ROBERT GORDON UNIVERSITY ABERDEEN

Contextual Lexicon-based Sentiment Analysis for Social Media

Muhammad, Aminu Bui

A thesis submitted in partial fulfilment
of the requirements of Robert Gordon University
for the degree of Doctor of Philosophy

May 2016

Abstract

Sentiment analysis concerns the computational study of opinions expressed in text. Social media domains provide a wealth of opinionated data, thus, creating a greater need for sentiment analysis. Typically, sentiment lexicons that capture term-sentiment association knowledge are commonly used to develop sentiment analysis systems. However, the nature of social media content calls for analysis methods and knowledge sources that are better able to adapt to changing vocabulary. Invariably existing sentiment lexicon knowledge cannot usefully handle social media vocabulary which is typically informal and changeable yet rich in sentiment. This, in turn, has implications on the analyser's ability to effectively capture the context therein and to interpret the sentiment polarity from the lexicons.

In this thesis we use SentiWordNet, a popular sentiment-rich lexicon with a substantial vocabulary coverage and explore how to adapt it for social media sentiment analysis. Firstly, the thesis identifies a set of strategies to incorporate the effect of modifiers on sentiment-bearing terms (local context). These modifiers include: contextual valence shifters, non-lexical sentiment modifiers typical in social media and discourse structures. Secondly, the thesis introduces an approach in which a domain-specific lexicon is generated using a distant supervision method and integrated with a general-purpose lexicon, using a weighted strategy, to form a hybrid (domain-adapted) lexicon. This has the dual purpose of enriching term coverage of the general purpose lexicon with non-standard but sentiment-rich terms as well as adjusting sentiment semantics of terms. Here, we identified two term-sentiment association metrics based on Term Frequency and Inverse Document Frequency that are able to outperform the state-of-the-art Point-wise Mutual Information on social media data. As distant supervision may not be readily applicable on some social media domains, we explore the cross-domain transferability of a hybrid lexicon. Thirdly, we introduce an approach for improving distant-supervised sentiment classification with knowledge from local context analysis, domain-adapted (hybrid) and emotion lexicons. Finally, we conduct a comprehensive evaluation of all identified approaches using six sentiment-rich social media datasets.

Keywords: Sentiment Analysis, SentiWordNet, Contextual analysis, Domain Adaptation, Hybrid Sentiment Lexicon, Distant Supervision, Emotion Features.

Declaration of Authorship

I declare that I am the sole author of this thesis and that all verbatim extracts contained in the thesis have been identified as such and all sources of information have been specifically acknowledged in the bibliography. Parts of the work presented in this thesis have appeared in the following publications:

- A. Muhammad, N. Wiratunga, R. Lothian, R. Glassey: Contextual Sentiment Analysis of social media Using High-Coverage Lexicon. In: Proc. of SGAI International Conference on Artificial Intelligence, BCS SGAI (2013)
  (**Chapter 4**)

- A. Muhammad, N. Wiratunga, R. Lothian: A Hybrid Sentiment Lexicon for Social Media Mining. In: Proc. of the 26th IEEE International Conference on Tools with Artificial Intelligence, IEEE ICTAI (2014)
  (**Chapter 5**)

- A. Muhammad, N. Wiratunga, R. Lothian and R. Glassey: Domain-based Lexicon Enhancement for Sentiment Analysis. In: Proc. SGAI Workshop on Social Media Mining. BCS SGAI (2013)
  (**Chapter 5**)

- A. Muhammad, N. Wiratunga, R. Lothian: Context-Aware Sentiment Analysis of Social Media. In: M.M. Gaber et. al (Eds), Advances in Social Media Analysis (2015). Springer
  (**Chapters 4 and 5**)

Acknowledgements

Firstly, I wish to express my profound appreciation to my supervisors, Dr Nirmalie Wiratunga and Dr Robert Lothian. You always went the extra mile to offer me support and advice whenever I needed them - thank you very much. I am also grateful to Dr Richard Glassey who used to form part of my supervisory team.

This research benefited from a generous funding by the Nigerian government (through the PTDF). I am very grateful for this. By the same token, I wish to thank my employer, Usmanu Danfodiyo University (UDU) for granting me a study fellowship for this research. Here, I wish to particularly thank Prof. T.M. Bande (former VC, UDU) and Mal. U.U. Bunza (former Registrar, UDU) for their kind efforts towards giving me all the assistance I needed from the university.

I also appreciate the insightful feedbacks I received from the SIS research group, IDEAS and SICSA research communities. I am also grateful to my colleagues at N426, the staff of CSDM and all those who played a part in making this research possible.

Finally, I would like to express my gratitude to my family for their unwavering love, encouragement and prayers.

Contents

List of Figures

List of Tables

Dedicated to my parents; to my beloved wife, Khadija; to my lovely sons, AbdurRahman and Al-Bashir; and to the memory of my brother, Usman (1987-2013) - may his soul rest in perfect peace, Ameen

Chapter 1

Introduction

Information on the Web has risen exponentially over the last decade due to the rapid increase in human activities enabled by the Web 2.0 technologies such as the social media platforms. It is now estimated that, in just 60 seconds, over 400,000 Twitter posts are shared, about 300,000 Facebook statuses updated, about 25,000 items purchased from Amazon, over 5 million Youtube videos viewed and about 2.7 million Google searches are made among many other things (see Figure 1.1). Opinions, being central to all human activities and key influencers of our behaviour, constitute a substantial amount of information posted or searched for, on the Web. This is evident from the fact that, in addition to opinion sites such as *Epinions.com*, *rottentomatoes.com*, and *cnet.com* which focus on collecting both professional and amateur reviews for numerous products and services; social media platforms such as Twitter, Facebook and Discussion forums enable virtually anyone to publish opinions on the Web.

Sentiment analysis or *opinion mining* concerns the study of opinions and related concepts such as evaluations, attitudes and affects (Liu, 2012). More specifically, the main tasks of sentiment analysis comprise the extraction of the five components of an opinion: the opinion polarity (positive or negative), the object and the specific aspects of the target to which the opinion refers to, the holder of the opinion and the time at which the opinion was expressed (Liu, 2010). Although the terms "sentiment analysis" and "opinion mining" are often used interchangeably especially in academia, sentiment analysis is the preferred term for industry practitioners (Liu, 2012). The two terms essentially

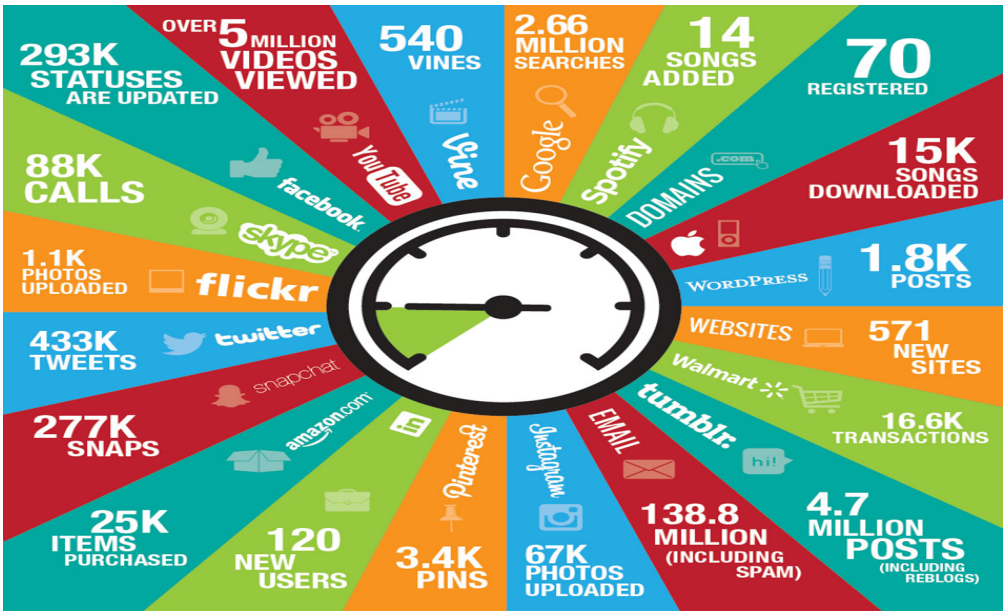

FIGURE 1.1: Human activities on the Web in 60 seconds[1]

represent the same field of study, although there exists a subtle difference between them (Liu, 2012). This difference is associated with the idea that sentiment is always polarised as 'positive' or 'negative' while an opinion may be unpolarised, for example, in *"I think he will go to Canada next year"* (Liu, 2015). In this thesis, we adopt the term sentiment analysis.

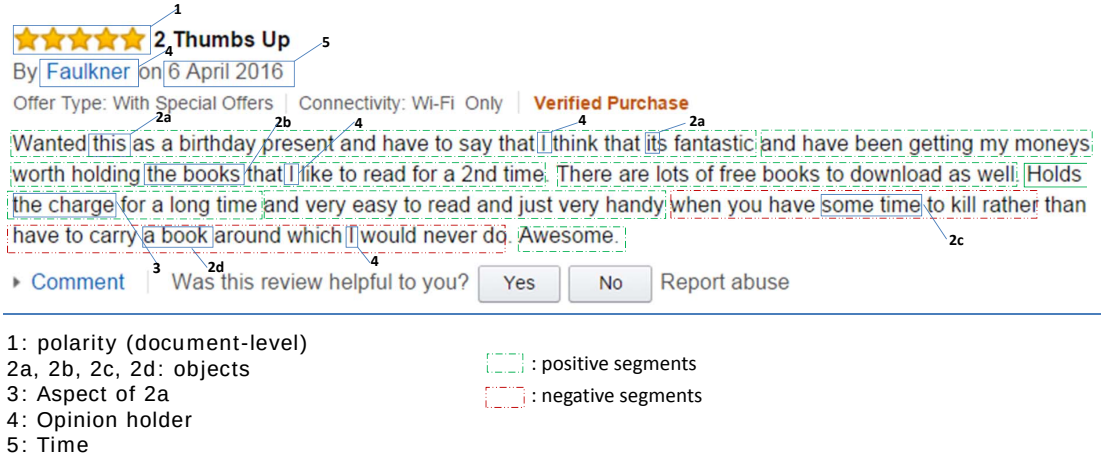

FIGURE 1.2: An Amazon Customer Review

Figures 1.2 and 1.3 show sentiments expressed in a review and a Twitter post (tweet) respectively. Both "documents" contain some components of an opinion that sentiment analysis aims to extract, such as polarities, objects, opinion holders and the times the

[1]source: http://blog.qmee.com/online-in-60-seconds-infographic-a-year-later/

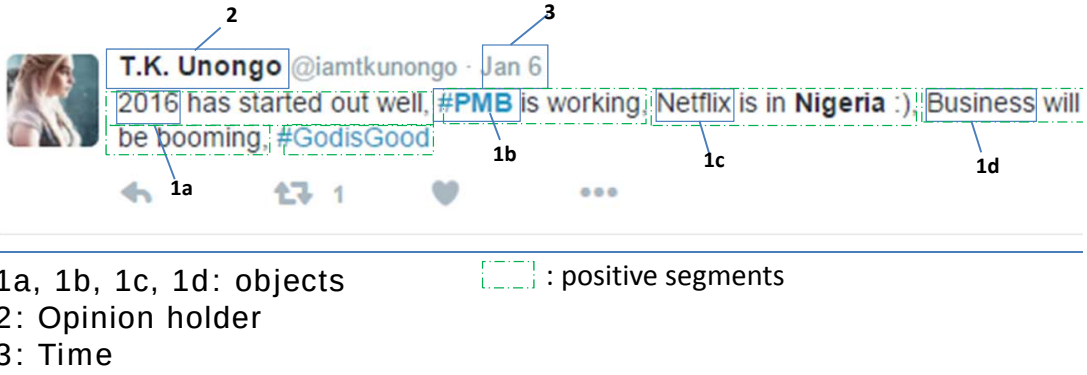

FIGURE 1.3: A Tweet

opinions were expressed. All the five components of an opinion are essential for a richer sentiment analysis (Liu, 2015), however, not all may be present in an opinion. For instance, whereas the review in Figure 1.2 contains an opinion on an aspect of the object being reviewed (*"holds charge"*), the opinions in the tweet (Figure 1.3) were expressed directly on objects. Furthermore, although the five-component definition in (Liu, 2012) introduced additional components over earlier definitions (such as in Wiebe (1994)) and covers most opinion expressions, it does not cover some complex opinion expressions. For instance, it does not cover the situation in *"The view finder and the lens are too close"* (i.e. opinion on the distance between two parts) (Liu, 2012). The two Figures also highlight some of the challenges of sentiment analysis. For instance, "holds charge" is an implicit mention of the aspect *"battery life"* of the tablet being reviewed and this needs to be resolved by a sentiment analysis system. The use of pronouns for entities also needs to be resolved.

The task of opinion polarity extraction, often referred to as sentiment classification, involves identifying the sentiment class (positive or negative) of an opinion. Such a sentiment class may be identified for the whole document (document-level). For instance, the sentiment class of both the review (Figure 1.2) and the tweet (Figure 1.3) is positive. However, as can observed in the review, polarity may change from one text segment to another with different objects/aspects being mentioned. This motivates a more fine-grained sentiment classification for individual sentences (sentence-level) and for individual object/aspect mentions (aspect-level). Nonetheless, document-level sentiment classification can reasonably support a variety of social media applications, working on

the assumption that there is an overall opinion, particularly for short texts (e.g in Figure 1.3). In this thesis, we concentrate on sentiment classification at the document-level.

1.1 Applications of Sentiment Analysis

The need to know other people's sentiment about objects has always been part of the human information needs. Traditionally, when an individual needs sentiment information, they ask friends and family; while organisations, companies and governments conduct surveys and opinion polls (Liu, 2015). These traditional channels of acquiring sentiment data tend to produce very limited and structured data that is manually manageable. The abundance of opinionated information about objects on the Web, too large to be managed manually, creates a suitable ground for automated sentiment analysis applications. For instance, a sentiment analysis system can be developed to determine consumer attitude on products/services from review data (e.g. Amazon customer reviews, Figure 1.2). Such a system is useful from the manufacturer's (or service provider's) perspective, to assess consumer perceptions, and from the consumer's perspective, to gain insights from other consumer opinions when making purchase decisions. This has been the target application of many sentiment analysis research works leading to the development of systems such as Opinion Observer (Liu et al., 2005), OpinionMiner (Jin et al., 2009) and OpinionFinder (Wilson et al., 2005).

Sentiment analysis is also abundantly employed in non-retail applications. Now, there is a proliferation of tools to quantify sentiment from platforms such as Twitter and Facebook. In some of these tools, typically objects are identified in social media and sentiment is extracted about these objects (e.g. sentiment140.com) while some assess general sentiment as an influencer of real-life activities such as stock market prediction (e.g. marketpsych.com). For instance, sentiment analysis was shown to complement and inform public opinion polling when several surveys on consumer confidence and political opinion over the 2008 to 2009 period were found to correlate with sentiment word frequencies in Twitter messages over the same period (O'Connor et al., 2010). Similarly, there is evidence that the moods of the nation, as measured by tweets, correlate with

changes in stock prices (Bollen et al., 2011). Also, sentiment analysis has been applied on tweets to forecast box-office revenue for movies (Asur and Huberman, 2010).

Industry Applications of Sentiment Analysis. Outside academic research, applications of sentiment analysis have spread to many different sectors and industries such as healthcare, tourism, hospitality, financial services, social events and political elections. In the healthcare industry, for example, a sentiment analysis system can be developed from patient satisfaction surveys or comments. Such responses can then be classified as positive, negative or neutral toward healthcare delivery topics and stakeholders such as a treatment received while in the care of Nurses and Doctors. Such an analysis can provide insights for hospitals to identify what is working and where there is a need for improvement. In tourism and hospitality, sentiment analysis plays an important role by providing summarised user sentiments about various stakeholders and their services, as users go on-line to book their travels, accommodations and tourism sites to visit.

There now exists hundreds of companies , start-ups and established corporations, that have built sentiment analysis capabilities either for themselves or for their clients. These companies include Google, Microsoft, Hewlett-Packard, Amazon, SAS, Oracle, Adobe, Bloomberg and Facebook (Liu, 2015). Other smaller, start-up, companies include Lexalytics, Semantria, Synapsify, ThriveMetrics, Etuma and MeshLabs. For example, the Facebook's Gross National Happiness interface provides estimated happiness of people on Facebook, by countries, by analysing the use of positive and negative words in the people's status updates (Cohen, 2009). Another company, *Sentex.com*, tracks sentiment in relation to specific politicians and political topics and provides sentiment analysis, positive negative and neutral, and the reasons for the sentiment. Other companies include *VivoText.com* which aims to develop a realistic text-to-speech tool that enables the portrayal of emotion and *crimsonhexagon.com* which developed a sentiment analysis tool studying biases in media coverage.

Despite the prolific systems already in existence, sentiment analysis still remains an open research field owing to its ever-expanding application domains, linguistic nuances,

differing contexts and even cultural factors making it challenging to automatically assess a piece of text for sentiment.

1.2 Related Research Fields

Sentiment analysis research has over the years been influenced by advances in Natural Language Processing (NLP), Information Retrieval (IR) and Text Classification. Generally, knowledge-rich representation and extraction strategies are drawn from NLP whilst knowledge-light and prediction strategies draw from IR and text classification respectively.

Natural Language Processing. NLP is the field of computer science concerned with the interaction between computers and human languages. Therefore, it is clearly relevant in sentiment analysis since sentiment is typically expressed in the form of unstructured free text. Sentiment analysis is closely related to some of the techniques developed in NLP such as the method of splitting text into individual words (tokenization), mapping words to their root forms (lemmatization) and the process of marking-up words corresponding to particular part-of-speech (PoS tagging). These techniques are typically available from standard NLP suites such as the GATE[2] and StanfordCoreNLP[3], but they need an extension to address peculiar challenges of sentiment analysis particularly applied to the informal/non-standard social media content. It can be noted, however, that such extensions are already underway in addition to new NLP tools developed specifically for social media platforms (e.g. TweetNLP[4]). Also, NLP draws from computational linguistics and statistics to develop rules to handle human language. Such rules are also essential for sentiment analysis, for instance, in lexicon generation and contextual analysis. However, existing NLP rules are often agnostic of social media requirements. This is an area we explore in this thesis.

[2]https://gate.ac.uk/
[3]http://nlp.stanford.edu/software/corenlp.shtml
[4]http://www.cs.cmu.edu/ark/TweetNLP/index.html

Information Retrieval. IR research (van Rijsbergen., 1979) has also had a significant impact on sentiment analysis research. It is concerned with the problem of identifying a set of documents, from amongst a larger collection, which are most relevant to a given query. IR has had its greatest impact on the web in the form of search engines such as Google or Bing. The basic text representation methods for IR are based on the vector space model employing feature weighting schemes such as the Term Frequency and Inverse Document Frequency. These schemes have been directly used for sentiment analysis. They also influence the development of other weighting schemes targeted at sentiment analysis. Likewise, sentiment analysis has influenced advances in IR in recent years.

Given the volume of opinionated content on the Web, it is not surprising that sentiment analysis feeds into IR in the indexing and retrieval of sentiment-rich information usually from social media platforms and review portals. The term *sentiment retrieval* has been introduced to signify document retrieval based on topic relevance as well as sentiment polarity criteria. In sentiment retrieval, the assumption is that the user's aim is to find relevant documents that contain opinions, for example, about a query such as *what do people think about the new iphone?*. Despite the advancement in IR technology, sentiment retrieval is still very challenging partly because IR systems are designed with the main objective of finding relevant documents to user's query typically based on the "bag-of-words" model rather than linguistic structures that can capture textual context. However, a drawback to using these structures is their dependence on language constructs which are problematic for multilingual systems. In the recent past, the IR community decided to take a number of measures to bridge the IR-to-sentiment analysis gap. These include the initiation of the opinion retrieval task as part of the Blog Track of the Text Retrieval Conference (TREC).

Text Classification. Text classification involves the task of automatically classifying a set of documents into a set of predefined classes. This is mostly done using supervised machine learning techniques (Mitchell, 1997). In the context of sentiment analysis, a supervised learning algorithm is trained on a set of sentiment labelled training documents.

Such documents are typically represented as vectors that lie within a space whose dimensions correspond to a sub-set of selected features[5] from the original training documents. Once training is complete, the algorithm would then be expected to correctly predict the class of a previously unseen test document that follows the same document-to-label distribution as the training set. A drawback in the applicability of supervised text classification is the need for labelled training data. Several solutions to this problem have been proposed for sentiment analysis, for instance, transfer learning and distant supervision. These solutions may also be useful in the context of lexicon-based sentiment analysis. This thesis explores their utility in lexicon-based methods.

1.3 Research Motivation

The task of sentiment classification involves labelling of text with a sentiment class. Several methods have been employed drawing from supervised/unsupervised machine learning and lexicon-based unsupervised strategies. Inspired by the field of text classification, supervised methods make use of machine learning algorithms trained with sentiment-labelled data to predict the sentiment class of unlabelled test documents. This approach becomes problematic when reliable and sufficient training data is difficult to obtain - a characteristic of non-review-based social media where content is not associated with ratings that could be exploited as "noisy" labels. Similarly, sentiment classifiers tend to be highly domain/genre specific performing well on the domain/genre of training but poorly on a different domain/genre. However, social media text is diverse in domains and genre ranging from political to lifestyle discussions with short messages (e.g. tweets) and lengthy posts (e.g blogs). Therefore, a system for analysing social media text needs to maintain consistent performance across domains/genres. This is a characteristic of the lexicon-based methods to sentiment classification.

The lexicon-based methods involve aggregation of sentiment polarity scores from a sentiment lexicon to classify opinionated text into sentiment classes. Many sentiment lexicons suffer from low term coverage and poor granularity of sentiment information. For example, General Inquirer (Stone et al., 1966) and Bing Liu (Hu and Liu, 2004) lexicons

[5] typically words contained in documents

contain only 4,216 and 6,789 unique sentiment-bearing terms respectively. Both lexicons do not distinguish for polarity strength and Bing Liu's lexicon does not distinguish between different parts of speech of the same term. In contrast, SentiWordNet (Baccianella et al., 2010) presents high term coverage of 28,431 unique sentiment-bearing terms distinguished by part-of-speech and contextual meaning (i.e. word sense). Furthermore, scores in this lexicon indicate sentiment strength in the range between 0 and 1. This allows for deeper linguistic analysis and score aggregation for sentiment prediction.

Despite the existence of high-coverage lexicons such as SentiWordNet, the performance accuracy of lexicon-based sentiment prediction remains lower when compared to the accuracies from machine learning methods Kolchyna et al. (2015). This is because the polarity with which a sentiment-bearing term appears in a piece of text (i.e. contextual polarity) can be different from its prior polarity offered by a lexicon. Two forms of semantic difference seem to contribute to this semantic gap. First, a difference in local context arising from the interaction of a term with sentiment modifiers. For example, the prior polarity of 'good' is positive, however, such polarity is changed in 'not good'. Second, the difference in domain semantics arising from the difference in the typical sentiment polarity of a term captured by a lexicon and the term's domain- or genre-specific polarity. For example, in the text 'the movie sucks', although the term 'sucks' seems to be rich in sentiment, this may not be reflected by a general purpose sentiment lexicon. Also, as sentiment lexicons are static resources, they need to be equipped with a strategy to adapt to changing vocabulary and sentiment over time, a characteristic of social media.

In addition to sentiment lexicons, there exists emotion lexicons that associate terms with emotion polarities such as *love*, *joy*, *surprise*, *sadness*, *anger* and *fear*. These resources are useful for sentiment analysis since most of the emotion polarities can be mapped onto positive and negative sentiment classes (Gonçalves et al., 2013). However, emotion knowledge may not be completely mapped to sentiment knowledge through emotion-to-sentiment lexicon mapping. For instance, the emotion class, surprise, is ambiguous as it could correspond to positive or negative sentiment. Also, although emotion and sentiment are inter-related, they are known to be theoretically different (Munezero et al.,

2014). Therefore, emotion knowledge should not be reduced to sentiment knowledge but when used carefully may help with sentiment analysis.

In order to address issues discussed above in relation to lexicon-based sentiment analysis, this thesis explores the following research questions:

1. Does the accuracy of lexicon-based sentiment analysis benefit from the integration of local context knowledge?
2. How can we evolve a static lexicon to dynamically adapt to vocabulary and domain-specific semantics in social media?
3. How does emotion knowledge captured in an emotion lexicon influence sentiment analysis?

1.4 Research Objectives

This thesis investigates the role of contextual analysis, domain adaptation and emotion knowledge for sentiment classification in social media employing a lexicon with rich sentiment information (SentiWordNet). Specifically, we address the following six objectives:

1. Conduct a comparative analysis of score extraction methods for SentiWordNet with a focus on using local context for word sense disambiguation.
2. Develop a lexicon-based classifier to integrate local context knowledge with sentiment content in SentiWordNet
3. Extend the classifier developed in 2 to address the continuously evolving vocabulary typical in social media streams
4. Investigate the utility of combining the local context analysis (in 2) and vocabulary adaptation (in 3) in the context of a hybrid sentiment classifier
5. Study the role of emotive concepts by integrating emotion knowledge into the classifier developed in 4.

6. Conduct a comprehensive evaluation of all developed classifiers/strategies.

1.5 Contributions

Figure 1.4 highlights the main contributions of this thesis within generic sentiment classification methods that employ lexicons (lexicon-based and hybrid). The figure focused on showing those components impacted by this research. In the lexicon-based, first, scores are extracted from a lexicon. These scores are then adjusted for local context, adapted to domain semantics and finally aggregated for sentiment prediction. The hybrid involves combining lexicon-based strategies with supervised machine learning. In this thesis, we concentrate on distant-supervised learning as it does not require hand-labelled data. It begins with unlabelled data on which distant supervision is applied to obtain labelled data. These labelled data is then represented in a format suitable for supervised learning and enriched with lexicon-based knowledge. It is expected that a contribution in any of the aforementioned stages will improve classification accuracy on test data.

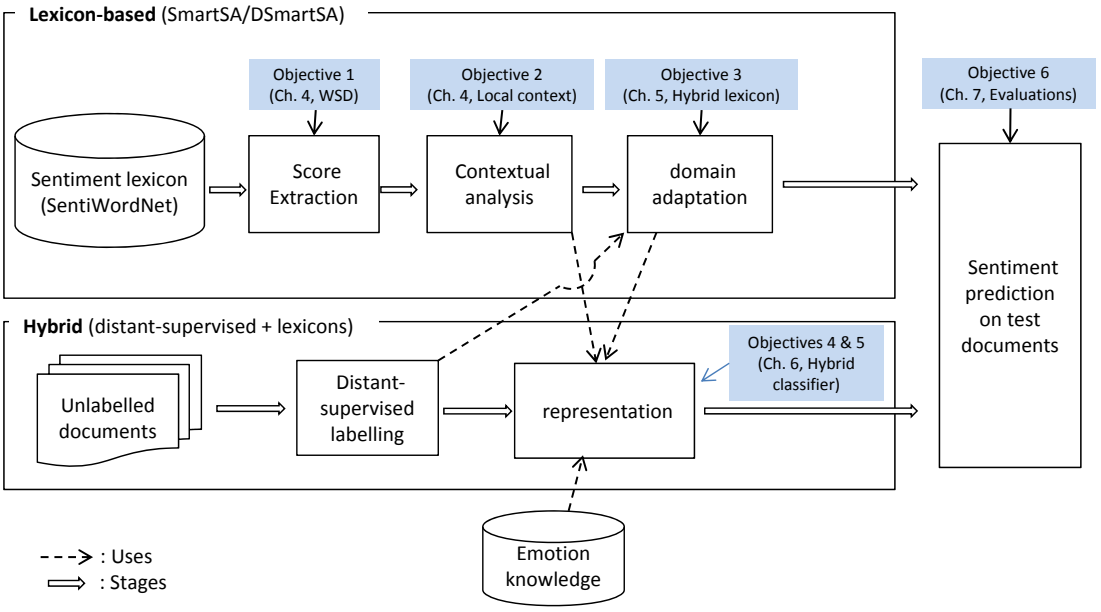

FIGURE 1.4: Objectives/Contributions Within Typical Classification Framework

The first significant contribution of this research is the development of a lexicon-based sentiment classifier (SMARTSA) that integrates contextual analysis strategies to adjust

prior polarities of terms in order to account for the effect of both standard and social media oriented sentiment modifiers as well as discourse structures. A key advantage of SmartSA is that it is an entirely heuristic-based unsupervised classifier that exploits the rich sentiment information from SentiWordNet. Thus, the system does not require any training data and, as it is developed using a general-purpose lexicon, its performance has a tendency to remain consistent across social media domains.

A second significant contribution is the development of an approach to dynamically improve lexical coverage and sentiment semantics of terms given a social media domain (DSmartSA). An important aspect of DSmartSA is that it combines sentiment knowledge from a general purpose lexicon and a target domain to create a hybrid lexicon. In doing so, it is able to capture non-standard but sentiment rich terms (i.e. improve coverage) and non-standard usage of terms for sentiment expression in social media. Another novel feature in DSmartSA is the introduction of two new term-sentiment association metrics inspired by Term Frequency and Inverse Document Frequency (TF, TFIDF). This is important because the state-of-the-art metrics, based on the Point-wise Mutual Information (PMI) do not work well on terms that have low frequencies in a collection (Sani, 2014), a characteristic of evolving terms in social media.

Our third major contribution is the development of a hybrid social media sentiment classifier that combines distant-supervised learning, contextual analysis, domain semantics and an emotion lexicon. This classifier benefits from the deeper analysis of supervised machine learning algorithms, local and domain context analysis without the overhead of requiring hand-labelled data. It also allows us to measure the extent to which our lexicon-based strategies and emotion knowledge are applicable in the hybrid sentiment classification setting.

Other secondary contributions of this research include the introduction of a word sense disambiguation algorithm for the extraction of sentiment scores from SentiWordNet. We conducted a detailed evaluation of this approach in comparison with the typical approaches used for the task. We also exploit transfer learning and assess the transferability of a hybrid lexicon (used by DSmartSA) on a social media domain different from the one it was generated from. This is important since distant-supervised data

(required to generate a hybrid lexicon) may not be available from some social media genres.

1.6 Thesis Overview

The rest of this thesis is outlined as follows: In Chapter 2 we present a review of literature related to sentiment analysis. We discuss three approaches to sentiment classification: machine learning, lexicon-based and hybrid. A more detailed discussion of lexicon-based methods is presented as these closely relate to the work presented in this thesis. In particular, we look at the importance of contextual analysis and the need for adaptability when applying a static lexicon such as SentiWordNet to social media content.

In Chapter 3, we present background details about the main sentiment lexicon and the baseline classification algorithms used in this research. We also provide details about the evaluation datasets, text pre-processing operations and performance metrics employed.

Chapter 4 presents SmartSA, a lexicon-based sentiment classification system for social media. SmartSA uses SentiWordNet as its lexicon. We start with the introduction of our word sense disambiguation algorithm in relation to existing approaches to sentiment score extraction from SentiWordNet. We then present the integration of contextual analysis with SmartSA. This includes lexical/non-lexical modifiers, social media oriented modifiers and discourse structures. the chapter closes after a presentation of the formal algorithm of SmartSA.

In Chapter 5, we present our hybrid lexicon approach aimed at dynamically extending vocabulary and sentiment context of terms in a general purpose lexicon. We begin this chapter with a discussion of our data labelling approach using distant supervision followed by the process of generating a domain-specific lexicon including our proposed term-sentiment association metrics. We then discuss the process of generating the hybrid lexicon combining the domain-specific lexicon with the general purpose lexicon. Thereafter, we present insights from social media data to illustrate that each of the lexicons (domain-specific and general purpose) can considerably contribute to the vocabulary

and sentiment context of terms in the hybrid lexicon. Finally, we conclude the chapter with a general discussion and summary.

The hybrid sentiment classifier is introduced in Chapter 6. It exploits local contextual analysis (introduced in Chapter 4) and social media adaptation (as captured by the hybrid lexicon introduced in Chapter 5). We also discuss a novel strategy for utilising knowledge from an emotion lexicon with focus on sentiment classification.

A comparative study of all relevant sentiment classification strategies discussed in Chapters 4, 5 and 6 together with baselines appear in Chapter 7. These include the evaluation of: SmartSA and an ablation test to study the contribution of each individual strategy integrated within the system; hybrid lexicon in comparison to a static- or domain-only lexicons; the transferability of the hybrid lexicon from one social media platform to another; and the performance of our distant-supervised hybrid classification approach that employs sentiment and emotion lexicons.

We conclude this thesis in Chapter 8 with a summary of our main contributions and desirable extensions for future work.

Chapter 2

Literature Review

In this chapter, we present a review of the existing literature related to the task of sentiment classification. Broadly, three methods have been adopted from supervised learning to lexicon-based unsupervised strategies and combined hybrid approaches. Research presented in this thesis focuses on the use of lexicon-based methods. We discuss all the three approaches and justify our preferences. We conclude with a discussion on the current research gap that this thesis will seek to address in relation to lexicon-based sentiment analysis for social media.

2.1 Machine Learning Methods

The vast majority of research in sentiment classification concentrates on the use of machine learning techniques, both *supervised* and *unsupervised*.

