Sentiment Analysis: Beyond Polarity

Thesis Proposal

Phillip Smith
pxs697@cs.bham.ac.uk
School of Computer Science
University of Birmingham

Supervisor:
Dr. Mark Lee

Thesis Group Members:
Professor John Barnden
Dr. Peter Hancox

October 2011

Contents

List of Figures

Abstract

Sentiment analysis has demonstrated that the computational recognition of emotional expression is possible. However, success has been limited to a number of coarse-grained approaches to human emotion that have treated the emotional connotations of text in a naive manner: as being either positive or negative. To overcome the problem, this research proposes that we use a fine-grained category system that is representative of more granular emotions. This research will explore the integration of emotional models into machine learning techniques that will attempt to go beyond the current state of the art. The effects of such a computational model of emotion, and how this can be implemented within machine learning approaches to sentiment analysis, will form the grounds for this investigation.

Chapter 1

Introduction

Old man - don't let's forget that the little emotions are the great captains of our lives, and that we obey them without knowing it.

-Vincent Van Gogh, *The Letters*

1.1 Beyond Polarity

Language is a naturally occurring phenomenon that utilises verbal channels to express intent and meaning. Part of this expressed meaning is the emotional connotations of a message. This affects both the speaker and the listener in its communication. One of the ways in which emotion is expressed during communication is through the words of a speaker, and this message is duly comprehended and interpreted by the recipient, whereby the message is the carriage for the emotional intent. In this process, cognitively, the speaker creates the message, and analogously, the recipient cognitively decodes this message, otherwise known as understanding. Herein lies the problem. The cognitive processes that operate to generate and interpret emotional meaning are not clear. Do the emotions motivate the actions, or actions the emotions? The assumption is made that a level of intelligence is required in the two separate cognitive actions that are performed here. Sentiment analysis aims to computationally model the cognitive tasks required in decoding the emotional intent of a message. By computationally modelling the conversion of a message to its emotional meaning we gain an insight into how human cognition can be defined in a mechanical format. Currently the modelling of this process has taken a naive approach to emotion, so this thesis proposes that we should go beyond polarity when modelling such cognitive processes.

For the above reasons, sentiment analysis is an important area in computational linguistics. The past decade has seen the rapid growth of this field of research due to two main factors: First, a significant increase in computational processing power. Second, through user-generated content on the world wide web, internet users have contributed a large number of textual documents with emotional connotations. In combination, researchers have been able to create, annotate and distribute corpora which otherwise would have restricted the reach of their work, and thereby the progress of the wider community. The availability of these corpora has led to research being carried out on the use of machine learning techniques to model the valence of emotions that are expressed in the documents. The research to date has tended to focus on polarity identification in text, rather than using a finer-grained category system for classification, such as human emotions. Where the literature has attempted to recognise emotion in text, the results indicate that this challenge is not a trivial task. Consequently, the purpose of this research is to develop a computational model that can reliably identify emotions in textual documents through the use of machine learning algorithms.

We spend ever more time interacting with computers, and quite often the routine is a slight rigmarole due to the lack of basic understanding a computer exhibits. Recognition of emotion is the first step in accomplishing such a goal, but sentiment analysis is not limited to just this. Due to the advent of an internet that has become quicker and easier to browse, we are now able to access a wealth of knowledge which can help guide some of our day to day decisions. Retrieval of information has become less of an issue, but the load that we are bombarded with has. We want to know what is relevant to us, in a clear and concise way. Sentiment analysis holds the key in bringing us this emotionally relative data.

This research will build upon current annotation schemes for emotion in text, and will optimise machine learning techniques in light of this improved annotation schema. The research will contribute knowledge towards a computational model of emotion that can recognise, understand and relay the emotional connotations of documents from a variety of domains in a robust manner.

1.2 Research Background

The aim of this research is to develop machine learning techniques to recognise the emotions that are interwoven into the text of a document. This problem originates as a combination of a Machine Learning, Information Retrieval, Cognitive Science and Natural Language Processing challenge. By combining these fields, the research problem of computationally identifying emotional expression within a given document set exists. In working towards a solution to this problem, the intention is to contribute knowledge to the creation of an intelligent, emotionally sensitive computer system, which could reason about problems such as the following one:

