |
| Random selection | Up this early for work. | 100 |

**Table 5.3: Corpus selection criteria and distribution**

### 5.2.2 Crowdsourcing of Sentiment Annotations

In order to annotate text documents with respect to their emotional content, we used CrowdFlower [1], a specific web platform which allows users to set up crowdsourcing jobs to distribute work to millions of online contributors. A bespoke annotation interface was designed, which consisted of three parts: input text (i.e. a document from the corpus described in Section 5.2.1), an annotation menu based on a classification framework, and, where appropriate, a graphical representation of the classification framework to serve as a visual aid (see Figure 5.2). To mitigate the complexity of WNA, we implemented autocomplete functionality, where matching items from the lexicon were automatically suggested as the annotator typed into a free text field.

[1] https://www.crowdflower.com/

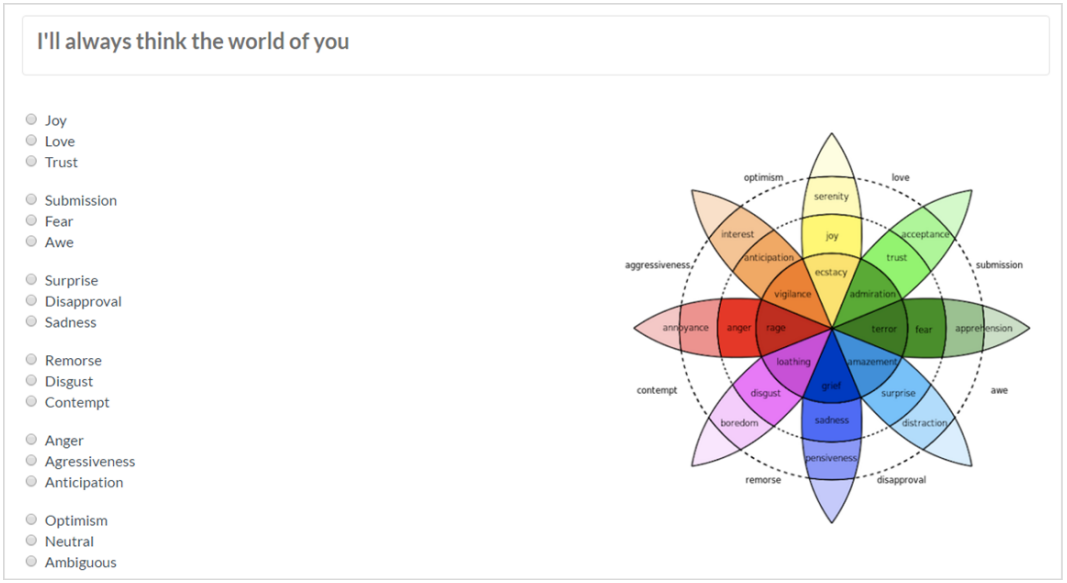


**Figure 5.2: Crowdflower's annotation platform interface**

We introduced a *neutral* category into all classification frameworks to allow for the annotation of flat or absent emotional responses. For example, "Fixing my iTunes library." was annotated as *neutral* by 23 of 30 annotators. Similarly, we introduced an *ambiguous* category to allow for the annotation of cases where an emotion is present, but is indeterminate in the absence of context, intonation, or body language. For example, the use of punctuation in "What a day!!!!!" clearly indicates an emotional charge, but is unclear whether the statement is positive or negative.

Having set up annotation jobs for each classification framework, contributors were asked to annotate each text document with a single class that best described its emotional content. A total of 189 annotators participated in the study. Given a classification framework, each document was annotated by 5 independent annotators. In total, 15,000 annotations (500 documents $\times$ 6 frameworks $\times$ 5 annotations) were collected. The distributions of annotations across each framework is shown in Figure 5.3, with WNA and free text charts displaying the distributions of the top 20 most frequently used annotations.

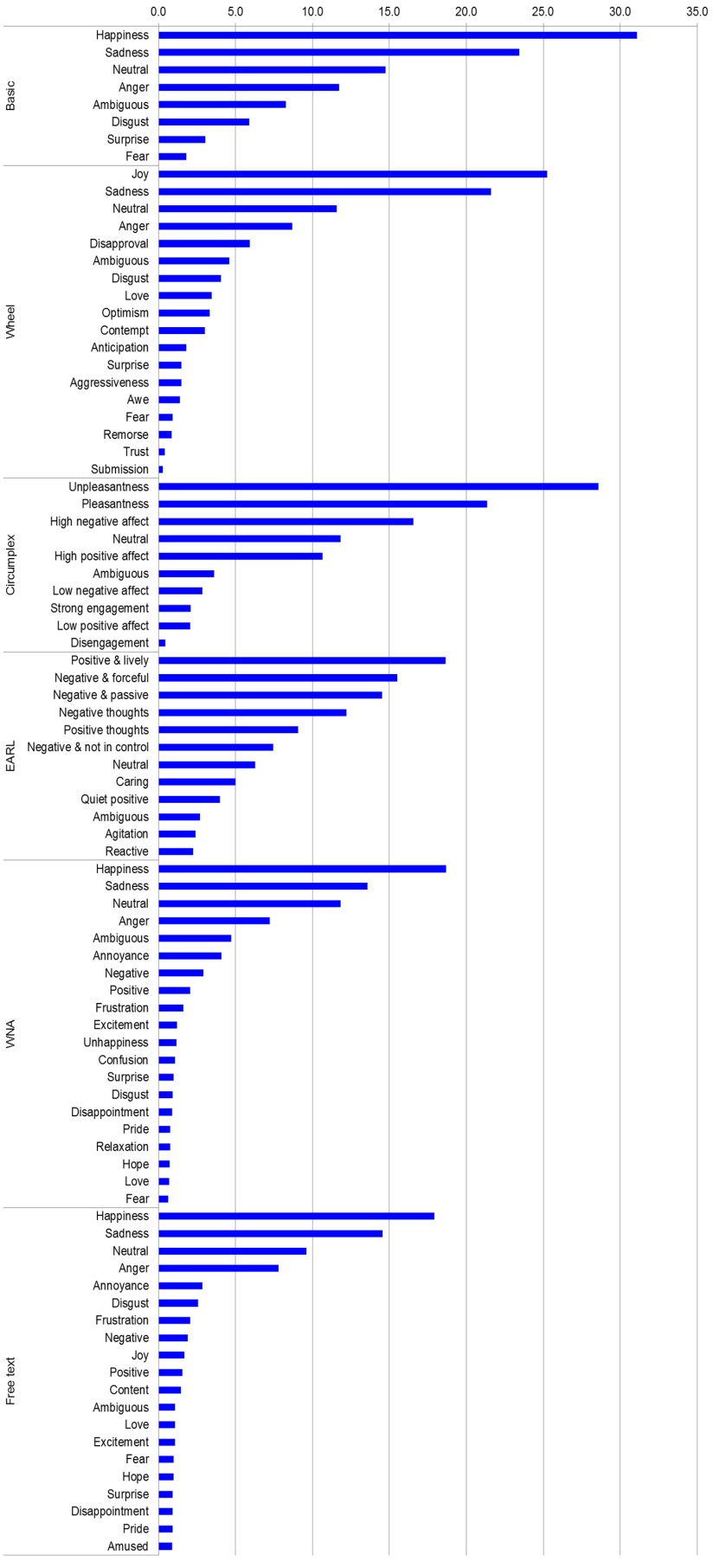


**Figure 5.3: Distribution (%) of annotations across each framework**

## 5.3 Utility Analysis: A Human Perspective

In the first half of this Chapter, we investigate the utility of our selected classification frameworks, by exploring their ease of use by human annotators.

### 5.3.1 Inter-Annotator Agreement

To quantitatively measure the utility of our selected classification frameworks from a human annotator's perspective, we measured IAA. Assuming that classification frameworks with a better balance between completeness and complexity are easier to interpret and use, we hypothesise that when a correct class is available, unambiguous and readily identifiable, the likelihood of independent annotators selecting that particular class increases, thus leading to higher IAA.

We used Krippendorff's alpha coefficient [99] to measure IAA. As a generalisation of known reliability indices, this measure was chosen as it applies to: (1) any number of annotators, not just two; (2) any number of categories; and (3) corrects for change expected agreement [100]. With the highest value of $\alpha$ = 0.483, the IAA results here (see Figure 5.4) are well below Krippendorff's recommended threshold ($\alpha$ = 0.667). This result is consistent with other studies on affective annotation (e.g. [43, 26, 9]).

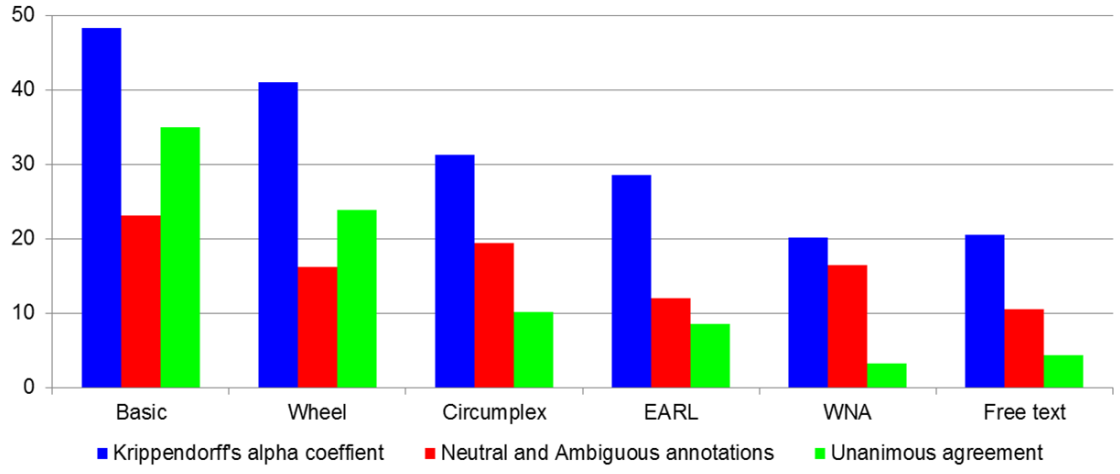


**Figure 5.4: Inter-annotator agreement results**

Annotation is a highly subjective process that varies with age, gender, experience, cultural location, and individual psychological differences [153]. Additionally, a text document may consist of multiple statements, which may convey different or competing emotional content. For example, there are two statements in the following sentence: "On train going skating :) Hate the rain :(" each associated with a different emotion, illustrated clearly by the use of emoticons. Using Plutchik's wheel of emotion, this sentence received the following five annotations: *sadness, sadness, joy, love, ambiguous*. It can be inferred that annotators 1 and 2 focused on the latter statement, whereas annotators 3 and 4 focused on the former statement. Annotator 5 acknowledged the presence of both positive and negative emotions by classifying the overall text as *ambiguous*. A genuine ambiguity occurs when the underlying emotion may be interpreted differently in different contexts (e.g. "Another week off," received two *ambiguous* and three *joy* annotations), which leads to inter-annotator disagreement. Other factors such as annotators' skills and focus, the clarity of the annotation guidelines and inherent ambiguity of natural language may have also contributed to low IAA. These factors may explain low IAA, but fail to explain the large variation in agreement across the annotation frameworks, which ranged from $\alpha = 0.202$ to 0.483, with a standard deviation of 11.2. Nonetheless, these results enable a comparison of each framework.

Unsurprisingly, given the smallest number of options, the highest IAA ($\alpha = 0.483$) and the highest number of unanimous agreements (175 out of 500, i.e. 35%) were recorded for the six basic emotions. An important factor to consider here is that this framework incurred by far the highest usage of *neutral* and *ambiguous* annotations (576 out of 2,500, i.e. 23%). This may imply that the six basic emotions has insufficient coverage of the emotion space.

Intuitively, one may expect IAA to be higher for frameworks with fewer classes, as seen in some empirical studies [9], as fewer choices offer fewer chances for disagreement. However, Krippendorff's alpha coefficient is a chance corrected measure of IAA, which suggests this may not necessarily be the case. Specifically, the IAA results

here shows higher agreement for a framework with 18 categories (Plutchik's wheel of emotion) than it does for frameworks of 10 or 12 classes (EARL and Circumplex). With $\alpha$ = 0.41, Plutchik's wheel of emotion recorded the second highest IAA. In comparison to the six basic emotions, annotators resorted less frequently to using *neutral* and *ambiguous* annotations. It also recorded the second highest number of unanimous agreements (119 out of 500, i.e. 24%).

Dimensional frameworks, Circumplex and EARL, both with a similar number of classes (12 and 10), recorded similar levels of IAA ($\alpha$ = 0.312 and 0.286 respectively). However, an important difference between these frameworks was the usage of *neutral* and *ambiguous* annotations. Circumplex incurred the second highest usage of these annotations. On the other hand, EARL had the second lowest usage of these annotations following free text annotations. This implies that with 10 generic categories, this framework provides better coverage of the emotion space.

Due to the ambiguity and polysemy of natural language, lexical frameworks, WNA and free text, recorded the lowest IAA ($\alpha$ = 0.202 and $\alpha$ = 0.205 respectively) and incurred the fewest unanimous agreements (16 and 22 out of 500, i.e. 3% and 4% respectively). The lower IAA for WNA may be explained by the difficulty of navigating through a large hierarchy. With 262 and 260 different annotations recorded respectively, WNA and free text covered a wide range of emotive expressions, which provided annotators with the means of referring to a specific emotion when a suitable generic category was not available in other frameworks, thus minimising the use of *ambiguous* annotations.