2.1.1 Supervised

Inspired by the field of topic-based text classification, supervised methods make use of machine learning algorithms trained with sentiment-labelled data to predict the sentiment class of unlabelled test documents. This is depicted in Figure 2.1. First, standard text pre-processing, feature engineering and vector-space representation are applied to

the training and test documents drawn from a problem domain. Thereafter, at the training phase, a machine learning algorithm is applied to learn a prediction model which is then used, at the testing/prediction phase, to classify documents that are previously unseen by the model. Of these components, feature engineering is perhaps the most crucial for classification. It is the process of using knowledge from the target problem domain to create features that make machine learning effective. This process involves the discovery of features on which to represent data (feature discovery), the removal of redundant features (feature selection) and the proposal of values to use in text representation (feature weighting).

A pioneer work employing supervised machine learning and with binary vector representation (presence or absence of individual words) has demonstrated that unlike with traditional text classifier, sentiment classifiers tended to result in lower accuracies (Pang et al., 2002). This indicates that the difficulty level of sentiment classification is more than that of topic classification. One of the reasons why there exists this disparity was that sentiment is expressed in a more subtle manner that the basic representation is unable to adequately capture. Accordingly, more advanced linguistic features were explored to enrich the representation, for instance: the addition of two consecutive words (bigram) features, the use of part-of-speech (PoS) tags to disambiguate between different usage of the same term and positional information (Pang et al., 2002). Surprisingly, these alternatives turned out to be less effective compared to the basic binary-valued unigram representation. Such findings have driven the need for more sophisticated feature engineering techniques.

Prominent contributions in feature discovery include the use of syntactic relations in addition to traditional features (Mullen and Collier, 2004, Xia and Zong, 2010), appraisal groups (e.g. *'very good'* or *'not terribly funny'*) (Whitelaw et al., 2005) and feature subsumption hierarchies (Riloff et al., 2006). Text representation for supervised machine learning typically employs the bag-of-words (BOW) model which disregards interdependencies between terms, some of which are crucial for sentiment classification

(e.g. negation and intensification). This problem is addressed by introducing appropriate features following contextual analysis (e.g. *'neg_good'* and *'int_good'* for a negated and intensified *'good'* respectively) (Kennedy and Inkpen, 2006).

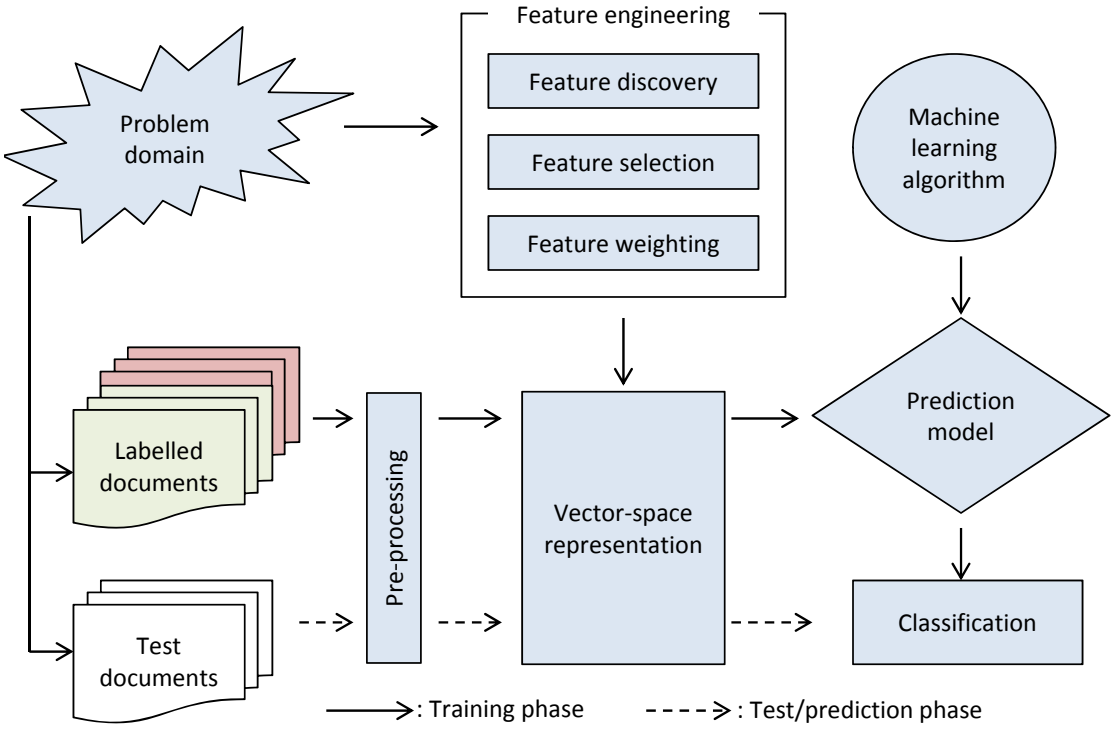

FIGURE 2.1: Supervised Machine Learning

The BOW model is unable to cope with variation in natural language vocabulary (e.g. synonymy and polysemy) which often requires semantic indexing approaches (Tsatsaronis and Panagiotopoulou, 2009). These approaches produce a generalisation of document representations away from low-level expressions (n-grams) to high-level semantic concepts. Several techniques have been proposed for transforming document representations from the space of individual terms to that of latent semantic concepts. These include Latent Semantic Indexing (LSI) which uses singular-value decomposition to exploit co-occurrence patterns of terms and documents in order to create a semantic concept space which reflects the major associative patterns in the corpus (Deerwester et al., 1990). Other simpler approaches use statistical measures of term co-occurrence within documents to infer semantic similarity (Wiratunga et al., 2004). However, representations produced using these approaches are not optimal for sentiment classification because they do not take into account class membership of documents (Chakraborti et al., 2007). To address this limitation, a technique called Supervised Sub-Spacing (S3) was proposed

for introducing supervision to term-relatedness extraction (Sani et al., 2013). S3 works by creating a separate sub-space for each class within which term relations are extracted. Evaluation results show S3 to outperform state-of-the-art classifiers that employed the BOW representation.

Feature selection is important in machine learning not only because it reduces feature space but because it can improve classification accuracy by providing a less redundant feature subset. Thus, both traditional and sentiment-oriented feature selection techniques have been explored for sentiment classification. A comparison of four traditional feature selection techniques: Information Gain, Mutual Information, Chi-squared Test and Document Frequency; shows Information Gain to perform best on sentiment categorization of Chinese documents (Tan and Zhang, 2008). Similar performance was also observed on English movie reviews (Sharma and Dey, 2012). Research also shows that feature selection based on the Fisher‘s discriminant ratio is further able to improve upon Information Gain (Wang et al., 2011). Notable work on feature selection specifically targeted towards sentiment analysis include the extension of Information Gain to address the fact that sentiment classes have ordinal relationships (Mukras et al., 2007) as opposed to having no obvious relationship as in text classification. Similarly, a genetic algorithm based feature selection was introduced for sentiment classification in different languages (Abbasi et al., 2008) and the use of a matrix factorization method to identify words with strong inter-sentiment distinction and intra-sentiment similarity (Liang et al., 2015).

Other supervised sentiment classification work concentrates on feature weighting schemes. The applicability of the existing TFIDF-based schemes from the field of Information Retrieval (IR) have been the focus of many studies, where the results show variants of this scheme to increase sentiment classification accuracy (Paltoglou and Thelwall, 2010). However, it can be noted that IR metrics such as TFIDF calculate term weights using statistics from the entire corpus and remain agnostic to class labels. This limitation has been addressed with Delta-TFIDF in which the calculation of TFIDF is restricted to documents from the same class (positive and negative) and the overall term weight is the difference between the two (Martineau and Finin, 2009).

User reviews form the main domain of choice on which supervised sentiment classification has been evaluated. This is because in addition to providing reliable opinionated content, this data typically includes star-rating information that can conveniently be used as sentiment labels for documents. For instance, a review (document) accompanied with a 1 or 2 star-rating denotes negative sentiment while that which is rated with 4 or 5 stars is positive. This approach has been extensively used to generate sentiment analysis datasets although it is not without problems. For instance, it is common for users to express their sentiment using just the stars without any accompanying text. It is also possible for the star information of a review to disagree from the review text (e.g. in sarcasm). Thus, some quality checks are useful to ensure the validity of using star-rating information in generating a gold-standard, training dataset for sentiment analysis. Increasingly opinion is also often expressed in non-review social media (such as in twitter posts, comments on news, and discussion forums). However in these settings access to labelled training data is a particular challenge for supervised sentiment classification. Several techniques have been proposed to overcome this challenge. These include the use of co-training algorithms that start with a few (human) labelled data alongside a large unlabelled sample from which relationships between the two datasets are explored to learn labels for the unlabelled data as well as perform classification (Blum and Mitchell, 1998, Li, Huang, Zhou and Lee, 2010, Liu et al., 2013). Transfer learning is also another alternative. Here, a classifier is trained in a domain where labelled training data is available or easy to obtain (e.g. product reviews) but adapted and tested in another domain where training data is difficult to obtain (e.g. discussion posts) (Pan et al., 2010, Pan and Yang, 2010). More recently, distant supervision has been proposed as a means to exploit reliable signals (e.g. emoticons and hashtags) within documents (usually twitter posts) as "noisy" class labels (Davidov et al., 2010, Go et al., 2009, Pak and Paroubek, 2010, Read, 2005). The class labels are noisy because they are automatically assigned using heuristics and thus may not be entirely correct. However, evaluation results from machine learning schemes trained with distant-supervised data but tested on hand-labelled data show the approach to be effective, attaining up to 83% accuracy on a combination of unigram and bigram features (Go et al., 2009).

Although the labelling problem is addressed by the afore-mentioned techniques, the

problem partly remains. Co-training still requires some initial labelled data, which, overall performance tends to improve in direct proportion to its size. Transfer learning, on the other hand, relies on labelled data from one domain to perform classification on another and, despite the adaptation overhead it incurs, it tends to produce lower accuracy compared to when within domain labelled data is employed. As for the distant supervision, it is applicable only when reliable and sufficient signals are available from a domain. These challenges make unsupervised sentiment classification, which does not require any labelled training data, an attractive alternative.

2.1.2 Unsupervised

The typical workflow of unsupervised machine learning sentiment classification is shown in Figure 2.2. It involves the use of probabilistic topic modelling methods to detect both topic and sentiment from a collection of unlabelled documents after a text pre-processing step. Prior knowledge in the form of seed sentiment-bearing terms is required to guide the process. Thereafter the sentiment class of a test document can be determined based on the topic/sentiments used to compose the document. Standard topic modelling approaches assume a three layered hierarchical framework, where topics are associated with documents, and words are associated with topics. For sentiment detection, this framework is extended with an additional sentiment layer in between documents and topics or with sentiment classes as additional topic models. In Mei et al. (2007), the probabilistic latent semantic indexing (pLSI) (Hofmann, 1999) was used to develop Topic-Sentiment Mixture (TSM) model which reveal latent topics including sentiment classes as additional topics. Figure 2.3 illustrates TSM whereby sentiment classes (+ and -) are modelled as topics from which d_1 and d_3 draw terms. The modelling also involves an additional filter layer which separates a background model (B) that capture general English words (e.g. 'the', 'a', 'of') from the more specific topic words. For instance, the probability of positive label given the document d_1 is 0.5 and is higher than the probability of negative label given the document (i.e 0). Thus, the document will be classified as positive.

The problem with the TSM model is that it was based on the pLSI framework which is known to suffer from overfitting the training data and thus has a weak inference

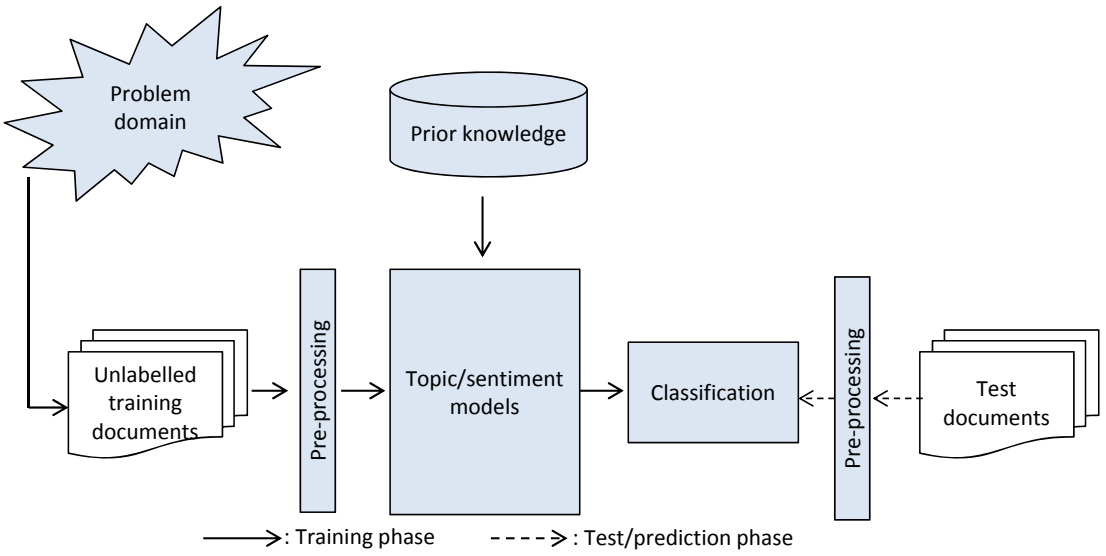

FIGURE 2.2: Unsupervised Machine Learning

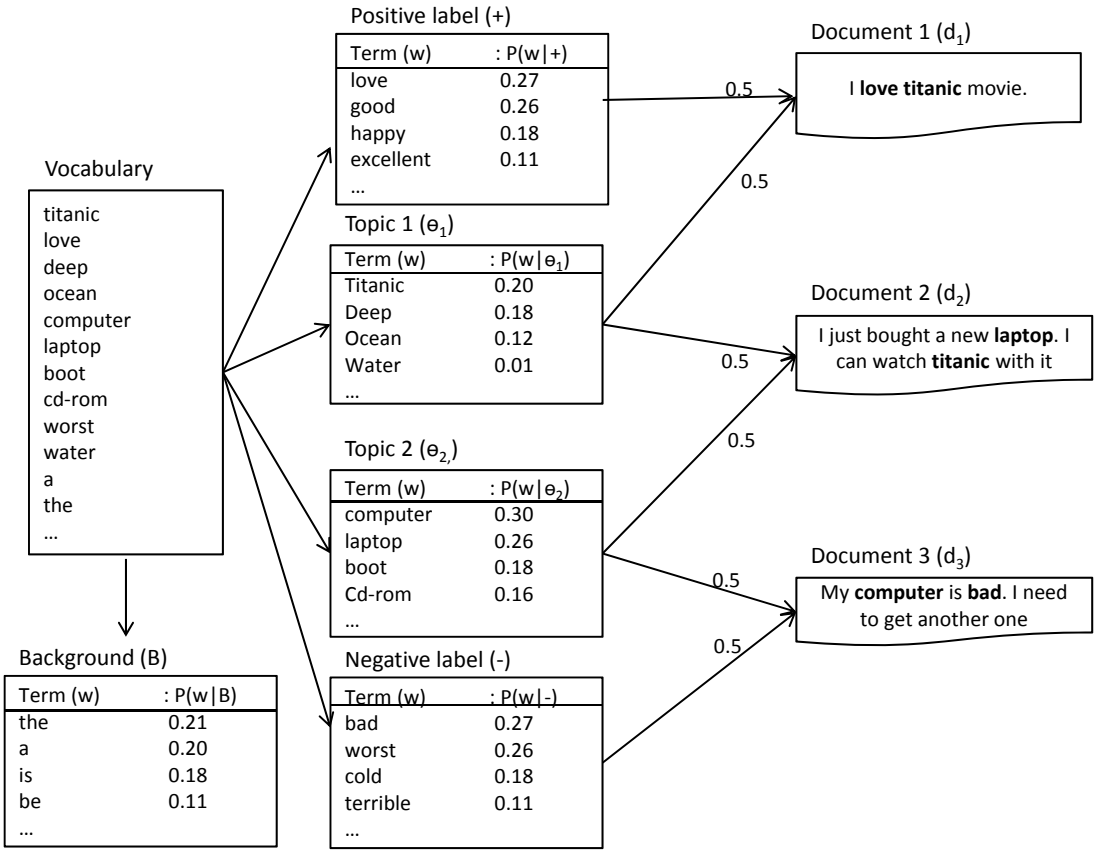

FIGURE 2.3: Topic Sentiment Model (TSM)

capability. To overcome this problem, Joint Sentiment/Topic model (JST) was proposed (Lin and He, 2009). JST is an extension of the topic detection model, the Latent Dirichlet Allocation (LDA) (Blei et al., 2003), with the capability to detect both topic

and sentiment simultaneously. LDA is essentially a generative probabilistic model for topic detection from a collection of documents. It is based on the intuition that when writing a document, the author typically thinks of a number of topics that are relevant to the document with different probabilities of relevance. The author then proceeds to draw terms from these topics in order to compose the document. For instance, the illustration in Figure 2.4 shows the document d_1 to draw terms entirely from the topic θ_1 while d_2 draws from both θ_1 and θ_2 with equal probabilities. Thus, given any document d with observed words w, the relevant topic distribution can be obtained by inferring the probability distribution of the words w over all topics (Steyvers and Griffiths, 2007). The JST extends LDA by adding a sentiment layer between the document and the topic layer. This produces a model in which sentiment labels are associated with documents, under which topics are associated with sentiment labels and words are associated with both sentiment labels and topics as depicted in Figure 2.5. The dynamic nature of social media data whereby sentiments and topics constantly change means that sentiment/topic models also need to be updated over time. This is addressed by the dynamic JST (Li, Huang and Zhu, 2010) which captures both topic and sentiment dynamics by assuming that the current sentiment-topic specific word distributions are generated according to the word distributions in the previous epoch.

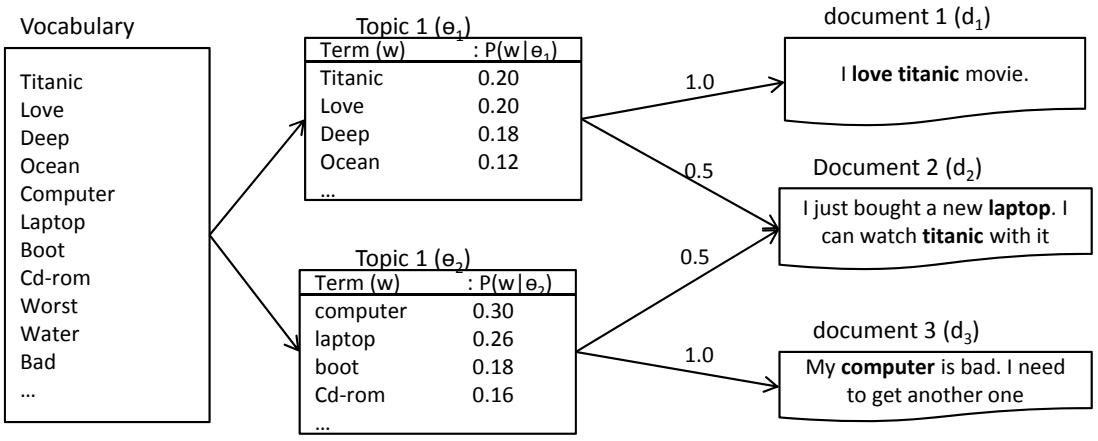

FIGURE 2.4: Latent Dirichlet Allocation Model (LDA)

Both TSM and JST are based on the 'bag of words' assumption that sentiment words are independent in a document. This is observed to be a limitation as sentiment orientation of each word is dependent on its local context (Li, Huang, Zhou and Lee, 2010).

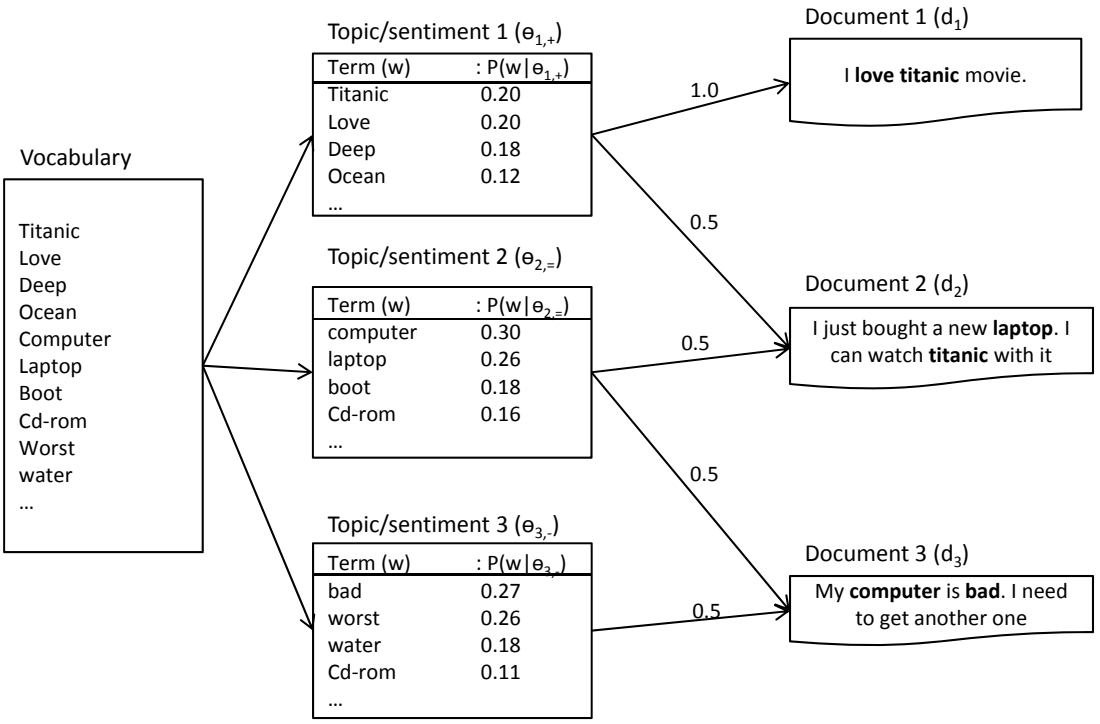

FIGURE 2.5: Joint Topic/Sentiment Model (JST)

Consequently the Dependency Sentiment-LDA model, which relaxes the sentiment independent assumption, was introduced (Li, Huang, Zhou and Lee, 2010). In this model the sentiments of the words in a document are viewed to form a Markov chain, where the sentiment of a word is dependent on the previous one.

Although topic modelling approaches to sentiment classification do not require labelled data, they still rely on sentiment lexicons as the source of prior sentiment knowledge. Like with purely lexicon-based methods, their performance was shown to be dependent on both the coverage and quality of the lexicons used (Lin and He, 2009). However, the lexicon-based methods offer greater flexibility to incorporate linguistically derived contextual knowledge making for a more transparent and accessible approach to sentiment classification.

2.2 Lexicon-based Methods

The lexicon-based methods also called linguistic approaches involve the extraction of terms' prior polarities from lexical resources and aggregation of such polarities based

on linguistic and natural language processing (NLP) rules to obtain sentiment conveyed by a piece of text. Two underlying assumptions underpin the lexicon-based methods to sentiment classification: terms have sentiment connotation independent of context (prior polarity) and such prior polarity can be expressed as a numerical value (Osgood et al., 1957). With these assumptions, general purpose look-up lists which associate terms with their prior polarities can be generated. Such lists are referred to as sentiment lexicons. Lexicon-based sentiment analysis begins with the creation of a sentiment lexicon or the adoption of an existing one, from which sentiment scores of terms are extracted and aggregated to predict sentiment of a given piece of text. Figure 2.6 illustrates the typical lexicon-based sentiment classification. The first step is the creation of a sentiment lexicon from a text collection or a lexical ontology. However because there is already a number of sentiment lexicons in existence, this step is typically the adoption of an existing lexicon. Thereafter, a document to be classified is pre-processed and each term in the document is associated with its prior polarity as given by the sentiment lexicon. Then these prior polarities are adjusted to reflect contextual polarities (contextual analysis, see Section 2.2.3) and aggregated to predict sentiment class.

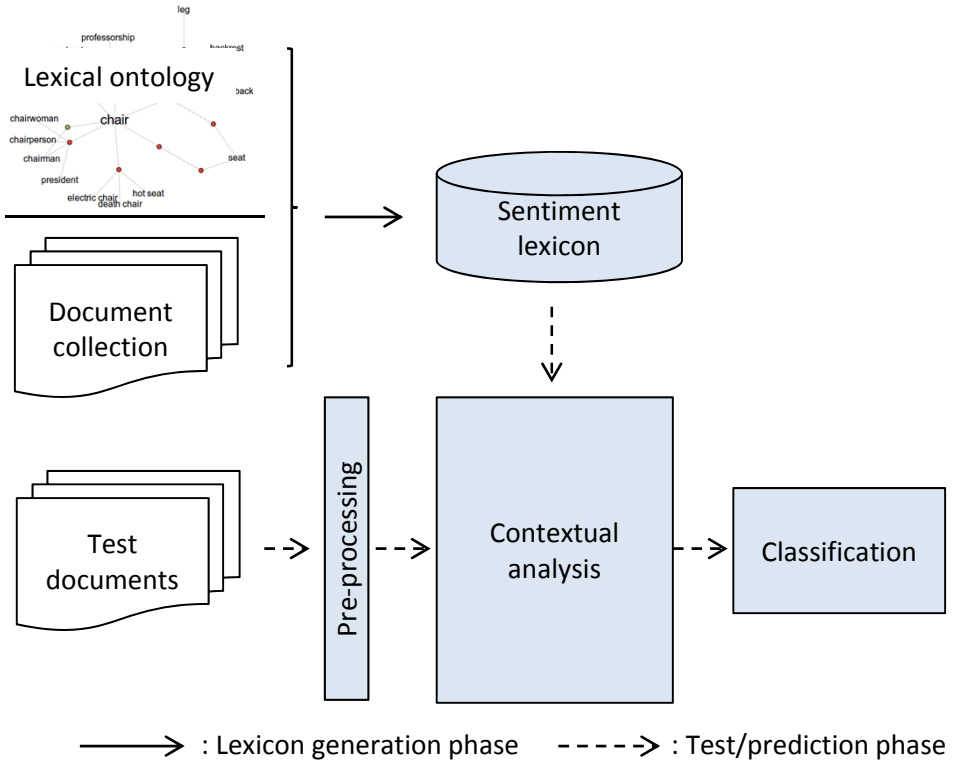

FIGURE 2.6: Lexicon-based Sentiment Classification

Lexicon	Number of Terms	Description
General Inquirer	4216	Manually generated from a corpus. Does not distinguish for polarity strength. It is often used as gold standard. Labelling is at part-of-speech level
Bing Liu's Opinion Lexicon	6789	Generated from General Inquirer. It includes mis-spellings, morphological variants, slang and social media mark-up. It does not distinguish for polarity strength. The labelling is at the word level.
MPQA Subjectivity Lexicon	8221	Contains both manually and automatically labelled terms. It does not distinguish for polarity strength. Labelling is at the part-of-speech level
LIWC	615 (Terms under affective or emotional process)	Automatically generated from a corpus. It does not distinguish for polarity strength. Labelling is at the word stem level
SentiWordNet	28431	Automatically generated from a dictionary. It distinguishes for polarity strength. Labelling is at the word-sense level

TABLE 2.1: Some widely used sentiment lexicons

Sentiment lexicons are either manually or semi-automatically generated from generic knowledge sources. Manually generated lexicons are obviously more accurate, however, they tend to have relatively low term coverage. In contrast, semi-automatically generated lexicons such as the corpus-based one in Mohammad et al. (2013) and the dictionary-based SentiWordNet (Baccianella et al., 2010) have high coverage (over 20,000 words). SentiWordNet is particularly interesting as it offers quantified positive and negative polarities for different senses of terms and at the deeper level of word sense.

2.2.1 Sentiment Lexicons

Sentiment lexicons are language resources that associate terms with sentiment polarity (positive, negative or neutral) usually by means of numerical scores that indicate sentiment dimension and strength. Table 2.1 describes some lexicons widely used for sentiment analysis. Sentiment lexicons can be categorised based on a number of factors. In this review, we organise the lexicons along three dimensions: method of generation (manual or automated); sentiment information (polarity, strength or both) and level of annotation (term, PoS or word sense).

Method of generation. Broadly, sentiment lexicon generation is manual or automated. With manually created lexicons such as General Inquirer (GI) (Stone et al., 1966) and Opinion Lexicon (OL) (Hu and Liu, 2004), sentiment polarity values are assigned by humans. Such lexicons tend to be of limited coverage owing to the cost of the manual effort required to develop them. As for the automatically generated lexicons, there are two semi-supervised approaches commonly adopted: *corpus-based* and *dictionary-based*. Both approaches begin with a small set of seed terms. For example a positive seed set could contain terms such as 'good', 'nice' and 'excellent' while a negative seed set could contain the terms such as 'bad', 'awful' and 'horrible'. They then leverage language resources and exploit relationships between terms to expand the sets. The two methods differ in that corpus-based uses a collection of documents as the language resource while the dictionary-based uses machine-readable dictionaries. Accordingly the relationship they exploit differs. In the corpus-based approach, co-occurrence relations are used to determine sentiment polarities of terms within a text collection using certain rules. This is based on the assumption that terms that have similar sentiment polarity tend to co-occur together. For instance, in Hatzivassiloglou and McKeown (1997), 657 and 679 adjectives were manually annotated as positive and negative seed sets respectively. Thereafter, the seed sets were expanded to conjoining adjectives based on connectives 'and' and 'but' where 'and' indicates that the conjoined adjectives have the same polarity and 'but' indicates a contrast in polarity. Also, a phrasal lexicon was generated from reviews collection (Turney, 2002). Here, two-word phrases were extracted based on some part-of-speech collocations and their polarity is inferred based on the strength of co-occurrence with the seed terms 'excellent' and 'poor'. Point-wise Mutual Information (PMI) is commonly used as a measure of co-occurrence strength:

$$Polarity(phrase) = PMI(phrase, \text{`}excellent\text{'}) - PMI(phrase, \text{`}poor\text{'}) \tag{2.1}$$

Where:

$$PMI(X, Y) = \log_2 \frac{P(x, y)}{P(x)P(y)} \tag{2.2}$$

Where $P(x, y)$ is the probability of x co-occurring together with y, $P(x)$ is probability of x occurring without y and $P(y)$ is the probability of y occurring without x.

Therefore,

$$Polarity(phrase) = \log_2 \frac{P(phrase, \text{`}excellent\text{'})}{P(phrase)P(\text{`}excellent\text{'})} - \log_2 \frac{P(phrase, \text{`}poor\text{'})}{P(phrase)P(\text{`}poor\text{'})} \quad (2.3)$$

These probabilities are typically estimated using Web/Internet search hits (e.g NEAR[1]) as follows:

$$Polarity(phrase) = \log_2 \frac{hits(phrase\, NEAR\, \text{`}excellent\text{'})}{hits(phrase)hits(\text{`}excellent\text{'})} - \log_2 \frac{hits(phrase\, NEAR\, \text{`}poor\text{'})}{hits(phrase)hits(\text{`}poor\text{'})} \quad (2.4)$$

$$= \log_2 \left(\frac{hits(phrase\, NEAR\, \text{`}excellent\text{'})hits(\text{`}poor\text{'})}{hits(phrase\, NEAR\, \text{`}poor\text{'})hits(\text{`}excellent\text{'})} \right) \quad (2.5)$$

With the corpus-based approach, sentiment polarity of domain-specific and non-standard words can be determined provided such words have some association with the (expanded) seed sets. This, however, affects the cross domain portability of such lexicons as some of the associations between words hold only within the domain from which the corpus is drawn.

The dictionary-based approach exploits structural relationships such as synonyms, antonyms and gloss from a dictionary to expand the seed sets (Esuli and Sebastiani, 2005, Hu and Liu, 2004, Kamps et al., 2004, Kim and Hovy, 2004). WordNet (Fellbaum, 1998) is one such dictionary that has been extensively used for generating dictionary-based sentiment lexicons such as WordNet Affect (Strapparava and Valitutti, 2004) and SentiWordNet (Esuli et al., 2010).

Sentiment Information. This refers to the amount of sentiment information the lexicons can offer about terms. Accordingly, lexicons can be broadly grouped into three categories. The first category includes those that can only offer categorical (positive or negative) polarity information about terms. These lexicons do not distinguish for polarity strength between terms within the same category. Thus, the terms *'good'* and

[1] Turney (2002) used Altavista with the NEAR operator

'excellent' are equal in sentiment from these lexicons as they belong to the same sentiment category (positive). Example of such lexicons include GI, OL and MPQA Subjectivity Lexicon (Wiebe and Cardie, 2005). Lexicons in the second category offer polarity strength of terms usually on a Likert scale (e.g. 1 to 5) in either the positive or negative dimension. A limitation with these lexicons is that when a term does not attain the maximum score it becomes ambiguous as to whether the remaining score indicates the term's leaning towards the opposite polarity or objectivity. Examples of these include the lexicon used in SentiStrength and SOCAL (Taboada et al., 2011, Thelwall et al., 2012). The final category of lexicons offer sentiment information about terms in both dimensions. These lexicons are rich in sentiment information, however, finding an optimal approach to utilising the information remains a challenge. An example of such lexicons is SentiWordNet.