1.2.1 Example Problem

The National Health Service plays a crucial role in our lives. For some, it is a fantastic experience, which aids them tremendously, whilst for others, the situation is unpleasant. Fortunately, websites such as Patient Opinion[1] enable patients to post feedback regarding their experiences with the UK health services. The people giving feedback are doing so because they feel inclined in some way to express their emotions about parts of the service which they have felt strongly about, and so they contribute content to a valid emotional dataset. Countless people are treated everyday, and a vast quantity of feedback is left for the health services to observe and act upon. For those processing this data, the task could be approached in a more efficient way by using sentiment analysis to categorize the comments into emotional categories, so they can be dealt with in an appropriate manner. Comments that express sadness could be prioritised over those expressing joy, as something has caused the patient to feel this way, and could be corrected before another patient is affected in such a way. If surprise was expressed alongside sadness, then an even higher priority could be assigned to such comments. By using sentiment analysis, the health services could ensure that comments were not missed that could make the difference to the running of their hospitals, and fundamentally, the well-being of patients.

[1] http://www.patientopinion.org.uk

The problem posed in sentiment analysis is the identification of emotional expressions, and the emotion that is being expressed. For example a comment giving feedback about the UK health services could say:

> *"I was not treated until late in the evening, and thought that the doctor would not come."*

The problem here is that there are no words which explicitly denote an emotion, but the emotions of fear and sadness can be attributed to this statement due to the situation which is described. Identifying the context to this problem can come through learning about similar sentences and the emotion that they expressed, and through this learning process, a solution to the problem can be created. This thesis will discuss possible solutions to this problem, but before we cover those, sentiment analysis must be formally defined, and background concerning linguistic expression given.

1.2.2 Definition of Sentiment Analysis

In the past there has been confusion surrounding the terminology of this field. Quite often the challenges of polarity recognition and emotion identification have been described using the same term, sentiment analysis. This thesis seeks to go beyond polarity-based identification, and focus on finer-grained emotional recognition. Therefore in this research, the term sentiment analysis will be used in a broader fashion.

The meaning of the term sentiment analysis is quite inclusive. From a non-computational viewpoint, reading a film review and deciding you want to see it because of what the reviewer has written is a form of sentiment analysis. However, for this work, it can be thought of in the following way:

> Sentiment Analysis is the computational evaluation of documents to determine the fine-grained emotions that are expressed.

or more formally:

> Given a document d from a document set D, computationally assign emotional labels, e, from a set of emotions E in such a way that e is reflective of d's expressed emotion or emotions at the appropriate level of expression.

It will be of use to first define what is meant by a document in this context. For a general text classification problem, Lewis (1998) describes a document as simply being a single piece of text. This is stored on a machine as a character sequence, where these characters in combination embody the text of a written natural language expression. Sentiment analysis builds upon the problem of text classification, which makes the definition of a document given by Lewis (*ibid*) relevant to this domain. It goes beyond naive text classification however by seeking to determine the expressed emotion in a document, which can occur at multiple expressive levels.

Historical definitions of sentiment analysis traditionally define it as recognising if the subjects of a text are described in a positive or negative manner. It is often referred to as determining the *polarity* of a text. By limiting categorisation through use of a small, closed set of possible classes that a document can be assigned to, this definition intelligently restricts the set of categories to either positive or negative (Turney, 2002), with the occasional use of neutrality. This differs drastically from the definition which will be used in this work, which concentrates on textual emotion recognition. In this the option of a variable set size is introduced. This is due to the range of possible emotions that can be linked to the text of a document. With a limited set to work with, it could be argued that polarity identification is a simpler task than textual emotion recognition. However, both areas struggle with the challenge posed by the written language of emotion, in particular its expressiveness.

1.2.3 Linguistic Expression of Emotion

In any form of written text that wishes to convey an emotion, there are two significant modes of expressing this phenomenon in language. The first is the *explicit* communication of emotion. Strapparava et al. (2006) refer to this as the direct affective words of a text. An example of this is:

> *"What a wonderful policy."*

This sentence explicitly describes through use of the word *'wonderful'* that the speaker's attitude towards the policy is positive. Therefore in sentiment analysis, if this sentence was regarded as the whole document for classification, with no external documents affecting its context, it could be assigned the positive label (more on document annotation will be discussed in Chapter 3). Whitelaw et al. (2005) demonstrate that only identifying the explicit features of a document yields favourable results, but by doing this the assumption is made that direct affective words are of more importance than other forms of linguistic expression in developing intelligent systems due to the favourable patterns of identification they yield. This should not be the case as *implicit* linguistic expressions can bear just as much, if not more emotional information:

> *"Jesus Christ!"*

Previous approaches to sentiment analysis may have suggested that due to the lack of emotional words the sentence here is inherently neutral. However, this sentence could refer to a number of scenarios, and is contextually ambiguous due to the emotional nature that this phrase can communicate when voiced. This sentence could provoke a positive emotion, as it could be uttered under the context of a positive event transpiring. However, this is not the only emotion that could be be associated with it, as can be revealed by further cogitating the sentence. This could also have been uttered under a negative context, whereby a tragedy may have occurred. Due to this, the desired emotional connotation would be a negative one. Strapparava et al. (2006) refer to this type of expression as containing the indirect affective words of a text.