To determine the significance of the differences in IAA across the frameworks, we constructed confidence intervals for the given values. Given an unknown distribution of the Krippendorff's alpha coefficient, we constructed confidence intervals by estimation using bootstrap [50]. We used 1,000 replicate re-samples from the 500 tweets. More specifically, we randomly selected instances from the original set of 500 tweets to be included in our sample. The sampling was performed with replacement and therefore when a single tweet was included in the same sample multiple times, we re-used the

same annotations. Subsequently, we used the percentage method [41] to construct 95% confidence intervals, by cutting 2.5% of the replicates on each end.

The confidence intervals were as follows: six basic emotions (0.4498, 0.5146), Plutchik's wheel of emotion (0.3809, 0.4372), Circumplex (0.2871, 0.3348), EARL (0.2602, 0.3073), WNA (0.1826, 0.2196), and free text (0.1842, 0.2250). Where there is no overlap between the confidence intervals (see Figure 5.5), we can assume that there is a statistically significant difference in the IAA between each pair of frameworks. Therefore, the six basic emotions framework has a significantly higher IAA in comparison to the remaining frameworks, and the wheel of emotion has a significantly larger agreement in comparison to Circumplex, EARL, WNA, and free text. Circumplex and EARL have similar agreement, but a significantly larger IAA than WNA and free text, which also have a similar IAA.

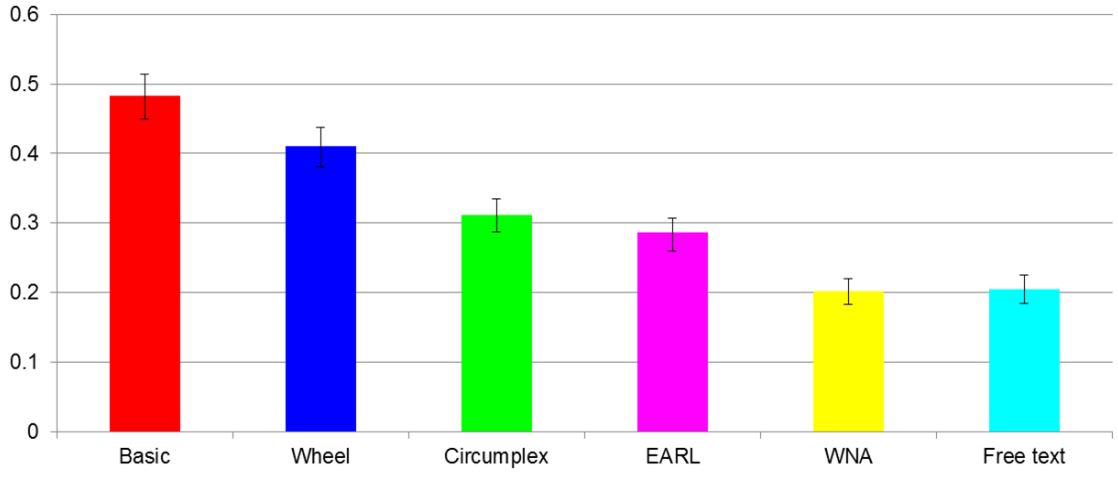


**Figure 5.5: Confidence intervals for the inter-annotator agreement**

### 5.3.2 Establishing the Ground Truth

In Section 5.2.2, each text document was annotated by 5 independent annotators per framework, using a crowdsourcing approach. In order to determine the ground truth, the most frequent annotation per data item was accepted, with an expectation for the automated system to behave as the majority of human annotators. For each framework,

each sentence with an annotation agreed upon by the relative majority of at least 50% of the annotators was adjudicated as the ground truth. For example, using the six basic emotions as the classification framework, the sentence "For crying out loud be quiet" was annotated with *anger* four times and once with *disgust*; thus *anger* was accepted as the ground truth.

When no majority annotation could be identified, a new independent annotator resolved the disagreement by adjudicating which annotation, from the 5 given, was the most appropriate to describe the emotion being expressed. Table 5.4, across the diagonal, provides the percentage of text instances that required disagreement resolution under each framework. The remaining values illustrate the overlap of such text instances across the frameworks. Overall, 18 instances (i.e. 3.6%) required disagreement resolution under all frameworks. Instances that required disagreement resolution under many frameworks are likely to be genuinely ambiguous. Otherwise, the ambiguity is likely to be related to a given annotation framework. See Figure 5.6 for distribution of ground truth annotations across each framework.

| | **Basic** | **Wheel** | **Circumplex** | **EARL** | **WNA** | **Free text** |
|---|---|---|---|---|---|---|
| **Basic** | 17.6 | 11.8 | 7.6 | 10.2 | 14.6 | 15.8 |
| **Wheel** | 11.8 | 29.6 | 11.8 | 13.0 | 22.4 | 23.0 |
| **Circumplex** | 7.6 | 11.8 | 22.0 | 10.2 | 16.4 | 15.4 |
| **EARL** | 10.2 | 13.0 | 10.2 | 37.2 | 22.0 | 23.4 |
| **WNA** | 14.6 | 22.4 | 16.4 | 22.0 | 48.0 | 36.4 |
| **Free text** | 15.8 | 23.0 | 15.4 | 23.4 | 36.4 | 46.0 |

**Table 5.4: The number (%) of instances that required disagreement resolution**

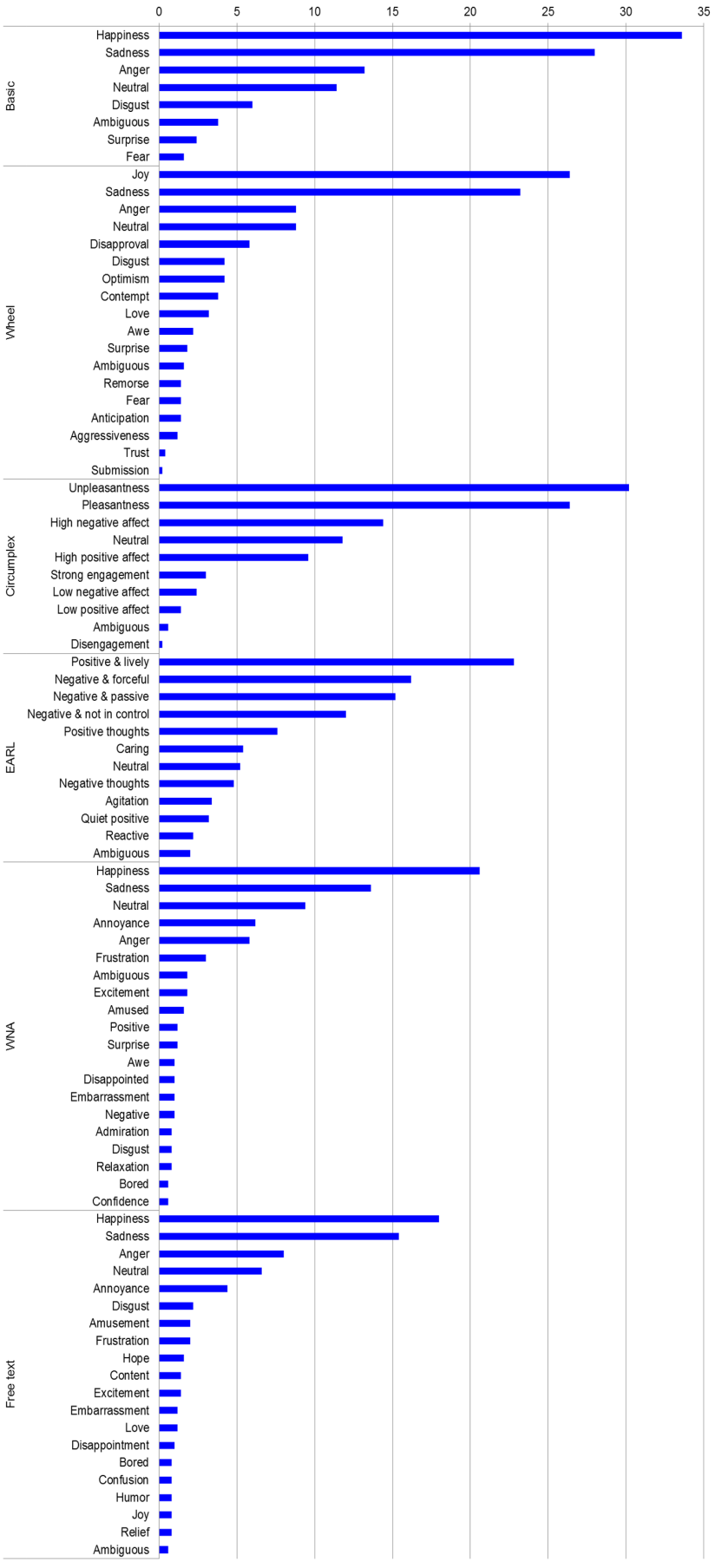


**Figure 5.6: Distribution (%) of ground truth annotations across each framework**

### 5.3.3 Correspondence Analysis

To address the differences in the coverage of each framework, let us consider annotations for the sentence "I'll always have a soft spot in my heart for this girl," (see Table 5.5). Despite the unanimous agreement under the six basic emotions, it is still difficult to interpret the given sentence as an expression of *happiness*. Where *love*, or related emotions are available, we can see the strong preference towards choosing such emotions in Plutchik's wheel, EARL, WNA, and free text. This point is reinforced in the case of the Circumplex, which similarly, lacks a category related to *love*.

| **Main emotion** | **Framework** | | | | | |
|---|---|---|---|---|---|---|
| | **Basic** | **Wheel** | **Circumplex** | **EARL** | **WNA** | **Free text** |
| *happiness* | 5 × *happiness* | 1 × *joy* | 5 × *pleasantness* (includes happy) | 1 × *positive & lively* (includes joy and happiness) | 1 × *happiness* | 2 × *happiness* |
| *love* | - - - | 4 × *love* | - - - | 4 × *caring* (includes affection and love) | 1 × *romantic*<br>1 × *soft-spot*<br>1 × *affection*<br>1 × *love* | 3 × *love* |

**Table 5.5: Examples of annotation preferences**

To generalise these observations and to explore the relationships between the classes across different frameworks, we conducted correspondence analysis [79], a dimension reduction method appropriate for categorical data. This analysis is used to present a graphical representation of the relationships between two sets of categories. The large number of classes in WNA and free text classification makes the graphical representation of the correspondence analysis involving either of these frameworks highly convoluted. We therefore only present the results involving the four remaining frameworks. For this analysis we used the majority annotations used to create our gold standard discussed in Section 5.3.2, and compared them between two frameworks at a time. Figures 5.8 - 5.12 show the first two dimensions in correspondence analysis between pairs of frameworks.

The correspondence analysis between the six basic emotions and Plutchik's wheel of emotion (Figure 5.7) demonstrates that the first dimension (along the x-axis) separates

the positive emotions (e.g. *happiness, love*) on the left, from the negative emotions (e.g. *anger, sadness*) on the right. One can claim that the second direction (along the y-axis) differentiates between aggressive emotions (e.g. *anger, aggressiveness*), from more passive emotions (e.g. *sadness, fear*). Further study into the distribution of emotions across the dimensions shows that four emotions from Plutchik's wheel (*submission, joy, love, awe*), correspond to a single basic emotion (*happiness*). Emotions that exist in both frameworks are located close together, e.g. *anger* in the basic emotions framework is close to *anger* in the Plutchik's wheel. On the other hand, emotions such as *remorse, anticipation, optimism, disapproval, trust* and *aggressiveness*, which only exist in Plutchik's wheel, do not correspond closely with a specific basic emotion. This evidence supports that these emotions are not redundant, i.e. they cannot be abstracted easily into a basic emotion. Moreover, further analysis demonstrates that the annotators often resorted to *happiness*, i.e. the only positive basic emotion, as a surrogate for a diverse range of emotions found in the Plutchik's wheel including *awe, submission* and *love*, which do not necessarily imply *happiness*.

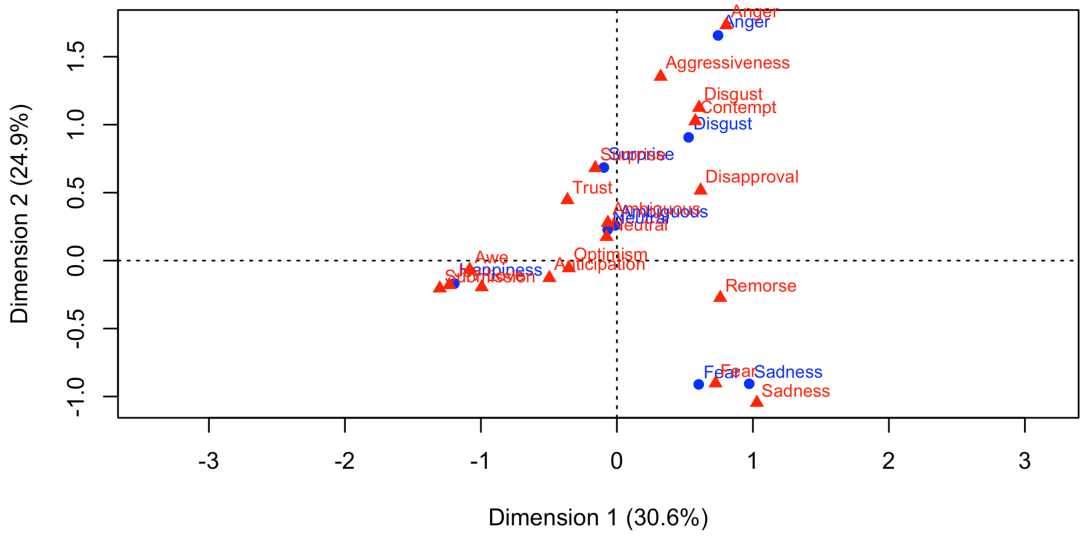


**Figure 5.7: Six basic emotions (blue) versus Pluchik's wheel of emotion (red)**

The analysis between the six basic emotions and Circumplex (Figure 5.8) demonstrates that two positive classes (*high positive affect* and *pleasantness*), correspond to the basic emotion of *happiness*. On the other hand, some classes from Circumplex, e.g. *strong engagement, low positive affect* and *low negative affect*, do not particularly correspond to any basic emotion.