Level of annotation. This refers to the linguistic properties that influence scores assignment in lexicons. For instance, a term can have multiple parts-of-speech and each part-of-speech can have multiple word senses and not necessarily connote the same polarity score. Accordingly, sentiment lexicons can be viewed at *term-level*, *PoS-level* and *word-sense-level*. Term-level annotation lexicons are the basic ones in which polarity is associated with terms. This is insufficient as polarities can change depending on part-of-speech or word sense. In the PoS-level lexicons the annotation is determined at the PoS-level while in word-sense-level, it is determined at the word sense level of terms. SentiWordNet is one of such lexicons where polarity annotation is at the word sense level. Here, a word such as 'like', for example, has three associated PoSs (adjective, noun and verb) and a total of eleven word senses each having its associated sentiment score. For instance, 'like' as an adjective at word sense number 1 (like#adjective#1) meaning "resembling or similar" has sentiment score: positive=0, negative=0.25 while like#verb#2 meaning "to find enjoyable or agreeable" has sentiment score: positive=1, negative=0.

2.2.2 Emotion Lexicons

Like with sentiment analysis, the field of emotion detection, concerned with the extraction of emotion-bearing text, has several lexicons developed for the task. Emotion is a concept that is closely related to sentiment. Scherer (2000) defines emotion as a "relatively brief episode of response to the evaluation of an external or internal event as being of major significance". Unlike sentiment, emotion is more fine-grained and can be classified into a larger number of different classes proposed in various emotion theories (Jurafsky and Martin, 2015). In a category of the theories, emotions are viewed as fixed atomic units, limited in number, and from which other basic emotions are generated (Plutchik, 1962, Tomkins, 1962). Perhaps the most popular emotion classes in this category are those proposed by Ekman (1999): *surprise*, *joy*, *anger*, *fear*, *disgust* and *sadness*. These emotion classes were derived from facial expressions and are likely to be present in all cultures (Ekman, 1999). A more elaborate set of emotion structure classes is Parrott (2001)'s grouped into high-level, emotion classes of *love*, *joy*, *surprise*, *sadness*, *anger* and *fear*.

While semi-supervised approaches are commonly employed in generating sentiment lexicons, the most common approach to build emotion lexicons is to have humans label the words through crowdsourcing: breaking the task into small pieces and distributing them to a large number of annotators (Jurafsky and Martin, 2015). Emotion lexicons generated in this manner include *Emolex* (Mohammad and Turney, 2013), a moderate-sized emotion lexicon of about 14000 words crowdsourced using the online service: Amazon Mechanical Turk[2]. Another manually generated emotion lexicon is the *WordNet-Affect* (Strapparava and Valitutti, 2004) which was derived from WordNet's synsets. Both Emolex and WordNet-Affect associate terms with the positive and negative sentiment classes in addition to emotion classes making them play the role of sentiment lexicons. In a different crowdsourcing approach, an emotion lexicon was developed from a corpus where each word was assigned a distribution over emotion classes using a maximum likelihood model (Rao et al., 2014). Recently, in contrast with manual crowdsourcing, emotion lexicons have been generated automatically from social media corpora. For

[2]https://www.mturk.com/mturk/welcome

instance, hashtags on Twitter such #sad, #joy, #surprised have been used to accordingly label tweets as sad, joy or surprise. Thereafter, emotion lexicons were developed from the labelled tweets using the PMI approach (Mohammad, 2012, Mohammad and Kiritchenko, 2015). Also, another lexicon which provides a superior classification performance over the PMI-based lexicon was developed using an expectation maximisation approach (Bandhakavi et al., 2014).

Sometimes, in addition to emotion labels, emotion lexicons also provide sentiment labels to terms. For instance, both Emolex and WordNet-Affect have positive and negative labels associated to terms making them readily useful for sentiment analysis. Even when emotion lexicons do not provide sentiment labels for terms, they are still useful for sentiment analysis since most of the emotion classes can be mapped onto positive and negative sentiment classes (Ghazi et al., 2010, Gonçalves et al., 2013, Poria et al., 2014). However, emotion knowledge may not be completely mapped to sentiment knowledge through emotion-to-sentiment lexicon mapping. For instance, the emotion class, surprise, is ambiguous as it could correspond to positive or negative sentiment (Alm, 2008) or even neutral (Ortony et al., 1990). For example, 'surprise' class was considered positive in Poria et al. (2014), it was considered negative in Ghazi et al. (2010). Also, although emotion and sentiment are inter-related, they are known to be theoretically different (Munezero et al., 2014). Therefore, emotion knowledge should not be reduced to sentiment knowledge but when used carefully may help with sentiment analysis.

.

2.2.3 Contextual Analysis

Early work in lexicon-based sentiment analysis involved the aggregation of individual polarities of terms irrespective of grammatical dependencies that may exist between them. This approach is incomplete and often gives the wrong results when implemented directly because a term's prior polarity changes due to the effect of other terms with which the term co-occurs. For example, the text *"I don't like the idea of smoking in general"* can be classified as positive because it is dominated by positive terms ('like' and 'idea'). However, the problem is that the appearance of the negation ('don't') renders

the text negative. This can be addressed by contextual analysis using valence shifters (Polanyi and Zaenen, 2004). Here, polarities of sentiment-bearing terms that are under the influence of negation (e.g. *'not'*, *'never'*,*'nothing'*) are inverted and those under the influence of terms that increase (i.e. intensifiers e.g. *'very'*, *'highly'*) and that decrease polarity strength (i.e. diminishers e.g. *'slightly'* and *'a-little-bit'*) are increased and decreased respectively. Negation analysis is a particular challenge as the polarity of negated terms do not always translate to its opposite. For instance, whereas "It is *not good*" is more or less the same as "It is *bad*", "It is *not excellent*" is more positive than "It is *horrible*". Consequently, shift approach was proposed as a preferred alternative to sentiment inversion for negation (Taboada et al., 2011). Here, prior polarity scores of sentiment terms that are under the influence of negation are reduced by a certain weight. A recent study suggests that negation terms are not just modifiers of sentiment but also indicators of sentiment (Potts, 2011*a*). For instance, it was found that the distribution of negation across reviews is as skewed towards negatively rated reviews as the word *'bad'* is. In SentiWordNet, negation terms are associated with polarity scores. Thus a strategy can be introduced to treat negation both as sentiment-bearing and as sentiment modifier for other terms

Another contextual analysis with a potential to influence sentiment analysis is based on discourse analysis. The main idea here is that different discourse segments have different level of importance thus sentiment scores for terms should reflect such importance. Discourse segments are often signalled by connectives such as 'but', 'and' and 'although' accordingly these were used to apply weights to various segments of Twitter posts (Mukherjee and Bhattacharyya, 2012). The results show improvement in sentiment classification. However, the work in (Mukherjee and Bhattacharyya, 2012) employed only a small number of discourse connectives.

2.2.4 Domain-Specific Vocabulary and Polarities

Sentiment lexicons are typically generated independent of any target application. Thus, they usually reflect general knowledge making them useful in diverse applications (i.e.

general purpose). However, the lexicons utility is reduced when a target application domain or genre deviates from the general sentiment knowledge.

The concept: Domain. Unfortunately, the concept 'domain' does not seem to have unambiguous definitions from the linguistics and sociolinguistics points of view. From a purely linguistic perspective, a domain has been defined as a *genre* attribute that describes the broad subject field that an instantiation of a certain genre deals with (Lee, 2001). A genre is defined as a category assigned to a text based on external, non-linguistic criteria such as intended audiences, purposes and activity type (Lee, 2001) as well as textual structure, form of argumentation and level of formality (Crystal, 2011). Based on this definition, for example, a text from the genre NEWSPAPER ARTICLE may belong to the domain of SCIENCE. Other domains may include ART, FINANCE, RELIGION, POLITICS, SPORTS and TECHNOLOGY. However, in sociolinguistics, a domain is viewed as a social setting that is likely to influence the use of language such as FAMILY, FRIENDSHIP, RELIGION, EDUCATION and EMPLOYMENT (Fishman, 1972). It can be observed that some categories in this latter definition may correspond to what can be called genre in the former definition (e.g. FRIENDSHIP and EMPLOYMENT). In fact, for socio-psychological analysis, social contexts such as INTIMATE, INFORMAL, FORMAL, and INTERGROUP are identified as domains (Fishman, 1972). Both notions of a domain have been used in sentiment analysis. For instance, it is common to refer to collections of documents grouped per subjects of discussion as domains (e.g. HOTELS, SPORTS, BOOKS and ELECTRONICS) (Du et al., 2010, Yoshida et al., 2011). It is also common to refer to the social setting in which documents are generated as domains (e.g TWITTER as a domain) (Kaur and Kumar, 2015, Kiritchenko and Mohammad, 2016, Reitan et al., 2015). In this thesis, we use the concept of a domain in a broader sense that encompasses both definitions. Specifically, we use the concepts to refer to any collection of documents that share certain characteristics that may influence the expression of sentiment. For instance, TWITTER with its informal setting, brief nature of communication and the general public as target audience forms a domain; so also MYSPACE with its severe informal communication between friends.

Differing Vocabulary and Polarities. The deviation between a lexicon and a target genre can be in terms of vocabulary coverage whereby the lexicon supplies insufficient sentiment-bearing terms for a target genre. This is particularly the case with social media genres where non-standard vocabulary is widely used to express sentiment. A potential remedy to the coverage problem is to generate a domain-specific lexicon. However, existing lexicon generation methods tend to result in lexicons with poor coverage for social media. For instance, the method in Hatzivassiloglou and McKeown (1997) has produced a lexicon based on the proximity of terms with adjectives and constrained by the occurrence of certain conjunctions. This is too restrictive for the informal social media content. A subsequent work has improved term coverage by relaxing the conjunction constraints and the use of a relatively larger corpus (the web) to measure terms co-occurrence (within a text window) with known seed terms (Turney, 2002). Nevertheless, coverage is still affected by the fact that the co-occurrence has to be with infrequent seed terms. Yet, to improve coverage, the concept of double propagation was introduced (Guang. et al., 2009). Here, co-occurrence with a product/service aspects was used to identify sentiment-bearing terms and vice-versa. This runs iteratively until no further sentiment-bearing term or aspect can be found. This method was meant for the domains of products/services reviews where aspects mentions in sentiment expression is common. Other methods employed supervised strategies whereby a lexicon is generated from sentiment-labelled data (Mohammad et al., 2013, Pang et al., 2002). The need for labelled data limits the utility of the supervised strategies. The use of a domain-specific lexicon alone for sentiment analysis is also problematic because although test instances are expected to be of similar composition to that of domain text, it is possible for a test instance to contain terms that never appear in the domain but which may be available from a general-purpose lexicon.

The deviation of a target domain from a general-purpose lexicon can also be in terms of sentiment polarities of terms. Sentiment-bearing property of terms is known to be domain-dependent such that the same term can have different sentiment semantics in different domains. For example, the adjective 'unpredictable' may indicate negative sentiment in a car review, as in "unpredictable steering" but a positive sentiment in a movie review, as in "unpredictable plot" (Liu, 2012). Indeed, a comparison of sentiment

analysis systems across different domains reveals that factors such as datasets size and domain/genre can significantly affect performance (Andreevskaia and Bergler, 2008).

The difference in polarities between a sentiment lexicon and a target domain has been addressed with techniques that produce domain-adapted lexicons. Choi and Cardie (2009) investigated the adaptation of a general-purpose lexicon to a domain specific one. Their approach adapts term polarities of a general-purpose lexicon by utilizing expression-level polarities from the domain. The polarity relationship between the terms and the expressions were modelled as a set of constraints that are solved using integer linear programming. This work relied on sentiment-labelled data to obtain the expression-level polarities. It was also limited to term polarity reversal (from one sentiment class to another) but unable to adjust polarity intensity within the same class. In a similar work, a domain-specific lexicon was adapted to another domain using the information bottleneck framework (Du et al., 2010). Here, the algorithm also assumes as input a set of in-domain sentiment-labelled documents. In another work, an approach was proposed to identify the most effective lexicon, from among several lexicons, for sentiment analysis in a target domain (Ohana et al., 2012). This approach employs the case-based reasoning methodology and extracts documents statistics and writing styles as features on which to represent the documents (cases). The solutions to a case are the lexicons that provide correct classification of the case document as checked against human judgment. Thus, given a domain containing new cases (documents), sentiment classification is performed by reusing lexicons from the most similar documents to those in the given domain. It can be noted that this approach does not attempt to adapt a lexicon to a target domain.

With social media domains, the idea of distant supervision can be leveraged to generate domain-specific lexicons that can capture evolving vocabulary. For instance, two Twitter-specific sentiment lexicons have been generated from tweets that are labelled based on the occurrence of certain emoticons and hashtags respectively (Kiritchenko et al., 2014, Mohammad et al., 2013), using the point-wise mutual information (PMI) approach (Turney, 2002). These lexicons are highly domain-specific and could miss general sentiment-bearing terms that may not be available in the tweets' vocabulary, a

limitation which can be addressed by a lexicon expansion strategy.

A lexicon expansion strategy begins with a standard lexicon whose polarities are propagated to domain-specific terms. This is similar to lexicon generation strategies except that a lexicon generation strategy begins with a very small set of seed terms known to have a high and stable sentiment connotation across domains. In Zhou et al. (2014), a standard lexicon has been expanded with terms from an emoticon-labelled Twitter dataset. Here, similar to Mohammad et al. (2013) and Kiritchenko et al. (2014), a Twitter-specific lexicon was generated using the PMI approach (Turney, 2002), however, unlike in Mohammad et al. (2013) and Kiritchenko et al. (2014), a negated co-occurrence of a term with a sentiment class was counted as co-occurrence of the term with the opposite sentiment class. For example, *"I don't like their online service :("* would be counted as a co-occurrence of 'like' and ':)'. In another lexicon expansion strategy, emoticon-labelled datasets were used to identify a suitable feature set on which to represent a set of seed terms, formed from a union of several general-purpose lexicons, for a supervised sentiment classification of unknown terms (Bravo-Marquez et al., 2015). The datasets were time-sorted and time-series were created for each term from the datasets' vocabulary. Then, the feature set was extracted from the location-based and dispersion properties of the time-series. A classifier learned from the representation was then used to classify every unknown term from the vocabulary as positive, negative or neutral.

Although a lexicon expansion strategy such as in Zhou et al. (2014) and Bravo-Marquez et al. (2015) is able to capture domain-specific terms, it is unable to adapt polarities of existing terms from the initial lexicon to domain-specific semantics. With distant supervision, a domain-specific lexicon can be generated for social media domains, and combining such a lexicon with a general-purpose lexicon will ensure domain adaptation as well as the acquisition of additional vocabulary available from the general-purpose lexicon (Muhammad et al., 2014, 2013*b*).

2.3 Hybrid

Increasingly term polarities from lexicons are used as additional features to train machine learning classifiers in a hybrid approach (Al-Mannai et al., 2014, Dang et al., 2010, Ikeda

et al., 2008, Mohammad et al., 2013, Ohana and Tierney, 2009). Sentiment classification was also observed to improve when multiple classifiers, formed from machine learning and lexicon-based methods, are used to classify a document (Prabowo and Thelwall, 2009).

The hybrid method also helps overcome certain limitations of the combined methods. For instance, in a system called *PSenti* lexicon knowledge was used to filter out non sentiment-bearing words from the feature set subsequently used for machine learning (Mudinas et al., 2012). Evaluation of *PSenti* shows the hybrid approach achieved better performance compared to pure lexicon-based, and better cross-domain portability compared to pure machine learning. In another work, a small amount of training data for machine learning was compensated with lexicon knowledge (Melville et al., 2009). This approach builds two generative models: one from a labelled corpus and a second from a sentiment lexicon. The distributions from the two models were then adaptively pooled to create a composite multinomial Naive Bayes classifier. The pooling approach employs a linear combination of conditional probabilities from the different generative models. The combined approach was compared to using Naive Bayes classifier built using only the labelled corpus with significant improvement in classification accuracy. In some other work, machine learning was applied to optimise sentiment scores prior to lexicon-based sentiment classification (Thelwall et al., 2012). This approach has the tendency to produce domain adapted lexicons which in turn improve sentiment classification.

It is noteworthy, however, that although the hybrid approach can help overcome certain limitations of either of the combined methods (lexicon-based or machine learning) alone, it can also combine challenges from both methods. For instance, it often requires both labelled data, which can be difficult to obtain, as well as a sentiment lexicon.

2.4 Sentiment Analysis using SentiWordNet

SentiWordNet is a general purpose sentiment lexicon generated from the WordNet (Fellbaum, 1998) dictionary. Each synset from WordNet (i.e. a group of synonymous terms

based on a particular usage or meaning) is associated with two numerical scores indicating the degree of association of the synset with the positive and negative sentiment polarities. A third score for objectivity or neutrality can be derived by subtracting the sum of positive and negative scores from 1.

Recently, SentiWordNet has become a popular resource for sentiment analysis given its high coverage of English terms and fine-grained sentiment information. It is being used with both pure lexicon-based and hybrid approaches. The baseline lexicon-based sentiment classification using SentiWordNet sums up respective positive and negative scores for all terms contained in the given test document. The dimension with the highest total score becomes the sentiment class for the document (Agrawal and Siddiqui, 2009, Denecke, 2008, Devitt and Ahmad, 2007, Hamouda and Rohaim, 2011, Heerschop et al., 2011, Muhammad et al., 2013*a*, Ohana and Tierney, 2009). Several approaches have been introduced to improve upon this baseline. For instance, polarity adjustments due to negation and intensification are introduced for sentiment classification of movie reviews (Agrawal and Siddiqui, 2009, Thet et al., 2009). However, negation and intensification terms are already associated with sentiment scores in SentiWordNet and are accounted for, to some extent, in the baseline approach. Therefore, further analysis is needed to ascertain the role of such terms and a strategy to appropriately account for them.

In another work, SentiWordNet was used for Sentiment polarity identification in financial news using a cohesion-based text representation algorithm (Devitt and Ahmad, 2007). Here, scores from the lexicon are overlaid onto the WordNet structure. Subsequently, a document to be classified for sentiment is represented as a graph within WordNet via term adjacency relation. The graph also expands to other nodes reachable via WordNet relation types (derived-from, see-also, hypernymy). Therefore, an aggregate score for a document can include not only the scores of terms within the document but also scores from other related terms. It can be noted, however, that from the manner in which SentiWordNet is generated, WordNet relations are already taken into account and the work in Devitt and Ahmad (2007) are likely to introduce some redundancy. Also, the approach does not allow for integrating sentiment modification from local context (e.g. negation, intensification).

SentiWordNet was also applied for multilingual sentiment classification (Denecke, 2008). Here, documents written in languages other than English are first translated to English and then classified for sentiment. This work does not attempt to extend the baseline approach as its focus was on addressing the problem of dealing with multilingual documents.

In a more recent work, the baseline approach is extended with score modification based on discourse structure in a system called *Pathos* (Heerschop et al., 2011). In movie reviews, it was observed that the conclusion which appears towards the end of the review tends to be more important than the introduction in the beginning. Thus, in *Pathos* it was shown that even simple variation of term weights from the beginning to the end of a review (in ascending order of importance) can improve a positive/negative sentiment classification. Further improvements were reported when a more sophisticated weighting approach based on the Rhetorical Structure Theory (RST) (Mann and Thompson, 1998) was introduced. However, such an approach, based on the use of standard RST parsers tends to be too brittle for short and highly informal social media content.

In the hybrid sentiment classification context, SentiWordNet is typically used to derive feature sets for supervised classification. Improvements in classification accuracy were shown using just the features derived from the lexicon (Ohana and Tierney, 2009) and in combination with other feature sets (Dehkharghani et al., 2012). In another work, the lexicon was used to filter out non sentiment bearing terms from n-gram feature set (Mudinas et al., 2012). Also, scores from the lexicon have been used as feature values in combination with frequency-based values (Sani et al., 2013).

Although sentiment scores are associated with word senses in SentiWordNet, this information is under-utilised in the existing literature as the afore-mentioned work adopt a strategy that avoids word sense disambiguation. Such strategies include using the first sense from WordNet because it is the most naturally occurring sense or some form of averaging over all senses. Similarly, SentiWordNet has so far largely been used for sentiment analysis of reviews. Given the high coverage of the lexicon, it will be interesting to explore its applicability in sentiment analysis of non-review social media.

2.5 Challenges of Sentiment Analysis

Sentiment analysis is a very challenging problem as it is highly domain and context dependent. A piece of text that is positive in one domain can be negative in another. For instance *"go read the book!"* is positive in the domain of books but negative in the domain of movies (Paltoglou, 2014). Sometimes only the context can help uncover sentiment expressed without the use of sentiment-bearing words. For example the comment *"After sleeping on the mattress for two days, a valley has formed in the middle"* can be understood to be very negative even though it does not seem to contain any obvious sentiment-bearing word (Liu, 2010).

Another challenge for sentiment analysis is the thwarted expectation phenomenon. It is a scenario in which the author sets up a contrasting introduction to the intended sentiment. This is particularly observed to be common in movie reviews, for example in *"This film should be brilliant. It sounds like a good plot, the actors first grade', and the supporting cast is good as well, and Stallone is attempting to deliver a good performance. However, it can't hold up"* (Pang et al., 2002). The overall sentiment of this review is negative despite many positive expressions at the beginning of the text. This challenge can potentially be addressed by the use of discourse structures for term weighting (Heerschop et al., 2011).

Social media text (e.g. tweets and discussion posts) presents peculiar challenges for sentiment analysis. Text from these platforms is typically short thus presenting high ambiguities (Maynard and Hare, 2015). It is also characterised by dynamic and diverse vocabulary use, although standard off-the-shelf lexicons remain static with fixed vocabulary and associated polarities. Similarly, in social media platforms users often employ non-standard spelling/grammar and sarcasm to express sentiment. In a recent work, lexicon-based sentiment analysis was extended to incorporate modification of term prior polarities based on non-lexical modifiers (Paltoglou and Thelwall, 2012, Thelwall et al., 2012, 2010). Such non-lexical modifiers include term elongation by repeating letter (e.g. 'haaappppyy' instead of 'happy'), capitalization of terms, and Internet acronyms. Both

repeated letter and capitalisation are identified and treated as intensification whilst Internet acronyms are expanded to their full meanings (e.g. *rotf* becomes *rolling on the floor*) or manually added into a lexicon. However, it can be observed that when Internet acronyms are expanded they could lose their sentiment connotations as in the previous example. Also, the approach needs to be extended to the phenomenon of repeating other characters not just letters (e.g. in *happy!!!*) and the use of emoticons in social media.

Sarcasm is a phenomenon that can have significant impact on sentiment expressions. It typically means to say the opposite of the true feelings in order to be funny or make a point[3]. From this definition, it can be observed that sarcasm is particularly a device for expressing sentiment and it is difficult to handle, as the literals affected by the use of the sarcasm have to be *detected* and be treated as their opposites. Existing research work on sarcasm detection typically focuses on detecting the presence or absence of sarcasm but not how to handle it for effective sentiment analysis (Maynard and Greenwood, 2014). Also, sarcasm has been treated mostly from a machine learning perspective in which text containing sarcasm is used to train algorithms for sarcasm detection. However, recently rules for sarcasm detection have been integrated into lexicon-based classification (Maynard and Hare, 2015). These rules were based on a strategy that detects sarcasm using hashtags as cues (e.g. #sarcasm, #lying and #notreally), identifies the scope of the sarcasm, which may not be the whole document, and applies a score reversal approach, similar to the effect of negation, to the scope of the sarcasm. However, more recent findings suggest that the #sarcasm hashtag is not a natural indicator of sarcasm expressed between friends, but rather serves an important communicative function of signalling the author's intent to an audience who may not otherwise be able to draw the correct inference about their message (as distinct from close friends who may be able to infer sarcasm without such labels) (Bamman and Smith, 2015). Therefore, relying on hashtags or similar explicit cues for sarcasm detection may have very limited utility particularly in the non-broadcast, highly contextualized social media communications such as between friends (e.g on MySpace or Facebook) or between members of a discussion forum.

[3]http://www.bbc.co.uk/worldservice/learningenglish/radio/specials/1210_how_to_converse/page13.shtml. As cited by Maynard and Greenwood (2014)

2.6 Conclusion from the Literature

The domains of non-review social media are characterised by the lack of training data, making lexicon-based approaches readily suitable for sentiment classification. These approaches also offer the advantage of better classification transparency, classifier flexibility and explanation of results. Previous lexicon-based methods concentrate on the use of low-coverage (often manually generated) lexicons that typically only provide coarse sentiment information for terms. SentiWordNet can potentially improve sentiment classification in social media given its high coverage of terms and the level of disambiguated polarity information it provides. However, effective integration of contextual analysis while also utilising the detailed polarity information offered by the lexicon remains under-explored. In this thesis, we address this research problem in relation to sentiment classification for social media text (Chapter 4).

The dynamic nature of social media characterised with an evolving vocabulary and sentiment semantics for terms means that sentiment lexicons need to be updated to reflect such changes. Similarly, as sentiment analysis is domain-dependent, a system for analysing social media needs to capture some characteristics of this genre while also maintaining some level of cross-domain portability to account for the diverse nature of social media. We address this research problem with a strategy to generate a hybrid lexicon which combines a domain-specific and general purpose lexicons thereby capturing the properties of both lexicons (see Chapter 5). Our approach to domain-specific lexicon generation does not rely on human annotation but rather sentiment signals (emoticons) within data. Thus, it is better able to capture the dynamic nature of social media.

2.7 Chapter Summary

In this chapter, we presented a review of the literature related to our work. We discussed the three broad methods for sentiment classification: machine learning, lexicon-based and hybrid. The discussion focused on the research progress made using each of the methods, their strengths and weaknesses. A more detailed discussion of lexicon-based methods was presented as these are closely related and motivate the work presented in

this thesis. In particular we explained the importance of contextual analysis and the need for adaptability when applying a static lexicon such as SentiWordNet to social media content.

Chapter 3

Background

In this chapter, we present background details about the main sentiment lexicon and the baseline sentiment classification algorithms used in the research. Similarly, we provide details about our evaluation datasets, text pre-processing operations and performance metrics employed.

3.1 SentiWordNet

Two versions of SentiWordNet exist publicly for research. Figure 3.1 illustrates the process of generating the first version of the lexicon, SentiWordNet 1.0 (Esuli and Sebastiani, 2006). Initial seed synsets (positive and negative) are expanded by exploiting the synonymy and antonymy relations in WordNet. The polarity of a seed synset is propagated to the synsets that are reachable through the synonymy relation while the opposite polarity is propagated in the case of antonymy. As there is no direct synonym relation between synsets in WordNet as synonymous terms are already grouped in synset, the relations: See-also, Similar-to, Pertains-to, Derived-from and Attribute are used to represent the synonymy relation. Thereafter, textual definitions of the expanded synsets (glosses) along with that of objective seed synsets are used as training data for eight diverse classifiers of positive, negative and objective classes. These classifiers assigned sentiment class to every synset and the proportion of classification for each class are deemed to be the sentiment scores for the synset.

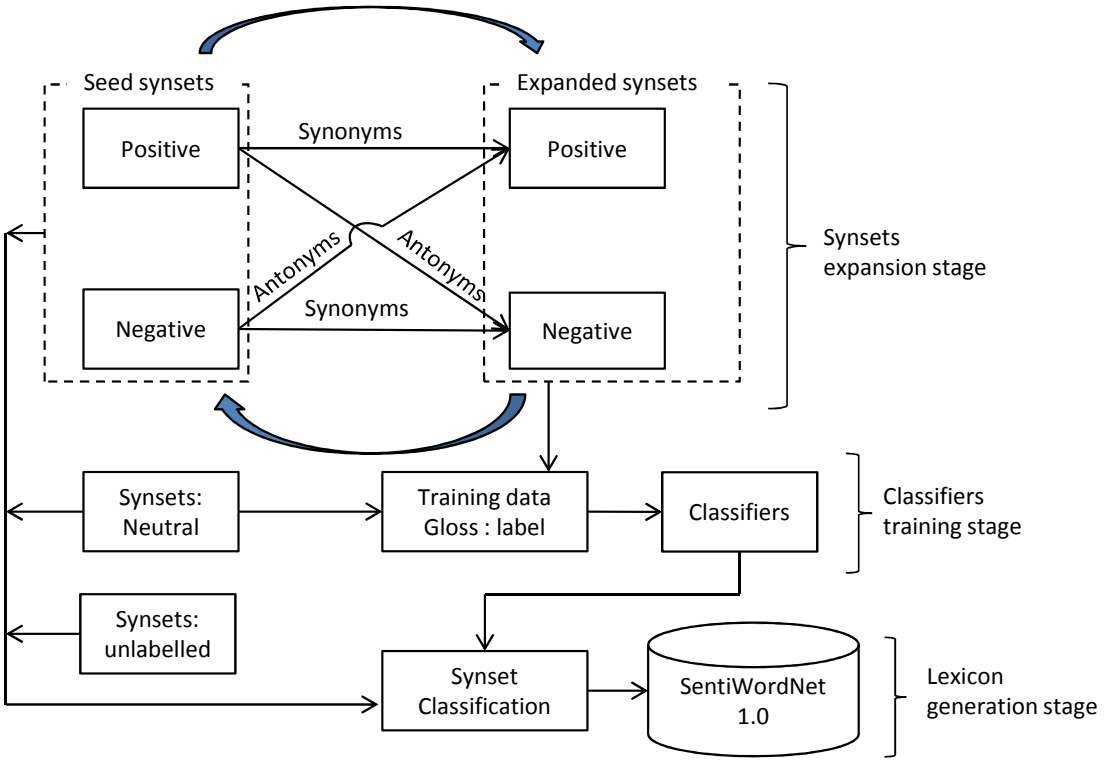

FIGURE 3.1: SentiWordNet 1.0

In the second and enhanced version of the lexicon, SentiWordNet 3.0 (Baccianella et al., 2010), sentiment scores are optimised via a random-walk using the PageRank approach (Brin and Page, 1998). This optimisation involves leveraging WordNet's graph structure, where a link is formed from one synset, S_1, to another, S_2, if a term from S_1 occurs in the gloss of S_2. This graph is illustrated in Figure 3.2 where the synsets {proper#1} and {satisfactory#1} connect to the target synset, {well#1,good#1}, because of the occurrence of the words 'proper' and 'satisfactory' in the gloss of the target synset. These form the *Backward neighbours* of the target synset. Likewise, the target synset connects to {exceptionally#1} and {fortunately#1,luckily#1} (its *Forward neighbours*). Starting with scores from SentiWordNet 1.0, the random walk iteratively adjusts scores using the relation in Equation 3.1 until convergence.

$$a_i^{(k)} \leftarrow \alpha \sum_{j \in B(i)} \frac{a_j^{(k-1)}}{|F(j)|} + (1-\alpha)e_i \tag{3.1}$$

where $a_i^{(k)}$ denotes the value of a target synset, a_i, at the k_{th} iteration, $F(i)$ is the set of forward neighbours of a_i, $B(i)$ is the set of backward neighbours of a_i, e_i is a constant

such that $\sum_i e_i = 1$ and $0 < \alpha < 1$ is a control parameter.

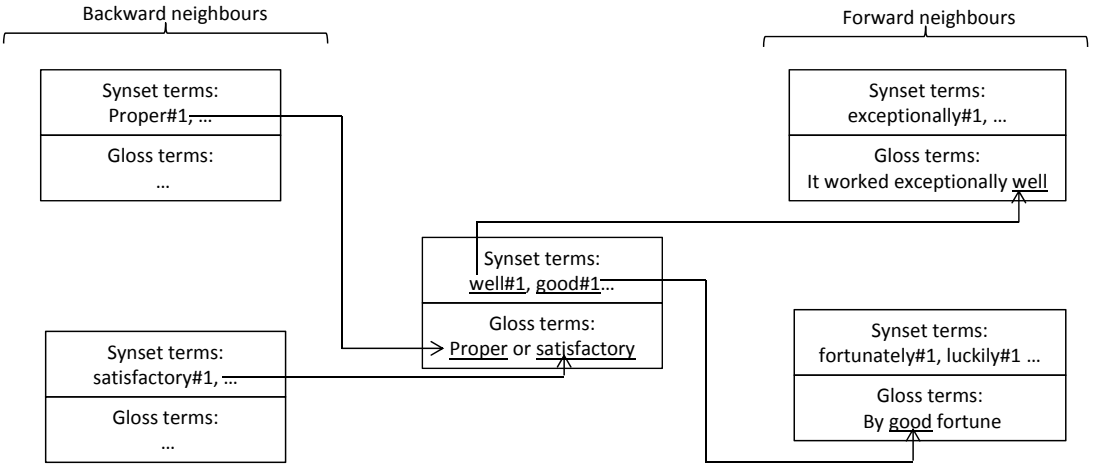

FIGURE 3.2: Graph Structure in WordNet

Figure 3.3 shows a fragment from SentiWordNet. Scores for a specific term within a synset can be extracted by specifying the synset's identification number (ID) or the term's lemma, part-of-speech (PoS) and sense number. For instance, the positive (+score) and negative (-score) for the first sense of the adjective 'scarce' can be extracted by specifying the ID: 00016756 or the three parameters: 'scarce' as the lemma, 'a' as the PoS and '1' as the sense number. Next we look at the baseline algorithms for sentiment classification using SentiWordNet and other lexicons.