The above example displays the difficulty of deducing an implied set of emotions from text, as either an a priori knowledge of the situation is required, or a mechanism to understand the underlying semantics of the document. If we take the two words of the sentence independently, no emotional information can trivially be deduced, and a religious reference could be associated with this utterance. Yet, if we take the words in combination, they probe the reader for a background knowledge that is crucial in deducing the sentence's emotional connotations. This phenomenon is common in natural language, especially English, where the emotional meaning of a document is subtler than it may first seem. This thesis must attempt to overcome the issue of identifying an implicit emotion, so research questions will be asked which aim to explore possible solutions to overcome this problem.

1.3 Research Questions

In light of the overview of this work, my proposed research will aim to address three major questions:

RQ1 Which model of emotion is best suited to sentiment analysis?

- (a) Are the emotions expressed in text suited to an ontology?

RQ2 How should documents be annotated with emotional information?

- (a) What spans of text should be annotated?
- (b) How will structural information be captured?
- (c) How will the different forms of expression be captured?

RQ3 Which machine learning techniques are most suited to the task of textual emotion recognition?

The first question is the motivating question of my research. It is a high level question, so it has been divided into a sub-question in order to produce a workable contribution to a solution. The first question (**RQ1**) must be asked as this thesis is not seeking to redefine the wheel of literature available concerning models of emotion. This thesis aims to critically assert which currently proposed model, if any, would be suitable to define the emotions that are held in text. This deviates from much of the scientific literature on emotion research, which tends to focus on modelling emotion given facial expressions (Picard, 1995; Russell, 1994) or speech data (Cowie et al., 2001; Dellaert et al., 1996; Murray & Arnott, 1996).

$RQ1_{(a)}$ questions whether a structure can be imposed on the emotions exhibited in text. If this is the case, it will be of interest to investigate whether a combination of emotions, or the combination of the relationships between them, could lead to a further emotion being derived, and if so how this system works. This investigation can only gauge so much from a literature review, so through experimentation this is a vital part of **RQ1**.

The following two research questions, **RQ2** and **RQ3**, branch from the main research question. They further expand on the issue of the emotions that are typically expressed in a document, and in doing so angle the research in a computational direction. **RQ3** considers the machine learning approaches to textual emotion recognition. There are two main classes of machine learning algorithms, supervised and unsupervised. To observe and thoroughly experiment with each approach that the two classes consist of is beyond the scope of this research, however a subset of the approaches will be considered in working towards a solution to this question.

RQ2 concerns the annotation framework which should be created in order to maximize the output of the algorithms. The question of how a document should be annotated with emotional information has been divided into specific questions where it is felt the literature does not provide a sufficient solution to the problem.

1.4 Hypotheses

The research questions raise the following hypotheses, which will form the basis for experimentation in this work:

Hypothesis 1 - (RQ1) Emotions can be structured as a tree, with valenced categories acting as the root node, and fine-grained emotional categories at the leaves.

Hypothesis 2 - (RQ2) Expressed emotion is not a sum of its parts, and therefore documents should be annotated across various levels to capture this expression.

Hypothesis 3 - (RQ3) Supervised machine learning techniques in combination with a dependency structure are most suited to sentiment analysis.

This introduction to the thesis has introduced a basis for the formation of these hypotheses, and the following chapters of this proposal will justify their inception.

1.5 Proposal Structure

These three major research questions and hypotheses are put forward on the basis that current work has not provided a solution to the problem of computationally identifying emotion contained in text, and their reasoning will be discussed in the remainder of this proposal. The remaining sections of this proposal are structured as follows:

Chapter 2 - Emotion will discuss the literature put forward regarding models of emotion, in particular highlighting those that are viewed as computational models of emotion.

Chapter 3 - Annotation will review the work on annotation of documents with emotional content, and highlight schema that are effective from the literature. This chapter will also review the datasets that have been used in various experiments, and discuss the outcomes of experimentation with a recently introduced corpus of suicide notes.