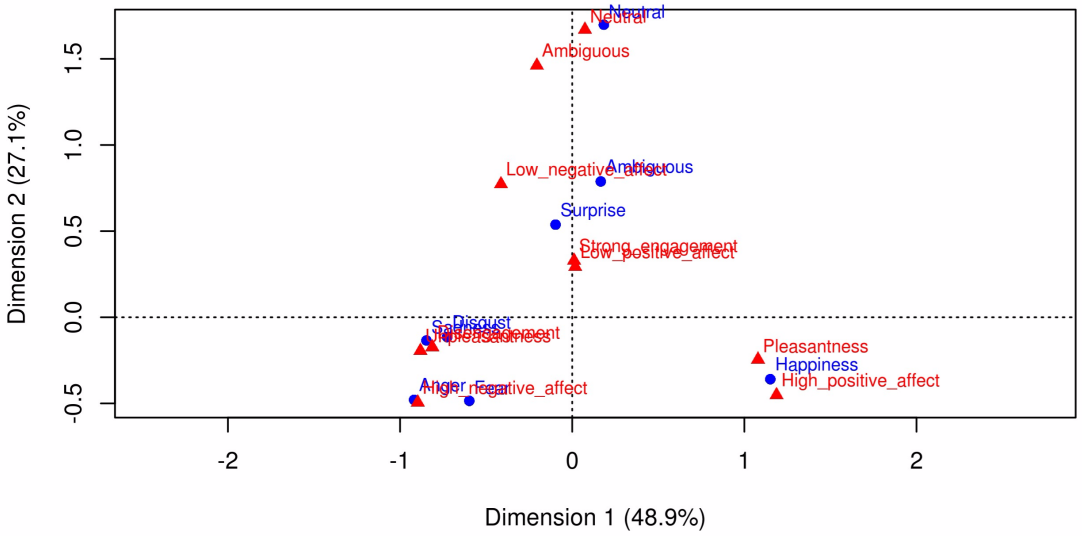


**Figure 5.8: Six basic emotions (blue) versus Circumplex (red)**

Similarly, the analysis between the six basic emotions and EARL (Figure 5.9), shows that all positive classes correspond to *happiness*, *negative thoughts* correspond to *fear*, *negative & forceful* corresponds to *anger*, and *negative & passive* and *negative & not in control* correspond to *sadness*.

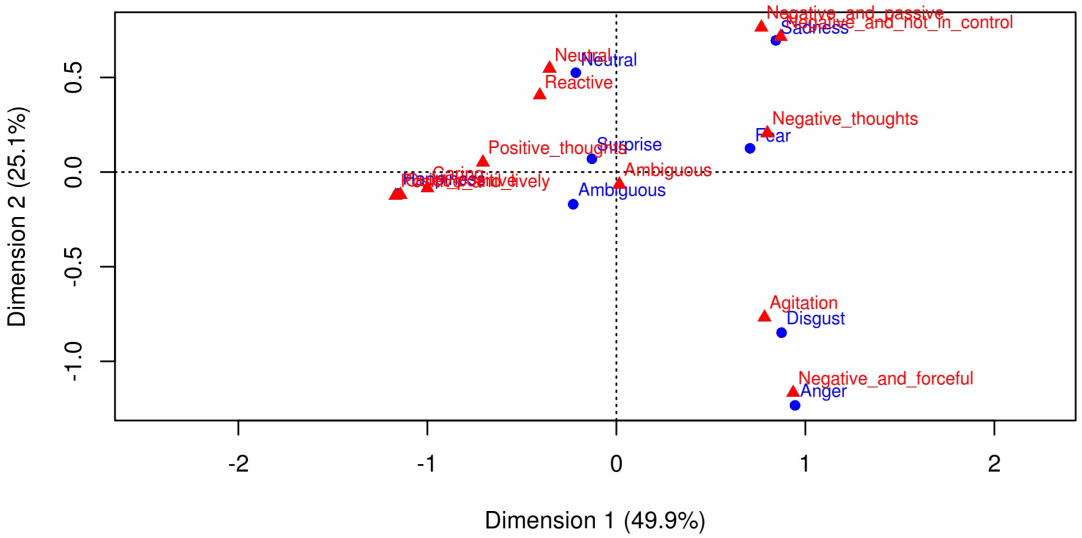


**Figure 5.9: Six basic emotions (blue) versus EARL (red)**

Figures 5.10 and 5.11 show how emotions from Plutchik's wheel relate to classes from Circumplex and EARL respectively. It is clear that, although Circumplex and EARL are richer than the six basic emotions, they still do not seem to completely and unambiguously model emotions from Plutchik's wheel. For example, we can see from Figure 5.10 that Circumplex does not have a class that corresponds to a number of emotions in Plutchik's wheel, e.g. *optimism, trust, anticipation, remorse* and *disapproval*. Similarly, from Figure 5.11 we can see that classes from EARL do not align well against *love, surprise, anticipation* and *aggressiveness*.

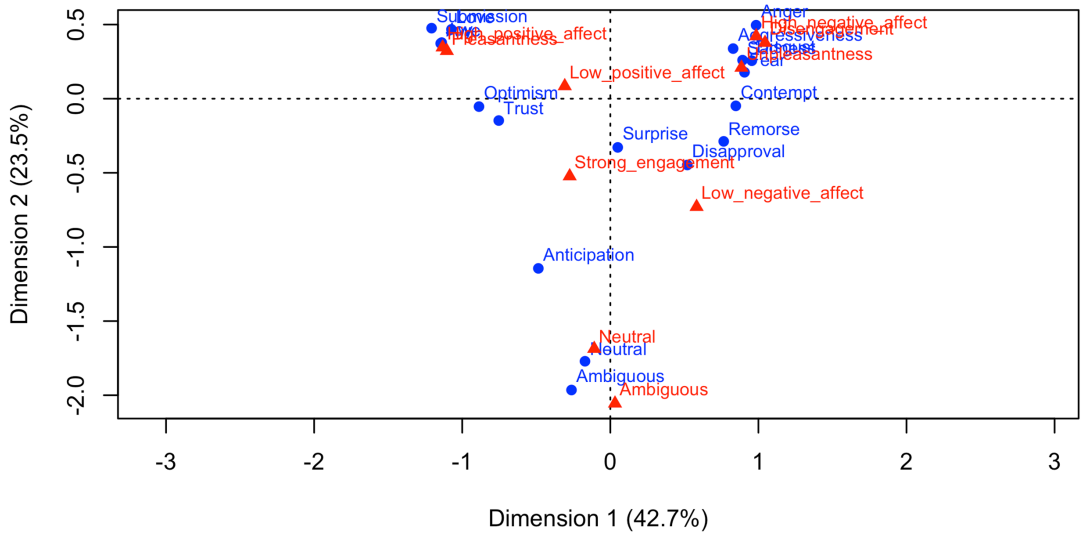


**Figure 5.10: Plutchik's wheel of emotion (blue) versus Circumplex (red)**

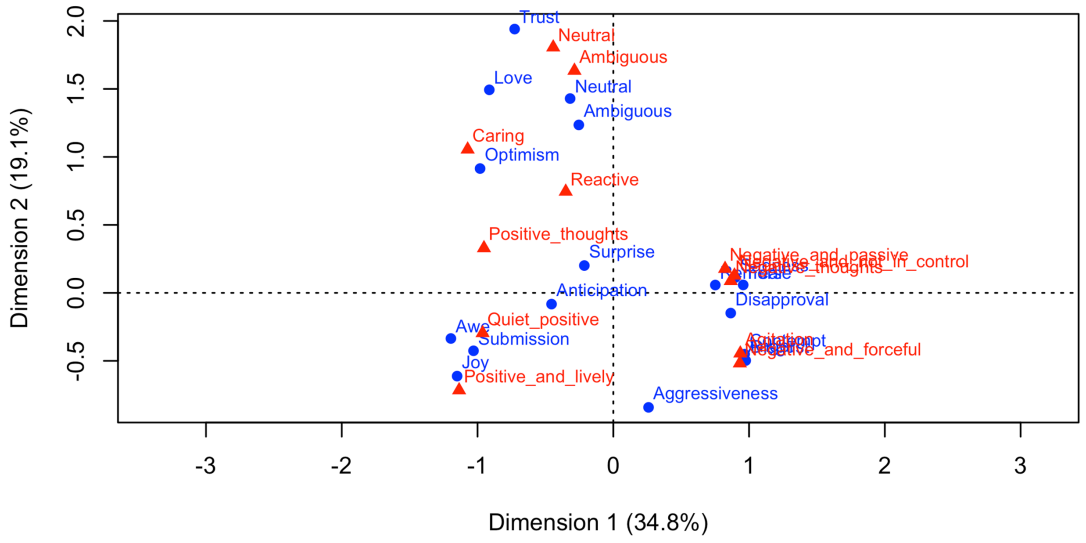


**Figure 5.11: Plutchik's wheel of emotion (blue) versus EARL (red)**

Finally, with few exceptions, Figure 5.12 illustrates a clear alignment between classes in Circumplex and EARL, suggesting that they cover, and partition, the semantic space of emotions in a similar way.

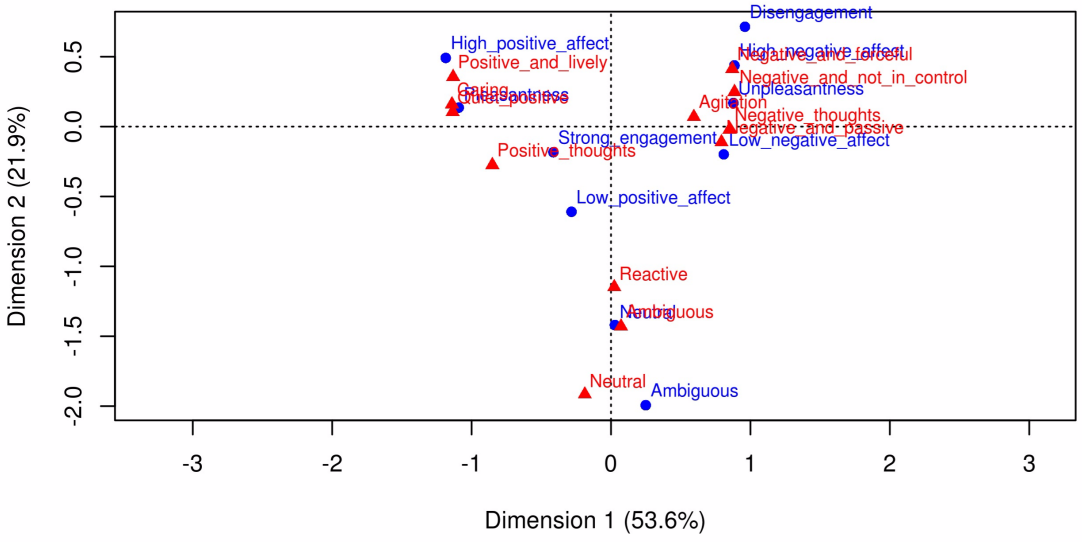


**Figure 5.12: Circumplex (blue) versus EARL (red)**

To summarise, both Circumplex and EARL provide generic classes, which align fairly well (see Figure 5.12). Unsurprisingly, they achieved similar IAA, which is not statistically different (see Figures 5.6). When compared with the frameworks that use specific emotions rather than generic classes, i.e. the six basic emotions and Plutchik's wheel of emotion, neither of the generic frameworks seem to model *surprise* well (see Figures 5.8 - 5.11). In addition, although EARL explicitly lists *love* as an example of the class *caring*, comparison with Plutchik's wheel of emotion shows no strong correspondence between the two (see Figure 5.11). The best results were seen with frameworks that use specific emotions, with the six basic emotions demonstrating significantly better IAA agreement, in comparison to Plutchik's wheel of emotion. However, the six basic emotions require a wider range of positive emotions. Figures 5.7 and 5.9 indicate that *happiness* is consistently used as a surrogate for *love*.

### 5.3.4 Annotator's Perception

In order to gain a qualitative insight into how human annotators interpret and use the frameworks, we conducted semi-structured interviews with 6 participants who had an academic background in social sciences. The annotation guidelines were explained to participants. Each participant was given a sample of 5 different text documents to annotate. They annotated the sample 6 times, once for each classification framework. Their experiences were then discussed in a semi-structured interview.

Table 5.6 provides the semi-structured interview guide. The interviews were recorded and transcribed verbatim. We conducted thematic analysis of the transcripts. The extracted themes (see Table 5.7) were related to annotators (subjectivity and certainty), data (context, ambiguity, and multiplicity), and the frameworks (coverage and complexity).

Generally, participants found the annotation task difficult, often not feeling confident about their choice. Annotators agreed that features such as punctuation (e.g. !), and words with strong sentiment polarity (e.g. beautiful, amazing, disgusting and horrible) were strong indicators of an emotion. The annotation choices for utterances that conveyed multiple emotions varied greatly across the annotators. For example, "My dress is so cute ugh. Praying no one wears the same one or else I will go ballistic," was interpreted to express both a positive and a negative emotion.