PoS	ID	+score	-score	synset	gloss
a	00016756	0	0.25	scarce#1	deficient in quantity or number compared with the demand; ...
n	00735936	0	0.625	misdeed#1 misbehaviour#1 misbehavior#1	improper or wicked or immoral behavior
r	00309249	0.125	0.125	despicably#1	in a despicable manner; "he acted despicably"
a	00017782	0.625	0	acceptable#1	worthy of acceptance or satisfactory; "acceptable levels of radiation" ...
n	04632063	0.75	0	chirpiness#1	cheerful and lively
v	02746140	0.625	0	beat#12	be superior; "Reading beats watching television"; "This sure beats work!"

FIGURE 3.3: A fragment from SentiWordNet 3.0

3.2 Lexicon-based Methods: The baseline Algorithms

Depending on the amount of sentiment information a lexicon can provide as discussed in Chapter 2, three baseline score aggregation strategies for sentiment classification can be derived from existing research: term counting, maximum score and aggregate-and-average.

The term counting simply counts the number of terms belonging to each sentiment class from the given document. Thereafter, the document is classified as the class with the majority count. Any ties are broken in favour of the class having the natural tendency to occur more often (typically, the positive class). It can be observed that this approach disregards term polarity intensities which are vital for sentiment expression. It is therefore not surprising that the term counting approach often gives poor results (Hamouda and Rohaim, 2011, Ohana and Tierney, 2009). With the maximum score approach, a given document is assigned to the sentiment class of its strongest sentiment-bearing term (Thelwall et al., 2012, 2010). Lastly, with the aggregate-and-average approach, sentiment class for a given document is determined by the average sentiment intensity of all its terms.

This thesis adopts the aggregate-and-average approach as the baseline sentiment classification algorithm. Using SentiWordNet, the approach is outlined in Algorithm 1. Positive (t^+) and negative (t^-) scores for each term are extracted from the lexicon. Thereafter, these scores are respectively summed for all terms contained in Doc (steps 4-7 in Algorithm 1). The sentiment class of the input text is deemed positive if the net positive score (Doc^+) exceeds the net negative score (Doc^-) and, negative, otherwise (steps 8-12).

3.3 Machine Learning Methods: The baseline Algorithms

Three benchmark classifiers are particularly used namely: Naïve Bayes, Support Vector Machines and Logistic Regression. Here we present a brief background about these classifiers.

Algorithm 1 Base

INPUT: Doc, document to be classified
S, Sentiment Lexicon
OUTPUT: Class, Sentiment class for Doc

Initialise: Doc^+, Doc^-
for all t∈ Doc **do**
  Retrieve t^+ and t^- from S
  if $t^+ + t^- > 0$ **then**
    $Doc^+ \leftarrow Doc^+ + t^+$; $Doc^- \leftarrow Doc^- + t^-$
  end if
end for
if $Doc^+ \geq Doc^-$ **then**
  Return Positive
else
  Return Negative
end if

3.3.1 Naïve Bayes Classifiers

Naïve Bayes is a probabilistic classifier that operates by building statistical models of classes from the training dataset. In order to describe the classifier, assume that documents from the training dataset are divided into m mutually exclusive classes, $C = \{c_1, c_2, ..., c_m\}$. Then, the parameters to the multinomial model for class $c \in C$ would be: $\theta_c = [\theta_{c1}, \theta_{c2}, ..., \theta_{cn}]$, where n is the number of features in the vocabulary, $\sum_j \theta_{cj} = 1$ and θ_{cj} is the conditional probability that feature j occurs in class c. The conditional probability θ_{cj} is typically smoothed by a Laplace count to avoid zero values. The class label for a test document $d = \{d_1, d_2, ..., d_n\}$, where d_j is the frequency of feature j in document d, is predicted using the Bayes rule (Rahman, 2009):

$$label(d) = \arg\max_c \left(P(c) \frac{P(d|c)}{P(d)} \right) \tag{3.2}$$

The probability $P(d|c)$ is estimated by using a multinomial distribution, i.e.

$$P(d|c) = \left(\sum_{d_1, d_2, ..., d_n} d_j \right) \prod_j (\theta_{cj})^{d_j} \tag{3.3}$$

The multinomial distribution assumes that the features in document d are independent of each other. This is known as the Naïve Bayes Assumption and only holds because of the stochastic nature in which words are used in language (Domingos and Pazzani, 1996). The multinomial coefficients in Equation 3.3 and the probability $P(d)$ in Equation 3.2 can be dropped as they are constant across all classes. This simplifies Equation 3.2 as

follows:

$$label(d) = \arg\max_{c} \left(P(c) \prod_{j} (\theta_{cj})^{d_j} \right) \tag{3.4}$$

The multiple products in Equation 3.4 have the tendency to lead to an arithmetic underflow thus, it is common practice to represent it in logarithm space:

$$label(d) = \arg\max_{c} \left(\log P(c) + \sum_{j} d_j \theta_{cj} \right) \tag{3.5}$$

Finally, the label of document d is taken as the class that yields the maximum value of the resultant Bayes rule formulation as shown in Equation 3.5

3.3.2 Maximum Entropy Classifiers

These are feature-based classifiers that work on the idea that the most uniform model that satisfies a given constraint should be preferred. In a two-class scenario, it is the same as using logistic regression to find a distribution over the classes. Unlike the Naïve Bayes, this classifier makes no feature independence assumptions thus features like bigrams and phrases can be added in building the classifier without overlap (Go et al., 2009). The model is represented by the following:

$$P(c|d, \lambda) = \frac{\exp\left(\sum_i \lambda_i f_i(c, d)\right)}{\sum_{c'} \exp\left(\sum_i \lambda_i f_i(c, d)\right)} \tag{3.6}$$

Where, c is the class, d is the document to be classified, and λ is a weight vector. The weight vectors decide the significance of a feature in classification. A higher weight means that the feature is a strong indicator for the class. The weight vector is found by numerical optimization of the lambdas to maximize the conditional probability. The Maximum Entropy classifier was found to perform best in sentiment classification compared to Naïve Bayes or Support Vector Machines (Go et al., 2009, Thelwall et al., 2012).

3.3.3 Support Vector Machines

The Support Vector Machine (SVM) belongs to a family of classifiers that perform classification by building a separating boundary between classes of interest. A special property of the SVM is that it simultaneously tries to minimise the generalisation error while maximising the geometric margin between the classes (Vapnik, 1998). Thus, it is also known as the maximum margin classifier. Figure 3.4 illustrates a simplified version of a linear SVM trained on instances from two classes. Here the SVM constructs a separating hyperplane and then maximises the margin between the two classes. In calculating the margin, the SVM constructs two parallel hyperplanes, one on each side of the initial one. These hyperplanes are then expanded perpendicularly away from each other until they are in contact with the closest training instances from either class. These instances are known as the support vectors and illustrated in bold in Figure 3.4. Intuitively, the best separation is the one with the largest margin between the two hyperplanes. Thus, the larger the margin; the lower the generalisation error.

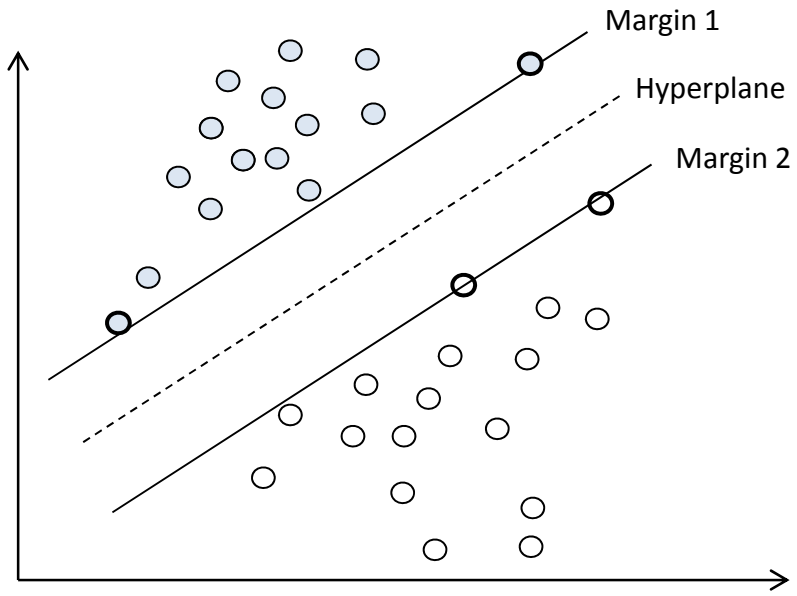

FIGURE 3.4: Support Vector Machines: Classification

SVM is a popular classifier for sentiment classification in particular and text classification in general and it is often considered to be the state-of-the-art classifier.

3.4 Datasets and Statistics

In this research, we perform evaluations on six publicly available benchmark datasets from different social media platforms. Some statistics from these are shown in Table 3.1 where a bracketed value shows the proportion of the corresponding marker (i.e. row) over the number of documents in the corresponding dataset (i.e. column). The datasets represent varying lengths of social media text, use of non-standard/informal terms and occurrence of one sentiment class (positive or negative) compared to the other. Thus enabling us to study the performance of our proposed algorithms on these criteria. The datasets are made available from two sources: cyberEmotions project[1] and SemEval 2014[2]. In this research, we use only the positive and negative documents from the datasets for the evaluation.

3.4.1 CyberEmotions datasets

This consists of 4 datasets labelled by three human annotators. Each document is assigned two scores by an annotator, each ranging from 1 to 5, indicating the strength of the positive and negative sentiment contained in the document. We use the maximum mean score for each sentiment class to label each document as positive or negative. The datasets are as follows.

Digg: This consists of comments crawled from the social news website: digg.com. Comments are extracted from discussion topics expected to contain expressions of sentiment, such as politics and lifestyle. It has 48% more negative than positive comments and are relatively lengthy in size, with an average of 6 sentences and 78 words per comment.

MySpace: This consists of message exchanges between a pair of Internet "friends" from myspace.com. Thus, it is mostly positive (68% more positive than negative). The messages are relatively shorter in size with an average of 2 sentences and 12 words per message.

[1] www.cyberemotions.com
[2] http://alt.qcri.org/semeval2014/index.php?id=tasks

Youtube: This is a collection of comments posted on Youtube. It has 36% more positive than negative comments that are relatively moderate in size, with average of 2 sentences and 18 words per comment.

RunnersW: This is a collection of comments from a specialised sports forum: runnersworld.com. It has 34% more positive than negative comments that are relatively lengthy with average of 5 sentences and 55 words per comment.

3.4.2 SemEval2014 datasets

These are two datasets introduced in SemEval 2014, in relation to task 10B (sentiment classification exercise). The datasets are manually labelled for sentiment classes of positive, negative and neutral via Amazon Mechanical Turk. We exclude neutral labelled documents for the current task. The datasets are as follows:

Twitter: This is a collection of 2587 positive and 843 negative Twitter posts (i.e. 50% more positive than negative documents). It has an average of 2 sentences and 18 words per document.

LiveJ: This collection includes responses to blogs on the social networking site, livejournal.com. It is the shortest of our datasets in document length with average of 1 sentence and 12 words per documentIt is also the least skewed in class composition with just 16% more positive than negative documents.

TABLE 3.1: Datasets and statistics

stats	Digg	LiveJ	MySpace	RunnersW	Twitter	Youtube
#Documents						
Positive	201	427	702	484	2587	1665
Negative	572	304	132	221	843	767
Statistics						
Avg. sentence	6	1	2	5	2	2
Avg. word	78	12	12	55	16	18
Negation	522(0.68)	31(0.04)	351(0.42)	987(1.40)	1227(0.36)	844(0.35)
Intensifiers/Dim	371(0.48)	240(0.33)	165(0.20)	541(0.77)	396(0.16)	448(0.18)
Discourse markers	743(0.96)	411(0.56)	543(0.65)	1231(1.75)	1161(0.34)	1238(0.51)
Capitalisation	95(0.12)	54(0.07)	84(0.10)	121(0.17)	669(0.20)	231(0.09)
Repeat letter	13(0.02)	23(0.03)	61(0.07)	16(0.02)	51(0.01)	61(0.03)
Emoticons	37(0.05)	91(0.12)	192(0.23)	180(0.26)	530(0.15)	341(0.14)

3.5 Text Pre-processing

Text pre-processing is an integral component of many NLP tasks including sentiment analysis. It usually involves a number of steps in a pipeline aimed at transforming raw text into a format suitable for input to an algorithm.

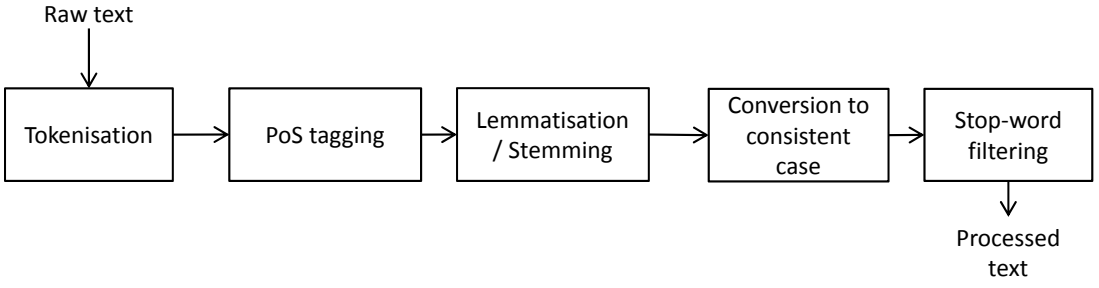

FIGURE 3.5: Text Pre-processing Steps

Figure 3.5 shows the typical text-preprocessing operations for sentiment analysis. First, an input text is broken into its unit constituents (tokenisation). This step is very important in sentiment analysis of social media text because sentiment information can be sparsely and unusually represented. For instance, a single cluster of punctuations

like *>:-(* might tell the whole story and should be kept together at tokenisation. Second, the resultant tokens are tagged with their respective PoS. In this research, PoS information is required for score extraction from SentiWordNet. This step is normally bypassed for lexicons that do not distinguish between PoSs and some machine learning approaches. Third, lemmatisation or stemming is performed on the PoS tagged tokens. The goal of both lemmatization and stemming is to reduce inflectional forms and sometimes derivationally related forms of a word to a common base form. However, the two differ in that stemming employs crude heuristics to achieve the goal to an extent that the resultant token may not be a valid language word while lemmatisation achieves the goal with reference to a standard dictionary thus the resultant token is not changed beyond meaningful words. For example, whereas when lemmatised *smoking* becomes *smoke*, it becomes *smok* when stemmed. In this research we employ lemmatisation as entries in SentiWordNet are in their base dictionary form (lemma). Fourth, the tokens are converted to a consistent case (usually lower case). This avoids algorithms from distinguishing between tokens such as "HERE" and "Here". However, it is possible that a sentiment classifier may benefit from capitalisation for emphasis such as "AWESOME," as compared to "awesome," would imply varying sentiment intensity. In our lexicon-based approach, capitalisation for emphasis is retained and used by our algorithm. Finally, stop-word filtering is typically performed in NLP tasks to remove words that are poor discriminators. Although this may be a form of feature reduction for machine learning, it is typically unnecessary for lexicon-based methods as stop-words are typically not included in a lexicon or are associated with zero values in high-coverage lexicons (e.g. SentiWordNet) thus cannot influence classification. Therefore, in this research, we do not perform stop-word filtering.

3.6 Evaluation Metrics

As typical with unbalanced datasets (Li et al., 2005), in this research we report evaluation results using precision, recall and F measure. The Contingency Table 3.2 illustrates the arrangement of classifier outcome given human judgment in a two-class problem (positive and negative). Where, TP is the number of positive documents correctly classified as

positive ("true positive"), FP is the number of negative documents falsely classified as positive ("false positive"), TN is the number of negative documents correctly classified as negative ("true negative") and FN is the number of positive documents falsely classified as negative (false negative).

TABLE 3.2: Contingency Table

Classification		Human Judgment	
		Positive	Negative
Classifier Judgment	Positive	TP	FP
	Negative	FN	TN

Precision for a given class c is the fraction of correctly classified documents out of documents classified as c. Thus, the precision values for the two classes, positive (P_{pos}) and negative (P_{neg}), are determined as follows:

$$P_{pos} = \frac{TP}{TP + FP}, \quad P_{neg} = \frac{TN}{TN + FN} \tag{3.7}$$

Recall is the fraction of documents correctly classified out of all documents from a given class c. Therefore, the recall values for the two class, positive (R_{pos}) and negative (R_{neg}), are determined as follows:

$$R_{pos} = \frac{TP}{TP + FN}, \quad R_{neg} = \frac{TN}{TN + FP} \tag{3.8}$$

The F Measure for a class c is given by the harmonic mean of the class' precision and recall as follows:

$$F_c = \frac{2P_cR_c}{P_c + R_c} \tag{3.9}$$

We combine F Measure from the two classes, positive (F_{pos}) and negative (F_{neg}), into a single value by taking their arithmetic mean as follows:

$$AvgF = \frac{F_{pos} + F_{neg}}{2} \tag{3.10}$$

Finally, we report statistical significance of results using the chi-square (χ^2) test for two proportions (Berenson et al., 2012), at 95% confidence level. This compares the contingency tables produced by any two competing systems.

3.7 Chapter Summary

In this chapter, we presented details about SentiWordNet, the main lexicon used in various stages of our research. This includes the lexicon generation process and the baseline sentiment classification algorithm using the lexicon. We also discussed the benchmark supervised machine learning algorithms for sentiment classification namely: Naïve Bayes, Maximum Entropy and Support Vector Machines. Finally, we provided details about the text pre-processing operations employed in the research, the datasets and metrics used for evaluation.

Chapter 4

SmartSA: A Contextual Sentiment Classifier for Social Media

Adopting the lexicon-based methodology, this chapter presents SmartSA, a sentiment classification system for social media text that leverages rich sentiment information in SentiWordNet for contextual analysis. We show how contextual adjustment of SentiWordNet scores for terms based on negation, intensification/diminishing, discourse structure and other non-lexical phenomena can significantly influence sentiment analysis of social media. Given that sentiment scores are associated to word senses in SentiWordNet, it is imperative to investigate the applicability of word sense disambiguation (WSD) in determining the right sense for terms in relation to other score extraction approaches that avoid WSD. To this end, we formalise score extraction approaches from the literature and introduce a Lesk-like algorithm for WSD (Lesk, 1986).

Being a high-coverage lexicon, SentiwordNet offers sentiment scores for typical sentiment modifying terms. In this chapter, we analyse the behaviour of these terms both as sentiment carriers and as sentiment modifiers of other terms. This informed our strategies for local contextual analysis in SmartSA. The main contribution of SmartSA is the use of contextual information to improve sentiment scores. We apply this information in two

places. Firstly, context is used to identify the correct sense when extracting scores from SentiWordNet. Secondly, the extracted scores are adjusted on the basis of contextual analysis.

Figure 4.1 shows the main components of the classifier. Sentiment classification of documents involves the extraction of scores from SentiWordNet. Thereafter, contextual analysis is applied to modify prior polarities of documents' terms. Here, we introduce strategies for negation, intensification/diminishing, discourse analysis, capitalisation, repeating letters/characters and emoticons. Sentiment class for a given document is determined by the maximum of the contextually modified scores. Details of these operations are presented next.

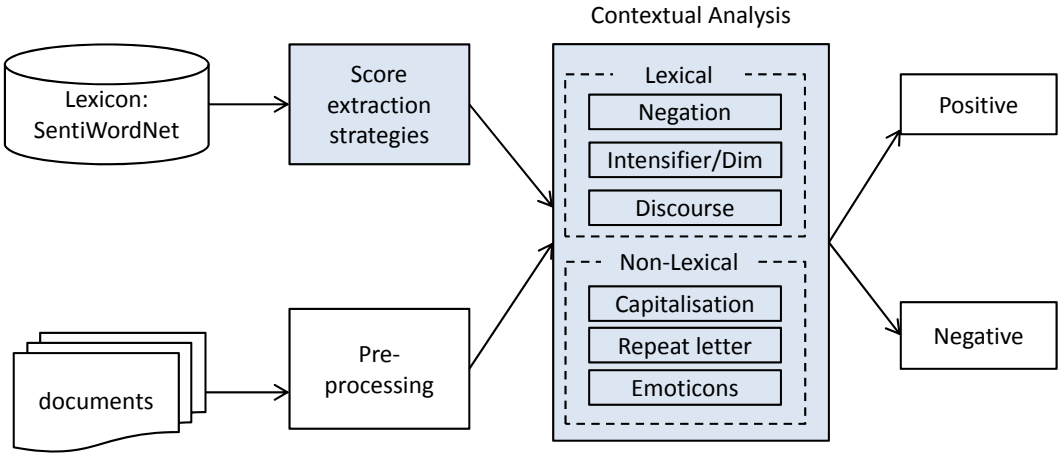

Figure 4.1: SmartSA

4.1 Score Extraction

Lexicon-based sentiment analysis involves the extraction of sentiment scores from a lexicon. Several score extraction approaches are possible with SentiWordNet given the detailed information it provides about terms. Typically, these approaches (presented in the next subsections) require part-of-speech (PoS) tags of terms to be determined prior to the extraction of the terms' positive and negative scores, (c={+, -}).

4.1.1 Most Frequent Word Sense (MFWS)

In WordNet (and also SentiWordNet), word senses for terms are ordered according to their natural usage frequency, with the first sense ($sense_1$) being the most frequent. This sense has higher chance to occur in a document than any other sense, thus, can be representative for the term. This approach is given in equation 4.1.

$$score(t|PoS)_c = score(t|PoS, sense_1)_c \tag{4.1}$$

4.1.2 Average of Word Senses (AWS)

In this approach, sentiment score of a term given PoS is determined by the average score over all the term's words senses as given in equation 4.2.

$$score(t|PoS)_c = \frac{\sum_{i=1}^{|sense(t|PoS)|} score(t|PoS, sense_i)_c}{|sense(t|PoS)|} \tag{4.2}$$

Where $|sense(t|PoS)|$ is the number of senses of the term, t, when occuring as the given part-of-speech, PoS.

4.1.3 Weighted Average of Word Senses (WAWS)

Here, frequency of word sense, as given by the sense order, i, in WordNet, is used to obtain a weighted average as follows.

$$score(t|PoS)_c = \frac{\sum_{i=1}^{|sense(t|PoS)|} \frac{1}{i} \times score(t|PoS, sense_i)_c}{|sense(t|PoS)|} \tag{4.3}$$

4.1.4 Average of Word Senses and Parts of Speech (APoS)

In equation 4.4, sentiment score for a given term is the average scores over all its word-senses across all PoS. Averaging in this way avoids word sense disambiguation as well as

PoS tagging which could be prone to error especially with informal social media content.

$$score(t)_c = \frac{\sum_{j=1}^{|PoS|} \left(\frac{\sum_{i=1}^{|sense(t|PoS)|} score(t|PoS, sense_i)_c}{|sense(t|PoS)|} \right)}{|PoS|} \tag{4.4}$$

4.1.5 Word Sense Disambiguation (WSD)

WSD involves the identification of the meaning evoked by words in context. It provides the ideal approach for the extraction scores from SentiWordNet since scores are associated to word senses rather than terms. We introduce a variant of the Lesk (1986) method for WSD (Algorithm 2). This method is based on the idea that similar adjacent terms imply similar sense. Thus, for each sense of a target term for which sense is to be disambiguated (step 3), its gloss is extracted from SentiWordNet (step 4) and similarity with the context (adjacent terms) of the term is measured. We use all terms that co-occur with the target term in a document as the context of the term, as opposed to sentence or text window, since documents in our domain of application (social media) are short in size. We use the cosine similarity metric to quantify the similarity between a term's gloss and the term's context. Finally, the sense with the highest similarity is returned as the adjudged word sense of the target term. The algorithm also ensures that in the case of a tie, the most frequent sense as specified by sense order is returned.

4.2 Contextual Analysis

In social media, two types of modifiers affect term polarity in context: *lexical* and *non-lexical* valence shifters. Lexical valence shifters are in the form of dictionary recognisable words whereas non-lexical valence shifters are other word inflections and artificial symbols that affect the expression of sentiment such as repeating a letter or character, capitalisation for emphasis and the use of emoticons. Crucial to implementing any score adjustment strategy is the identification of the terms affected by modifiers in text (scope of modifiers). This can be the immediate term succeeding the modifier (e.g. in *"I didn't*

Algorithm 2 WSD

INPUT: t, term to be disambiguated
D, document containing t
S, SentiWordNet
OUTPUT: Sense, Adjudged word sense of t

```
sense ← sense_1
tempScore ← 0
for all sense_i ∈ senses(t) do
    gloss_i ← ExtractGloss(sense_i) from S
    score_i ← CoSim(gloss_i, D)
    if score_i > tempScore then
        tempScore ← score_i
        sense ← sense_i
    end if
end for
Return sense
```

enjoy it") or a term farther away from the modifier *"I don't think I will enjoy it"*. The modified term can also be before the modifier (e.g. *I enjoy it very much*) or after (e.g. *I very much enjoy it*). Ideally, it is the task of a dependency parser to identify modifiers in text and the terms they modify. However, with the attendant non-standard spelling and grammar of social media, standard parsers often fail to produce satisfactory results (Liu et al., 2011, Ritter et al., 2011). For instance, with the omission of the apostrophe in *"I dont like sausages"*, the Stanford parser[1] fails to recognise the negation. Therefore, instead of using the standard parsers, we adopt the window-based approaches, whereby modifiers are assumed to affect terms within a specific text window (Hogenboom et al., 2011, Thelwall et al., 2012, 2010).

4.2.1 Lexical Valence Shifters

Lexical valence shifters are typically used to increase sentiment (i.e. intensifiers e.g. 'very', 'highly'); decrease sentiment (i.e. diminishers e.g. 'slightly', 'somewhat') or negate sentiment (i.e. negation terms, e.g. 'not', 'never'). These terms are associated with sentiment scores in SentiWordNet. For example, the positive and negative scores of the adverb 'very' are 0.25 and 0.0 respectively, thus, the term always contributes positively. However, this term can also contribute negatively, for example in 'very bad'.

[1] http://nlp.stanford.edu:8080/parser/

Therefore, it is important to determine the polarity contribution likely to be made and modify scores accordingly.

4.2.1.1 Negation

Negation is a common linguistic phenomenon that affect sentiment expressions in a profound way. Taking into account the positive and negative scores for terms in SentiWordNet, we propose and investigate the following implementation of the *switch* (Taboada et al., 2011) and *shift* (Taboada et al., 2011) approaches of handling negation in SmartSA. The choice of a window size as the scope of negation should be guided by two requirements: the need to capture the affected words despite a long-distance effect of the negation, and the need to constrict the size so as not to capture other terms that are not affected by the negation. Existing literature suggests several text window sizes as the scope for negation ranging from one to five words following the negation word or on both sides of the negation word (i.e. a radius). For instance, Polanyi and Zaenen (2004) and Kennedy and Inkpen (2006) assume the word following a negation word as its scope while Paltoglou and Thelwall (2012) assume a radius of five words from a negation word to be its scope. However, a recent studies show that there is no significant difference in performance between various radii (between 1 and 5) as the scope of negation (Dadvar et al., 2011, Paltoglou and Thelwall, 2012). Indeed, Dadvar et al. (2011) found that the performance in sentiment classification remains the same with a three, four, or five term window, which was slightly better than using a two or one term window. This shows that a three term window is more appropriate than the other alternatives, as it attains the best performance with the least number of terms to search. Therefore, in this work, we use a radius of three terms from a negation term as the scope of the negation. Our negation detection is based on a list of negation terms by Thelwall et al. (2012) extended to include scenarios when apostrophe is omitted or misplaced for terms such as in *don't*, *wouldn't*, *couldn't* and *can't*.

Switch. Involves the swap of positive and negative scores of terms that are under the influence of negation. This will have the same effect as switch approaches implemented

with single score lexicons. Consider examples (a) and (b) in Figure 4.2 where positive and negative prior polarities of *"not good / not excellent"* are swapped after the switch operation. This reflects the contextual polarity (negative) of the phrases. However, switch tends to produce an undesired effect of making negated high sentiment-bearing terms more negative than negated low sentiment-bearing terms. For instance "not excellent" is overall more negative (-1.625) than "not good" (-1.138). The shift approach is supposed to mitigate against this limitation.

	Before Switch				After Switch			
(a)	not	good	→		not	good	:	sum
pos:	0.000	0.638		pos:	0.000	0.125	=	0.125
neg:	0.625	0.125		neg:	0.625	0.638	=	1.263
							pos-neg=-1.138	
(b)	not	excellent	→		not	excellent		
pos:	0.000	1.000		pos:	0.000	0.000	=	0.000
neg:	0.625	0.000		neg:	0.625	1.000	=	1.625
							pos-neg=-1.625	

Figure 4.2: Switch negation

Shift. With this approach, negation is considered as a sentiment diminisher rather than complete inverter of sentiment. With single score lexicons, this involves reducing a term polarity score by a certain weight. Considering that, negation seems to affect the dominant polarity of terms, we implement the shift approach in SmartSA by focusing on this polarity dimension. When a term is negated, its dominant polarity is ignored. For instance, in Figure 4.3 examples (c) and (d), the contextual polarity of the phrases 'not good' and 'not excellent' becomes negative after the shift operation. The relative intensities of their polarity are also maintained (i.e. 'not good' is more negative than 'not excellent').

It can be noted, from examples (a)-(c), that it is possible to remove sentiment scores of the negation term 'not' from the aggregation process without changing the contextual polarities of the phrases. However, recent literature suggest that negation terms are

	Before Shift				After Shift			
(c)	not	good	→		not	good	:	sum
pos:	0.000	0.638		pos:	0.000	~~0.638~~	=	0.000
neg:	0.625	0.125		neg:	0.625	0.125	=	0.750
							pos-neg=-0.750	
(d)	not	excellent	→		not	excellent		
pos:	0.000	1.000		pos:	0.000	~~1.000~~	=	0.000
neg:	0.625	0.000		neg:	0.625	0.000	=	0.625
							pos-neg=-0.625	

Figure 4.3: Shift negation

sentiment carriers of their own (Potts, 2011*a*). This is further evident from the high sentiment scores associated with such terms in SentiWordNet. Therefore, in SmartSA we include scores of negation terms in the aggregation. Thus, by doing so we implement the concept that negation terms are both modifiers of sentiment and sentiment-bearing. An exception arises with negation of negatively dominant terms (terms that are more negative than positive). In such a case, including the scores of negation terms will produce undesired result because sentiment scores for negation terms from SentiWordNet are very negative and typically more negative than many negative terms such as *'angry', 'bad' or 'worry'*. Thus, the scores of the negation may dominate the aggregate leading to the incorrect assessment of phrases like *"not angry", "not bad" or "don't worry"*. For instance, *'not angry'* still remains overall negative after the shift operation as shown in Figure 4.4, example e. Therefore, in the case of negation of negatively dominant terms, we exclude the scores of the negation terms from the aggregation as shown in Figure 4.4, example f.

4.2.1.2 Intensification/Diminshing

Intensifiers and diminishers are linguistic constructs used to increase and decrease sentiment or emotional charge of terms. In SmartSA, the dominant polarity of sentiment-bearing terms within the scope of an intensifier is increased (or decreased in the case of a diminisher) relative to the strength of the intensifier (or diminisher) as illustrated in

	Before Shift			After Shift			
(e)	not	angry	→	not	angry	:	sum
pos:	0.000	0.307	pos:	0.000	0.307	=	0.307
neg:	0.625	0.500	neg:	0.625	~~0.500~~	=	0.625
						pos-neg=-0.318	
	Before Shift			After Shift (without scores)			
(f)	not	angry	→	not	angry		
pos:	0.000	0.307	pos:	~~0.000~~	0.307	=	0.307
neg:	0.625	0.500	neg:	~~0.625~~	~~0.500~~	=	0.000
						pos-neg=0.307	

Figure 4.4: Modified shift negation

Figure 4.5. We use a lexicon of intensifiers and diminshers where each term is assigned a strength score of 1 or 2 indicating the degree to which the term increases or decreases sentiment (Thelwall et al., 2012). For instance, the intensification strength of 'extremely' is 2 while that of 'very' is 1. However, Taboada et al. (2011) argue that rather than absolute values, modifiers should have scores relative to the sentiment strength of the term they modify. Thus, they proposed assigning a modifier a percent score of the term they modify. Adapting this approach, we convert the strengths assigned to modifiers by Thelwall et al. (2012) to a percentage increase or decrease in dominant polarity of terms (50% for 1 and 100% for 2). This approach also ensures that the score of a modifier does not exceed the score of the term being modified. Notice that, similar to negation terms, intensifiers and diminishers are associated with sentiment scores in SentiWordNet. Thus, the scores could be incorporated into the aggregation or nullified. We investigate both options in the evaluation of SmartSA.