Chapter 4 - Machine Learning will include a review of the literature of how machine learning techniques have been applied to sentiment analysis.

Chapter 5- Methodology & Evaluation will describe the work to be carried out, and how this work will contribute to knowledge.

Chapter 6 - Timetable will outline the work plan for this research.

Chapter 2

Emotion

In researching how a machine learning system can successfully identify emotions in a document, we must first grasp the essence of what an emotion is, and how it can be defined. This chapter will outline the relevant literature, and cast light on the range of definitions that have been put forward over the years. This will show that while emotions come as second-nature to many of us, requiring little thought in their production, the models that have been proposed vary greatly in two main ways. The first is the types of emotion which the model consists of. Despite emotions being part of our everyday lives, there is little agreement on what the types of emotion that we express are. The second is the dimensionality of emotions. Some believe that emotions can differ in intensity, and therefore can be seen as dimensional notions which can be assigned scalable values. With these models in mind, this chapter will conclude with a summary of the emotional models presented, and highlight which ones should be investigated further in this research.

2.1 Differences in Definition

The definition of emotion is one which is unclear, despite being a phenomenon which occurs frequently in our lives. The problem with defining emotion is that mentally it is experienced by so many, yet physically and verbally the forms of expression vary. This can make recognition challenging if common forms of expression are to be relied on. If however we are to successfully recognise emotions, particularly in text, a universal model must exist. Unfortunately this is something that is difficult to define. Researchers have put forward definitions in an attempt to pinpoint what emotion is, and to describe the sequence of events that could combine in the inception of one. By doing this they have also attempted to describe just what types of emotion there are, and as we will note, consensus on this issue is lacking.

Kleinginna & Kleinginna (1981) identify this lack of agreement within the literature. They attempt to summarize and narrow down the various definitions into a more concise description of what an emotion is. By drawing from 92 definitions and 9 skeptical comments in the literature, they observe the themes of the statements, and group them into eleven distinct categories.

- *Affect*: The feelings of excitement/depression or pleasure/displeasure and the arousal levels that are invoked.

- *Cognition*: Appraisal and labelling processes.

- *External emotional stimuli*: Triggers of emotion.
- *Physiological mechanisms*: These align the dependence of emotions on biological functions.
- *Emotional behaviour*: Expressions of emotion.
- *Disruptive*: Disorganizing attributes of emotion.
- *Adaptive*: Functional attributes of emotion.
- *Multi-aspect nature of emotion*: The combination of a number of these categories within a definition.
- *Differences from other processes*: The differences highlighted were between emotions and other affective processes.
- *Overlap between emotion and motivation*: Affect is central to our primary motivations
- *Skeptical*: Definitions that highlight a dissatisfaction with the lack of agreement.

In this research, the objective is to adapt and apply a model of emotion that can be implemented within a computational model. Owing to this we can dismiss a number of these themes as not being contributing factors of this work, despite the fact that some may fit well in other domains such as psychology or biology. The groupings that are of particular interest to this study are those of affect and external stimuli.

The work of Plutchik (1980a) argues that external stimuli is a pivotal factor in defining emotion. He believes that emotion can be defined in the following way:

1. *Emotions are generally aroused by external stimuli*
2. *Emotional expression is typically directed toward the particular stimulus in the environment by which it has been aroused.*
3. *Emotions may be, but are not necessarily or usually, activated by a physiological state.*
4. *There are no 'natural' objects in the environment (like food or water) toward which emotional expression is directed.*
5. *An emotional state is induced after an object is seen or evaluated, and not before.*

While some of these points are pertinent to this research, there are some which are irrelevant. It may be the case that physiological states play an important role in the activation of emotion, as Plutchik (1980a) notes in point 3 of his definition, but this study will not observe the role of bodily organs in emotional functions. It is an interesting problem, with many biological implications, but unfortunately it is beyond the scope of this research. Furthermore, point 4 may have contained relevance at the time of writing, but we now live in an age where people express and share the most mundane of opinions and emotions towards seemingly sentient objects. Food is one example of this, with countless documents being returned when searching social media sites such as Twitter for content regarding this topic.

Nonetheless, this description introduces a directional concept into the definition of emotion, which must be upheld in its computational modelling. By stating that emotion must have an external stimuli or source, a paradigm is created that is suited to computation. It implies that emotion can be attributed to a cause, therefore making it a reaction. This gives emotion a context for existence. Accordingly, in the verbal expression of emotion, a context will be communicated, which will aid in the recognition of the emotional utterance, and just what emotion is being communicated.