When context was absent (e.g. "Please stay away"), participants required more time to find an appropriate annotation. Annotators found themselves reading the text with different intonation in order to re-contextualise the underlying emotion. Upon failing to identify the context, annotators doubted their original annotation choice, claiming they may have over-compensated for the lack of context.

| Question | Prompt |
|---|---|
| How much effort was required to annotate whilst using this framework? | What factors of this framework made annotation easy/difficult? |
| | What factors of this framework did you like/dislike? |
| How did the number of classes affect your annotation choice? | Do you feel restricted with the number of classes on offer? |
| | Did it affect how much time you took when making a decision? |
| | Do you think this affected the annotation accuracy? |
| How accurate do you think your annotates were? | Were you annotating with a class most similar to the emotion you had interpreted? |
| | Is it fair to say without neutral and ambiguous, you'd be misclassifying? |
| | What was your thought process when an emotion was not available? |
| Were visual aids helpful during the exercise? | Did the colour framework/ framework structure influence your annotation or mean anything to you? |
| | How could it be improved? |
| | Why do you think this would improve it? |
| | What in particular was confusing about...? |
| What was your thought process when the text reflected multiple emotions? | Were you torn between two emotions? |
| | Were you certain about your annotation? |
| | Can you provide an example? |
| | Did you resort to using neutral and ambiguous? |
| | Would allowing you to choose two or more emotions be more helpful? |

**Table 5.6: Semi-structured interview guide**

| Theme | Definition | Examples |
|---|---|---|
| Subjectivity | Interpreting text differently | "*Gonna hate being home alone tonight...* I'm imagining it as an everyday situation, she's not scared... it's not sincere, it's not a very strong emotion." |
| | | "To me, 'father' is quite a distant term, so I'm not sure how the person feels about,their dad..." |
| Certainty | Doubting their choices | "*Happiness* just doesn't do it. It's not quite there. But if I were to re-annotate, it'd probably be *ambiguous*..." |
| | | "I was torn between *negative thoughts* and *negative and not in control*..." |
| Context | Having insufficient information | "I've put *sadness*, but it could also be *love*. It depends on the tone and when it was said..." |
| | | "The sentence would be different if there was an exclamation mark at the end... the full stop to me means *anger*..." |
| Ambiguity | Multiple possible interpretations | "I've got *ambiguous*. If it's about a job she could be anxious, or she could be ambitious and raring to go..." |
| Multiplicity | A range of emotions associated with a single sentence | "There were more options, but I wanted to choose two or three top ones or rate them in order..." |
| | | "I was torn between *disgust* and *contempt*..." |
| Coverage | How well the framework covers the emotion space | "I picked what I thought was the best, but I didn't think they fitted that well..." |
| | | "I wanted *excited* or a similar emotion..." |
| Complexity | The perception of complexity of the framework including its presentation | "It'd have to be explained to me before annotation, otherwise I'd just gloss over it..." |
| | | "I thought it was a positive statement, but I wasn't sure what kind, so I looked at the words under each category which helped me decide why it was *quiet positive*..." |

**Table 5.7: The summary of thematic analysis**

One significant factor affecting annotation was the number of classes available in a framework. In particular, for the six basic emotions, annotators found that the classes were meaningful, or relevant for distinguishing the polarity of the text, but not the types of emotion expressed. This is consistent with the results of the correspondence analysis described in Section 5.3.3, which decried *happiness* as a poor surrogate for *love*. The insufficient coverage of the emotion space in this framework significantly restricted the choices, resulting in a poor capture of the primary emotion conveyed. It became "unsatisfying" for participants to annotate with a class that was not fit for purpose, i.e. classes that do not map easily onto the emotional content. For example, "So proud of myself right now :D," was annotated as *happiness*, but also construes *pride*, an emotion that is distinct from *happiness* [197].

When faced with Plutchik's wheel of emotion, the annotators found it difficult to choose between related emotions. For example, annotators debated whether "I'll always think the world of you," expressed *love, trust* or *awe*. In this case, annotators agreed that having multiple options, in terms of intensity or similarity, would be more appropriate. The structure of the wheel received a negative response. Annotators agreed that it contained too much information and was quite complex to understand without additional explanation. There was debate that some emotions in the wheel (e.g. *trust*) are not necessarily emotions, but states, and questioned some emotion combinations (e.g. *anticipation* + *joy* = *optimism, sadness* + *surprise* = *disapproval*). Annotators felt that Plutchik's wheel, in comparison to the six basic emotions, provides better coverage, but lacks the ability to encode some emotions. For example, annotators required an emotion to represent *discomfort* for "My throat is killing me," but annotated it with *disapproval, sadness* and *neutral* instead.

Annotators felt the categories in both EARL and Circumplex were not distinct. An overlap between some classes (e.g. *negative & not in control* and *negative thoughts*) was named as one of the reasons for annotators' disagreement. Although conceptually similar, the dimensional structure of Circumplex and its choice of emotions caused

more resistance among annotators, as they misinterpreted the mapping of emotions onto their categories (e.g. they disagreed that *dull, sleepy* and *drowsy* were positive affects). For both EARL and Circumplex, very little attention was paid to the categories themselves. Annotators were in favour of the examples of emotions in each category, and felt that "once they had distinguished" the nuance of emotion being represented, they "had a general feeling as to which category it belonged to." Annotators appreciated having similar emotions clustered together into a generic category. This provided them with useful cues when classifying the general mood of the text, which may be easier than choosing a specific emotion. However, they acknowledged that some information would be lost when annotating with generic categories.

When faced with WNA, annotators were able to freely decide on a specific emotion in the hierarchy (e.g. "I don't feel well ugh," received *sick, miserable, unhappy* and *fed-up* annotations). Yet, annotators felt "restricted," as some of the emotions that they had interpreted in the text were not available (e.g. "I was certain it was *relief*, but it only had *relieve*, and they are not the same thing..."). The autocomplete functionality proved insufficient, as annotators continually searched for emotions that were not present in the lexicon. This increased the time spent in completing the annotation task. Emotional content was often described in complex terms (e.g. "I chose *aggravated*, because it's stronger than *annoyance*"), or could not be pinpointed (e.g. "You know what it means and you feel it, but you can't find the right word to describe it"). In these situations, annotators would search for the synonyms of their original choices, or search for the basic form of the emotion, until an appropriate substitute was found (e.g. "I wanted *exhausted*, but had to settle for *tired*..."). This may imply that WNA is somewhat incomplete. A recommendation for improving the navigation of this framework is to have similar emotions suggested in addition to the autocomplete functionality.

The free text classification framework diminished the confidence in choosing an appropriate annotation (e.g. "There is too much choice now. I think of a word, and doubt. Is this what I really mean? Because there is no guideline I doubt. When there is a group,

I think 'it definitely fits here"'). Annotators described this framework as "resembling what we do in everyday life when we read a piece of text. We read something and we feel it." However, when asked to describe an emotion using a particular word, annotators could often not articulate it (e.g. "I had multiple emotions, but could not find a word to describe them all"). Free text annotations accrued a range of lexical representations of emotions with similar valences (e.g. "My girlfriend disapproves of me :(" received *self-disgust, shame* and *disapproval* annotations), and intensities (e.g. "Don't want to see the hearse coming down my road today. RIP Anna," received *trepidation, dread, sadness* and *upset* annotations). For both lexical frameworks, annotators acknowledged that "regardless of the terms we use, we are all in agreement of the general feeling expressed" in the text.

## 5.4 Utility Analysis: A Machine Perspective

The second half of this investigation aims to measure the utility of our selected frameworks from a machine's perspective of automated sentiment analysis. In Section 5.4.1, we evaluate the performance of supervised machine learning when the corresponding annotations were used to train the classification algorithms.

### 5.4.1 Cross-Validation Experiments

To explore how well text classification algorithms can learn to differentiate between the classes within a given framework, we evaluated the performance of such algorithms when the corresponding annotations were used to train the classification model.

The ground truth annotations discussed in Section 5.3.2 were used to create a set of gold standards per classification framework, and were used with Weka [74] to perform 10-fold cross-validation experiments. All text documents from the corpus of 500

tweets described in Table 5.3 were converted into feature vectors using a BOW representation.

We tested a wide range of supervised learning methods included in Weka. SVM consistently outperformed other methods. We, therefore, report the results achieved by this method (see Figure 5.13). Classification performance was measured using precision, recall, and F-measure. Classification performance can be negatively affected by class imbalance and the degree of overlapping among the classes [161].

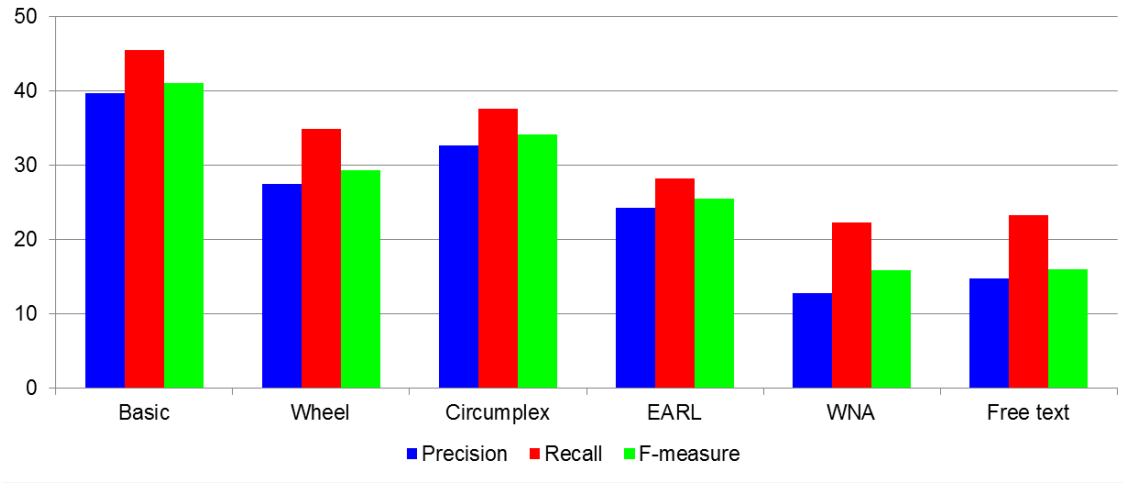


**Figure 5.13: The results of cross-validation experiments**

With respects to F-measure, the ranking of the classification frameworks is similar to the ranking with respect to the IAA, with the exception of the wheel of emotion and Circumplex. F-measure ranged from 15.9% to 41.0% with a standard deviation of 9.5. Notably, there was less variation in classification performance across the frameworks in comparison to IAA. Intuitively, the classification results are expected to be inversely proportional to the number of classes in the framework. Unsurprisingly, given the smallest number of options, the highest value F-measure (41.0%) was recorded for six basic emotions. However, EARL (10 classes) is ranked behind the wheel of emotion (16 classes). WNA and free text demonstrated an almost identical F-measure, and were found to have poorer classification performances.

To gain a better insight into the classification performance across the frameworks, we analysed the confusion matrices, which show how the automatically predicted classes compare against the actual classes from the gold standard. For each framework, confusion often occurred between opposite emotions, *happiness* and *sadness* (see Table 5.8). These confusions may be explained by the limitations of the BOW approach, which ignores the text structure, hence disregarding compositional semantics. Specifically, negation, which can reverse the sentiment of a text expression, was found to contribute to confusion. For example, "Why do you not love me? Why? :(" was automatically classified as *pleasantness, caring* or *happiness*, whereas it was annotated as *unpleasantness, negative & passive* and *depression* in the gold standard. Such predictions were largely based on the use of the word 'love,' which represents a text feature highly correlated with the positive classes in the training set. For example, out of 14 mentions of the word 'love,' 10 were used in a positive context (e.g. "I love my manager she is so sweet!"). The remaining 4 instances were used in a negative context, with 3 negated mentions (e.g. "Sitting in Tesco's cafe by myself because nobody loves me :("), and sarcasm (e.g. "That moment when a UNM shuttle drops people off and and then leaves without picking up the rest of us at the bus stop! Gotta love UNM!"). These examples illustrate the need to include negation as a salient feature in sentiment analysis.

| **Main emotion** | **Framework** | | | | | |
|---|---|---|---|---|---|---|
| | **Basic** | **Wheel** | **Circumplex** | **EARL** | **WNA** | **Free text** |
| *happiness* | 39 × *sadness* | 34 × *sadness* | 42 × *unpleasantness* (includes sad) | 20 × *negative & passive* (includes sadness) | 20 × *sadness* | 21 × *sadness* |
| *sadness* | 44 × *happiness* | 40 × *joy* | 42× *pleasantness* (includes happy) | 25 × *positive & lively* (includes joy and happiness) | 34 × *happiness* | 17 × *happiness* |

**Table 5.8: Misclassification of opposite emotions**

The second largest consistently occurring confusion was related to the *neutral* category (see Table 5.9), which, in the absence of discriminative features, was typically misclassified as one of two largest classes in the gold standard, i.e. either *happiness* or *sadness* (see Figure 5.6). Another trend found across all frameworks, was the misclassification of active negative emotions (e.g. *anger, annoyance,* and *disgust* were classified as *sad-*

*ness*). Again, because this behaviour is recorded consistently across all frameworks, this phenomenon may be explained by the limitations of the BOW approach. Further investigation is needed to determine whether a richer feature set (e.g. additional syntactic features to differentiate between active and passive voice) would help to better discriminate between these classes.

| **Main emotion** | **Framework** | | | | | |
|---|---|---|---|---|---|---|
| | **Basic** | **Wheel** | **Circumplex** | **EARL** | **WNA** | **Free text** |
| *happiness* | 26 | 18 | 22 | 11 | 22 | 11 |
| *sadness* | 22 | 19 | 24 | 6 | 8 | 6 |