	As modifier			
(h)	really	awful	:	sum
pos:	0.438	0.250	=	0.688
neg:	0.065	0.542×(100%+50%)	=	0.878
				pos-neg=-0.19

Figure 4.5: Intensifier as modifier

Table 4.1: Grouping of Discourse Structures

Group		Markers	Example	Effect
1	Concession	Admitting, Albeit, Allowing that, Although	[**although** I don't like the series,]$_S$ [I really enjoyed this episode]$_N$	No effect on nucleus, decrease satellite
	Background	X earlier, X later, Over X, From X to Y, But X after, But X later, Between X and Y	[I was happy the laptop was working]$_S$ [**but 3 days later** it stopped]$_N$	
2	Condition	As because, As far as, As long as, Assuming that, Conceding that	[**if** the world ends on december2,]$_S$ [i'm gonna be so disappointed]$_N$	Decrease Nucleus, Decrease Satellite
	Circumstance	When, After, Following, Once, Before, While, And then, And when, now that,	[The animal is dangerous]$_N$ [**when** left in hunger]$_S$,	
	Purpose	So that, So as	[the quality of the food should be improved]$_N$ [**so as** to improve sales]$_S$	
3	Elaboration	And, In fact, In addition, Also, By verb-ing, For example	[**in addition** to the location,]$_N$ [the food also tastes good]$_S$,	No effect on Nucleus, Increase Satellite
	Evaluation	It (is|was) (our|my) (opinion|understanding) (that), In (our|my) opinion, it (seems|seemed) to (us|me) (that)	[Now **it seems** action of Yadav]$_N$ [have back fired]$_S$	
	Re-statement	Or, For instance		
	Summary	In any case, In sum, To sum up, In summary, In a nutshell		
	Cause/Result	So that, In case, Because, Since, After all, On the grounds that, Given that, Therefore	[I always eat in that restaurant]$_N$ [**because** of its friendly staff]$_S$	

4.2.1.3 Discourse structure

Discourse structure is concerned with how text units (discourse segments) are organised to convey meaning. This structure is determined through discourse analysis involving the identification of discourse segments of text, their structural arrangement and the relation that may exist among them. A popular theory for discourse analysis is the rhetorical structure theory (RST) (Mann and Thompson, 1998). It posits that text can be broken into non-overlapping spans in a tree-like structure with relations that

may exist between any two adjacent spans. Each text span can either have the status of the central focal point of the writer's message (i.e. nucleus) or a supporting message that help in understanding the nucleus (i.e. satellite). Mann and Thompson (1998) highlighted 24 relation types which include: the introduction of additional information (elaboration), conflicting statements (concession) and conditional statements (condition). These relations can either hold between 2 adjacent nuclei (paratactic), or between a nucleus and a satellite (hypotactic) spans.

The major challenges of automatic discourse analysis are: to split a piece of text into discourse segments, to identify applicable relations, their spans and the statuses of the spans (nucleus or satellite); and the construction of a valid RST tree. There exists a large body of work on these, focused on supervised or unsupervised methods. Supervised methods often use Penn Discourse Treebank (PDTB [2]), a human annotated corpus for discourse structure, to train machine learning algorithms which in turn predict the structure of unseen documents (Hernault et al., 2010, Soricut and Marcu, 2003). Previous work on sentiment classification of reviews has employed the supervised discourse analysis parsers (Heerschop et al., 2011, Taboada et al., 2008). However, considering that the PDTB corpus is made of documents from the Wall Street Journal, which are fairly well-written in terms of the use standard spellings for terms, punctuations and grammar, one can expect parsers trained on this data to perform poorly on informal social media data.

The unsupervised discourse parsing relies on insights from corpus studies to generate rule-based parsers. Existing rule-based algorithms are formulated using insights from fairly formal text, similar to PDTB. However, unlike machine learning models, these algorithms are flexible and can be extended to incorporate insights from social media data. For instance, the terms that signal the occurrence of discourse relations (discourse markers) can be shortened in social media. For example, 'because' can as well be written as *'cos'*, *'bcos'* or *'bc'*. Therefore, it is useful to incorporate these variations.

The main idea behind harnessing discourse structure for sentiment analysis is that, since discourse structure of a text can specify segments of the text that are more (or

[2] https://www.seas.upenn.edu/pdtb/

less) important to the writer's message, it can also be exploited to associate weights to the segments. Consequently, sentiment terms that occur within the important segments will have higher weights. This should lead to an improved sentiment analysis. Working with this notion, in SmartSA we use regular expressions to identify occurrences of discourse markers and apply a weight to their scope. Here, the scope of a discourse marker is the two text segments (nucleus/satellite) involved in the relation that the marker represents. We use the rule-based algorithm in Marcu (2000) to split a text into discourse segments using lists of discourse markers per relation, which we extend to include social media variation the markers (Das, 2010). Next, we need to determine the nucleus/satellite (or nucleus/nucleus, for paratactic relations). To this end, we utilise the contextual information derived from corpus study of distributional environments for discourse markers (Das, 2010). This information specifies the nucleus/satellite of a relation in reference to a given segment containing a discourse marker from the specified relation (this can be the segment before or after). This is usually influenced by the position of the discourse marker within its segment (beginning, middle or end).

After the discourse segmentation and the identification of nucleus/satellite segments, we apply a weight corresponding to the potential effect of each segment for sentiment analysis. Considering that, similar to the role of intensifiers/diminshers, the effect of discourse increases/decreases sentiment, we mapped this on the effect of typical intensifier/diminisher (i.e. 50% increase/decrease). Although, Mann and Thompson (1998) identified 24 generic discourse relations, not all are relevant for sentiment analysis. Thus, here we concentrate on the subset of 11 relations identified to be useful for sentiment analysis (Das, 2010). Although Das (2010) identifies the discourse relations that are important for sentiment analysis and the distributional information of their markers, which enables the identification of the nucleus and satellite for each marker, they did not utilise such information for sentiment analysis. Hence, we introduce the groupings and weights in order to utilise the relations for sentiment analysisWe heuristically group the discourse relations according to their potential effect, with respect to sentiment expression, to their nucleus or satellite. Table 4.1 shows the groupings, some discourse markers (or constructs) for each relation, and some example sentences that illustrate

the behaviour of the relations. A full list of these constructs can be found in (Das, 2010). The groupings are discussed in more details next.

Group 1: No Effect on Nucleus, Decrease Satellite. These are the relations of *concession* and *background*. Concession holds between conflicting information present in nucleus and satellite segments whereby the writer clearly favours the nucleus, though not denying the satellite. Therefore, it is worthwhile for a sentiment analysis system to concentrate on the sentiment expressed within the nucleus of this relation while suppressing the satellite. For example, in [**although** I don't like the series,]$_{\text{S}}$ [I really enjoyed this episode]$_{\text{N}}$, the writer seems to promote the positive sentiment (really enjoy) within the nucleus segment (denoted by the subscript N) despite the negative sentiment (don't like) of the satellite segment (denoted by the subscript S). In this example, the relation is signalled by the discourse marker *although* (denoted in bold font). For background, the satellite provides a context based on which the information provided in the nucleus can be better understood. The sentiment expressed in this context can be the same or different from that expressed in the nucleus. However, since the nucleus is the focal point of the relation, it is more reliable to concentrate on the sentiment it conveys and suppress the sentiment in the satellite which can be tangential to the sentiment expressed in the nucleus. For example, in [I was happy the laptop was working]$_{\text{S}}$ [**but 3 days later** it stopped]$_{\text{N}}$, the focus is on the negative sentiment within the nucleus (stopped) despite the positive sentiment in the satellite (happy).

Group 2: Decrease Nucleus, Decrease Satellite. These are the relations of *condition*, *circumstance* and *purpose*. Condition presents a hypothetical future whereby the realisation of the nucleus depends on the realisation of the satellite. However, both nucleus and satellite are unrealised. Thus, for the purpose of sentiment analysis, such situation can be given low weight. For instance, in [**if** the world ends on december 2,]$_{\text{S}}$ [i'm gonna be so disappointed]$_{\text{N}}$, despite the negatively charged terms in both segments (world ends, so disappointed), the text still seems to remain largely neutral. For circumstance, the satellite sets the framework within which the reader is expected to interpret the nucleus. It tends to soften both the nucleus and the satellite. For example, the

statement: [The animal is dangerous]$_{\mathrm{N}}$ [**when** left in hunger]$_{\mathrm{S}}$, though dominated by negative terms (dangerous, hunger) is still of mild sentiment. Similarly, in purpose, the satellite presents a situation to be realised through the activity in the nucleus, as in the example: [the quality of the food should be improved]$_{\mathrm{N}}$ [**so as** to improve sales]$_{\mathrm{S}}$.

Group 3: No effect on Nucleus, Increase Satellite. These are *elaboration*, *evaluation*, *re-statement*, *summary* and *cause/result*. Elaboration exists between a nucleus and a satellite when the satellite presents additional information to better understand the nucleus. Thus, the sentiment expressed in the satellite tends to be supportive of the nucleus. It also tends to be more verbose, increasing the chance of containing sentiment-bearing terms. For example, in [**in addition** to the location,]$_{\mathrm{N}}$ [the food also tastes good]$_{\mathrm{S}}$, the sentiment expressed within the satellite (good) also applies to the nucleus. Re-statement tends to function similar to elaboration. The satellite is the paraphrase of the nucleus. Thus, sentiment within the satellite is important as it is also applicable to the nucleus. In evaluation, the satellite tends to contain an opinion regarding the nucleus. This is directly relevant for sentiment analysis as it signals a reliable location for opinions. For example, [Now **it seems** action of Yadav]$_{\mathrm{N}}$ [have back fired]$_{\mathrm{S}}$, the evaluation marker (it seems) signals the appearance of the sentiment-charged term (back fired) in the satellite. In the summary relation, the satellite provides concise and overall information the writer meant to convey from an often lengthier nucleus. The opinion expressed in the satellite is thus representative of the text and can be given high weights. Finally, the cause/result signifies relation between satellite and nucleus whereby the information given in the satellite is the cause of the information present in the nucleus. Both segments tend to present the same sentiment orientation, with satellite being central to believing the nucleus. For example, in the text: [I always eat in that restaurant]$_{\mathrm{N}}$ [**because** of its friendly staff]$_{\mathrm{S}}$, the positive justification in the satellite (friendly staff) adds strength to the overall sentiment of the text.

4.2.2 Non-lexical Modifiers

In addition to lexical valence shifters, non-lexical modifiers are also commonly used to increase sentiment in social media. These modifiers manifest in the form of term inflection with a sequence of repeating characters/letters, capitalization and the occurrence of emoticons.

4.2.2.1 Capitalisation

The informal social media communication presents the convention of term capitalisation for emphasis. This is often used to emphasise sentiment or emotion expressions. Therefore, we introduce an approach in which capitalisation is treated as the intensification of the capitalised term. This adjustment is applied only if the rest of the sentence is not capitalised because in such cases the capitalisation may not be for emphasis but writing style. We use the intensification strength of 'very', being an average and the most occurring lexical intensifier in our datasets. For example, the sentence "saw this last night...AMAZING!" becomes "saw this last night...very amazing!". We do not extend the intensification to the neighbouring terms because capitalisation is also often used for abbreviations and acronyms.

4.2.2.2 Repeated Letter/Character

A repeat of the same letter or character is another phenomenon used to express emphasis in social media. In SmartSA, when a sequence(s) of three or more letters is detected, the target term is identified by first reducing the number of the letter to a maximum of two and checked with SentiWordNet. If the intermediate word is not found, the repeated letters are further reduced to one letter, one sequence at a time. We consider a sequence of repeated letters as an intensification of not just the affected term but also its context. This is because, unlike with capitalisation, a sequence of repeated letter is mainly for emphasis and sometimes the affected term is not sentiment-bearing (e.g. "Mannnnnn, I loved this show"). The occurrence of three or more consecutive exclamation or question

marks or a mixture of both is also treated as sentiment intensification context using the intensification weight of the word 'very'.

4.2.2.3 Emoticons

In the informal social media, emoticons are often used to express sentiment for either the whole document or individual sentences. In SmartSA, we identify occurrence of emoticons based on the emoticon list in Thelwall et al. (2010). If one or more positive (or negative) emoticons are found in a sentence, the sentence is simply assigned the scores of the emoticon (i.e. pos=1.0, neg=0.0 for positive emoticon; pos=0.0 and neg=1.0 for negative emoticon). We restrict the context of emoticons to sentence level as sentiment can change from one sentence to another (Andreevskaia et al., 2015).

4.2.3 SmartSA Algorithm

The classifier is shown in Algorithm 3. It takes as input, the document to be classified, SentiWordNet and lists of lexical valence shifters and emoticons. Each sentence contained in the document is checked for the occurrence of an emoticon. If present, the sentence carries sentiment scores of the emoticon without further analysis of the sentence's text (steps 3-4). Otherwise, the sentence's text is scanned for terms that contain repeating letters or characters of question/exclamation marks. These are converted to their dictionary equivalents (step 8) and appended with the intensifier 'very' (step 9). Next, sentiment scores for each term are extracted from SentiWordNet. Terms that are selectively capitalised within the sentence are intensified using the intensification weight of a typical intensifier (i.e. 50%). Thereafter, score adjustments based on the occurrence of lexical valence shifters are applied to the context of the term (i.e. its neighbourhood) in steps 16-22. Each sentence is assigned the total adjusted scores of its terms. Likewise, each document is assigned the total scores of its sentences. Lastly, the document class is returned as positive, if its total positive score is greater than or equal to its total negative score. Otherwise, the class is returned as negative. Notice that, we can choose to or not to apply any of the contextual adjustment strategies by blocking the applicable steps (for non-lexical valence shifters) or excluding the applicable list from the input (for

lexical valence shifters). This makes for an easy ablation test in order to find out the contribution of each strategy incorporated into SmartSA.

Algorithm 3 SmartSA

INPUT: S, SentiWordNet
LexValShifters{} list of Negation, Intensifiers/Diminishers and discourse markers
Emoticons{} List of positive and negative emoticons
Doc, Document to be classified
OUTPUT: Class, Sentiment class for Doc

Initialise Doc^+, Doc^-, $Sent^+$, $Sent^-$
for all Sentence ∈ Doc **do**
  if ContainSingleType(Emoticon{}) **then**
    $Sent^+ +\leftarrow EmoticonType^+$; $Sent^- +\leftarrow EmoticonType^-$
  else
    for all t ∈ Sentence **do**
      if t.hasRepeatCharacter **then**
        convertStandard(t, SentiWordNet)
        sentence.replace(t, t+" very ")
      end if
      Retrieve t^+ and t^- from S
      if t.isCaps AND ¬sentence.isCaps **then**
        applyAdjustment(50%, t)
      end if
    end for
    for all mod ∈ LexValShifters{} **do**
      if mod ∈ sentence **then**
        modType ← getType(mod)
        context ← getContext(mod, modType, sentence)
        ApplyAdjustment(modType, context)
      end if
    end for
    $Sent^+ +\leftarrow$ sum ($t^+ \in$ sentence), $Sent^- +\leftarrow$ sum ($t^- \in$ sentence)
  end if
  $Doc^+ +\leftarrow Sent^+$, $Doc^- +\leftarrow Sent^-$
end for
if $Doc^+ \geq Doc^-$ **then**
  Return Positive
else
  Return Negative
end if

4.3 Chapter Summary

In this chapter, we introduced a lexicon-based sentiment classification system (SmartSA) for social media domains. The novel feature of SmartSA is that it incorporates social media oriented contextual analysis, that exploits the rich sentiment information for terms in SentiWordNet. First, we formalise various score extraction approaches from the lexicon that are often used in the literature and introduced a new WSD algorithm for the same purpose. Second, we introduced new strategies to handle negation, intensification/diminishing and discourse structure in social media. And Third, Considering the various phenomena often used to emphasise sentiment in social media, we introduced non-lexical contextual analysis based on term capitalisation, elongation by repeating letters/characters and the use of emoticons.

Evaluation of SmartSA is presented in Chapter 7. We conduct ablation experiments to establish contributions of each contextual analysis component of the system. Thereafter, we compare the performance of the system against a baseline (Bag-of-words) aggregation and a state-of-the-art sentiment classification system.

Chapter 5

Hybrid Sentiment Lexicon

SmartSA (Chapter 4) implements several context-aware strategies for sentiment analysis. However, as the system employs a static lexicon, it needs to be extended to address the dynamic nature of social media. The use of a static general-purpose lexicon is insufficient for sentiment analysis of social media because of the following limitations:

- *Dynamic vocabulary*: General purpose lexicons are static resources with fixed vocabulary. Such vocabulary usually does not include non-standard but often sentiment loaded terms found in social media text (e.g. *'lol', 'arrrgh', 'xoxo', thx etc*).
- *Dynamic polarity*: For some terms, though their sentiment scores might be obtained from a general purpose lexicon, such scores may not adequately represent domain-specific semantics. For instance, the dominant polarity of *'sucks'* in SentiWordNet is positive, even though it is typically used to express negative sentiment in social media.

In this chapter, we introduce dynamic SmartSA, DSmartSA, a sentiment classification system that integrates an approach to ascertain sentiment of domain-specific terms and modify sentiment polarities from the general-purpose lexicon according to domain specific semantics. We achieve this by generating a hybrid lexicon that combines sentiment knowledge from a general-purpose static lexicon and a domain-specific dynamic

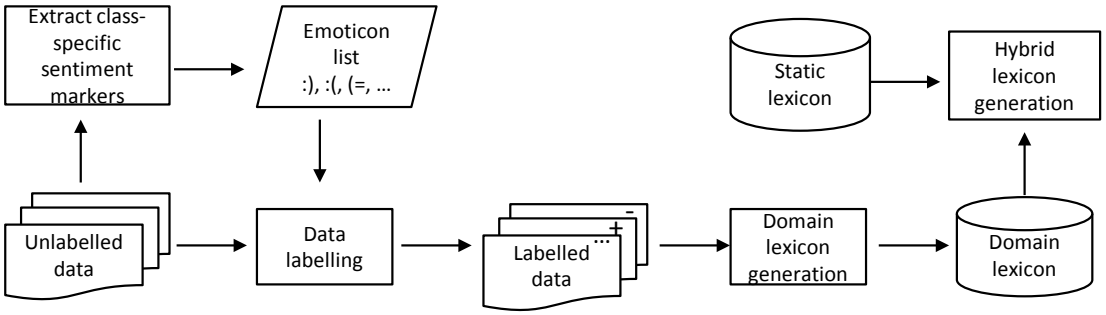

FIGURE 5.1: Hybrid lexicon stages

lexicon. Here, we leverage the idea of distant supervision to learn the domain-specific lexicon. Distant supervision offers an automated strategy to generate sentiment labelled data. Subsequently, each term from the labelled data can be associated with sentiment scores. One of the most popular, and arguably the state-of-the-art metric for associating terms with sentiment scores is based on the Point-wise Mutual Information (PMI) (Turney, 2002). Considering the fact that PMI does not work well on low frequency terms (Sani, 2014), which is an inherent characteristic of newly inducted vocabulary terms, we introduce two metrics inspired by the Term Frequency and Inverse Document Frequency (TF, TFIDF). Further, we present a weighted strategy to integrate scores from the domain-specific with the static lexicon to generate a hybrid lexicon.

The main contribution of this chapter is two-fold. First, we introduce an automated approach to generating a hybrid sentiment lexicon for social media. Second, we introduce two novel term-sentiment association metrics for generating a domain-specific lexicon from social media.

The process of generating a hybrid lexicon from a target domain is shown in Figure 5.1. First, a domain-specific lexicon is generated from data labelled using distant supervision. Next the hybrid lexicon is generated by combining the sentiment scores (learnt for domain terms) in the domain-specific lexicon with existing scores in the general-purpose static lexicon. Sentiment scores in this hybrid lexicon capture general sentiment knowledge as well as domain vocabulary and context.

5.1 Data Labelling: Distant Supervision

Distant supervision offers an automated approach to assigning sentiment class labels to documents. It uses the presence of class specific emoticons in a document as evidence for its true class. For example, a smiley-face emoticon according to distant supervision would express positive sentiment and as such suggest a positive class label for the underlying document content. Accordingly, given a dataset and a lexicon of class-specific emoticons, we can assign such 'noisy' labels to all documents that contain them in order to generate a labelled dataset for supervised learning tasks. This approach provides the positive and negative datasets that we require for a positive/negative sentiment classification. However, for a subjectivity classification, a neutral class dataset may be required. Such a dataset have been gathered from tweets generated by the mainstream media organisations (Go et al., 2009). In order to minimise the level of potential noise, a reasonable strategy is needed to process documents containing emoticons from both positive and negative classes. We noticed from our datasets that less than 1% of documents contain emoticons from both classes. Thus, we remove such documents from the datasets.

We generate distant-supervised datasets on three domains: Twitter, Digg and MySpace (Table 5.2). Twitter distant-supervised data (DsTwitter) consists of 20,000 sentiment labelled tweets based on the appearance of positive and negative emoticons, selected from a larger dataset made available by Sentiment140[1]. Although more data could be selected from twitter, we use the proportionate amount of 20,000 as we intend to investigate the effect of combining data from different platforms to complement for domains where the use of emoticons is not pervasive (e.g. Digg) or is generally scarce (e.g. MySpace). Furthermore, the computational cost of processing a large dataset is a concern in developing dynamic systems that usually iterate over some time intervals. As for the distant-supervised data from Digg (DsDigg) and MySpace (DsMySpace), we extract sentences that contain one or more emoticons of the same sentiment polarity (positive or negative) from Digg and MySpace respectively, using the collections harnessed by the CyberEmotions project[2]. Unlike Twitter which has a character limit, we confine the

[1] www.sentiment140.com
[2] www.cyberemotions.eu

TABLE 5.1: List of emoticons

Positive	Negative
:) ;) ;-) :-)	:(;(;-(:-(
:] ;] ;-] :-]	:[;[;-[:-[
:D =)	

labelling to sentences rather than documents. Such a sentence-level labelling is more intuitive since emoticons often apply only to the sentence in which they appear. This means that multiple micro-documents are generated from a single document ensuring each micro-document is labelled according to one or more emoticons belonging to the same sentiment class. With both collections of Digg and MySpace comments, there were many more positive (almost 80%) compared to negative emoticons present. It may be proper to allow this imbalance, if it is as a result of the natural class distributions in the datasets, however, we observed that the imbalance may not be due to the natural tendency of one class to occur more often than the other, but, due to the manner in which emoticons are used. Accordingly, we select balanced samples from the skewed distributions for the distant-supervised datasets. The main difference between the DsDigg and DsMySpace is in their size (DsDigg with 10,444 and MySpace with 604 documents).

Table 5.1 shows the list of emoticons used for the distant-supervised labelling. These emoticons are carefully selected to balance a trade-off between the reliability of the sentiment connotation of the emoticons and the size of the lists to improve recall of labelled documents. Still, to improve the recall we use regular expressions that ignore spaces in between the emoticon characters. All distance-supervised datasets are preprocessed to a reduced feature space using the approach introduced by Go et al. (2009), whereby, all user names (preceded by the @ symbol) are replaced with the token 'USERNAME' and URLs (e.g. "http://tinyurl.com/cvvg9a") are replaced with the token 'URL'. Moreover, words consisting of a sequence of three or more repeated character (e.g. "haaaaapy") are normalised to contain a maximum of two adjacent character repetition.

TABLE 5.2: Distant-supervised datasets

Dataset	Collection	Labelled documents (pos/neg)	Selected documents (pos/neg)	Example text
Twitter	-	800,000/800,000	10,000/10,000	*-What tragedy and disaster in the news this week :(* *-YAY ! found a new cuddle buddy*
Digg	1,646,153	21,214/5,222	5,222/5,222	*-Those bots are pretty bad :(* *-Glad you like it guys=)*
MySpace	2,867	2,824/302	302/302	*-am bored aswel:(* *-That's great because I love you too :)*

5.2 Domain-Specific Lexicon

The domain-specific lexicon associates a positive and a negative score to each unique term from the distant-supervised dataset. Crucial to this process is the pre-processing of documents to obtain their individual terms. This is particularly difficult with informal social media text. We use TweetNLP (Gimpel et al., 2011), a recently developed Twitter-oriented API for text tokenisation and part-of-speech tagging. Each word and its part-of-speech (i.e. a lexeme) forms an entry into the domain-specific lexicon. Although reducing words to their root form (stem) using standard stemming algorithms is often considered harmful for sentiment analysis (Potts, 2011*b*) (as words with completely different sentiment connotations can be mapped to the same stem), term variations prevalent in informal communications are also likely to have a negative impact on the task. For example, the term "cannot" may as well be written as "cant", "ca'nt" or "can't", thus resulting in an undesired variation for statistics purposes. Also, non-standard words can be written differently as such words may not have a specific generally accepted spelling. For instance, the words "argh", "arghh" and "arrgh" can be used deliberately interchangeably. Word shortening is also likely to be common in informal text, for example "exam" for "examination" or "fab" for "fabulous". We introduce a preprocessing step to address these problems. First, all uniquely identified lexemes that are candidate terms in the domain-specific lexicons are sorted alphabetically. Thereafter, if any two adjacent terms are known from a dictionary, both terms are retained.

Otherwise, if the difference between the two terms is:

- an apostrophe (‘) or a possessive form (‘s), the term with no apostrophe or possessive form is retained (e.g *would’nt* and *wouldnt* are merged in *wouldnt*)
- a repeated letter, the term with no repeating letter is retained (e.g. *arghh* and *argh* are merged in *argh*)
- that one is the substring of the other, the term with the full word is retained. This is meant to collapse variation such as between *exam* and *examination*. To avoid the undesired effect of merging words having different meanings a minimum overlap threshold is required. We set this threshold to 4 characters, after a preliminary investigation with different values
- that one is the plural form of the other, the term with the singular form is retained

In all these cases, the retained term takes the statistics from the two terms. This approach helps reduce unwanted term variability without the adverse effect of stemming.

5.2.1 Term-Sentiment Association

Key to the generation of the domain-specific lexicon is to capture association of a term t_i to a class c_j given a set of distant-supervised documents, D. Let D_{c_j} be the subset of D labelled as class c_j. Similarly, let the notation $\mathrm{TF}(t_i, X)$ represent term frequency of t_i in a set of documents X and $ds(t_i, c_j)$ be the domain score of association of t_i with c_j. We investigate three weighting metrics from which normalised sentiment polarity scores (for positive and negative classes) are computed and used to populate the domain-specific lexicon.

Supervised TF

Term frequency (TF) is the number of times a specific term appears in a document. It is a well-established quantifier of association between documents in many text analysis tasks (e.g. Information Retrieval). We propose supervised TF (sTF) to associate terms

with sentiment classes (positive and negative). Here, the association of term t_i with class c_j is measured as the ratio of frequency of t_i in documents labelled as class c_j to the total frequency of t_i in all documents (Equation 5.1). This produces association scores that are bound to [0,1] and sum to 1 for both positive and negative classes. Thus, the scores are directly compatible with SentiWordNet scores.

$$ds(t_i, c_j) = \frac{\text{TF}(t_i, D_{c_j})}{\text{TF}(t_i, D)} \tag{5.1}$$

Supervised TFIDF

TFIDF is the combination of term's TF with inverse document frequency (IDF). Originally designed for IR, IDF measures the popularity of a term across all documents. Terms that appear in many documents have less weight than terms that appear in a smaller number of documents. The TFIDF metric is designed to operate at the document level because, in IR, a document needs to be distinguished from all other documents by virtue of its relevance to a given query and ranking purposes. In contrast, it is the discriminative power between classes that is required of a metric for sentiment classification. Also, IDF does not incorporate class knowledge about documents as such information is unavailable in IR tasks. We propose supervised TFIDF (sTFIDF) to associate a score to a term given a sentiment class. In sTFIDF, IDF calculation is restricted to documents of the same class as shown in Equation 5.2.

$$ds(t_i, c_j) = \text{TF}(t_i, D_{c_j}) \times \log \frac{|D_{c_j}|}{|d \in D_{c_j} : t_i \in d|} \tag{5.2}$$

Where $|Dc_j|$ is the number of documents labelled c_j and $|d \in D_{c_j} : t_i \in d|$ is the number of documents in Dc_j that contain term t_i. Unlike sTF, terms are weighted by their distribution across documents within the target class in sTFIDF. Accordingly the strength of association of terms with class is reduced as they become more evenly distributed across classes.

Point-wise Mutual Information

Point-wise mutual information (PMI) can be used to associate terms with sentiment classes as follows:

$$ds(t_i, c_j) = PMI(t_i, c_j) = \log \frac{P(t_i, c_j)}{P(t_i) \times P(c_j)} \tag{5.3}$$

Where $P(t_i, c_j)$ is the probability of t_i and c_j, $P(t_i)$ is the propability of t_i and $P(c_j)$ is the probability of c_j. Calculating the probabilities from term frequencies, Equation 5.3 is re-written as follows (Mohammad et al., 2013).

$$ds(t_i, c_j) = \log \frac{\text{TF}(t_i, D_{c_j}) \times |T|}{\text{TF}(t_i, D) \times |T_{c_j}|} \tag{5.4}$$

Where $|T|$ and $|T_{cj}|$ are the number of terms in the corpus and in the documents of class c_j respectively. When a term does not occur in a class the association given by Equation 5.4 is deemed to be 0 avoiding the log(0). Similarly, negative associations are converted to 0, resulting in positive point-wise mutual information (pPMI) (Niwa and Nitta, 1994, Turney and Pantel, 2010). Unlike sTF or sTFIDF, PMI has a theoretical basis in probability theory and is arguably the most common approach to associating terms with sentiment scores (Mohammad et al., 2013, Turney, 2002, Turney and Pantel, 2010). However, it has a tendency to produce very low values for low frequency terms (Sani, 2014).

Table 5.3 shows the top ranking positive and negative terms from a Twitter domain-specific lexicon. Each term is disambiguated by its parts-of-speech that in the vocabulary of the distant-supervised dataset (_N for noun, _V for verb, _R for adverb, _J for adjective, and _O for other). Both sTF and pPMI have a similar ranking for terms. However, they differ in aggregate scores for terms as is evident from their formulae. For instance, the positive/negative scores of 'welcome_O' are 0.573/0.0 and 0.966/0.034 from pPMI and sTF respectively. Therefore, although both pPMI and sTF provide a similar ranking for terms, their document-level classification could be different, as it is the aggregate of

TABLE 5.3: Top ranking terms from Twitter domain-specific lexicons

sTF		sTFIDF		pPMI	
+	-	+	-	+	-
welcome_O	worst_O	good_O	sad_O	welcome_O	worst_O
thx_N	stomach_N	thank_N	ugh_O	thx_N	stomach_N
heyy_O	snowing_V	url_N	not_R	yum_O	snowing_V
yum_O	gah_O	great_O	sick_O	heyy_O	sad_O
smile_N	sad_O	love_V	no_O	vote_V	gah_O
vote_V	lonely_O	haha_O	work_N	smile_N	lonely_O
adorable_O	messed_V	nice_O	miss_V	adorable_O	messed_V
proud_O	earthquake_N	thank_V	why_R	luv_V	earthquake_N
luv_V	shitty_O	happy_O	hate_V	proud_O	shitty_O
interested_O	sandra_O	awesome_O	sorry_O	yah_O	sandra_O

term level scores. In contrast, sTFIDF gives a different set of top ranking terms which are more akin to the standard vocabulary. This is because standard terms are likely to have more distribution over documents of one class compared to their distribution over the whole corpus.

5.3 Static Lexicon

We use SentiWordNet (Baccianella et al., 2010) as the static lexicon, from which generic sentiment scores are obtained for terms. Given a tokenised term with its part-of-speech (PoS) tag, sentiment scores (positive and negative) are retrieved from the lexicon as a weighted average of scores attached to all word senses of the term as follows:

$$gs(t_i, c_j) = \frac{\sum_{k=1}^{|sense(t|PoS)|} \frac{1}{i} \times score(t|PoS, sense_k)_{c_j}}{|sense(t|PoS)|} \tag{5.5}$$

Where $gs(t_i, c_j)$ is the general-purpose score of term t_i with the sentiment class of c_j (c_j is either positive or negative) and $score(t_i|PoS, sense_i)_{c_j}$ is the sentiment score of the term t_i given the part-of-speech (PoS) at sense k for the sentiment class c_j. Finally, $|sense(t|PoS)|$ is the number of word senses for the given part-of-speech (PoS) of term t_i.

5.4 Hybrid Lexicon Generation

Scores from static, S, and domain-specific, D, lexicons for each term t_i are combined to form the hybrid score for the term (see Algorithm 4). When t_i appears in both lexicons, a weighted average of the positive and the negative scores supplied by both lexicons is calculated using α and β as mixing parameters for positive and negative scores respectively. This weighting favours scores from one lexicon over the other. So $\alpha = 0.5$ would lead to equal weighting of positive scores from S and D whilst $\alpha = 0$ will ignore positive score from SentiWordNet lexicon (see steps 3 and 4). The use of different mixing parameters is likely to address possible bias towards a sentiment dimension (usually positive) due to the observation that people tend to use positive terms in a more frequent and diverse manner (Pollyanna hypothesis) (Boucher and Osgood, 1969). We determine optimal values for the mixing parameters, α and β as the combination that produces the highest performance on an optimisation dataset. We envisage as this optimisation dataset is relatively small in size, it is typically available as part of test data.

When only one lexicon (SentiWordNet or domain-specific) contains scores for t_i, such scores are fully used without an aggregation (see steps 6 and 8). Thereafter, the new scores for t_i (i.e. t_i^+ and t_i^-) are added to the hybrid lexicon, H (step 11). Finally, H is returned as the output.