Other descriptions of emotion such as those given by Ekman (1992) also share the view that external stimuli play an important role in how it is defined.

2.2 Basic Emotions

An important link between the articles of Plutchik (1980a) and Ekman (1992), is the idea that there exists a set of basic emotions which dictate the fundamental reactions that we should exhibit. The idea of a small set of fundamental emotions is a frequently used concept in the literature (Mowrer, 1960; Oatley & Johnson-Laird, 1987; Weiner & Graham, 1984). However, just as there is a lack of agreement in the definition of emotion, there is also a lack of agreement as to what emotions should form this basic set; which poses the question of its role and existence.

In their paper questioning just how *'basic'* a basic emotion is, Ortony & Turner (1990) highlight similarities and differences that exist in the literature on this topic. The table in Appendix A from Ortony & Turner (*ibid*) highlights the idiosyncrasies of what researchers over the past two centuries have believed are included in the fundamental set of emotions. The differences are clear. From the work of Weiner & Graham (1984), postulating that happiness and sadness are the only basic emotions, and the proposal of Mowrer (1960) that pain and pleasure constitute the primary set, to the argument from Oatley & Johnson-Laird (1987) that includes anger, disgust and anxiety in the basic set; little consistency is exhibited.

In the work of Ortony et al. (1988), the nature of basic emotions is challenged. The notion of the universality of basic emotions is disputed, and from this the question of whether emotions can blend to form more complex, secondary emotions is brought to light. These explorative questions bring doubt upon the concept of basic emotions, and Ortony & Turner (1990) voice this concern succinctly by making a comparison between emotions as a whole, and natural languages. They argue that while there are many human languages, with the possibility to create many more languages, linguists do not seek to define language as a whole by using a few languages which they view as *fundamental*. By arguing the notion that there are basic structures in all natural languages, such as syntax and phonology, Ortony & Turner (*ibid*) hypothesise that emotions themselves are not basic, but can be constructed from basic elements.

With this hypothesis in mind, Ortony et al. (1988) reduce the first step in recognising emotions, and thereby its definition, by stating the following:

> *Valenced reactions are the essential ingredients of emotions in the sense that all emotions involve some sort of positive or negative reaction to something or other.*

This conjecture moves away from the idea that emotions are basic, to the notion that emotions are differentiated forms of two high level categories, positive and negative. Just as Plutchik's description attributes emotions to a reaction, this also implies that emotions start life as simple affective response to an event or object, and differentiate in such a way that an identifiable emotion is formed. Therefore, instead of viewing emotions as being either basic or non-basic, the emphasis has now shifted onto how different an emotion is from a set of valenced reactions. Another implication of the theory that Ortony et al. (1988) outline is that emotions must be linked to an initial valenced reaction, and that if this is not the case, the emotion in question is not genuine. By assuming this, the theory dismisses possible emotions that are merely a description of some real world state, and have no emotional connotations within their model. This sits well with the research questions of my work, as over the past decade research has focused on the identification of these valenced categories in sentiment analysis (Blitzer et al., 2007; Dave et al., 2003; Turney, 2002). However this research aims to go beyond valenced categories, and test this hypothesis in a computational domain.

2.3 Secondary Emotions

Although Ortony & Turner (1990) argue against the notion of basic emotions, it could be seen that the difference between an emotion and its valenced reaction is comparable to what Plutchik (1997) describes as *dyadic*, or secondary emotions. The model of emotion proposed by Plutchik (*ibid*) is the circumplex model of emotion. This used a set of eight basic emotions (listed in Appendix A) that were represented in a dimensional way, such that in combination with one another dyadic emotions are created. Figure 2.1, created by Drews (2007) displays a selection of proposed combinations and their resulting dyads. Computationally this has a number of implications due to the fact that if we hypothesise that emotions are multi-faceted, and more than one emotion can be attributed to a text, this can reveal implicit emotions that common feature selection techniques may not have realised were present.

Taking the idea of dyadic emotions, in combination with the previously outlined claim of Ortony et al. (1988), presents an interesting amalgamation of ideas to which a computational model of emotion can be applied. It leads to the hypothesis that if we are presented with a positive or negative reaction in a document, we can automatically, using machine learning techniques, determine a fine-grained emotion associated with it. This in turn, is the basis for an emotional ontology.