**Table 5.9: Misclassification of the neutral category**

| **Main emotion** | **Framework** | | | | | |
|---|---|---|---|---|---|---|
| | **Basic** | **Wheel** | **Circumplex** | **EARL** | **WNA** | **Free text** |
| *anger, annoyance* or *disgust* | 18 | 29 | 23 | 11 | 17 | 26 |
| *sadness* | 7 | 5 | 11 | 6 | 4 | 9 |

**Table 5.10: Misclassification of active and passive negative emotions**

Whereas the classification confusions discussed above were common across all frameworks, it was notable that both dimensional frameworks, Circumplex and EARL, demonstrated relatively more confusion across a wider range of classes (see Tables 5.11 - 5.12). This suggests that their generic categories may not be sufficiently distinctive, and therefore, are not the best suited for emotion analysis.

| | | | Predicted | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **a** | **b** | **c** | **d** | **e** | **f** | **g** | **h** | **i** | **j** |
| **Actual** | Pleasantness | **a** | 69 | 42 | 7 | 6 | 0 | 0 | 0 | 0 | 8 | 0 |
| | Unpleasantness | **b** | 42 | 87 | 5 | 11 | 0 | 0 | 0 | 0 | 6 | 0 |
| | High positive affect | **c** | 26 | 11 | 3 | 4 | 0 | 0 | 0 | 0 | 4 | 0 |
| | High negative affect | **d** | 15 | 23 | 3 | 23 | 1 | 1 | 0 | 0 | 6 | 0 |
| | Low positive affect | **e** | 3 | 3 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |
| | Low negative affect | **f** | 4 | 6 | 1 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Strong engagement | **g** | 3 | 6 | 2 | 1 | 0 | 0 | 0 | 0 | 3 | 0 |
| | Disengagement | **h** | 0 | 0 | 0 | 1 | 0 | 0 | 0 | 0 | 0 | 0 |
| | Neutral | **i** | 22 | 24 | 4 | 3 | 0 | 0 | 0 | 0 | 6 | 0 |
| | Ambiguous | **j** | 0 | 2 | 0 | 0 | 0 | 0 | 0 | 0 | 1 | 0 |

**Table 5.11: Confusion matrix for the classification predictions against Circumplex classes.**

| | | | Predicted | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | **a** | **b** | **c** | **d** | **e** | **f** | **g** | **h** | **i** | **j** |
| **Actual** | Positive & lively | **a** | 55 | 12 | 4 | 11 | 4 | 2 | 2 | 20 | 0 | 2 |
| | Negative & forceful | **b** | 22 | 28 | 3 | 6 | 2 | 4 | 0 | 11 | 0 | 2 |
| | Caring | **c** | 5 | 3 | 14 | 2 | 1 | 1 | 0 | 1 | 0 | 0 |
| | Negative & not in control | **d** | 22 | 10 | 2 | 14 | 3 | 1 | 0 | 8 | 0 | 0 |
| | Positive thoughts | **e** | 14 | 4 | 1 | 5 | 5 | 1 | 0 | 7 | 0 | 0 |
| | Negative thoughts | **f** | 5 | 8 | 0 | 1 | 0 | 0 | 0 | 7 | 0 | 1 |
| | Quiet positive | **g** | 9 | 2 | 2 | 0 | 0 | 0 | 0 | 2 | 1 | 0 |
| | Negative & passive | **h** | 25 | 6 | 3 | 13 | 3 | 1 | 0 | 23 | 0 | 0 |
| | Reactive | **i** | 3 | 1 | 0 | 2 | 0 | 1 | 0 | 4 | 0 | 0 |
| | Agitation | **j** | 3 | 8 | 0 | 2 | 0 | 0 | 0 | 3 | 0 | 1 |

**Table 5.12: Confusion matrix for the classification predictions against EARL classes.**

## 5.5 Summary

The goal of this Chapter, was to identify an appropriate classification framework in terms of completeness and complexity. We selected six emotion classification frameworks and investigated their utility for sentiment analysis by exploring their ease of use by human annotators, as well as the performance of supervised machine learning algorithms when they were used to annotate training data. For this purpose, we assembled a corpus of 500 emotionally charged documents that were manually annotated under each framework using a crowdsourcing approach.

In order to quantitatively measure their utility from a human annotator perspective, we measured IAA using Krippendorff's alpha coefficient, according to which the frameworks were ranked as follows: (1) six basic emotions ($\alpha$ = 0.483), (2) wheel of emotion ($\alpha$ = 0.410), (3) Circumplex ($\alpha$ = 0.312), (4) EARL ($\alpha$ = 0.286), (5) free text ($\alpha$ = 0.205), and (6) WNA ($\alpha$ = 0.202). The six basic emotions framework was found to have a significantly higher IAA in comparison to other frameworks. However, correspondence analysis of annotations across the frameworks highlighted that basic emotions are oversimplified representations of such complex phenomena, and as such, are likely to lead to invalid interpretations, which are not necessarily reflected by high IAA. Specifically, the basic emotion of *happiness* was mapped to classes distinct to *happiness* in other frameworks, namely *submission, love* and *awe* in Plutchik's wheel, *high positive affect* (e.g. enthusiastic, excited, etc.) in Circumplex, and all positive classes in EARL including *caring* (e.g. love, affection, etc.), *positive thoughts* (e.g. hope, pride, etc.), *quiet positive* (e.g. relaxed, calm, etc.), and *reactive politeness* (e.g. interest, surprise, etc). Semi-structured interviews with the annotators also highlighted this issue. The framework of six basic emotions was perceived as having insufficient coverage of the emotion space, forcing annotators to resort to inferior alternatives, e.g. using *happiness* as a surrogate for *love*. Therefore, further investigation is needed into ways of better representing basic positive emotions, e.g. by considering *love, hope* and *pride*.

In the second part of this study, we explored how well text classification algorithms can

learn to differentiate between the classes within a given framework. Classification performance can be negatively affected by class imbalance and the degree of overlapping among the classes. In terms of feature selection, poorly defined classes may not be linked to sufficiently discriminative text features that would allow them to be identified automatically. To measure the utility of different frameworks in this sense, we created training datasets for each framework, and used them in cross-validation experiments to evaluate the classification performances. According to the F-measure, the classification frameworks were ranked as follows: (1) six basic emotions (F = 0.410), (2) Circumplex (F = 0.341), (3) wheel of emotion (F = 0.293), (4) EARL (F = 0.254), (5) free text (F = 0.159), and (6) WNA (F = 0.158). Unsurprisingly, the smallest framework, the six basic emotions, achieved a significantly higher F-measure in comparison to all other frameworks.

For each framework, confusion often occurred between opposite emotions (e.g. *happiness* and *sadness*) and equivalent categories (e.g. *positive & lively* and *positive thoughts*). These confusions may be explained by the limitations of the BOW approach, which ignores text structure, hence disregarding compositional semantics. Specifically, negation, which can reverse the sentiment of a text expression, was found to contribute to confusion. Another trend found across all frameworks was misclassification of active and passive negative emotions (e.g. *anger* vs. *sadness*). Again, this phenomenon may be explained by the limitations of the BOW approach. Further investigation is needed to determine whether a richer feature set (e.g. syntactic features) would help to better discriminate between related classes. The classification confusions in this investigation were commonly found across all frameworks and, as suggested, represent the effects of a document representation choice rather than specific classification frameworks. However, it was notable that both dimensional frameworks, Circumplex and EARL, demonstrated higher confusion across a wider range of classes. This suggests that their categories may not be sufficiently distinctive, and, therefore, are not the best suited for sentiment analysis.

To conclude, given our investigation, the six basic emotions emerged as the most suitable classification framework for sentiment analysis. Nonetheless, further investigation is needed into ways of extending basic emotions to encompass a variety of positive emotions, as *happiness*, the only representative of positive emotion, is forcibly used as a surrogate for a wide variety of distinct positive emotions, such as *love*.

# *Chapter 6*

# Conclusions & Future Work

*"Collecting one's thoughts."*

The research in this thesis was motivated by the fact that idioms, despite their significance, are underrepresented as features in sentiment analysis. As idioms often express an affective stance towards an entity or an event, we hypothesised that the inclusion of idiom-based features would reduce misclassification of sentiment, when such features are present. To estimate the significance of idioms as features of sentiment analysis, we used them alongside traditional sentiment analysis approaches and evaluated the classification performance. Our experiments provided strong evidence that the use of idiom-based features significantly improves sentiment classification results.

These results can stand to be improved in two ways. First, having had no prior knowledge of the significance of idioms for the sentiment classification task, we simply concatenated all features into a single vector. While sufficient to demonstrate the significance of idioms in sentiment analysis, such a brute-force approach does not guarantee an optimal performance. The latter was beyond the scope of the proposed research, but the lexico-semantic resources developed herein enable the community to conduct further research into different ways of utilising idioms as features of sentiment analysis. They include a comprehensive collection of almost 600 idioms manually annotated with sentiment polarity, with a reliable IAA ($\alpha$ = 0.662). To our knowledge, this dataset represents the largest lexico-semantic resource of this kind to be utilised

in sentiment analysis. Additionally, we implemented a set of local grammars that can be used to recognise occurrences of these idioms in text. We also assembled a corpus of over 2,500 sentences with a wide range of idioms used in context. Similarly to the idioms themselves, this corpus was also annotated with sentiment polarity, and as such, can be used in systematic evaluation of sentiment analysis approaches that claim to use idioms as features.

Another method for improving the performance of sentiment analysis is to simply increase the range of idioms covered by the aforementioned lexico-semantic resources. Rather than manually extending these resources, we wanted to investigate ways of automating this step, which would in turn improve the generality of our method and enable its portability to other languages. By doing so, we also wanted to address the main limitation of our original approach - the knowledge-engineering overhead involved in hand-crafting lexico-semantic patterns for the recognition of idioms in text, as well as the manual effort associated with acquiring the sentiment polarity of idioms. To minimise the bottleneck associated with the acquisition of lexico-semantic resources, in Chapter 4, we scaled up our original approach in order to be able to consider an arbitrary set of idioms by automating the engineering of such resources. In order to automate the recognition of idioms, we hypothesised that the canonical form of an idiom can be used to automatically derive rules that are robust enough to recognise its lexico-syntactic variations in a discourse. In order to automate the acquisition of idioms' sentiment polarity, we hypothesised that it is possible to automatically extract sentiment from their dictionary definitions. To evaluate the feasibility of this approach and, in doing so, test our hypothesis, we replaced the manually engineered lexico-semantic resources with their automatically generated counterparts and repeated the same classification experiments. As before, the classification performance of sentiment analysis was improved when such idiom-based features were present, which confirms our hypothesis. Despite the results being poorer than those achieved with manually engineered features, the advantage of the fully automated approach is that existing idiom dictionaries can be re-purposed, allowing an arbitrary lexicon of idioms to be explored as

part of sentiment analysis. Moreover, one of the alternative methods proposed in this part demonstrated that relevant idioms can be mapped to specific affects and thereby address a significant limitation of state-of-the-art sentiment analysis approaches by focusing on a full range of emotions as opposed to mere sentiment polarity.

This observation leads us to the investigation presented in Chapter 5, which concerns advancing the research area of sentiment analysis from sentiment classification to emotion classification. Such research is hindered by the lack of consensus among researchers on a standardised framework for classifying emotions. Our goal was to identify an appropriate classification framework, in terms of completeness and complexity, for sentiment analysis. We considered six emotion classification frameworks, and systematically investigated their utility from a human perspective, as well as from a supervised machine learning perspective. Our experiments provided evidence that the six basic emotions are best suited for sentiment analysis. However, both quantitative and qualitative analyses highlighted its major shortcoming of oversimplifying positive emotions. Nonetheless, the resources developed herein enable the community to conduct further research into fine-grained approaches to emotion classification and sentiment analysis, as well as ways of extending basic emotions to encompass a variety of positive emotions, as *happiness*, the only representative of positive emotion, is forcibly used as a surrogate for a wide variety of distinct positive emotions, such as *love*. The aforementioned resources include a corpus of 500 emotionally charged text documents manually annotated with emotions from our selected comprehensive frameworks.

## 6.1 Future Work

In this Section, we discuss the various ways in which the research in this thesis can be extended further in future work.

An important point to consider in the evaluation in Chapter 4, is that the performance of a fully automated approach to using idioms as features in sentiment analysis can

still be improved. To support the compatibility with the original study, we re-used the Bayesian Network classifier, which outperformed alternative machine learning algorithms in the cross-validation experiments performed on the original gold standard. The fact that the distribution of the training data changed when manually engineered features were replaced with those that are automatically engineered (see Tables 4.4 and 4.5), opens the possibility that another machine learning algorithm may produce a better performing classification model. In this case, we would perform further classification experiments, where the included idiom-based features are automatically generated, and investigate the performance of a variety of classifiers distributed as part of Weka.

A closer inspection of the confusion matrices in Chapter 4 (see Table 4.8 and Table 4.9), reveals that the use of idioms as features, whether manually or automatically engineered, improves the sensitivity of classification with respect to *positive* and *negative* polarities. Whereas manually engineered features seem to be more biased towards *positive* polarities, automatically engineered features were demonstrated to be more biased towards *negative* ones. This may be explained by the way in which the idiom polarities were encoded. The originally crowdsourced idiom polarities allowed for a fuzzy representation of polarities, by distributing the number of annotations across the three coordinates (positive, negative, and other). For example, the idiom *mind someone's own business* was originally represented by the polarity vector (0, 60, 40), which allows for different interpretations of the given idiom, depending on the context. On the other hand, the automatically extracted idiom polarities did not allow for such representation. For example, *mind someone's own business* was represented by the polarity vector (0, 100, 0), which indicates that its sentiment is strictly *negative*. This may be remedied by incorporating the notion of ambiguity, and/or intensity, into the idiom polarity representation. Off-the-shelf sentiment analysis tools, such as those used in our experiments, output the strength of the sentiment expressed in a text segment. A future hypothesis is that this information can be used to support fuzzy representations of automatically extracted sentiment polarities. The inclusion of this information may

also further improve the performance of a fully automated approach to using idioms as features in sentiment analysis.