5.4.1 Preliminary Insight: Difference in Coverage and Polarities

We conduct a preliminary study to gain insight into the variability between static and domain lexicons in vocabulary coverage and sentiment polarities of terms. We use Twitter data for this study (distant-supervised, for domain-specific lexicon generation and human-labelled, for test). We use this data as it is the largest, thus, we can experiment with its small and large subsets. Figure 5.2 shows the distribution of unique terms (vocabulary) from the static and domain-specific lexicons. As expected, with small data sizes (horizontal axis), the domain-specific lexicon has a very limited vocabulary (vertical axis). Therefore, static lexicon makes the most contribution in vocabulary (for hybrid

Algorithm 4 Generate Hybrid Lexicon

INPUT: S, Static lexicon
D, domain-specific Lexicon
α, β Mixing parameters
OUTPUT: H, Hybrid lexicon

for all $t_i \in (S \cup D)$ **do**
  if $t_i \in S \cap D$ **then**
    $t_i^+ \leftarrow \alpha \times (t_i^+ \in S) + (1 - \alpha) \times (t_i^+ \in D)$
    $t_i^- \leftarrow \beta \times (t_i^- \in S) + (1 - \beta) \times (t_i^- \in D)$
  else if $t_i \in S$ **then**
    $t_i^+ \leftarrow (t_i^+ \in S)$
    $t_i^- \leftarrow (t_i^- \in S)$
  else
    $t_i^+ \leftarrow (t_i^+ \in D)$
    $t_i^- \leftarrow (t_i^- \in D)$
  end if
  $H.AddEntry(t_i^+, t_i^-)$
end for
Return H

lexicon) with smaller domain data. The vocabulary intersection increases with increase in domain data size. However, still there is a considerable difference in vocabulary coverage between the two lexicons even with larger dataset sizes (e.g. 20000). With regard to sentiment polarities for terms, both lexicons tend to agree on the same polarity more than they differ (see Figure 5.3). However, there is also a considerable difference, which tends to be independent of dataset size, as shown by the figure.

These differences in vocabulary coverage and sentiment polarities of terms suggest that each individual lexicon is lacking in vocabulary and polarity representations. The static lexicon is more likely to capture general sentiment knowledge that may not become available to the domain-specific lexicon while the domain specific lexicon is more likely to capture new sentiment knowledge evolving in social media. Our hybrid lexicon approach harnesses the strengths from both lexicons for potential improvement of sentiment classification accuracy.

5.4.2 Transferability Across Social Media Platforms

The occurrence of (particularly negative) emoticons is not very common on some social media platforms. For instance, out of about 1.6 million discussion posts from *Digg.com*

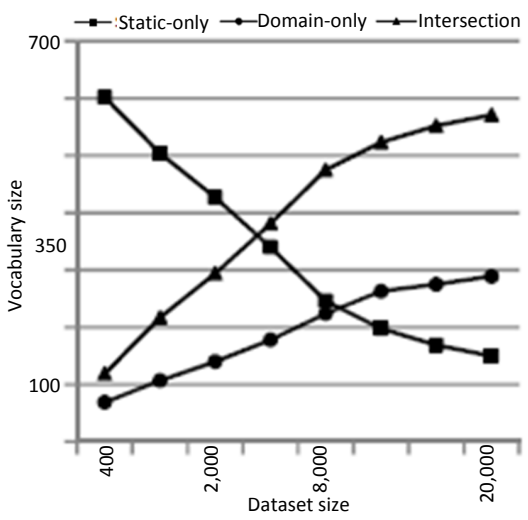

FIGURE 5.2: Lexicon Coverage

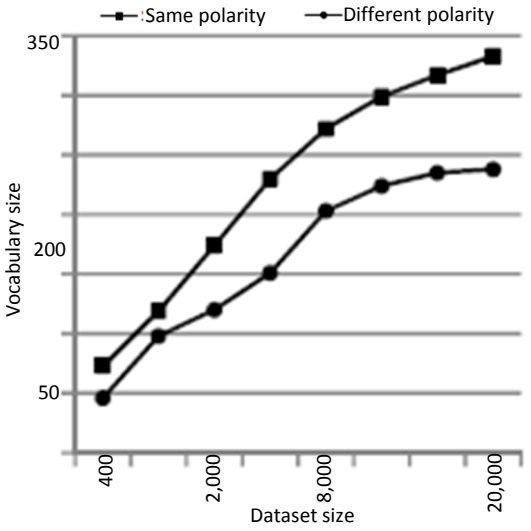

FIGURE 5.3: Polarity Difference

only about 5,000 have a negative emoticon (see Table 5.2). This is despite the fact that the posts are extracted from topics that are likely to be rich in sentiment. Therefore, it is imperative to investigate whether distant-supervised data obtained from one social media platform could be used to generate a hybrid lexicon for another platform. This falls within the realm of transfer learning.

In machine learning, transfer learning involves learning a model from training data obtained from one domain, adapting and testing the model on another domain. This is based on the following two assumptions:

- labelled training data, which is a magnitude larger than test data, is available or easily obtained from one domain (in-domain) but is unavailable or difficult to obtain from another domain (out-of-domain)
- test data is available from both domains

These, therefore, suit our problem at hand in that the in-domain is the social media platform which has an abundance of distant-supervised data and on which we can generate the domain-specific lexicon. The out-of-domain will be the platform with little or no distant-supervised data and on which we wish to generate a hybrid lexicon and subsequently perform sentiment classification. This is illustrated in Figure 5.4. The difference between learning the hybrid lexicon as described earlier and the transfer learning

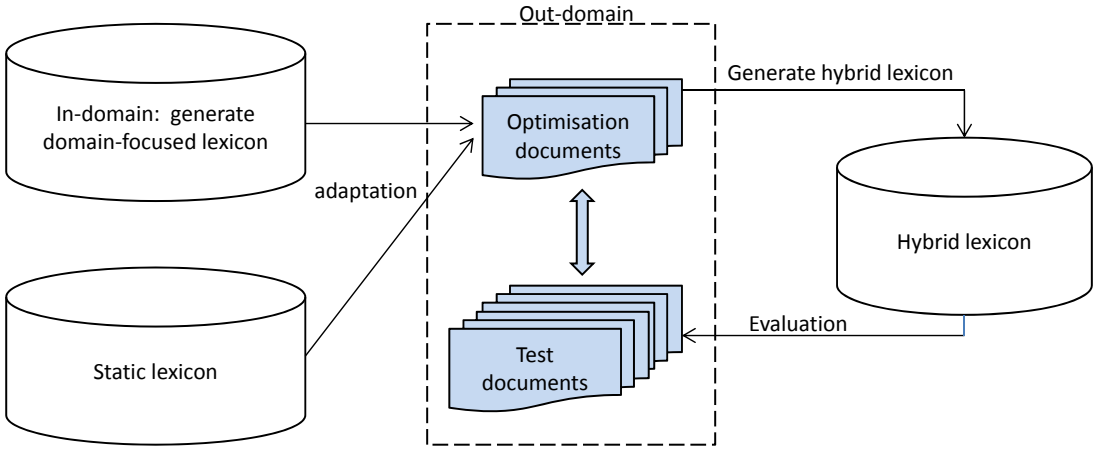

FIGURE 5.4: Transfer learning hybrid lexicon

of the lexicon is in the use of out-of-domain data to adapt the combination of static and domain-specific lexicons and the evaluation of the hybrid lexicon on the out-of-domain test data. Transfer learning has been extensively studied in various NLP tasks including sentiment analysis. For instance, in Blitzer et al. (2006), a transfer learning framework has been proposed, which identifies features that have high mutual information with polarity labels (i.e. pivot features). Thereafter, the pivot features are connected to domain-specific words to guide the transfer learning. In Daume III and Marcu (2006), a maximum entropy genre adaptation model (MEGA) was proposed, motivated by the notion that the distribution of test data may not be identical with that of the training data in some applications. MEGA is a simple mixture model with a hidden variable that indicates whether the data is drawn from the in-domain distribution, the out-of-domain distribution, or the general domain distribution. Also, in Yoshida et al. (2011), transfer learning was performed using a Bayesian probabilistic model that handles multiple sources and multiple target domains. Here, each word is associated with three characteristics, indicating the domain in which it is extracted, whether its polarity is domain dependent, and its polarity label. In our work, we achieved transfer learning of a hybrid lexicon by utilising very limited optimisation data from the target domain and the sentiment information from the general-purpose lexicon and the domain-specific lexicon.

5.5 Chapter Summary

This chapter presents a novel approach to generating a hybrid lexicon by combining a domain generated lexicon and a static lexicon using a weighted strategy. We demonstrated how distant supervision can be exploited for this purpose. Also, to address the drawback of PMI applied to low frequency terms, we introduced new sentiment scoring metrics inspired by the term frequency and inverse document frequency. We also showed how transfer learning can be exploited in generating a hybrid lexicon for domains that have scarce distant-supervised data.

The evaluation of the hybrid lexicon approach is presented in Chapter 7. It involves testing the main hypothesis of this chapter, that is, performance in sentiment classification improves with a hybrid lexicon compared to either a domain-specific or a general purpose lexicon. Also, as distant supervision has until now been used for machine learning methods to sentiment classification, we compare our lexicon-based method with machine learning classifiers. We also study the effect of transfer learning for a hybrid lexicon generation.

Chapter 6

Leveraging Local, Domain and Emotion Features for Sentiment Classification

In this chapter, we introduce a hybrid sentiment classifier that exploits local contextual analysis (introduced in Chapter 4) and domain semantics captured by a hybrid lexicon (introduced in Chapter 5). Although sentiment analysis and emotion detection are inter-related fields, research in sentiment analysis has typically ignored resources from emotion detection or has assumed certain emotion classes are equivalent to sentiment classes (Ghazi et al., 2010, Gonçalves et al., 2013, Poria et al., 2014). In our hybrid classifier, we introduce a novel strategy for utilising knowledge from an emotion lexicon for sentiment classification. Since emotion and sentiment are different by definition (Munezero et al., 2014), we do not collapse emotion classes into sentiment classes, thus, we are able to explicitly demonstrate the contribution of emotion detection for sentiment analysis. We present this hybrid classifier next.

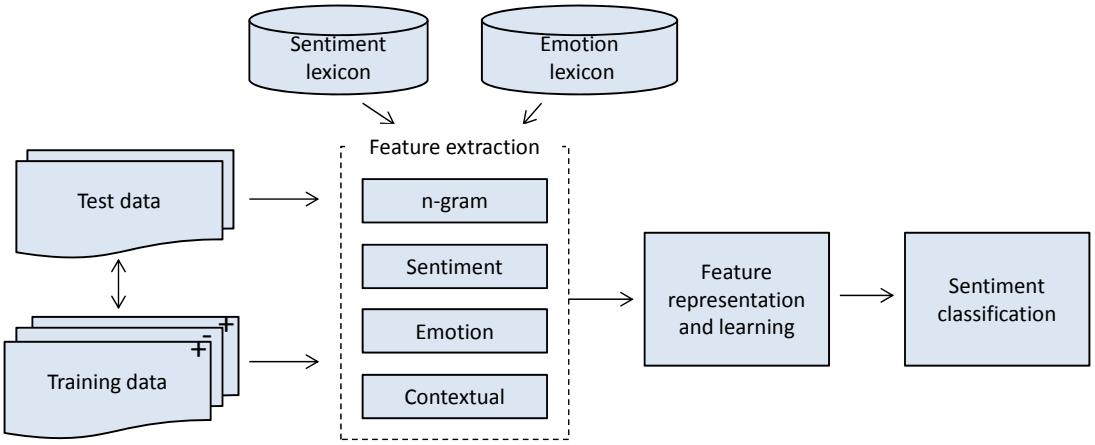

FIGURE 6.1: The Supervised Classifier

6.1 The Hybrid Classifier

The classifier ϕ is trained on a collection of training documents $D = \{d_1, d_2, ..., d_N\}$ where each document $d_j \in D$ is associated with a class label (positive or negative). The documents D are represented on feature sets $F = \{f_1, f_2, ..., f_K\}$ extracted from the documents' vocabulary, sentiment and emotion lexicons. Thus, given a new document d_q with an unknown class, the classifier ϕ is applied to the document to determine its sentiment class. Figure 6.1 shows the architecture of the classifier. It comprises of feature sets grouped as n-gram, sentiment, emotion and contextual features.

6.1.1 n-gram Features

An n-gram is a contiguous sequence of n tokens from a given piece of text. Typically, n-grams are the basic features used in supervised sentiment classification. We extract 1-, 2- and 3-gram from training documents as n-gram features for representation after a pre-processing step similar to that discussed in Section 5.2. That is, we use TweetNLP for tokenisation and part-of-speech tagging after which we apply lemmatization and social media oriented feature merging rules. We then use a binary-valued representation for the n-gram features.

6.1.2 Sentiment Features

As we mentioned in the literature review section, previous research shows that lexicon-based features improve sentiment classification. However, previously explored lexicons were mainly domain-independent. Here we explore the benefit of using the hybrid lexicon introduced in Chapter 5 which adapts its vocabulary to social media domains. We consider the following feature sets for which we extract values from the hybrid lexicon:

1. *Total_sentiment_score*: This is the sum of the sentiment scores for all the terms contained in a document. We calculate this value for each polarity class $c \in \{positive, negative\}$.

$$Total\_score(d)_c = \sum_{t \in d} score(t)_c \tag{6.1}$$

2. *Max_score*: This is the score of the highest sentiment-bearing term in the given document. We determine the score for each polarity class as follows:

$$Max\_score(d)_c = \max_{t \in d} \big(Score(t)_c\big) \tag{6.2}$$

3. *Total_sentiment_count*: This is the number of terms from a document that have dominant polarity of a particular sentiment class $c \in \{positive, negative\}$. We determine the dominant polarity as the sentiment dimension having the maximum score as shown in Equation 6.3. Thus, this feature also has values for both positive and negative polarities.

$$Count(d)_c = |\{t \in d : Score(t)_c > Score(t)_{\bar{c}}\}| \tag{6.3}$$

   Where $\bar{c}$ is the opposite class from c.

4. *Graded_score*: The occurrence of high sentiment-bearing terms is indicative of sentiment class of the document regardless of the average score for the document Thelwall et al. (2012, 2010). In this feature (and subsequent related features), we aim to capture the influence of high sentiment-bearing terms for classification. We

graded the polarity spectrum of the hybrid lexicon into strong negative (-1 to -0.5), negative (-0.49 to -0.01), positive (0.01 to 0.49) and strong positive (0.5 to 1). Then, for each term in a given document, we calculate its overall sentiment score (difference between positive and negative scores) and add to the respective grade of the term as shown in Equation 6.4. Thus we have four values per document for this feature.

$$Graded\_score(d)_{(a,b)} = \sum_{t \in ds} Score\_diff(t) \tag{6.4}$$

Where $ds = t \in d : a \leq Score\_diff(t) \leq b$; a and b are the lower and upper bounds of a score interval; and $Score\_diff$ is the difference in positive and negative score of the term t given by $Score(t)_{positive} - Score(t)_{negative}$.

5. *Graded_count*: This is similar to the previous feature except that in this case we count the number of terms belonging to a polarity grade instead of sum, as shown in Equation 6.5

$$Graded\_count(d)_{(a,b)} = |\{t \in d : a \leq Score\_diff(t) \leq b\}| \tag{6.5}$$

6. *PoS_score*: This is the total sentiment score for each part-of-speech $PoS \in \{noun, verb, adjective, adverb, other\}$. It is aimed at capturing the relative importance of parts-of-speech in sentiment expression. We calculate the scores for each polarity dimension (positive and negative), thus, we have ten values per document for this feature.

$$PoS\_score(d)_{c,PoS} = \sum_{t \in dp} Score(t) \tag{6.6}$$

Where $dp = t \in d : Pos(t) = PoS$.

6.1.3 Emotion Features

The field of emotion analysis from text concerns the detection of emotive text and the corresponding emotion class. Several emotion classes have been proposed in the literature including the Parrott's emotion taxonomy which comprises of six basic emotion classes: *love*, *joy*, *surprise*, *sadness*, *anger*, and *fear* (Parrott, 2001). These emotion

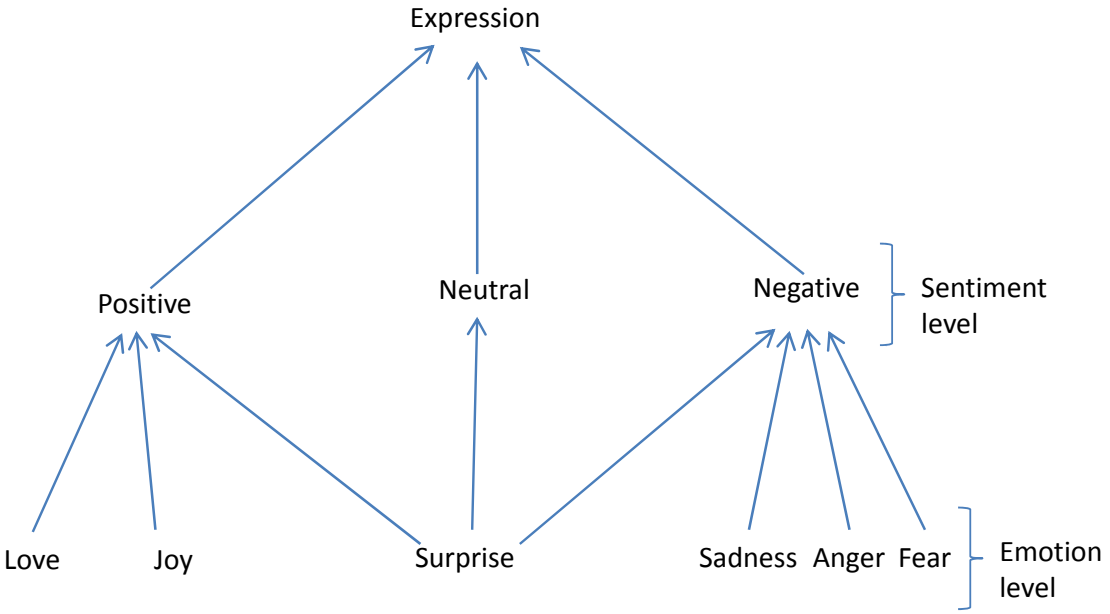

FIGURE 6.2: Typical Emotion-to-Sentiment Relationship

classes can be detected from text with the use of emotion lexicons (unsupervised setting) or a training dataset (supervised setting). An obvious relation between sentiment and emotion is that emotion classes can be mapped onto sentiment classes (Gonçalves et al., 2013). For instance, love and joy correspond to the positive sentiment while sadness, anger and fear correspond to negative sentiment as illustrated in Figure 6.2. However, the emotion class of 'surprise' is ambiguous and typically does not exclusively map to a particular sentiment class. Therefore, emotion knowledge as cannot be completely mapped to sentiment knowledge. This is one of the reasons why we introduce emotion features different from sentiment features. Also, as emotion is more fine-grained than sentiment, there is the potential that the details offered by emotion classes will help in a more accurate sentiment detection. Therefore, in this work, we do not map emotion classes to sentiment but use them as additional features.

Our objective is to leverage emotion knowledge from lexicons for sentiment prediction. To this end, we adopt features similar to those used to represent sentiment knowledge (the previous sub-section) and extract values for the emotion classes. Although there are a number of different emotion schemes (as highlighted in Section 2.2.2.), here, we use the Parrott (2001)'s scheme. Our choice of this scheme is motivated by three reasons. Firstly, because it provides a more balanced sets of positive and negative emotions.

For instance, out of the six Parrott (2001) emotion classes, two are positive, three are negative, and one is ambiguous; while out the of the six Ekman (1999) emotion classes, only one is positive, four are negative, and one is ambiguous. Secondly, Parrott (2001) scheme has been argued to be more appropriate for the social media text because of its inclusion of the 'love' emotion class, which is common in social media, and is not present in the Ekman (1999) set of emotions (Bandhakavi et al., 2014). Thirdly, because there exists a twitter-oriented emotion lexicon based on the Parrott (2001) scheme, which has been shown to produce a state-of-the-art performance (Bandhakavi et al., 2014), and which we can conveniently re-use in this work.

6.1.4 Contextual Features

We introduce the following feature sets to integrate local context into the classifier. These include features that capture word-based sentiment modification (lexical valence shifters) and modification based on the use of social media oriented symbols (non-lexical modifiers).

6.1.4.1 Lexical Valence Shifters

These are modifiers based on the explicit use of standard sentiment-modifying terms (negation, intensifiers, diminishers and discourse markers). Following the in-depth analysis we conducted about these modifiers in Chapter 4, here we derive the following feature sets to capture the influence of such modifiers

1. *Negation*: This feature records the number of times negation occur in the given document. Negation also affects the n-gram features: a term t becomes t_NEG in a negated context.

2. *Intensifiers*: The number of times intensification occurs in the given document.

3. *Diminishers*: The number of times diminishers occur in the given document.

4. *Discourse*: The number of times each of the discourse relations introduced in Section 4.2.1.3 occurs in the given document. Again, we group the discourse relations according to their effect, with respect to sentiment, on nucleus and satellite segments.

6.1.4.2 Non-Lexical Valence Shifters

Similar to the lexical valence shifters, non-lexical valence shifters are often used to modify sentiment in social media (e.g. by capitalisation or repeating a letter/character). They are also used directly to express sentiment (e.g. emoticons). We derive the following feature sets based on the non-lexical valence shifters for integration into the hybrid classifier:

1. *Capitalisation*: Here we record the number of terms that have all their characters in uppercase. Where all the terms in the document are in capital letters, we set the value for this feature to zero, as in such situation the capitalisation is unlikely for emphasis.

2. *Repeat letter*: This is the number of elongated words by repeating a letter (e.g. haaaaaappy).

3. *punctuation*: In this feature set we record the number of contiguous sequences of two or more exclamation marks or question marks or combination of both exclamation and question marks (e.g. !!!, ???, !?!).

4. *emoticon*: In this feature set, we record the number of positive and negative emoticons from a given document. We determine sentiment class (positive or negative) of emoticons using an emoticon lexicon (Thelwall et al., 2012). Also, we introduce a feature that captures whether the last token in the document is a positive or negative emoticon.

5. *hashtag*: This records the number of hashtags from the given document.

6.2 Chapter Summary

The main hypotheses presented are that the lexicon-based strategies introduced in previous chapters can improve a hybrid method (combining supervised learning and lexicon-based knowledge) to sentiment classification. Similarly, emotion knowledge can improve classification accuracy. To investigate these hypotheses, we proposed feature sets to be extracted from training data, local context analysis as well as from the hybrid and emotion lexicons. These are integrated into a hybrid sentiment classifier.

Evaluation of the hybrid classifier is discussed in Chapter 7. It involves testing the hypotheses of this chapter using distant-supervised datasets for training and human-labelled datasets for testing.

Chapter 7

Evaluations

In this chapter we present evaluations of our sentiment classification strategies discussed in Chapters 4, 5 and 6. These include the evaluation of SmartSA and the contribution of each strategy integrated into the system. We also study the performance of the hybrid lexicon approach in comparison to the static or the domain-specific lexicon and compare the transferability of a hybrid lexicon from one social media platform to another. Finally, we investigate the performance of the hybrid sentiment classifier that exploits features extracted from local contextual analysis, sentiment and emotion lexicons.

7.1 Evaluation of SmartSA and Related Strategies

The aim of this evaluation is to study the performance of sentiment analysis strategies introduced in Chapter 4. First, we conduct an experiment to ascertain the performance of our WSD approach in comparison to the existing approaches to score extraction from SentiWordNet. Next, we investigate the performance of each score adjustment strategy proposed for SmartSA as well as the overall performance of the system in comparison with the baseline and the state-of-the-art systems. These experiments are designed to provide evidence towards addressing our first research question: *Does the accuracy of lexicon-based sentiment analysis benefit from the integration of local context knowledge?* Accordingly, we investigate the following classifier settings:

- Score extraction strategies: These are the approaches of Most Frequent Sense (MFWS), Word Sense Disambiguation (WSD), Averaging at Word Sense level (AWS), Weighted Averaging at Word Sense (WAWS) and averaging at Part-of-speech level (APoS). Details of these strategies are discussed in Section 4.1.

- Base: The baseline lexicon-based sentiment classification approach (see Algorithm 1)

- Contextual score adjustment strategies: These are the lexical strategies of switch negation (Switch), shift negation (Shift), intensification/diminishing (IntDim) and discourse markers (Disc). With all these approaches, sentiment scores of the involved modifiers can be included into the aggregation process. Also, we investigate the performance of non-lexical strategies: capitalisation (Caps), repeated letter, exclamation or question mark (Rp) and appearance of emoticons (Em).

- SmartSA: Our sentiment classification algorithm that integrates contextual score adjustment strategies (Algorithm 3).

- SentiStrength: a state-of-the-art, lexicon-based sentiment classifier designed for the social media text (Thelwall et al., 2012). The system takes a piece of text as input, assesses its sentiment content, and produces a number of outputs; including the binary (positive or negative) class, trinary (positive, negative, or neutral) class, dual (both positive and negative scores) for the text. It uses a sentiment lexicon derived from the Linguistic Inquiry and Word Count (LIWC) software (Pennebaker et al., 2007), and extended with social media slang such as *lol*, *lolol* and *lmao*. Also, it uses an additional manually created lexicon of emoticons, and performs a number of contextual adjustments of prior polarities based on both the lexical and non-lexical modifiers. An extensive evaluation on social media datasets shows the system to produce state-of-the-art performance, better than many existing approaches; hence, our decision to use this system as a state-of-the-art classifier in our evaluations.

The differences between our system and SentiStrength are, first, in the lexicons used by the systems. While SentiStrength uses a relatively smaller lexicon derived from LIWC, we use a lexicon with high term coverage (SentiWordNet). Second, unlike in SentiStrength, we use sentiment scores of negation in SmartSA, and incorporate a strategy for discourse analysis. Finally, instead of the manual addition of social media oriented terms in a lexicon, as is the case with SentiStrength; with a hybrid lexicon, we introduced an automated approach to extending a lexicon with social media terms in SmartSA. SentiStrength can be used with a number of different settings. In our evaluations, we used its default setting, as this is the setting used in the evaluation of the system (Thelwall et al., 2012).

7.1.1 Results of Score Extraction Strategies

As expected, WSD performed better than the rest on 3 (Digg, RunnerW and Youtube) out of the 6 datasets as shown in Table 7.1. These datasets have the highest document sizes (as shown in Table7.4), and so, are likely to provide sufficient context for effective word sense disambiguation. Also, it is noteworthy that these datasets contain the least proportion of non-lexical valence shifters - an indication of being relatively more formal/standard. This might have helped obtain more overlap between term context and dictionary glosses which might have influenced the better performance of WSD. However, it can be noted that even on these 3 datasets, the WAWS performed only marginally worse than WSD. Nevertheless, the results show that despite the challenges of social media data, the proposed WSD approach is quite competitive with the existing approaches. The WAWS consistently outperformed AWS on all datasets. This shows the importance of using sense order (and by extension, sense frequency) as weights for averaging scores. This is further demonstrated by the performance of MFWS which, in some datasets (MySpace and LiveJ), performs better than WSD.

We also observed that although averaging at word sense level resulted in better performance, such an approach applied at the PoS level (APoS) resulted in the worst

TABLE 7.1: Results from score extraction strategies on test datasets

Algorithm	Positive			Negative			Avg F1
	P	R	F1	P	R	F1	
Digg							
WAWS	37.44	75.24	50.00	85.56	53.85	66.10	58.05
WSD	38.65	76.19	51.28	86.41	55.59	67.66	**59.47**
AWS	35.65	70.95	47.46	83.24	52.97	64.74	56.10
MFWS	36.34	72.86	48.49	84.21	53.15	65.17	56.83
APoS	33.97	68.10	45.33	81.44	51.40	63.02	54.18
RunnersW							
WAWS	81.60	76.03	78.72	54.33	62.44	58.10	68.41
WSD	82.00	78.10	80.00	56.56	62.44	59.35	**69.68**
AWS	81.43	75.21	78.20	53.49	62.44	57.62	67.91
MFWS	81.26	74.38	77.67	52.67	62.44	57.14	67.41
APoS	78.03	71.90	74.84	47.49	55.66	51.25	63.05
Youtube							
WAWS	79.33	86.91	82.95	64.14	50.85	56.73	69.84
WSD	79.77	86.91	83.19	64.72	52.15	57.76	**70.48**
AWS	78.72	86.01	82.20	61.99	49.54	55.07	68.64
MFWS	77.86	84.50	81.04	58.72	47.85	52.73	66.89
APoS	76.50	82.10	79.20	53.80	45.24	49.15	64.18
MySpace							
WAWS	88.67	82.48	85.46	32.04	43.94	37.06	**61.26**
WSD	86.94	79.63	83.12	25.13	36.36	29.72	56.42
AWS	88.44	81.58	84.87	29.26	41.67	34.38	59.63
MFWS	87.00	80.06	83.39	25.53	36.36	30.00	56.70
APoS	86.32	78.21	82.07	22.73	34.09	27.27	54.67
LiveJ							
WAWS	81.95	76.58	79.17	69.88	76.32	72.96	**76.07**
WSD	80.65	70.26	75.10	64.62	76.32	69.98	72.54
AWS	81.58	72.60	76.83	66.67	76.97	71.45	74.14
MFWS	80.84	72.13	76.24	66.00	75.99	70.64	73.44
APoS	75.91	68.62	72.08	61.16	69.41	65.02	68.55
Twitter							
WAWS	84.90	75.88	80.14	44.19	58.60	50.38	**65.26**
WSD	84.04	75.11	79.32	42.40	56.23	48.35	63.84
AWS	84.47	75.49	79.73	43.29	57.41	49.36	64.55
MFWS	83.17	74.33	78.50	40.61	53.86	46.31	62.41
APoS	83.61	74.72	78.92	41.50	55.04	47.32	63.12

performance. This finding seems to reflects the higher ambiguity in the APoS setting compared to the rest which further supports the usefulness of sense disambiguation.

7.1.2 Results of Score Adjustment Strategies

Tables 7.2 and 7.3 show results of sentiment classification using various score adjustment strategies introduced in this research. Bold font indicates the best performance on a dataset under categories of lexical and non-lexical valence shifters respectively, as well as between SmartSA and SentiStrength. Asterisk (*) indicates significant difference from the Base.

All the proposed strategies improve the Base classification except Switch and Shift. These are the only cases where sentiment scores attached to negation terms are not included in the aggregation process. However when scores are included (Switch+scores and Shift+scores), performance increases above the Base. This clearly shows that negation terms are sentiment-bearing terms as well as sentiment modifiers of other terms. This finding supports the findings in Potts (2011*a*) where negation terms were found to align with negatively labelled documents. A different result pattern is seen for score adjustment based on intensifiers/diminishers and discourse markers. Here scores associated with the markers do not contribute to better classification as IntDim and Disc performed better than IntDim+scores and Disc+scores in most of the datasets. Therefore, although sentiment scores for intensifiers/diminshers and discourse markers might be obtained from SentiWordNet, such terms tend to function more as modifiers than as bearing sentiment of their own.