Our primary interest regarding future work concerns expanding upon our existing research to tackle the more complex problem of emotion classification. In an attempt to automate the acquisition of the emotion of an idiom, we can utilise our existing approach in Chapter 4 to map idioms to nodes in WNA, which distributes the six basic emotions, along with other primary emotions, as part of its second hierarchical level (see Figure 2.6), coinciding with the results of our investigation in Chapter 5. However, we are still faced with the labour-intensive task of manually annotating the overall sentiment of idioms used in context. In this case, we can investigate the possibility of using a combination of POS, Word Sense Disambiguation (WSD) technologies, which identify the sense of a word in a given context, and our idiom recognition rules, to map contextual examples of idioms to nodes in WNA. We would compare these automatically acquired emotions to those collected in Chapter 3, where both idioms and contextual examples of idioms were manually annotated with categories from EARL.

# Bibliography

[1] Ahmed Abbasi, Hsinchun Chen, and Arab Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM Transactions on Information Systems (TOIS)*, 26(3):12, 2008.

[2] Apoorv Agarwal, Boyi Xie, Ilia Vovsha, Owen Rambow, and Rebecca Passonneau. Sentiment analysis of twitter data. In *Proceedings of the workshop on languages in social media*, pages 30–38. Association for Computational Linguistics, 2011.

[3] Khurshid Ahmad, David Cheng, and Yousif Almas. Multi-lingual sentiment analysis of financial news streams. In *Proceedings of the 1st International Conference on Grid in Finance*, 2006.

[4] Farrokh Alemi, Manabu Torii, Laura Clementz, and David C Aron. Feasibility of real-time satisfaction surveys through automated analysis of patients' unstructured comments and sentiments. *Quality Management in Healthcare*, 21(1):9–19, 2012.

[5] S Ali. Customer reviews: Sennheiser hd201 closed dynamic stereo headphones. *Amazon.co.uk*, 2006. [Online. Accessed 13 May. 2015].

[6] Cecilia Ovesdotter Alm and Richard Sproat. Emotional sequencing and development in fairy tales. In *International Conference on Affective Computing and Intelligent Interaction*, pages 668–674. Springer, 2005.

[7] Saima Aman. *Recognizing emotions in text*. PhD thesis, University of Ottawa (Canada), 2007.

[8] Saima Aman and Stan Szpakowicz. Identifying expressions of emotion in text. In *International Conference on Text, Speech and Dialogue*, pages 196–205. Springer, 2007.

[9] Jean-Yves Antoine, Jeanne Villaneau, and Anaïs Lefeuvre. Weighted krippendorff's alpha is a more reliable metrics for multi-coders ordinal annotations: experimental studies on emotion, opinion and coreference annotation. In *EACL 2014*, pages 10–p, 2014.

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

[11] Sitaram Asur and Bernardo A Huberman. Predicting the future with social media. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 492–499. IEEE, 2010.

[12] Akshat Bakliwal, Jennifer Foster, Jennifer van der Puil, Ron O'Brien, Lamia Tounsi, and Mark Hughes. Sentiment analysis of political tweets: Towards an accurate classifier. Association for Computational Linguistics, 2013.

[13] Alexandra Balahur, Ralf Steinberger, Mijail Kabadjov, Vanni Zavarella, Erik Van Der Goot, Matina Halkia, Bruno Pouliquen, and Jenya Belyaeva. Sentiment analysis in the news. *Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC 2010)*, 2010.

[14] Carmen Banea, Rada Mihalcea, and Janyce Wiebe. Multilingual sentiment and subjectivity analysis. *Multilingual natural language processing*, 6:1–19, 2011.

[15] Eugene Yuta Bann. Discovering basic emotion sets via semantic clustering on a twitter corpus. 2012.

[16] Luciano Barbosa and Junlan Feng. Robust sentiment detection on twitter from biased and noisy data. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*, pages 36–44. Association for Computational Linguistics, 2010.

[17] Jeremy Claude Barnes. Comparing the performance of knowledge-based and machine-learning approaches for the detection of emotions in an english text. 2015.

[18] Farah Benamara, Carmine Cesarano, Antonio Picariello, Diego Reforgiato Recupero, and Venkatramana S Subrahmanian. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. In *ICWSM*. Citeseer, 2007.

[19] BNC-Consortium. British national corpus. *URL: http://www.natcorp.ox.ac.uk/*, 3 (BNC XML Edition), 2014.

[20] Samuel A Bobrow and Susan M Bell. On catching on to idiomatic expressions. *Memory & Cognition*, 1(3):343–346, 1973.

[21] Erik Boiy, Pieter Hens, Koen Deschacht, and Marie-Francine Moens. Automatic sentiment analysis in on-line text. In *ELPUB*, pages 349–360, 2007.

[22] Johan Bollen, Huina Mao, and Alberto Pepe. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *ICWSM*, 11:450–453, 2011.

[23] Aleksander Buczynski and Aleksander Wawer. Shallow parsing in sentiment analysis of product reviews. In *Proceedings of the Partial Parsing workshop at LREC*, volume 2008, pages 14–18, 2008.

[24] Türkay Bulut and İlkay Çelik-Yazici. Idiom processing in l2: Through rose-colored glasses. *The reading matrix*, 4(2), 2004.

[25] Cristina Cacciari and Patrizia Tabossi. The comprehension of idioms. *Journal of memory and language*, 27(6):668–683, 1988.

[26] Zoraida Callejas and Ramon Lopez-Cozar. Influence of contextual information in emotion annotation for spoken dialogue systems. *Speech Communication*, 50(5):416–433, 2008.

[27] Erik Cambria. An introduction to concept-level sentiment analysis. In *Mexican International Conference on Artificial Intelligence*, pages 478–483. Springer, 2013.

[28] Erik Cambria, Amir Hussain, Catherine Havasi, Chris Eckl, and James Munro. Towards crowd validation of the uk national health service. *WebSci10*, pages 1–5, 2010.

[29] Erik Cambria, Andrew Livingstone, and Amir Hussain. The hourglass of emotions. In *Cognitive behavioural systems*, pages 144–157. Springer, 2012.

[30] Hsia-Ching Chang. A new perspective on twitter hashtag use: Diffusion of innovation theory. *Proceedings of the American Society for Information Science and Technology*, 47(1):1–4, 2010.

[31] François-Régis Chaumartin. Upar7: A knowledge-based system for headline sentiment tagging. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 422–425. Association for Computational Linguistics, 2007.

[32] Yejin Choi and Claire Cardie. Learning with compositional semantics as structural inference for subsentential sentiment analysis. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 793–801. Association for Computational Linguistics, 2008.

[33] Jacob Cohen. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46, 1960.

[34] Jacob Cohen. Weighted kappa: Nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[35] William W Cohen. Minorthird: Methods for identifying names and ontological relations in text using heuristics for inducing regularities from data. 2004.

[36] Roddy Cowie, Ellen Douglas-Cowie, Susie Savvidou, Edelle McMahon, Martin Sawey, and Marc Schröder. Feeltrace: An instrument for recording perceived emotion in real time. In *ISCA tutorial and research workshop (ITRW) on speech and emotion*, 2000.

[37] Antonio R Damasio. *The feeling of what happens: Body and emotion in the making of consciousness*. Houghton Mifflin Harcourt, 1999.

[38] Charles Darwin. The expression of the emotions in man and animals. *London, UK: John Marry*, 1872.

[39] Dipankar Das and Sivaji Bandyopadhyay. Sentence level emotion tagging. In *2009 3rd International Conference on Affective Computing and Intelligent Interaction and Workshops*, pages 1–6. IEEE, 2009.

[40] Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528. ACM, 2003.

[41] Anthony Christopher Davison and David Victor Hinkley. *Bootstrap methods and their application*, volume 1. Cambridge university press, 1997.

[42] Daantje Derks, Arjan Bos, and Jasper Von Grumbkow. Emoticons and online message interpretation. *Social Science Computer Review*, 2007.

[43] Laurence Devillers, Laurence Vidrascu, and Lori Lamel. Challenges in real-life emotion annotation and machine learning based detection. *Neural Networks*, 18(4):407–422, 2005.

[44] Ann Devitt and Khurshid Ahmad. Sentiment polarity identification in financial news: A cohesion-based approach. In *ACL*, volume 7, pages 1–8, 2007.

[45] Nicholas A Diakopoulos and David A Shamma. Characterizing debate performance via aggregated twitter sentiment. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, pages 1195–1198. ACM, 2010.

[46] Ernesto Diaz-Aviles, Claudia Orellana-Rodriguez, and Wolfgang Nejdl. Taking the pulse of political emotions in latin america based on social web streams. In *Web Congress (LA-WEB), 2012 Eighth Latin American*, pages 40–47. IEEE, 2012.

[47] Thomas G Dietterich. Approximate statistical tests for comparing supervised classification learning algorithms. *Neural computation*, 10(7):1895–1923, 1998.

[48] Xiaowen Ding, Bing Liu, and Philip S Yu. A holistic lexicon-based approach to opinion mining. In *Proceedings of the 2008 international conference on web search and data mining*, pages 231–240. ACM, 2008.

[49] Pedro Domingos and Michael Pazzani. On the optimality of the simple bayesian classifier under zero-one loss. *Machine learning*, 29(2-3):103–130, 1997.

[50] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[51] Paul Ekman. Universals and cultural differences in facial expressions of emotions. *J, Cole*, pages 207–283, 1972.

[52] Ahmed Esmin, Roberto L De Oliveira Jr, and Stan Matwin. Hierarchical classification approach to emotion recognition in twitter. In *Machine Learning and Applications (ICMLA), 2012 11th International Conference on*, volume 2, pages 381–385. IEEE, 2012.

[53] Andrea Esuli and Fabrizio Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *Proceedings of LREC*, volume 6, pages 417–422. Citeseer, 2006.

[54] Brian S. Everitt. The analysis of contingency tables, 1977.

[55] Ethan Fast, Pranav Rajpurkar, and Michael S Bernstein. Text mining emergent human behaviors for interactive systems. In *Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2265–2270. ACM, 2015.

[56] Chitra Fernando. *Idioms and idiomaticity*. Oxford University Press, USA, 1996.

[57] The Association for the Advancement of Affective Computing. Emotion annotation and representation language, 2014. [Version 0.4.0, 30 June 2006].

[58] Virginia Francisco and Pablo Gervás. Automated mark up of affective information in english texts. In *International Conference on Text, Speech and Dialogue*, pages 375–382. Springer, 2006.

[59] Bruce Fraser. Idioms within a transformational grammar. *Foundations of language*, pages 22–42, 1970.

[60] Michael Gamon. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *Proceedings of the 20th international conference on Computational Linguistics*, page 841. Association for Computational Linguistics, 2004.

[61] J Geertzen. Inter-rater agreement with multiple raters and variables. *URL: https://mlnl. net/jg/software/ira/*, 2012.

[62] Michel Genereux and Roger Evans. Distinguishing affective states in weblog posts. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 40–42, 2006.

[63] Michel Genereux and Roger Evans. Towards a validated model for affective classification of texts. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 55–62. Association for Computational Linguistics, 2006.

[64] Diman Ghazi, Diana Inkpen, and Stan Szpakowicz. Hierarchical approach to emotion recognition and classification in texts. In *Canadian Conference on Artificial Intelligence*, pages 40–50. Springer, 2010.

[65] Aniruddha Ghosh, Guofu Li, Tony Veale, Paolo Rosso, Ekaterina Shutova, John Barnden, and Antonio Reyes. Semeval-2015 task 11: Sentiment analysis of figurative language in twitter. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 470–478, 2015.

[66] Raymond W Gibbs and Herbert Colston. Figurative language. *Handbook of psycholinguistics, 2nd edn. Elsevier, Amsterdam*, pages 835–860, 2006.

[67] Raymond W Gibbs and Nandini P Nayak. Psycholinguistic studies on the syntactic behavior of idioms. *Cognitive psychology*, 21(1):100–138, 1989.

[68] Alec Go, Richa Bhayani, and Lei Huang. Twitter sentiment classification using distant supervision. *CS224N Project Report, Stanford*, 1(12), 2009.

[69] Lynn E Grant. A corpus-based investigation of idiomatic multiword units. 2003.

[70] Lynn E Grant. Frequency of core idioms in the british national corpus (bnc). *International Journal of Corpus Linguistics*, 10(4):429–451, 2005.

[71] Claire Greaves. Mental illness talk. URL: https://mentalillnesstalk.wordpress.com/, 2017. [Online. Accessed 02 Feb. 2017].

[72] Maurice Gross. The construction of local grammars. *Finite-state language processing*, page 329, 1997.

[73] Emma Haddi. *Sentiment analysis: text, pre-processing, reader views and cross domains*. PhD thesis, Brunel University London, 2015.

[74] 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.

[75] Maryam Hasan, Elke Rundensteiner, and Emmanuel Agu. Emotex: Detecting emotions in twitter messages. 2014.

[76] Vasileios Hatzivassiloglou and Janyce M Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of the 18th conference on Computational linguistics-Volume 1*, pages 299–305. Association for Computational Linguistics, 2000.

[77] Stefanie Haustein, Timothy D Bowman, Kim Holmberg, Andrew Tsou, Cassidy R Sugimoto, and Vincent Larivière. Tweets as impact indicators: Examining the implications of automated "bot" accounts on twitter. *Journal of the Association for Information Science and Technology*, 67(1):232–238, 2016.

[78] Kate Hevner. Experimental studies of the elements of expression in music. *The American Journal of Psychology*, 48(2):246–268, 1936.

[79] Hermann O Hirschfeld. A connection between correlation and contingency. In *Mathematical Proceedings of the Cambridge Philosophical Society*, volume 31, pages 520–524. Cambridge Univ Press, 1935.

[80] Lars E Holzman and William M Pottenger. Classification of emotions in internet chat: An application of machine learning using speech phonemes. *Retrieved November*, 27(2011):50, 2003.

[81] 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*, pages 168–177. ACM, 2004.

[82] Xiao Hu, J Stephen Downie, and Andreas F Ehmann. Lyric text mining in music mood classification. *American music*, 183(5,049):2–209, 2009.

[83] Zhiheng Huang, Wei Xu, and Kai Yu. Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*, 2015.

[84] Gordon Hughes. On the mean accuracy of statistical pattern recognizers. *IEEE transactions on information theory*, 14(1):55–63, 1968.

[85] Hossam S Ibrahim, Sherif M Abdou, and Mervat Gheith. Idioms-proverbs lexicon for modern standard arabic and colloquial sentiment analysis. *International Journal of Computer Applications*, 11:26–31, 2015.

[86] Carroll E Izard. *Human emotions.* New York: Plenum Press, 1977.

[87] Ray Jackendoff and Steven Pinker. The nature of the language faculty and its implications for evolution of language (reply to fitch, hauser, and chomsky). *Cognition*, 97(2):211–225, 2005.

[88] Nitin Jindal and Bing Liu. Mining comparative sentences and relations. In *AAAI*, volume 22, pages 1331–1336, 2006.

[89] David John, Anthony C Boucouvalas, and Zhe Xu. Representing emotional momentum within expressive internet communication. In *EuroIMSA*, pages 183–188, 2006.

[90] Alistair Kennedy and Diana Inkpen. Sentiment classification of movie reviews using contextual valence shifters. *Computational intelligence*, 22(2):110–125, 2006.

[91] Fazel Keshtkar and Diana Inkpen. A hierarchical approach to mood classification in blogs. *Natural Language Engineering*, 18(01):61–81, 2012.

[92] Soo-Min Kim and Eduard Hovy. Determining the sentiment of opinions. In *Proceedings of the 20th international conference on Computational Linguistics*, page 1367. Association for Computational Linguistics, 2004.

[93] Suin Kim, JinYeong Bak, and Alice Haeyun Oh. Do you feel what i feel? social aspects of emotions in twitter conversations. In *ICWSM*, 2012.

[94] Beata Beigman Klebanov, Jill Burstein, and Nitin Madnani. Sentiment profiles of multiword expressions in test-taker essays: The case of noun-noun compounds. *ACM Transactions on Speech and Language Processing (TSLP)*, 10(3):12, 2013.

[95] Manfred Klenner, Angela Fahrni, and Stefanos Petrakis. Polart: A robust tool for sentiment analysis. In *Proceedings of the 17th Nordic Conference on Computational Linguistics (NODALIDA 2009)*, pages 235–238, 2009.

[96] Moshe Koppel and Itai Shtrimberg. Good news or bad news? let the market decide. In *Computing attitude and affect in text: Theory and applications*, pages 297–301. Springer, 2006.

[97] Barry Kort, Rob Reilly, and Rosalind W Picard. An affective model of interplay between emotions and learning: Reengineering educational pedagogy-building a learning companion. In *Advanced Learning Technologies, 2001. Proceedings. IEEE International Conference on*, pages 43–46. IEEE, 2001.

[98] Efthymios Kouloumpis, Theresa Wilson, and Johanna D Moore. Twitter sentiment analysis: The good the bad and the omg! *Icwsm*, 11(538-541):164, 2011.

[99] Klaus Krippendorff. Content analysis: An introduction to its methodology, 1980.

[100] Klaus Krippendorff. Reliability in content analysis. *Human communication research*, 30(3):411–433, 2004.

[101] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[102] Peter J Lang. Behavioral treatment and bio-behavioral assessment: Computer applications. 1980.

[103] Michael Laver, Kenneth Benoit, and John Garry. Extracting policy positions from political texts using words as data. *American Political Science Review*, 97(02):311–331, 2003.

[104] Timothy Leary. *Interpersonal diagnosis of personality: A functional theory and methodology for personality evaluation*. Wipf and Stock Publishers, 2004.

[105] Joseph LeDoux and Jules R Bemporad. The emotional brain. *Journal of the American Academy of Psychoanalysis*, 25(3):525–528, 1997.

[106] Geoffrey Leech. 100 million words of english. *English Today*, 9(01):9–15, 1993.

[107] Nan Li and Desheng Dash Wu. Using text mining and sentiment analysis for online forums hotspot detection and forecast. *Decision support systems*, 48(2):354–368, 2010.

[108] Yu-Ru Lin, James P Bagrow, and David Lazer. More voices than ever? quantifying media bias in networks. *In proceedings of ICWSM*, 2011.

[109] Bing Liu. Sentiment analysis and subjectivity. *Handbook of natural language processing*, 2:627–666, 2010.

[110] Bing Liu. Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167, 2012.

[111] Changsheng Liu and Rebecca Hwa. Phrasal substitution of idiomatic expressions. In *Proceedings of NAACL-HLT*, pages 363–373, 2016.

[112] Dilin Liu. The most frequently used spoken american english idioms: A corpus analysis and its implications. *Tesol Quarterly*, pages 671–700, 2003.

[113] 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*, pages 125–132. ACM, 2003.

[114] Yue Lu, Malu Castellanos, Umeshwar Dayal, and ChengXiang Zhai. Automatic construction of a context-aware sentiment lexicon: an optimization approach. In *Proceedings of the 20th international conference on World wide web*, pages 347–356. ACM, 2011.

[115] 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.

[116] Sunghwan Mac Kim. *Recognising Emotions and Sentiments in Text*. University of Sydney, 2011.

[117] Karen A Machleit and Sevgin A Eroglu. Describing and measuring emotional response to shopping experience. *Journal of Business Research*, 49(2):101–111, 2000.

[118] Isa Maks and Piek Vossen. A lexicon model for deep sentiment analysis and opinion mining applications. *Decision Support Systems*, 53(4):680–688, 2012.

[119] Ryan McDonald, Kerry Hannan, Tyler Neylon, Mike Wells, and Jeff Reynar. Structured models for fine-to-coarse sentiment analysis. In *Annual Meeting-Association For Computational Linguistics*, volume 45, page 432. Citeseer, 2007.

[120] Albert Mehrabian. Silent messages: Implicit communication of emotions and attitudes. 1972.

[121] Albert Mehrabian. Pleasure-arousal-dominance: A general framework for describing and measuring individual differences in temperament. *Current Psychology*, 14(4):261–292, 1996.

[122] Yelena Mejova. Sentiment analysis: an overview. *Comprehensive exam paper, available on http://www. cs. uiowa. edu/˜ ymejova/publications/CompsYelenaMejova. pdf [2010-02-03]*, 2009.

[123] Yelena Mejova. *Sentiment analysis within and across social media streams*. University of Iowa, 2012.

[124] Rada Mihalcea and Hugo Liu. A corpus-based approach to finding happiness. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 139–144, 2006.

[125] George Miller. Wordnet: a lexical database for english communications of the acm 38 (11) 3941. *Niemela, I*, 1995.

[126] 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.

[127] Saif M Mohammad. # emotional tweets. In *Proceedings of the First Joint Conference on Lexical and Computational Semantics-Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation*, pages 246–255. Association for Computational Linguistics, 2012.

[128] Saif M Mohammad and Peter D Turney. Emotions evoked by common words and phrases: Using mechanical turk to create an emotion lexicon. In *Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text*, pages 26–34. Association for Computational Linguistics, 2010.

[129] Andrés Montoyo, Patricio MartíNez-Barco, and Alexandra Balahur. Subjectivity and sentiment analysis: An overview of the current state of the area and envisaged developments. *Decision Support Systems*, 53(4):675–679, 2012.

[130] Rosa E Vega Moreno. *Creativity and convention: The pragmatics of everyday figurative speech*, volume 156. John Benjamins Publishing, 2007.

[131] Andrius Mudinas, Dell Zhang, and Mark Levene. Combining lexicon and learning based approaches for concept-level sentiment analysis. In *Proceedings of the First International Workshop on Issues of Sentiment Discovery and Opinion Mining*, page 5. ACM, 2012.

[132] Tony Mullen and Nigel Collier. Sentiment analysis using support vector machines with diverse information sources. In *EMNLP*, volume 4, pages 412–418, 2004.

[133] Tony Mullen and Robert Malouf. A preliminary investigation into sentiment analysis of informal political discourse. In *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs*, pages 159–162, 2006.

[134] Myriam D Munezero, Calkin Suero Montero, Erkki Sutinen, and John Pajunen. Are they different? affect, feeling, emotion, sentiment, and opinion detection in text. *IEEE transactions on affective computing*, 5(2):101–111, 2014.

[135] Jin-Cheon Na, Haiyang Sui, Christopher Khoo, Syin Chan, and Yunyun Zhou. Effectiveness of simple linguistic processing in automatic sentiment classification of product reviews. *Advances in Knowledge Organization*, 9:49–54, 2004.

[136] 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: Volume 1-Volume 1*, pages 180–189. Association for Computational Linguistics, 2009.

[137] Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. Recognition of fine-grained emotions from text: An approach based on the compositionality principle. In *Modeling Machine Emotions for Realizing Intelligence*, pages 179–207. Springer, 2010.

[138] Marilyn A Nippold and Stephanie Tarrant Martin. Idiom interpretation in isolation versus context: A developmental study with adolescents. *Journal of Speech, Language, and Hearing Research*, 32(1):59–66, 1989.

[139] Geoffrey Nunberg, Ivan A Sag, and Thomas Wasow. Idioms. *Language*, pages 491–538, 1994.

[140] Bruno Ohana and Brendan Tierney. Sentiment classification of reviews using sentiwordnet. In *9th. IT & T Conference*, page 13, 2009.

[141] Neil O'Hare, Michael Davy, Adam Bermingham, Paul Ferguson, Páraic Sheridan, Cathal Gurrin, and Alan F Smeaton. Topic-dependent sentiment analysis of financial blogs. In *Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 9–16. ACM, 2009.

[142] Tim O'Keefe and Irena Koprinska. Feature selection and weighting methods in sentiment analysis. In *Proceedings of the 14th Australasian document computing symposium, Sydney*, pages 67–74. Citeseer, 2009.

[143] Alvaro Ortigosa, José M Martín, and Rosa M Carro. Sentiment analysis in facebook and its application to e-learning. *Computers in Human Behavior*, 31:527–541, 2014.

[144] Andrew Ortony. Understanding figurative language. *Handbook of reading research*, 1:453–470, 1984.

[145] Andrew Ortony, Gerald L Clore, and Allan Collins. *The cognitive structure of emotions*. Cambridge university press, 1988.

[146] Andrew Ortony and Terence J Turner. What's basic about basic emotions? *Psychological review*, 97(3):315, 1990.

[147] Charles Egerton Osgood, William H May, and Murray S Miron. *Cross-cultural universals of affective meaning*. University of Illinois Press, 1975.

[148] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, 2010.

[149] Bo Pang and Lillian Lee. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. In *Proceedings of the 42nd annual meeting on Association for Computational Linguistics*, page 271. Association for Computational Linguistics, 2004.

[150] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and trends in information retrieval*, 2(1-2):1–135, 2008.

[151] Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. Thumbs up? sentiment classification using machine learning techniques. In *Proceedings of the ACL-02 conference on Empirical methods in natural language processing-Volume 10*, pages 79–86. Association for Computational Linguistics, 2002.

[152] W Gerrod Parrott. *Emotions in social psychology: Essential readings*. Psychology Press, 2001.

[153] Rebecca J Passonneau, Tae Yano, Tom Lippincott, and Judith Klavans. Relation between agreement measures on human labeling and machine learning performance: Results from an art history image indexing domain. *Computational Linguistics for Metadata Building*, page 49, 2008.

[154] James W Pennebaker, Martha E Francis, and Roger J Booth. Linguistic inquiry and word count: Liwc 2001. *Mahway: Lawrence Erlbaum Associates*, 71:2001, 2001.

[155] John P Pestian, Pawel Matykiewicz, Michelle Linn-Gust, Brett South, Ozlem Uzuner, Jan Wiebe, Kevin B Cohen, John Hurdle, and Christopher Brew. Sentiment analysis of suicide notes: A shared task. *Biomedical informatics insights*, 5(Suppl. 1):3, 2012.

[156] Scott SL Piao, Paul Rayson, Dawn Archer, Andrew Wilson, and Tony McEnery. Extracting multiword expressions with a semantic tagger. In *Proceedings of the ACL 2003 workshop on Multiword expressions: analysis, acquisition and treatment-Volume 18*, pages 49–56. Association for Computational Linguistics, 2003.

[157] Rosalind Picard. *Affective computing*, volume 252. MIT press Cambridge, 1997.

[158] Robert Plutchik. A general psychoevolutionary theory of emotion. *Theories of emotion*, 1:3–31, 1980.

[159] Soujanya Poria, Erik Cambria, Gregoire Winterstein, and Guang-Bin Huang. Sentic patterns: Dependency-based rules for concept-level sentiment analysis. *Knowledge-Based Systems*, 69:45–63, 2014.