Score adjustment based on negation provides the most improvement on a majority of the datasets (4 out 6) with larger margin on lengthier datasets, 4.32% on Digg and 3.76% on RunnersW. It can also be observed, from the datasets statistics, that these datasets have the highest proportion of negation terms even though one is composed mostly of positive documents and the other, mostly of negative. This shows that in addition to having a correlation with negative documents (Potts, 2011*a*), negation is also a characteristic of lengthy documents. IntDim gives more performance improvement than Disc on a majority of the datasets (5 out of 6) even though the occurrence of discourse

TABLE 7.2: Results from SMARTSA and related strategies on test datasets

Algorithm	Positive			Negative			Avg F1
	P	R	F1	P	R	F1	
Digg							
BASE	37.44	75.24	50.00	85.56	53.85	66.10	58.05
SWITCH	35.37	80.00	49.05	86.32	46.33	60.30	54.68
SHIFT	35.88	80.48	49.63	86.82	47.20	61.15	55.39
SWITCH+scores	40.24	79.52	53.44	88.28	56.64	69.01	61.23
SHIFT+scores	41.26	80.95	54.66	89.19	57.69	70.06	**62.36**
INTDIM	40.27	84.76	54.60	90.59	53.85	67.55	61.08
INTDIM+scores	38.28	76.19	50.96	86.26	54.90	67.10	59.03
DISC	40.79	83.33	54.77	90.08	55.59	68.75	61.76
DISC+scores	40.32	83.33	54.34	89.94	54.72	68.04	61.19
CAPS	38.17	77.62	51.17	86.76	53.85	66.45	58.81
RP	37.44	75.24	50.00	85.56	53.85	66.10	58.05
EM	39.09	77.62	51.99	86.94	55.20	67.53	**59.76**
SMARTSA	43.00	83.33	56.73	90.67	59.44	71.81	64.27*
SENTISTRENGTH	45.60	81.90	58.68	90.60	64.20	75.15	**66.87***
LiveJ							
BASE	81.95	76.58	79.17	69.88	76.32	72.96	76.07
SWITCH	77.18	76.81	76.99	67.65	68.09	67.87	72.43
SHIFT	78.10	76.81	77.45	68.17	69.74	68.95	73.20
SWITCH+scores	82.00	76.81	79.32	70.09	76.32	73.07	76.20
SHIFT+scores	82.41	76.81	79.51	70.27	76.97	73.47	**76.49**
INTDIM	82.04	77.05	79.47	70.30	76.32	73.19	76.33
INTDIM+scores	81.80	76.81	79.23	70.00	75.99	72.87	76.05
DISC	81.80	76.81	79.23	70.00	75.99	72.87	76.05
DISC+scores	81.80	76.81	79.23	70.00	75.99	72.87	76.05
CAPS	81.95	76.58	79.17	69.88	76.32	72.96	76.07
RP	81.95	76.58	79.17	69.88	76.32	72.96	76.07
EM	82.29	77.28	79.71	70.61	76.64	73.50	**76.61**
SMARTSA	82.50	77.28	79.80	70.69	76.97	73.70	**76.75**
SENTISTRENGTH	73.10	93.70	82.13	85.30	51.60	64.30	73.33
RunnersW							
BASE	81.60	76.03	78.72	54.33	62.44	58.10	68.41
SWITCH	78.79	83.68	81.16	58.64	50.68	54.37	67.77
SHIFT	79.26	83.68	81.41	59.28	52.04	55.42	68.42
SWITCH+scores	79.56	82.85	81.17	58.71	53.39	55.92	68.55
SHIFT+scores	82.15	83.68	82.91	62.74	60.18	61.43	**72.17**
INTDIM	82.17	78.10	80.08	56.73	62.90	59.66	69.87
INTDIM+scores	81.68	76.45	78.98	54.76	62.44	58.35	68.67
DISC	81.76	76.86	79.23	55.20	62.44	58.60	68.92
DISC+scores	81.60	76.03	78.72	54.33	62.44	58.10	68.41
CAPS	81.58	76.86	79.15	55.02	61.99	58.30	68.73
RP	81.64	76.24	78.85	54.55	62.44	58.23	68.54
EM	82.20	77.27	79.66	56.00	63.35	59.45	**69.56**
SMARTSA	83.06	84.09	83.57	64.19	62.44	63.30	**73.44***
SENTISTRENGTH	81.00	73.80	77.23	51.90	62.00	56.50	66.87

Table 7.3: Results from SmartSA and related strategies on test datasets

Algorithm	Positive			Negative			Avg F1
	P	R	F1	P	R	F1	
Twitter							
Base	84.90	75.88	80.14	44.19	58.60	50.38	65.26
Switch	83.26	78.43	80.77	43.81	51.60	47.39	64.08
Shift	83.22	78.24	80.65	43.59	51.60	47.26	63.96
Switch+scores	85.15	76.50	80.59	45.03	59.07	51.10	65.85
Shift+scores	86.02	78.24	81.95	47.73	60.97	53.54	**67.75**
IntDim	85.59	77.85	81.54	46.80	59.79	52.50	67.02
IntDim+scores	85.13	76.11	80.37	44.67	59.19	50.92	65.65
Disc	85.42	77.70	81.38	46.43	59.31	52.09	66.74
Disc+scores	84.94	76.07	80.26	44.38	58.60	50.51	65.39
Caps	85.52	77.85	81.50	46.70	59.55	52.35	66.93
Rp	85.04	76.27	80.42	44.68	58.84	50.79	65.61
Em	86.11	78.82	82.30	48.40	60.97	53.96	**68.13***
SmartSA	87.93	80.29	83.94	52.25	66.19	58.40	**71.17***
SentiStrength	86.20	84.20	85.19	54.70	58.60	56.58	70.87*
MySpace							
Base	88.67	82.48	85.46	32.04	43.94	37.06	61.26
Switch	87.35	82.62	84.92	28.24	36.36	31.79	58.36
Shift	87.56	83.19	85.32	29.34	37.12	32.77	59.05
Switch+scores	88.15	82.62	85.30	30.68	40.91	35.06	60.18
Shift+scores	89.47	84.88	87.11	37.28	47.37	41.72	**64.42**
IntDim	89.28	83.05	86.05	34.25	46.97	39.61	62.83
IntDim+scores	88.96	82.62	85.67	32.97	45.45	38.22	61.95
Disc	88.99	82.91	85.84	33.33	45.45	38.46	62.15
Disc+scores	88.69	82.62	85.55	32.22	43.94	37.18	61.37
Caps	88.75	83.19	85.88	32.95	43.94	37.66	61.77
Rp	88.72	82.91	85.72	32.58	43.94	37.42	61.57
Em	89.31	83.33	86.22	34.64	46.97	39.87	**63.05**
SmartSA	89.31	83.33	86.22	35.00	47.37	40.26	63.24
SentiStrength	91.80	90.50	91.15	52.80	56.80	54.73	**72.94***
YouTube							
Base	79.33	86.91	82.95	64.14	50.85	56.73	69.84
Switch	76.94	88.77	82.43	63.41	42.24	50.70	66.57
Shift	77.14	88.77	82.55	63.76	42.89	51.28	66.92
Switch+scores	79.53	88.65	83.84	67.19	50.46	57.64	70.74
Shift+scores	79.82	88.83	84.08	67.88	51.24	58.40	**71.24**
IntDim	79.66	89.37	84.24	68.62	50.46	58.16	71.20
IntDim+scores	79.37	89.43	84.10	63.64	44.32	52.25	68.18
Disc	79.60	89.07	84.07	68.01	50.46	57.94	71.01
Disc+scores	79.22	86.79	82.83	63.88	50.65	56.50	69.67
Caps	79.33	86.91	82.95	64.14	50.85	56.73	69.84
Rp	79.33	86.91	82.95	64.14	50.85	56.73	69.84
Em	79.99	89.07	84.29	68.51	51.63	58.88	**71.59**
SmartSA	80.55	89.74	84.90	70.36	52.93	60.41	72.66
SentiStrength	83.30	91.10	87.03	75.70	60.20	67.07	**77.05***

markers is more than that of intensifiers/diminishers (Table 7.4). This suggests that score adjustment based on the occurrence of intensifiers/diminishers is more beneficial than discourse markers for sentiment analysis of social media. However, the consistent improvement observed with DISC over BASE shows that sentiment analysis can benefit from discourse analysis.

For the non-lexical valence shifters, score adjustment based on emoticons (EM) performed best.

TABLE 7.4: Datasets statistics and average F scores

	Digg	LiveJ	MySpace	RunnersW	Twitter	Youtube
#Documents						
Positive	201	427	702	484	2587	1665
Negative	572	304	132	221	843	767
Statistics						
Avg. sentence	6	1	2	5	2	2
Avg. word	78	12	12	55	16	18
Negation	522(0.68)	31(0.04)	351(0.42)	987(1.40)	1227(0.36)	844(0.35)
Intensifiers/Dim	371(0.48)	240(0.33)	165(0.20)	541(0.77)	396(0.16)	448(0.18)
Discourse markers	743(0.96)	411(0.56)	543(0.65)	1231(1.75)	1161(0.34)	1238(0.51)
Capitalisation	95(0.12)	54(0.07)	84(0.10)	121(0.17)	669(0.20)	231(0.09)
Repeat letter	13(0.02)	23(0.03)	61(0.07)	16(0.02)	51(0.01)	61(0.03)
Emoticons	37(0.05)	91(0.12)	192(0.23)	180(0.26)	530(0.15)	341(0.14)
Average F scores						
BASE	58.05	76.07	68.41	65.26	61.26	69.84
SWITCH	54.68	72.43	67.77	64.08	58.36	66.57
SHIFT	55.39	73.20	68.42	63.96	59.05	66.92
SWITCH+scores	61.23	76.20	68.55	65.85	60.18	70.74
SHIFT+scores	62.36	76.49	72.17	67.75	64.42	71.24
INTDIM	61.08	76.33	69.87	67.02	62.83	71.20
INTDIM+scores	59.03	76.05	68.67	65.65	61.95	68.18
DISC	61.76	76.05	68.92	66.74	62.15	71.01
DISC+scores	61.19	76.05	68.41	65.39	61.37	69.67
CAPS	58.81	76.07	68.73	66.93	61.77	69.84
RP	58.05	76.07	68.54	65.61	61.57	69.84
EM	59.76	76.61	69.56	68.13	63.05	71.59
SMARTSA	64.27	76.75	73.44	71.17	63.24	72.66
SENTISTRENGTH	66.87	73.33	66.87	70.87	72.94	77.05

This might be explained by the fact that emoticons are the most common non-lexical modifiers in 4 (out of 6) of our datasets. It could also be because emoticons are more

discriminative between positive and negative documents. The other non-lexical strategies (CAPS and RP) provide marginal but consistent improvement over the BASE. This marginal improvement might be because of the limited occurrence of the non-lexical valence shifters in our datasets which limits the opportunity for the strategies to be extensively utilised.

SMARTSA integrates best-performing options for lexical score adjustment strategies (SHIFT+scores, INTDIM and DISC) and the non-lexical strategies. Its performance, on all the datasets, is consistently better than any of the individual contextual score adjustment strategies and significantly better than the BASE on 4 datasets. This shows that the contextual score adjustment strategies tend to provide complementary improvement for sentiment classification. It was particularly observed that SHIFT+scores tends to improve classification of negative documents; INTDIM, DISC and EM tend to improve both positive and negative classification; and CAPS and RP tend to improve positive classification. The inability of SMARTSA to attain significant improvement on LiveJ and MySpace could be attributed to the relatively small occurrence of lexical valence shifters (especially negation) in these datasets. It can also be observed that these datasets have the shortest average document lengths, thus, rendering some of our strategies (e.g. DISC) less relevant.

Compared to the state-of-the-art system, SENTISTRENGTH, SMARTSA performed better on 3 datasets (LiveJ, RunnersW and Twitter) while SENTISTRENGTH was better on the other 3 datasets (Digg, MySpace and Youtube). The negation analysis integrated with SMARTSA could have especially helped in its better accuracy on RunnersW as this dataset has a relatively high proportion of negation terms. Whereas the high coverage of SentiWordNet might have influenced the better performance of SMARTSA on LiveJ and Twitter, the unavailability of certain, social media prolific, sentiment-bearing terms from the lexicon could have affected the performance of SMARTSA on MySpace. Such terms (e.g. *'lol', 'xoxo'*, e.t.c) were manually included into the SENTISTRENGTH lexicon. We address this problem in DSMARTSA by adapting the lexicon to the vocabulary of social media.

7.2 Evaluation of DSmartSA and Related Strategies

The aim of this study is three-fold. Firstly, to investigate whether or not combining the two lexicons (static and domain-specific) is better than using each individually. Secondly, to investigate the performance of our approach compared to the performance of machine learning algorithms that are trained on the distant-supervised datasets. Lastly, to assess the transferability of a hybrid lexicon from one social media domain to another. A secondary aim is to study the suitability of the term-class association metrics (sTF, sTFIDF and pPMI) as the means to quantify the sentiment polarity scores for the domain-specific lexicon. Accordingly, we investigate the following classifier settings:

1. STATIC: Lexicon-based sentiment classification using static lexicon (SentiWordNet) as implemented in SMARTSA.

2. DOMAIN_{X}: Lexicon-based sentiment classification using a domain-specific lexicon. Here, x refer to the term-sentiment association metric used.

3. HYBRID_{X}: Lexicon-based sentiment classification using a hybrid lexicon (DSMARTSA). Again, x denotes the term-sentiment association metric used in generating the domain-specific lexicon integrated into the hybrid lexicon. For each dataset we report results using a setting for the parameters α and β ($0 \leq \{\alpha, \beta\} \leq 1$) that produces the best classification accuracy on the distant-supervised data. The parameter values for the setting, on each dataset, are shown in Table 7.5.

TABLE 7.5: Mixing parameter values

Dataset	α	β
Twitter	0.5	0.3
Digg	0.3	0.3
MySpace	0.7	0.5

4. Machine Learning algorithms: These are the Support Vector Machines (SVM), Naïve Bayes (NB) and Logistic Regression or Maximum Entropy (LR). Comparison of DSMARTSA with these machine learning algorithms enable us to test the

TABLE 7.6: Results from DSMARTSA and related strategies on test datasets

Algorithm	Positive			Negative			Avg F1
	P	R	F1	P	R	F1	
Twitter							
Machine Learning							
SVM	67.40	33.20	44.49	55.60	82.70	66.49	55.49
NB	65.60	67.30	66.44	65.20	63.50	64.34	65.39
LR	71.70	78.00	74.72	75.20	68.40	71.64	73.18
Lexicon-based							
STATIC	87.64	79.51	83.38	51.06	65.60	57.42	70.40
$\text{DOMAIN}_{\text{sTF}}$	68.20	74.50	71.21	76.20	70.20	73.08	72.15
$\text{DOMAIN}_{\text{sTFIDF}}$	71.50	70.80	71.15	69.50	70.20	69.85	70.50
$\text{DOMAIN}_{\text{PPMI}}$	66.50	70.30	68.35	71.20	67.50	69.30	68.83
$\text{HYBRID}_{\text{sTF}}$	74.80	79.00	76.84	79.70	75.60	77.60	**77.22**
$\text{HYBRID}_{\text{sTFIDF}}$	75.40	67.20	71.06	61.50	70.30	65.61	68.34
$\text{HYBRID}_{\text{PPMI}}$	75.90	68.70	72.12	63.80	71.60	67.48	69.80
Digg							
Machine Learning							
SVM	35.10	49.70	41.14	69.90	55.00	61.56	51.35
NB	35.30	49.70	41.28	70.10	55.50	61.95	51.62
LR	45.80	72.20	56.05	81.70	58.20	67.98	62.02
Lexicon-based							
STATIC	43.00	83.33	56.73	90.67	59.44	71.81	64.27
$\text{DOMAIN}_{\text{sTF}}$	81.50	44.60	57.65	53.30	89.20	66.73	62.19
$\text{DOMAIN}_{\text{sTFIDF}}$	81.40	45.60	58.45	55.50	90.00	68.66	63.56
$\text{DOMAIN}_{\text{PPMI}}$	84.30	43.60	57.47	48.30	89.10	62.64	60.06
$\text{HYBRID}_{\text{sTF}}$	87.10	49.00	62.72	59.20	95.10	72.97	67.85
$\text{HYBRID}_{\text{sTFIDF}}$	84.30	49.00	61.98	61.00	93.70	73.89	**67.94**
$\text{HYBRID}_{\text{PPMI}}$	87.10	48.80	62.55	58.70	95.00	72.56	67.56
MySpace							
Machine Learning							
SVM	79.20	100	88.40	0.00	0.00	0.00	44.20
NB	86.90	43.30	57.8	26.60	73.50	39.06	48.43
LR	91.00	70.80	79.64	37.50	67.80	48.29	63.97
Lexicon-based							
STATIC	89.31	83.33	86.22	35.00	47.37	40.26	63.24
$\text{DOMAIN}_{\text{sTF}}$	61.90	86.60	72.20	58.80	29.10	38.93	55.57
$\text{DOMAIN}_{\text{sTFIDF}}$	48.40	88.80	62.65	74.20	28.30	40.97	51.81
$\text{DOMAIN}_{\text{PPMI}}$	53.10	88.50	66.38	70.30	29.00	41.06	53.72
$\text{HYBRID}_{\text{sTF}}$	77.20	90.30	83.24	61.70	40.20	48.68	**65.96**
$\text{HYBRID}_{\text{sTFIDF}}$	54.90	91.50	68.63	77.00	31.60	44.81	56.72
$\text{HYBRID}_{\text{PPMI}}$	62.40	61.10	61.74	72.20	33.80	46.04	53.89

utility of distant supervision in DSMARTSA against its standard use in supervised machine learning.

7.2.1 Results and Discussion

Table 7.6 shows sentiment classification results on Twitter, Digg and MySpace test datasets. Overall, the hybrid approach performs better than all supervised machine learning algorithms (SVM, NB and LR) on all the three datasets; 77.26% Vs 73.18% on Twitter, 67.94% Vs 62.02% on Digg and 65.96.9% Vs 63.97% on MySpace (when compared with the best-performing supervised machine learning classifier, LR). This confirms the superiority of our lexicon approach using distant-supervised learning over the machine learning approaches to sentiment classification. These results also show that the DSmartSA has achieved improvements over SmartSA outperforming the state-of-the-art, SentiStrength, on one more dataset (Digg). As for the weighting metrics, sTF performed overall best on 2 out of the 3 datasets (Twitter and MySpace). sTFIDF performed best on the remaining dataset, Digg. Documents in this dataset tend to have higher chance for the appearance of standard vocabulary due to their verbosity. This could have influenced the good performance of sTFIDF because, as we observe and mentioned previously, sTFIDF tends to favour standard terms. pPMI has the lowest performance on all the 3 datasets, however, we note that its performance is quite competitive with the newly introduced metrics.

The comparison of the lexicon based approaches to sentiment analysis shows that the hybrid lexicon does perform significantly better than alternative approaches. Next, we look at the performance of the hybrid lexicon against the individual lexicons it combined.

7.2.2 Hybrid Vs Individual Lexicons

Results from Table 7.6 show that, as expected, on the Twitter dataset the hybrid approach ($\text{Hybrid}_{\text{sTF}}$) performs better than Static and the best-performing Domain_{X} approach (77.22% Vs 70.40% and 72.15% respectively). Also, Domain_{X} performs better than Static, indicating the inability of the static lexicon, which is generated from fairly standard texts, to capture certain sentiment expressions from non-standard texts. Similar results are also observed on the Digg dataset. However, although best results are obtained with a hybrid lexicon, the Static lexicon has out performed the best Domain

TABLE 7.7: Transferability of hybrid lexicon across social media domains

Algorithm	Positive			Negative			Avg F1
	P	R	F1	P	R	F1	
Twitter as Distant-supervised dataset:							
Digg	70.90	58.80	64.29	77.10	85.70	81.17	72.73^{+}
MySpace	63.40	93.60	75.60	79.00	36.30	49.74	62.67^{-}
Digg as Distant-supervised dataset:							
MySpace	86.20	90.40	88.25	56.80	48.50	52.32	70.29^{+}
Twitter	74.30	64.10	68.82	56.40	67.40	61.41	65.12^{-}
MySpace as Distant-supervised dataset:							
Twitter	46.10	73.30	56.60	84.30	61.10	70.85	63.73^{-}
Digg	44.50	55.40	49.36	84.80	77.40	80.93	65.15^{-}
All genres as source							
Twitter	73.40	76.10	74.73	76.40	73.80	75.08	74.91^{-}
Digg	70.40	73.10	71.72	73.40	70.60	71.97	71.85^{+}
MySpace	90.40	93.00	91.68	68.40	51.20	58.56	75.12^{+}

lexicon. Although this difference is marginal it does raise two interesting questions: either distant-supervised labelling is more suitable for Tweets than Digg sentences or the smaller distant-supervised data size in Digg, compared to Twitter, has affected the reliability of the domain-specific lexicon generated from Digg. It is also interesting to note that unlike on the Twitter dataset, all machine learning algorithms have performed extremely poorly on the Digg dataset. Given that they rely heavily on the distant-supervised labelled data (just as the DOMAIN_X algorithms) it is likely that considerable noise has been introduced by relying on sentiment markers from a poorly representative sample of data. This observation is further supported by the results from MySpace (the smallest of the three datasets for distant supervision). Once again we see poor accuracy with machine learning algorithms and STATIC performing better than DOMAIN_X and comparable to HYBRID_X. This is more likely to be caused by the very limited data from which the domain-specific lexicon is generated for MySpace (see Table 5.2). This suggests the need to establish minimum dataset requirements below which a domain-specific lexicon becomes unreliable due to the small datasets size and/or atypical usage of emoticons such as when used to express sarcasm or to soften the intensity of their opposite sentiment. This then begs the question of can we augment smaller distant-supervised datasets that are likely to be less representative of the underlying emoticon usage behaviour with larger datasets that are easier to obtain from a different domain. This issue brings us conveniently onto the next topic of transferability.

7.2.3 Transferability Across Social Media Domains

As distant supervision relies on certain sentiment markers to label documents which may not be very common in some social media platforms, it is imperative to assess the performance of a hybrid lexicon on a platform different from the one it was initially generated on (i.e. transferability of the lexicon). We use HYBRID with sTF for this experiment as it has overall best performance. Sentiment classification results from this experiment are shown in Table 7.7 (the plus sign, +, indicates improvement while the minus sign, −, indicates a decline over using within platform/domain distant-supervised data).

For Twitter, using its own domain for distant supervision (i.e. within platform) is better than either using Digg posts or MySpace messages (77.22 Vs 65.12 and 63.73). However with the other smaller distant-supervised datasets (Digg and MySpace) we see significant improvements when they are augmented or replaced with the larger Twitter distant-supervised dataset. For instance, with Digg, an increase of over 5% is observed when using a distant-supervised Twitter dataset. Whilst with MySpace an impressive 10% improvement is observed with a distant-supervised dataset formed by combining data from all platforms. This performance surpasses the performance from SENTISTRENGTH, making our dynamic hybrid lexicon approach significantly better than a state-of-the-art system on the three datasets used in the evaluation of the hybrid lexicon approach. These results indicate that when a within platform dataset is small or unavailable, using data from a different platform is advantageous. However, the results on MySpace raise the question of what platform is compatible with another, considering that the Digg generated lexicon compares favourably over Twitter lexicon even though the size of the distant-supervised Twitter dataset is a magnitude larger than the Digg dataset.

7.3 Evaluation of the Hybrid Classifier

We conduct experiments to evaluate our hybrid approach to sentiment classification that combines knowledge from a training dataset as well as from a hybrid sentiment lexicon and an emotion lexicon. We use distant-supervised datasets (discussed in Section 5.1)

for training and human-labelled data for testing. Based on the performance of machine classifiers in the previous experiments, here we concentrate on the maximum entropy classifier which had best results over support vector machines and naïve bayes classifiers. Similarly, we use the sTF approach to quantify term-sentiment association in the hybrid lexicon from which we extract values for sentiment features. The hybrid classifier is aimed at combining distant-supervised (rather traditional supervised) learning with knowledge from lexicons, thus, experiments presented here concentrate on the three social media platforms for which we have distant-supervised data (i.e. Twitter, Digg and MySpace)

Our main objective in this evaluation is to determine whether the novel feature sets introduced from local contextual analysis, hybrid sentiment lexicon and emotion lexicon can improve sentiment classification accuracy. To this end, we run experiments with the following classifier settings:

- NGRAM: A baseline classifier that uses just n-gram features from the training data
- NGRAM+LC: A classifier that uses n-gram and local context features
- NGRAM+LC+SENT: A classifier that uses n-gram, local context and a hybrid sentiment lexicon features
- NGRAM+LC+SENT+EMO: A classifier that uses n-gram, local context, hybrid sentiment lexicon and emotion lexicon features

With the above experimental setting, our expectation is that sentiment classification accuracy will increase with additional feature sets. We expect features from local context (LC) to capture linguistics aspects and writing style in social media; sentiment features (SENT) to capture sentiment-bearing properties of terms utilising a hybrid lexicon, and emotion features (EMO) to capture emotive aspects utilising knowledge from an emotion lexicon.

TABLE 7.8: Results from the hybrid classifier on test datasets

Algorithm	Positive			Negative			Avg F1
	P	R	F1	P	R	F1	
Twitter							
NGRAM	71.70	78.00	74.72	75.20	68.40	71.64	73.18
NGRAM+LC	73.07	78.56	75.72	76.53	68.73	72.42	74.07
NGRAM+LC+SENT	73.66	80.11	76.75	78.21	70.40	74.10	75.42*
NGRAM+LC+SENT+EMO	74.70	81.00	77.72	78.20	71.40	74.65	**76.18***
Digg							
NGRAM	41.47	85.71	55.9	91.38	55.59	69.13	62.52
NGRAM+LC	42.45	85.71	56.78	91.50	56.97	70.22	63.50
NGRAM+LC+SENT	44.23	85.71	58.35	92.00	60.31	72.86	65.61
NGRAM+LC+SENT+EMO	45.23	85.71	59.21	92.19	61.89	74.06	**66.64***
MySpace							
NGRAM	91.00	70.80	79.64	37.50	67.80	48.29	63.96
NGRAM+LC	88.80	69.28	77.83	37.27	67.34	47.98	62.91
NGRAM+LC+SENT	93.20	72.32	81.44	37.73	68.26	48.60	65.02
NGRAM+LC+SENT+EMO	93.24	72.79	81.76	38.38	68.35	49.16	**65.46**

7.3.1 Results and Discussion

As expected, the use of all feature sets (NGRAM+LC+SENT+EMO) provides best classification accuracy on all the 3 datasets (Table 7.8, bold font indicates the best performance on a dataset and asterisk, *, indicates significant difference from the baseline, NGRAM). The approach (NGRAM+LC+SENT+EMO) achieves significantly better performance on Twitter and Digg datasets. However, the improvement is not significant on MySpace dataset due to the very limited training data from this platform.

The local contextual features consistently provide performance improvement over pure n-gram features on Twitter (0.89%) and Digg (0.98%). However, a marginal degradation is observed on the MySpace dataset. Again, this can be explained by the limited training data from this platform, resulting in a sparser representation.

The sentiment feature set provides the most performance improvement over the baseline on all 3 datasets. This is not surprising given the high-coverage of the hybrid lexicon used and the domain adaptation involved. Similarly, the addition of emotion-based feature sets provides moderate but consistent performance improvement on all the 3 datasets. This confirms the usefulness of emotion knowledge in addition to sentiment knowledge for sentiment classification.

7.4 Chapter Summary

In this chapter, we presented evaluations of our sentiment classification strategies as discussed in Chapters 4, 5 and 6. For SmartSA, the evaluations involved ablation tests to investigate the performance of each strategy (and its various options where applicable) as well as the combined effect of strategies integrated into the algorithm. This is further compared with results from a state-of-the-art system for sentiment analysis (SentiStrength). The results show SmartSA to significantly outperform the baseline. The results also reveal that negation is most beneficial of lexical score adjustments while emoticons are the most useful of non-lexical adjustments. However, each of the strategies integrated into SmartSA contribute to success in sentiment classification. The comparison with SentiStrength shows SmartSA to be competitive even though SentiStrength uses a lexicon that is manually extended with social media oriented vocabulary.

In DSmartSA, we investigated whether or not combining the two lexicons (static and domain-specific) to form a hybrid lexicon is better than using each individually. Likewise, we investigated the performance of our hybrid lexicon approach compared with machine learning algorithms trained with distant-supervised data. We also evaluate the performance of the two introduced term-sentiment associations metrics (sTF and sTFIDF) in relation to the state-of-the-art metric pPMI where both metrics performed better than pPMI with sTF performing overall best. Lastly we assess the transferability of the hybrid lexicon from one social media domain to another. The results show that, as DSmartSA incorporates dynamic vocabulary and polarities of social media domains, it improves on SmartSA and outperforms SentiStrength in sentiment classification. The results also show the hybrid lexicon approach to outperform each of the combined lexicons.

Finally, we presented evaluations of our hybrid classification approach combined distant-supervised training data; features from local context analysis, the hybrid sentiment lexicon and emotion lexicon. The results demonstrated that each of our newly introduced feature sets improves sentiment classification performance.

Chapter 8

Conclusions

In this thesis, we addressed the problem of determining contextual polarity when lexicon-based sentiment analysis is applied to social media content. We modelled the problem from two perspectives: the interaction of terms with their neighbouring terms (local context) and the interpretation of meaning specific to domain usage (domain context). Accordingly, we set out to achieve six research objectives. In this chapter, we revisit these objectives drawing conclusions and also propose future extensions to our work.

8.1 Objectives Revisited

1. **Conduct a comparative analysis of score extraction methods for SentiWordNet with focus on using local context for word sense disambiguation.** In Chapter 4, we formalise existing score extraction approaches from SentiWordNet and introduced a word sense disambiguation (WSD) algorithm that exploits local context of terms in order to determine the appropriate sense from the lexicon. We evaluate performance of this algorithm in comparison with the existing approaches (in Chapter 7). The results confirm that WSD is useful on social media domains that have relatively longer documents (e.g. discussion posts).

2. **Develop a lexicon-based classifier to integrate local context knowledge with sentiment content in SentiWordNet.** In line with this objective, in

Chapter 4 we introduced SmartSA, a sentiment classification system that integrates linguistic contextual analysis for sentiment prediction using SentiWordNet, a popular lexicon with a high term-coverage and rich sentiment information. In SmartSA, we show how contextual adjustment of SentiWordNet scores for terms based on negation, intensification/diminution, discourse structure and other non-lexical phenomena can significantly influence sentiment analysis of social media. Evaluation results also show that sentiment classification of social media significantly benefit from the contextual score adjustment introduced in SmartSA. A further comparison with a state-of-the-art system (SentiStrength) shows SmartSA to be competitive.

Being a high-coverage lexicon, SentiWordNet offers sentiment scores for typical sentiment modifying terms such as negation and intensifiers. Thus, we investigated the behaviour of such terms when they are treated as modifiers or as sentiment-bearing. Our results show that negation terms are sentiment-bearing in addition to being modifiers. This confirms the previous work that shows negation to be indicative of sentiment (Potts, 2011*a*). However, such is not the behaviour of intensifiers, diminishers and discourse markers. Sentiment classification is better when these terms are decoupled from their sentiment scores and treated as modifiers.

This and the previous objective provide insights to discuss our first research question: *"Does the accuracy of lexicon-based sentiment analysis benefit from the integration of local context knowledge?"*. Unlike in the existing research, we investigate this question using a lexicon with a more fine-grained sentiment information. Our results provide evidence that shows the benefit of several strategies for local context analysis. However, to achieve such a benefit, there is the need for careful assessment of each type of modifier in relation to its modification and sentiment-bearing characteristics. For social media domains, local context analysis should always go beyond the lexical modifiers. Results in this research provide evidence that non-lexical, social media oriented, sentiment modifiers consistently improve sentiment classification accuracy. Local context knowledge is also useful in determining the correct senses (meanings) of terms in a given document. Using only the information associated with the correct senses improves classification accuracy.

However, the short document length nature of social media seems to make the task of determining the correct senses (i.e. WSD) error-prone.

3. **Extend the classifier developed in 2 to address the continuously evolving vocabulary typical in social media streams.** In Chapter 5, we presented a novel approach to generating a hybrid lexicon that adapts a general purpose sentiment lexicon (SentiWordNet) to the context of social media domains. We achieved this by first generating a domain-specific lexicon and subsequently combining the lexicon with the general-purpose lexicon. This approach has the dual effect of capturing domain-specific terms which are otherwise unavailable in the general-purpose lexicon as well as providing a strategy to modify sentiment polarities of terms depending on domain-specific usage. We demonstrated how distant supervision can be exploited for this purpose. In order to quantify term-sentiment association, we introduced metrics (sTF and sTFIDF) that are able to produce better results for low frequency terms compared to a state-of-the-art metric, pPMI. This is important since many non-standard terms encountered in social media tend to have relatively low frequencies. Evaluation results show that when a hybrid lexicon is used in SmartSA (i.e. DSmartSA), further classification improvements are gained over SentiStrength. As distant supervision is typically employed in machine learning approaches to sentiment classification, we compared our classifier with three state-of-the-art algorithms (Support Vector Machines, Naïve Bayes and Maximum Entropy). Our lexicon-based classifier performed better than all three machine learning algorithms on all our evaluation datasets. Social media datasets vary considerably in the use of sentiment markers that can be exploited for distant supervision ranging from platforms having an abundance to those having very few. Thus, we introduced transfer learning in which a domain adapted (hybrid) lexicon is generated for a domain using distant-supervised data from another domain. By doing so, we were able to discuss compatibility between social media domains as well as open up a future research direction.

   With regards to the research question: *"How can we evolve a static lexicon to dynamically adapt to vocabulary and domain-specific semantics in social media?"*; our evaluation results provide evidence that the hybrid lexicon approach is capable

of evolving a static lexicon to reflect the vocabulary and polarities of social media domains.

4. **Investigate the utility of combining the local context analysis (in 2) and vocabulary adaptation (in 3) in the context of a hybrid sentiment classifier.** To achieve this objective we introduced a sentiment classifier following the hybrid sentiment classification approach combining machine learning and lexicon-based methods (Chapter 6). We introduced a method to derive feature sets from our local context strategies (Chapter 4) and a domain-adapted hybrid lexicon (Chapter 5). Evaluation results show that each feature category (local context and sentiment) improves sentiment classification accuracy over the baseline of pure n-gram features.

5. **Study the role of emotive concepts by integrating emotion knowledge into the classifier developed in 4.** In Chapter 6, we introduced an approach to incorporating emotion knowledge derived from a lexicon for sentiment analysis. Evaluation results (in Chapter 7) show that emotive features provide moderate but consistent accuracy improvement in sentiment classification. The emotive feature sets, in combination with feature sets from local context and domain-adapted lexicon, provide statistically significant improvement over the baseline.

   Regarding the research question: *"How does emotion knowledge captured in an emotion lexicon influence sentiment analysis?"*; our results show that emotion knowledge can provide improvement in classification accuracy. Although emotion knowledge can be viewed to be an alternative to sentiment knowledge, this research shows that a sentiment lexicon does not exclude the utility of an emotion lexicon for sentiment analysis.

6. **Conduct a comprehensive evaluation of all developed classifiers/strategies.** We conducted evaluations to ascertain the effectiveness of each of the classifiers developed in this research: SmartSA, DSmartSA and the hybrid classifier (Chapter 7). We compare the performance of these classifiers against the baseline

and the state-of-the-art classifiers. The evaluations also involved testing the contribution of each strategy integrated with the classifiers (and its alternative where applicable).

In summary, the main contributions to knowledge from this research are, first, the introduction of a word sense disambiguation (WSD) algorithm for the extraction of sentiment scores from SentiWordNet and the detailed evaluation of this approach in comparison with the typical approaches used for WSD. Second, the introduction of a lexicon-based sentiment classifier (SmartSA) that integrates contextual analysis strategies to adjust prior polarities of terms in order to account for the effect of both standard and social media oriented sentiment modifiers as well as discourse structures. Third, the development of an approach to dynamically improve lexical coverage and sentiment semantics of terms given a social media domain (DSmartSA). This approach combines sentiment knowledge from a general purpose lexicon and a target domain to create a hybrid lexicon that captures the non-standard, sentiment rich terms; and non-standard usage of terms for sentiment expression in social media. Another novel feature in DSmartSA is the introduction of two new term-sentiment association metrics inspired by Term Frequency and Inverse Document Frequency (TF, TFIDF). This is important because the state-of-the-art metrics, based on the Point-wise Mutual Information (PMI) do not work well on terms that have low frequencies in a collection (Sani, 2014), a characteristic of evolving terms in social media. The fourth contribution is the development of a hybrid social media sentiment classifier that combines distant-supervised learning, contextual analysis, domain semantics, and an emotion lexicon. This classifier benefits from the deeper analysis of supervised machine learning algorithms, local and domain context analysis without the overhead of requiring hand-labelled data. It also allows us to measure the extent to which our lexicon-based strategies and emotion knowledge are applicable in the hybrid sentiment classification setting. Lastly, the assessmecnt of the[21 transferability of a hybrid lexicon (used by DSmartSA) on a social media domain different from the one from which it was generated. This is important since distant-supervised data (required to generate a hybrid lexicon) may not be available from some social media domains.

8.2 Future Work

In this section we highlight some of the limitations of the work we presented in this thesis and also point out some desirable future extensions. Firstly, given the focus of sentiment classification involving positive and negative classes, a natural extension to this work is subjectivity detection whereby a piece of text is classified as "objective" or "subjective". It will be interesting to investigate whether a zero aggregate score is indicative of the objective class. This is because, with high-coverage lexicons, many terms are associated with non-zero sentiment scores including those that would appear to be objective. Secondly, the work presented in this thesis, aimed at accounting for contextual polarities of terms, can go beyond lexicons and domain-specific knowledge to leverage "common-sense" knowledge. This is particularly useful in sarcasm detection. Similar to the existing research, we notice sarcasm is quite a characteristic of social media text and its detection is likely to have significant impact on sentiment analysis. We note that recently resources are being developed to capture common-sense knowledge (e.g. SenticNet). However, it remains a research problem to investigate the extent to which such resources are useful for reasoning with social media text given the informal/non-standard nature of its genres.

Thirdly, the hybrid lexicon approach presented for adapting a general purpose lexicon to social media domains (hybrid lexicon) should also be considered in the context of "big data" since volume and veracity are typical characteristics of social media platforms (e.g. Twitter). Therefore, the approach needs to be extended to large-scale data streams using big data processing methods. Also, the transfer learning aspect of the hybrid lexicon revealed that there is more to compatibility between domains than what is expected (document lengths or tendency towards having similar non-standard contents). Therefore, it will be useful to investigate the characteristics that govern the affinity between social media domains for the purpose of transfer learning. Finally, as we investigated the role of emotion for sentiment detection, it is also imperative to investigate the role of sentiment for emotion detection. In particular, does the sentiment-based features help improve emotion analysis or classification, as is the case when emotion features are used for sentiment analysis.

Appendix A

Publications

- A. Muhammad, N. Wiratunga, R. Lothian: Contextual Sentiment Analysis for Social Media Genres. Knowledge-based Systems (2016). Elsevier
- A. Muhammad, N. Wiratunga, R. Lothian: Context-Aware Sentiment Analysis of Social Media. In: M.M. Gaber et. al (Eds), Advances in Social Media Analysis (2015). Springer
- A. Muhammad, N. Wiratunga, R. Lothian: A Hybrid Sentiment Lexicon for Social Media Mining. In: Proc. of the 26th IEEE International Conference on Tools with Artificial Intelligence, IEEE ICTAI (2014)
- A. Muhammad, N. Wiratunga, R. Lothian, R. Glassey: Contextual Sentiment Analysis of social media Using High-Coverage Lexicon. In: Proc. of SGAI International Conference on Artificial Intelligence, BCS SGAI (2013)
- A. Muhammad, N. Wiratunga, R. Lothian and R. Glassey: Domain-based Lexicon Enhancement for Sentiment Analysis. In: Proc. SGAI Workshop on Social Media Mining. BCS SGAI (2013)

Bibliography

Abbasi, A., Chen, H. and Salem, A. (2008), 'Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums', *ACM Transactions on Information Systems (TOIS)* **26**(3), 12.

Agrawal, S. and Siddiqui, T. (2009), Using syntactic and contextual information for sentiment polarity analysis, *in* 'Proceedings of the 2nd International Conference on Interaction Sciences: Information Technology, Culture and Human', ICIS '09, ACM, New York, NY, USA, pp. 620–623.

Al-Mannai, K., Alshikhabobakr, H., Wasi, S. B., Neyaz, R., Bouamor, H. and Mohit, B. (2014), Cmuq-hybrid: Sentiment classification by feature engineering and parameter tuning, *in* 'Proceedings of Semantic Evaluation Workshop (SemEval2014)', p. 181.

Alm, E. C. O. (2008), *Affect in text and speech*, ProQuest.

Andreevskaia, A. and Bergler, S. (2008), When specialists and generalists work together: Overcoming domain dependence in sentiment tagging., *in* 'ACL', pp. 290–298.

Andreevskaia, A., Bergler, S. and Urseanu, M. (2015), All blogs are not made equal: Exploring genre differences in sentiment tagging of blogs.

Asur, S. and Huberman, B. A. (2010), Predicting the future with social media, *in* 'Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on', Vol. 1, IEEE, pp. 492–499.

Baccianella, S., Esuli, A. and Sebastiani, F. (2010), Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, *in* 'Proceedings of the Annual Conference on Language Resouces and Evaluation'.

Bamman, D. and Smith, N. A. (2015), Contextualized sarcasm detection on twitter, *in* 'Proceedings of the 9th International Conference on Web and Social Media', AAAI Menlo Park, CA, pp. 574–77.

Bandhakavi, A., Wiratunga, N., P, D. and Massie, S. (2014), Generating word-emotion lexicon from #emotional tweets, *in* 'Proceedings of the 3rd Joint Conference on Lexical and Computational Semantics'.

Berenson, M., Levine, D., Szabat, K. A. and Krehbiel, T. C. (2012), *Basic business statistics: Concepts and applications*, Pearson Higher Education AU.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003), 'Latent dirichlet allocation', *the Journal of machine Learning research* **3**, 993–1022.

Blitzer, J., McDonald, R. and Pereira, F. (2006), Domain adaptation with structural correspondence learning, *in* 'Proceedings of the 2006 conference on empirical methods in natural language processing', Association for Computational Linguistics, pp. 120–128.

Blum, A. and Mitchell, T. (1998), Combining labeled and unlabeled data with co-training, *in* 'Proceedings of the eleventh annual conference on Computational learning theory', ACM, pp. 92–100.

Bollen, J., Mao, H. and Zeng, X. (2011), 'Twitter mood predicts the stock market', *Journal of Computational Science* **2**(1), 1–8.

Boucher, J. and Osgood, C. E. (1969), 'The pollyanna hypothesis', *Journal of Verbal Learning and Verbal Behavior* **8**(1), 1–8.

Bravo-Marquez, F., Frank, E. and Pfahringer, B. (2015), Positive, negative, or neutral: learning an expanded opinion lexicon from emoticon-annotated tweets, *in* 'Proceedings of the 24th International Conference on Artificial Intelligence', AAAI Press, pp. 1229–1235.

Brin, S. and Page, L. (1998), The anatomy of a large-scale hypertextual web search engine, *in* 'Seventh International World-Wide Web Conference (WWW 1998)'.

Chakraborti, S., Wiratunga, N., Lothian, R. and Watt, S. (2007), Acquiring word similarities with higher order association mining, *in* 'Case-Based Reasoning Research and Development', Springer, pp. 61–76.

Choi, Y. and Cardie, C. (2009), Adapting a polarity lexicon using integer linear programming for domain-specific sentiment classification, *in* 'Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 2-Volume 2', Association for Computational Linguistics, pp. 590–598.

Cohen, N. (2009), 'Is it a day to be happy? check the index', The New York Times, Online, Accessed 08/04/2016.
URL: *http://www.nytimes.com/2009/10/12/technology/internet/12link.html?_r=0*

Crystal, D. (2011), *Dictionary of linguistics and phonetics*, Vol. 30, John Wiley & Sons.

Dadvar, M., Hauff, C. and De Jong, F. (2011), 'Scope of negation detection in sentiment analysis'.

Dang, Y., Zhang, Y. and Chen, H. (2010), 'A lexicon-enhanced method for sentiment classification: An experiment on online product reviews', *IEEE Intelligent Systems* **25**, 46–53.

Das, D. (2010), Computational analysis of text sentiment: A report on extracting contextual information about the occurrence of discourse markers, Technical report, Simon Fraser University.

Daume III, H. and Marcu, D. (2006), 'Domain adaptation for statistical classifiers', *Journal of Artificial Intelligence Research* pp. 101–126.

Davidov, D., Tsur, O. and Rappoport, A. (2010), Enhanced sentiment learning using twitter hashtags and smileys, *in* 'Proceedings of the 23rd International Conference on Computational Linguistics: Posters', Association for Computational Linguistics, pp. 241–249.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W. and Harshman, R. A. (1990), 'Indexing by latent semantic analysis', *JAsIs* **41**(6), 391–407.

Dehkharghani, R., Yanikoglu, B., Tapucu, D. and Saygin, Y. (2012), Adaptation and use of subjectivity lexicons for domain dependent sentiment classification, *in* 'Data Mining Workshops (ICDMW), 2012 IEEE 12th International Conference on', IEEE, pp. 669–673.

Denecke, K. (2008), Using sentiwordnet for multilingual sentiment analysis, *in* 'ICDE Workshop'.

Devitt, A. and Ahmad, K. (2007), Sentiment polarity identification in financial news: A cohesion-based approach, *in* 'Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics', Prague, Czech Republic, pp. 984–991.

Domingos, P. and Pazzani, M. (1996), Beyond independence: Conditions for the optimality of the simple bayesian classifier, *in* 'Procecedings of ICML', pp. 105–112.

Du, W., Tan, S., Cheng, X. and Yun, X. (2010), Adapting information bottleneck method for automatic construction of domain-oriented sentiment lexicon, *in* 'Proceedings of the third ACM international conference on Web search and data mining', ACM, pp. 111–120.

Ekman, P. (1999), 'Basic emotions', *Handbook of cognition and emotion* **98**, 45–60.

Esuli, A., Baccianella, S. and Sebastiani, F. (2010), Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining, *in* 'Proceedings of the Seventh conference on International Language Resources and Evaluation (LREC10)'.

Esuli, A. and Sebastiani, F. (2005), Determining the semantic orientation of terms throush gloss classi?cation, *in* 'Proceedings of CIKM', pp. 617–624.

Esuli, A. and Sebastiani, F. (2006), Determining term subjectivity and term orientation for opinion mining, *in* 'Proceedings of EACL-06, 11th Conference of the European Chapter of the Association for Computational Linguistics, Trento, IT.'.

Fellbaum, C. (1998), *WordNet: An Electronic Lexical Database*, MIT Press.

Fishman, J. A. (1972), *The sociology of language*, Newbury House Rowley, MA.

Ghazi, D., Inkpen, D. and Szpakowicz, S. (2010), Hierarchical versus flat classification of emotions in text, *in* 'Proceedings of the NAACL HLT 2010 workshop on computational approaches to analysis and generation of emotion in text', Association for Computational Linguistics, pp. 140–146.

Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J. and Smith, N. A. (2011), Part-of-speech tagging for twitter: annotation, features, and experiments, *in* 'Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2', HLT '11, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 42–47.

Go, A., Bhayani, R. and Huang, L. (2009), 'Twitter sentiment classification using distant supervision', *Processing* pp. 1–6.

Gonçalves, P., Araújo, M., Benevenuto, F. and Cha, M. (2013), Comparing and combining sentiment analysis methods, *in* 'Proceedings of the first ACM conference on Online social networks', ACM, pp. 27–38.

Guang., Q., Liu., B., Bu, J. and Chen, C. (2009), Expanding domain sentiment lexicon through double propagation, *in* 'Proceedings of IJCAI'.

Hamouda, A. and Rohaim, M. (2011), Reviews classification using sentiwordnet lexicon, *in* 'World Congress on Computer Science and Information Technology'.

Hatzivassiloglou, V. and McKeown, K. R. (1997), Predicting the semantic orientation of adjectives, *in* 'Proceedings of the 35th Annual Meeting of the ACL and the 8th Conference of the European Chapter of the ACL', New Brunswick, NJ, pp. 174–181.

Heerschop, B., Goossen, F., Hogenboom, A., Frasincar, F., Kaymak, U. and de Jong, F. (2011), Polarity analysis of texts using discourse structure, *in* 'Proceedings of the 20th ACM international conference on Information and knowledge management CIKM'11', ACM, Glasgow UK, pp. 1061–1070.

Hernault, H., Prendinger, H., duVerle, D. and Ishizuka, M. (2010), 'Hilda: A discourse parser using support vector machine classification', *Dialogue and Discourse* **1**(3), 1–33.

Hofmann, T. (1999), Probabilistic latent semantic indexing, *in* 'Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval', ACM, pp. 50–57.

Hogenboom, A., van Iterson, P., Heerschop, B., Frasincar, F. and Kaymak, U. (2011), Determining negation scope and strength in sentiment analysis, *in* 'Proceedings of the IEEE International Conference on Systems, Man and Cybernetics, Anchorage, Alaska, USA, October 9-12, 2011', pp. 2589–2594.

Hu, M. and Liu, B. (2004), Mining and summarizing customer reviews, *in* 'Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining', pp. 168–177.

Ikeda, D., Takamura, H., Ratinov, L.-A. and Okumura, M. (2008), Learning to shift the polarity of words for sentiment classification., *in* 'IJCNLP', pp. 296–303.

Jin, W., Ho, H. H. and Srihari, R. K. (2009), Opinionminer: a novel machine learning system for web opinion mining and extraction, *in* 'Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 1195–1204.

Jurafsky, D. and Martin, J. H. (2015), 'Lexicons for sentiment and affect extraction', Online. Draft Chapter to appear in Speech and Language Processing 3rd edition.
URL: *https://web.stanford.edu/ jurafsky/slp3/21.pdf*

Kamps, J., Marx, M., Mokken, R. J. and de Rijke, M. (2004), Using wordnet to measure semantic orientations of adjectives, *in* 'Proceedings of LREC'.

Kaur, H. J. and Kumar, R. (2015), Sentiment analysis from social media in crisis situations, *in* 'Computing, Communication & Automation (ICCCA), 2015 International Conference on', IEEE, pp. 251–256.

Kennedy, A. and Inkpen, D. (2006), 'Sentiment classification of movie reviews using contextual valence shifters', *Computational Intelligence* **22**, 2006.

Kim, S.-M. and Hovy, E. (2004), Determining the sentiment of opinions, *in* 'Proceedings of COLING', pp. 1367–1373.

Kiritchenko, S. and Mohammad, S. M. (2016), Capturing reliable fine-grained sentiment associations by crowdsourcing and best–worst scaling, *in* 'Proceedings of The 15th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL), San Diego, California'.

Kiritchenko, S., Zhu, X. and Mohammad, S. M. (2014), 'Sentiment analysis of short informal texts', *Journal of Artificial Intelligence Research* pp. 723–762.

Kolchyna, O., Souza, T. T., Treleaven, P. and Aste, T. (2015), 'Twitter sentiment analysis: Lexicon method, machine learning method and their combination', *arXiv preprint arXiv:1507.00955* .

Lee, D. Y. (2001), 'Genres, registers, text types, domains and styles: Clarifying the concepts and nevigating a path through the bnc jungle'.

Lesk, M. (1986), Automatic sense disambiguation using machine readable dictionaries: how to tell a pine cone from an ice cream cone, *in* 'Proceedings of the 5th annual international conference on systems documentation', pp. 24–26.

Li, F., Huang, M. and Zhu, X. (2010), Sentiment analysis with global topics and local dependency., *in* 'AAAI', Vol. 10, pp. 1371–1376.

Li, S., Huang, C.-R., Zhou, G. and Lee, S. Y. M. (2010), Employing personal/impersonal views in supervised and semi-supervised sentiment classification, *in* 'Proceedings of the 48th annual meeting of the association for computational linguistics', Association for Computational Linguistics, pp. 414–423.

Li, Y., Bontcheva, K. and Cunningham, H. (2005), Using uneven margin svm and perceptron for information extraction, *in* 'Proceedings of the Conference on Natural Language Learning (CONLL'05)', pp. 72–79.

Liang, J., Zhou, X., Guo, L. and Bai, S. (2015), Feature selection for sentiment classification using matrix factorization, *in* 'Proceedings of the 24th International Conference on World Wide Web', WWW '15 Companion, International World Wide Web Conferences Steering Committee, Republic and Canton of Geneva, Switzerland, pp. 63–64. **URL:** *http://dx.doi.org/10.1145/2740908.2742741*

Lin, C. and He, Y. (2009), Joint sentiment/topic model for sentiment analysis, *in* 'Proceedings of the 18th ACM conference on Information and knowledge management', ACM, pp. 375–384.

Liu, B. (2010), *Sentiment Analysis and Subjectivity*, second edn, Chapman and Francis, chapter Handbook of Natural Language Processing, pp. 627–666.

Liu, B. (2012), *Sentiment Analysis and Opinion Mining*, Morgan and Claypool Publishers.

Liu, B. (2015), *Sentiment analysis: Mining opinions, sentiments, and emotions*, Cambridge University Press.

Liu, B., Hu, M. and Cheng, J. (2005), Opinion observer: analyzing and comparing opinions on the web, *in* 'Proceedings of the 14th international conference on World Wide Web', WWW '05, ACM, New York, NY, USA, pp. 342–351.
URL: *http://doi.acm.org/10.1145/1060745.1060797*

Liu, S., Li, F., Li, F., Cheng, X. and Shen, H. (2013), Adaptive co-training svm for sentiment classification on tweets, *in* 'Proceedings of the 22nd ACM international conference on Conference on information & knowledge management', ACM, pp. 2079–2088.

Liu, X., Zhang, S., Wei, F. and Zhou, M. (2011), Recognizing named entities in tweets, *in* 'Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1', Association for Computational Linguistics, pp. 359–367.

Mann, W. C. and Thompson, S. A. (1998), 'Rhetorical structure theory: Toward a functional theory of text organization.', *Text* **8**(3), 243–281.

Marcu, D. (2000), 'The rhetorical parsing of unrestricted texts: A surface-based approach', *Computational linguistics* **26**(3), 395–448.

Martineau, J. and Finin, T. (2009), Delta tfidf: An improved feature space for sentiment analysis., *in* 'ICWSM'.

Maynard, D. and Greenwood, M. A. (2014), Who cares about sarcastic tweets? investigating the impact of sarcasm on sentiment analysis., *in* 'LREC', pp. 4238–4243.

Maynard, D. and Hare, J. (2015), Entity-based opinion mining from text and multimedia, *in* 'Advances in Social Media Analysis', Springer, pp. 65–86.

Mei, Q., Ling, X., Wondra, M., Su, H. and Zhai, C. (2007), Topic sentiment mixture: modeling facets and opinions in weblogs, *in* 'Proceedings of the 16th international conference on World Wide Web', ACM, pp. 171–180.

Melville, P., Gryc, W. and Lawrence, R. D. (2009), Sentiment analysis of blogs by combining lexical knowledge with text classification, *in* 'Proceedings of the 15th ACM SIGKDD international conference on Knowledge discovery and data mining', ACM, pp. 1275–1284.

Mitchell, T. M. (1997), *Machine Learning*, McGraw Hill.

Mohammad, S. (2012), #emotional tweets, *in* '*SEM 2012: 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 (SemEval 2012)', Association for Computational Linguistics, Montréal, Canada, pp. 246–255.
URL: *http://www.aclweb.org/anthology/S12-1033*

Mohammad, S. M. and Kiritchenko, S. (2015), 'Using hashtags to capture fine emotion categories from tweets', *Computational Intelligence* **31**(2), 301–326.

Mohammad, S. M., Kiritchenko, S. and Zhu, X. (2013), Nrc-canada: Building the state-of-the-art in sentiment analysis of tweets, *in* 'Proceedings of the seventh international workshop on Semantic Evaluation Exercises (SemEval-2013)', Atlanta, Georgia, USA.

Mohammad, S. M. and Turney, P. D. (2013), 'Crowdsourcing a word-emotion association lexicon', **29**(3), 436–465.

Mudinas, A., Zhang, D. and Levene, M. (2012), 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', WISDOM

'12, ACM, New York, NY, USA, pp. 5:1–5:8.
URL: *http://doi.acm.org/10.1145/2346676.2346681*

Muhammad, A., Wiratunga, N. and Lothian, R. (2014), A hybrid sentiment lexicon for social media mining, *in* 'Tools with Artificial Intelligence (ICTAI), 2014 IEEE 26th International Conference on', IEEE, pp. 461–468.

Muhammad, A., Wiratunga, N., Lothian, R. and Glassey, R. (2013*a*), Contextual sentiment analysis in social media using high-coverage lexicon, *in* 'Research and Development in Intelligent Systems XXX', Springer, pp. 79–93.

Muhammad, A., Wiratunga, N., Lothian, R. and Glassey, R. (2013*b*), Domain-based lexicon enhancement for sentiment analysis., *in* 'SMA@ BCS-SGAI', pp. 7–18.

Mukherjee, S. and Bhattacharyya, P. (2012), Sentiment analysis in twitter with lightweight discourse analysis, *in* 'Proceedings of the 24th International Conference on Computational Linguistics (COLING 2012)'.

Mukras, R., Wiratunga, N. and Lothian, R. (2007), Selecting bi-tags for sentiment analysis of text, *in* 'Proceedings of the Twenty-seventh SGAI International Conference on Innovative Techniques and Applications of Artificial Intelligence'.

Mullen, T. and Collier, N. (2004), Sentiment analysis using support vector machines with diverse information sources, *in* 'Proceedings of the Conference on Empirical Methods on Natural Language Processing', pp. 412–418.

Munezero, M., Montero, C. S., Sutinen, E. and Pajunen, J. (2014), 'Are they different? affect, feeling, emotion, sentiment, and opinion detection in text', *Affective Computing, IEEE Transactions on* **5**(2), 101–111.

Niwa, Y. and Nitta, Y. (1994), Co-occurrence vectors from corpora vs. distance vectors from dictionaries, *in* '15th International Conference On Computational Linguistics'.

O'Connor, B., Balasubramanyan, R., Routledge, B. R. and Smith, N. A. (2010), 'From tweets to polls: Linking text sentiment to public opinion time series.', *ICWSM* **11**(122-129), 1–2.

Ohana, B., Delany, S. J. and Tierney, B. (2012), A case-based approach to cross domain sentiment classification, *in* 'Case-Based Reasoning Research and Development', Springer, pp. 284–296.

Ohana, B. and Tierney, B. (2009), Sentiment classification of reviews using sentiwordnet, *in* '9th IT&T Conference, Dublin, Ireland'.

Ortony, A., Clore, G. L. and Collins, A. (1990), *The cognitive structure of emotions*, Cambridge university press.

Osgood, C. E., Suci, G. J. and Tannenbaum, P. H. (1957), *The Measurement of Meaning*, University of Illinois Press, Urbana, IL.

Pak, A. and Paroubek, P. (2010), Twitter as a corpus for sentiment analysis and opinion mining., *in* 'LREC', Vol. 10, pp. 1320–1326.

Paltoglou, G. (2014), Sentiment analysis in social media, *in* 'Online Collective Action', Springer, pp. 3–17.

Paltoglou, G. and Thelwall, M. (2010), A study of information retrieval weighting schemes for sentiment analysis, *in* '48th Annual Meeting of the Association for Computational Linguistics (ACL 2010)', pp. 1386–1395.

Paltoglou, G. and Thelwall, M. (2012), 'Twitter, myspace, digg: Unsupervised sentiment analysis in social media', *ACM Transactions on Intelligent Systems and Technology* **3**(4).

Pan, S. J., Ni, X., Sun, J.-T., Yang, Q. and Chen, Z. (2010), Cross-domain sentiment classification via spectral feature alignment, *in* 'Proceedings of the 19th international conference on World wide web', ACM, pp. 751–760.

Pan, S. J. and Yang, Q. (2010), 'A survey on transfer learning', *Knowledge and Data Engineering, IEEE Transactions on* **22**(10), 1345–1359.

Pang, B., Lee, L. and Vaithyanathan, S. (2002), Thumbs up? sentiment classification using machine learning techniques, *in* 'Proceedings of the Conference on Empirical Methods on Natural Language Processing'.

Parrott, W. G. (2001), *Emotions in social psychology: Essential readings*, Psychology Press.

Pennebaker, J. W., Booth, R. J. and Francis, M. E. (2007), 'Linguistic inquiry and word count: Liwc [computer software]', *Austin, TX: liwc. net* .

Plutchik, R. (1962), 'The emotions: Facts, theories, and a new model.', *Random House* .

Polanyi, L. and Zaenen, A. (2004), *Contextual valence shifters*, Vol. 20, Springer, Dordrecht, The Netherlands.

Poria, S., Gelbukh, A., Cambria, E., Hussain, A. and Huang, G.-B. (2014), 'Emosenticspace: A novel framework for affective common-sense reasoning', *Knowledge-Based Systems* **69**, 108–123.

Potts, C. (2011*a*), On the negativity of negation, *in* N. Li and D. Lutz, eds, 'Proceedings of Semantics and Linguistic Theory 20', CLC Publications, Ithaca, NY, pp. 636–659.

Potts, C. (2011*b*), 'Sentiment symposium tutorial: Stemming', Online.
URL: *http://sentiment.christopherpotts.net/stemming.html*

Prabowo, R. and Thelwall, M. (2009), 'sentiment analysis: A combined approach', *Journal of Informetrics* **3**(2), 143–157.

Rahman, M. (2009), Representation and Learning Schemes for Sentiment Analysis, PhD thesis, The Robert Gordon University.

Rao, Y., Li, Q., Mao, X. and Wenyin, L. (2014), 'Sentiment topic models for social emotion mining', *Information Sciences* **266**, 90–100.

Read, J. (2005), Using emoticons to reduce dependency in machine learning techniques for sentiment classification, *in* 'Proceedings of the ACL Student Research Workshop', ACLstudent '05, Association for Computational Linguistics, Stroudsburg, PA, USA, pp. 43–48.

Reitan, J., Faret, J., Gambäck, B. and Bungum, L. (2015), Negation scope detection for twitter sentiment analysis, *in* '6TH WORKSHOP ON COMPUTATIONAL APPROACHES TO SUBJECTIVITY, SENTIMENT AND SOCIAL MEDIA ANALYSIS WASSA 2015', p. 99.

Riloff, E., Patwardhan, S. and Wiebe, J. (2006), Feature subsumption for opinion analysis, *in* 'Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing (EMNLP-06)'.

Ritter, A., Clark, S., Mausam and Etzioni, O. (2011), Named entity recognition in tweets: an experimental study, *in* 'Proceedings of the Conference on Empirical Methods in Natural Language Processing', Association for Computational Linguistics, pp. 1524–1534.

Sani, S. (2014), The Role of Semantic Indexing for Text Classification, PhD thesis, Robert Gordon University.

Sani, S., Wiratunga, N., Massie, S. and Lothian, R. (2013), Sentiment classification using supervised sub-spacing, *in* 'Research and Development in Intelligent Systems XXX', Springer, pp. 109–122.

Scherer, K. R. (2000), 'Psychological models of emotion', *The neuropsychology of emotion* **137**(3), 137–162.

Sharma, A. and Dey, S. (2012), 'Performance investigation of feature selection methods and sentiment lexicons for sentiment analysis', *IJCA Special Issue on Advanced Computing and Communication Technologies for HPC Applications* **3**, 15–20.

Soricut, R. and Marcu, D. (2003), Sentence level discourse parsing using syntactic and lexical information, *in* 'Human Language Technology and North American Association for Computational Linguistics Conference (HLT/NAACL 2003)', Association for Computational Linguistics, 149-156.

Steyvers, M. and Griffiths, T. (2007), 'Probabilistic topic models', *Handbook of latent semantic analysis* **427**(7), 424–440.

Stone, P. J., Dexter, D. C., Marshall, S. S. and Daniel, O. M. (1966), *The General Inquirer: A Computer Approach to Content Analysis*, MIT Press, Cambridge, MA.

Strapparava, C. and Valitutti, A. (2004), Wordnet-affect: an affective extension of wordnet, *in* 'In Proceedings of the 4th International Conference on Language Resources and Evaluation', pp. 1083–1086.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. and Stede, M. (2011), 'Lexicon-based methods for sentiment analysis', *Computational Linguistics* **37**, 267–307.

Taboada, M., Voll, K. and Brooke, J. (2008), Extracting sentiment as a function of discourse structure and topicality, Technical report, Simon Fraser University,.

Tan, S. and Zhang, J. (2008), 'An empirical study of sentiment analysis for chinese documents', *Expert Systems with Applications* **34**(4), 2622–2629.

Thelwall, M., Buckley, K. and Paltoglou, G. (2012), 'Sentiment strength detection for the social web', *Journal of the American Society for Information Science and Technology* **63**(1), 163–173.

Thelwall, M., Buckley, K., Paltoglou, G., Cai, D. and Kappas, A. (2010), 'Sentiment strength detection in short informal text', *Journal of the American Society for Information Science and Technology* **61**(12), 2444–2558.

Thet, T. T., Na, J.-C., Khoo, C. S. and Shakthikumar, S. (2009), Sentiment analysis of movie reviews on discussion boards using a linguistic approach, *in* 'Proceedings of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion', TSA '09, ACM, New York, NY, USA, pp. 81–84.
URL: *http://doi.acm.org/10.1145/1651461.1651476*

Tomkins, S. S. (1962), 'Affect, imagery, consciousness: Vol. i. the positive affects.'.

Tsatsaronis, G. and Panagiotopoulou, V. (2009), A generalized vector space model for text retrieval based on semantic relatedness, *in* 'Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics: Student Research Workshop', Association for Computational Linguistics, pp. 70–78.

Turney, P. D. (2002), Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews., *in* 'Proceedings of the Annual Meeting of the Association for Computational Linguistics', pp. 417–424.

Turney, P. D. and Pantel, P. (2010), 'From frequency to meaning: Vector space models of semantics', *Journal of Artficial Intelligence Research* **37**, 141–188.

van Rijsbergen., C. J. (1979), *Information Retrieval*, Butterworth-Heinemann.

Vapnik, V. N. (1998), *Statistical learning theory*, Vol. 1, Wiley New York.

Wang, S., Li, D., Song, X., Wei, Y. and Li, H. (2011), 'A feature selection method based on improved fisher's discriminant ratio for text sentiment classification', *Expert Systems with Applications* **38**(7), 8696–8702.

Whitelaw, C., Garg, N. and Argamon., S. (2005), Using appraisal groups for sentiment analysis, *in* '14th ACM International Conference on Information and Knowledge Management (CIKM 2005)', pp. 625–631.

Wiebe, J. (1994), 'Tracking point of view in narrative', *Computational Linguistics* **20**(2), 233–287.

Wiebe, J. and Cardie, C. (2005), Annotating expressions of opinions and emotions in language. language resources and evaluation, *in* 'Language Resources and Evaluation (formerly Computers and the Humanities', p. 2005.

Wilson, T., Hoffmann, P., Somasundaran, S., Kessler, J., Wiebe, J., Choi, Y., Cardie, C., Riloff, E. and Patwardhan, S. (2005), Opinionfinder: A system for subjectivity analysis, *in* 'Proceedings of hlt/emnlp on interactive demonstrations', Association for Computational Linguistics, pp. 34–35.

Wiratunga, N., Koychev, I. and Massie, S. (2004), Feature selection and generalisation for retrieval of textual cases, *in* 'Advances in Case-Based Reasoning', Springer, pp. 806–820.

Xia, R. and Zong, C. (2010), A pos-based ensemble model for crossdomain sentiment classification, *in* 'in Proceedings of the 5th International Joint Conference on Natural Language Processing'.

Yoshida, Y., Hirao, T., Iwata, T., Nagata, M. and Matsumoto, Y. (2011), Transfer learning for multiple-domain sentiment analysis - identifying domain dependent/independent word polarity, *in* 'Twenty-Fifth AAAI Conference on Artificial Intelligence'.

Zhou, Z., Zhang, X. and Sanderson, M. (2014), Sentiment analysis on twitter through topic-based lexicon expansion, *in* 'Databases Theory and Applications', Springer, pp. 98–109.