[160] Soujanya Poria, Alexander Gelbukh, Erik Cambria, Amir Hussain, and Guang-Bin Huang. Emosenticspace: A novel framework for affective common-sense reasoning. *Knowledge-Based Systems*, 69:108–123, 2014.

[161] Ronaldo C Prati, Gustavo Batista, and Maria Carolina Monard. Class imbalances versus class overlapping: an analysis of a learning system behavior. In *Mexican international conference on artificial intelligence*, pages 312–321. Springer, 2004.

[162] 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*, pages 482–491. Association for Computational Linguistics, 2012.

[163] James Pustejovsky and Amber Stubbs. *Natural language annotation for machine learning*. " O'Reilly Media, Inc.", 2012.

[164] Baojun Qiu, Kang Zhao, Prasenjit Mitra, Dinghao Wu, Cornelia Caragea, John Yen, Greta E Greer, and Kenneth Portier. Get online support, feel better–

sentiment analysis and dynamics in an online cancer survivor community. In *Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom)*, pages 274–281. IEEE, 2011.

[165] J. Ross Quinlan. Induction of decision trees. *Machine learning*, 1(1):81–106, 1986.

[166] Randolph Quirk, Sidney Greenbaum, Geoffrey Leech, Jan Svartvik, and David Crystal. *A comprehensive grammar of the English language*, volume 397. Cambridge Univ Press, 1985.

[167] Jonathon Read. Recognising affect in text using pointwise-mutual information. *Unpublished M. Sc. Dissertation, University of Sussex, UK*, 2004.

[168] Jonathon Read. Using emoticons to reduce dependency in machine learning techniques for sentiment classification. In *Proceedings of the ACL student research workshop*, pages 43–48. Association for Computational Linguistics, 2005.

[169] Vassiliki Rentoumi, George Giannakopoulos, Vangelis Karkaletsis, and George A Vouros. Sentiment analysis of figurative language using a word sense disambiguation approach. In *RANLP*, pages 370–375, 2009.

[170] Vassiliki Rentoumi, George A Vouros, Vangelis Karkaletsis, and Amalia Moser. Investigating metaphorical language in sentiment analysis: A sense-to-sentiment perspective. *ACM Transactions on Speech and Language Processing (TSLP)*, 9(3):6, 2012.

[171] Susanne Z Riehemann. *A constructional approach to idioms and word formation*. PhD thesis, stanford university, 2001.

[172] Ellen Riloff and Janyce Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical methods in nat-*

*ural language processing*, pages 105–112. Association for Computational Linguistics, 2003.

[173] David C Rubin and Jennifer M Talarico. A comparison of dimensional models of emotion: Evidence from emotions, prototypical events, autobiographical memories, and words. *Memory*, 17(8):802–808, 2009.

[174] Victoria L Rubin, Jeffrey M Stanton, and Elizabeth D Liddy. Discerning emotions in texts. In *The AAAI Symposium on Exploring Attitude and Affect in Text (AAAI-EAAT)*, 2004.

[175] James A Russell. Affective space is bipolar. *Journal of personality and social psychology*, 37(3):345, 1979.

[176] James A Russell. A circumplex model of affect. *Journal of Personality and Social Psychology*, 1980.

[177] James A Russell and Lisa Feldman Barrett. Core affect, prototypical emotional episodes, and other things called emotion: dissecting the elephant. *Journal of personality and social psychology*, 76(5):805, 1999.

[178] Franco Salvetti, Stephen Lewis, and Christoph Reichenbach. Automatic opinion polarity classification of movie. *Colorado research in linguistics*, 17(1):2, 2004.

[179] Klaus R Scherer. Emotion as a multicomponent process: A model and some cross-cultural data. *Review of Personality & Social Psychology*, 1984.

[180] Harold Schlosberg. The description of facial expressions in terms of two dimensions. *Journal of experimental psychology*, 44(4):229–237, 1952.

[181] Marc Schröder, Hannes Pirker, and Myriam Lamolle. First suggestions for an emotion annotation and representation language. In *Proceedings of LREC*, volume 6, pages 88–92, 2006.

[182] Fabrizio Sebastiani. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47, 2002.

[183] Julian Sedding and Dimitar Kazakov. Wordnet-based text document clustering. In *proceedings of the 3rd workshop on robust methods in analysis of natural language data*, pages 104–113. Association for Computational Linguistics, 2004.

[184] Satoshi Sekine and Kapil Dalwani. Ngram search engine with patterns combining token, pos, chunk and ne information. In *LREC*, 2010.

[185] Yong-Soo Seol, Dong-Joo Kim, and Han-Woo Kim. Emotion recognition from text using knowledge-based ann. In *ITC-CSCC: International Technical Conference on Circuits Systems, Computers and Communications*, pages 1569–1572, 2008.

[186] Lokendra Shastri, Anju G Parvathy, Abhishek Kumar, John Wesley, and Rajesh Balakrishnan. Sentiment extraction: Integrating statistical parsing, semantic analysis, and common sense reasoning. In *IAAI*, 2010.

[187] 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, 1987.

[188] Phillip Smith, Mark Lee, John Barnden, and Peter Hancox. *Sentiment analysis: beyond polarity*. PhD thesis, Thesis Proposal, School of Computer Science, University of Birmingham, UK, 2011.

[189] Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, Christopher Potts, et al. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the conference on empirical methods in natural language processing (EMNLP)*, volume 1631, page 1642. Citeseer, 2013.

[190] Swapna Somasundaran, Josef Ruppenhofer, and Janyce Wiebe. Detecting arguing and sentiment in meetings. In *Proceedings of the SIGdial Workshop on Discourse and Dialogue*, volume 6, 2007.

[191] Irena Spasic, Pete Burnap, Mark Greenwood, and Michael Arribas-Ayllon. A naïve bayes approach to classifying topics in suicide notes. *Biomedical informatics insights*, 5(Suppl. 1):87, 2012.

[192] Irena Spasic, Lowri Williams, and Andreas Buerki. Scaling up the extraction of idiom-based features in sentiment analysis. *Submitted to Transactions on Affective Computing*, 2017.

[193] Caroline Sporleder, Linlin Li, Philip Gorinski, and Xaver Koch. Idioms in context: The idix corpus. In *LREC*, 2010.

[194] Philip J Stone, Dexter C Dunphy, and Marshall S Smith. The general inquirer: A computer approach to content analysis. 1966.

[195] 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, 2008.

[196] Carlo Strapparava and Alessandro Valitutti. Wordnet affect: an affective extension of wordnet. In *LREC*, volume 4, pages 1083–1086, 2004.

[197] Gavin B Sullivan. Wittgenstein and the grammar of pride: The relevance of philosophy to studies of self-evaluative emotions. *New ideas in psychology*, 25(3):233–252, 2007.

[198] Maite Taboada, Julian Brooke, Milan Tofiloski, Kimberly Voll, and Manfred Stede. Lexicon-based methods for sentiment analysis. *Computational linguistics*, 37(2):267–307, 2011.

[199] Luke Kien-Weng Tan, Jin-Cheon Na, Yin-Leng Theng, and Kuiyu Chang. Sentence-level sentiment polarity classification using a linguistic approach. In

*International Conference on Asian Digital Libraries*, pages 77–87. Springer, 2011.

[200] Luke Kien-Weng Tan, Jin-Cheon Na, Yin-Leng Theng, and Kuiyu Chang. Phrase-level sentiment polarity classification using rule-based typed dependencies and additional complex phrases consideration. *Journal of Computer Science and Technology*, 27(3):650–666, 2012.

[201] Songbo Tan and Jin Zhang. An empirical study of sentiment analysis for chinese documents. *Expert Systems with applications*, 34(4):2622–2629, 2008.

[202] Robert E Thayer. *The origin of everyday moods: Managing energy, tension, and stress*. Oxford University Press, USA, 1997.

[203] Mike Thelwall. Sentistrength. URL: http://sentistrength.wlv.ac.uk/, 2014.

[204] 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.

[205] Matt Thomas, Bo Pang, and Lillian Lee. Get out the vote: Determining support or opposition from congressional floor-debate transcripts. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 327–335. Association for Computational Linguistics, 2006.

[206] S Tomkins. Affect theory. a kr scherer i p. ekman (eds.), approaches to emotion, 1984.

[207] Mikalai Tsytsarau and Themis Palpanas. Survey on mining subjective data on the web. *Data Mining and Knowledge Discovery*, 24(3):478–514, 2012.

[208] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G Sandner, and Isabell M Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. *ICWSM*, 10:178–185, 2010.

[209] Piyoros Tungthamthiti, Kiyoaki Shirai, and Masnizah Mohd. Recognition of sarcasms in tweets based on concept level sentiment analysis and supervised learning approaches. In *PACLIC*, pages 404–413, 2014.

[210] 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*, pages 417–424. Association for Computational Linguistics, 2002.

[211] 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.

[212] Frederik Vaassen, Jeroen Wauters, Frederik Van Broeckhoven, Maarten Van Overveldt, Walter Daelemans, and Koen Eneman. delearyous: Training interpersonal communication skills using unconstrained text input. In *European Conference on Games Based Learning*, page 505. Academic Conferences International Limited, 2012.

[213] Maarten van Gompel and Antal van den Bosch. Efficient n-gram, skipgram and flexgram modelling with colibri core. *Journal of Open Research Software*, 4(1), 2016.

[214] Bao-Khanh Vo and Nigel Collier. Twitter emotion analysis in earthquake situations. *International Journal of Computational Linguistics and Applications*, 4(1):159–173, 2013.

[215] Xiaolong Wang, Furu Wei, Xiaohua Liu, Ming Zhou, and Ming Zhang. Topic sentiment analysis in twitter: a graph-based hashtag sentiment classification approach. In *Proceedings of the 20th ACM international conference on Information and knowledge management*, pages 1031–1040. ACM, 2011.

[216] David Watson and Auke Tellegen. Toward a consensual structure of mood. *Psychological bulletin*, 98(2):219, 1985.

[217] Cynthia M Whissel. The dictionary of affect in language, emotion: Theory, research and experience: vol. 4, the measurement of emotions, r. *Plutchik and H. Kellerman, Eds., New York: Academic*, 1989.

[218] Cynthia Whissell. Using the revised dictionary of affect in language to quantify the emotional undertones of samples of natural language. *Psychological reports*, 105(2):509–521, 2009.

[219] Casey Whitelaw, Navendu Garg, and Shlomo Argamon. Using appraisal groups for sentiment analysis. In *Proceedings of the 14th ACM international conference on Information and knowledge management*, pages 625–631. ACM, 2005.

[220] Janyce Wiebe. Tracking point of view in narrative. *Computational Linguistics*, 20(2):233–287, 1994.

[221] Janyce Wiebe, Rebecca F Bruce, and Thomas P O'Hara. Development and use of a gold-standard data set for subjectivity classifications. In *Proceedings of the 37th annual meeting of the Association for Computational Linguistics on Computational Linguistics*, pages 246–253. Association for Computational Linguistics, 1999.

[222] Janyce Wiebe, Theresa Wilson, Rebecca Bruce, Matthew Bell, and Melanie Martin. Learning subjective language. *Computational linguistics*, 30(3):277–308, 2004.

[223] Janyce Wiebe, Theresa Wilson, and Claire Cardie. Annotating expressions of opinions and emotions in language. *Language resources and evaluation*, 39(2-3):165–210, 2005.

[224] Michael Wiegand, Alexandra Balahur, Benjamin Roth, Dietrich Klakow, and Andrés Montoyo. A survey on the role of negation in sentiment analysis. In *Proceedings of the workshop on negation and speculation in natural language processing*, pages 60–68. Association for Computational Linguistics, 2010.

[225] Lowri Williams, Michael Arribas-Ayllon, Andreas Artemiou, and Irena Spasić. Comparing the utility of different emotion classification schemes for emotive language analysis. *Submitted to Journal of Classification*, 2017.

[226] Lowri Williams, Christian Bannister, Michael Arribas-Ayllon, Alun Preece, and Irena Spasić. The role of idioms in sentiment analysis. *Expert Systems with Applications*, 2015.

[227] 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*, pages 347–354. Association for Computational Linguistics, 2005.

[228] David H Wolpert. The lack of a priori distinctions between learning algorithms. *Neural computation*, 8(7):1341–1390, 1996.

[229] Song-xian Xie and Ting Wang. Construction of unsupervised sentiment classifier on idioms resources. *Journal of Central South University*, 21:1376–1384, 2014.

[230] Changhua Yang, Kevin Hsin-Yih Lin, and Hsin-Hsi Chen. Emotion classification using web blog corpora. In *Web Intelligence, IEEE/WIC/ACM International Conference on*, pages 275–278. IEEE, 2007.

[231] Hong Yu and Vasileios Hatzivassiloglou. Towards answering opinion questions: Separating facts from opinions and identifying the polarity of opinion sentences. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 129–136. Association for Computational Linguistics, 2003.

[232] Gunel Izzaddin Yusifova. Syntactic features of english idioms. *International Journal of English Linguistics*, 3(3):133, 2013.

[233] Li Zhang, John A Barnden, Robert J Hendley, and Alan M Wallington. Exploitation in affect detection in open-ended improvisational text. In *Proceedings of the Workshop on Sentiment and Subjectivity in Text*, pages 47–54. Association for Computational Linguistics, 2006.

[234] Xu Zhe and AC Boucouvalas. Text-to-emotion engine for real time internet communication. In *Proceedings of International Symposium on Communication Systems, Networks and DSPs*, pages 164–168. Citeseer, 2002.