2.4 The Role of Affect

Affect is an essential condition of emotion that Kleinginna & Kleinginna (1981) define as the *"feelings of arousal level and pleasure/displeasure"*. They note that of the definitions of emotion that they reviewed, 23 mention affect as the focus of their definition, while 44 used it as a secondary emphasis, which outweighs the emphasis of other themes significantly. Even though affect is a recurring theme in many of the definitions that were reviewed, they still question if it should be thought of as the prevalent feature of emotion.

Picard (1995) argues that affect is the central emphasis of a definition of emotion, going so far as to name a field of computing that should behave in an emotional way, *Affective Computing.* Picard (1995) presents the argument that a simple model of emotion, with a basic set of categories is fitting for the task of emotion recognition. The comparison is made with the simple emotions that a baby displays. It does not seem appropriate however to compare the intelligence of a young human with that of a questionably sentient machine. Despite this Picard goes on to describe an affective state model, that is illustrated in the figure on the following page (Figure 2.2).

This figure shows the attempts made in defining a transitional model of emotion that is fit for computation. This model can be viewed as an evolutionary model, with transitions between emotional states denoted by the vertices connecting the nodes. Conditional probabilities, similar to those seen in Bayes' theorem, are assigned to each vertex, which relates to how one emotional state can permute into another. This could be a fairly robust model of emotion, due to the way in

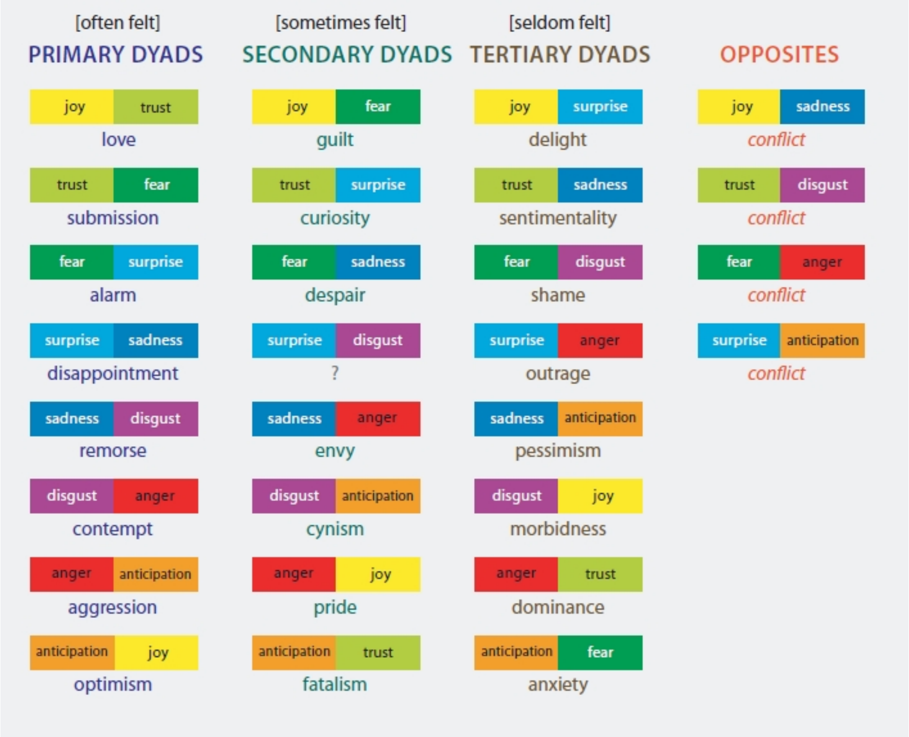


Figure 2.1 Dyadic implications (Drews, 2007)

which it can be implemented computationally by a machine learning algorithm. One issue with this approach is that it requires sufficient empirical data to train the model with, so state transitions can be efficiently approximated. A second issue is which emotions should be used, as this is not defined by Picard (1995). This however is a point which my research will develop.

2.5 Summary

This chapter has looked at the definition of emotion, and has highlighted certain models that are suitable for computation. The diversity in definitions has been emphasized, which has shown that what we take for granted is extremely difficult to define. A number of models use a set of basic emotions as a central argument, yet the work of Ortony & Turner (1990) has demonstrated that this is a questionable position to take. Despite this, we must still give weight to models which work on a dimensional scale, and view secondary emotions being produced as a result of various basic models being exhibited. Finally this chapter has looked at the work of Picard (1995) in defining the field of affective computing, which has laid the way for computational emotion research. It is hoped that by introducing these models, researchers are able to use these approaches to annotation in machine learning algorithms. An essential part of supervised machine learning systems is the role of annotated data, which forms the input that teaches the system what traits of a dataset

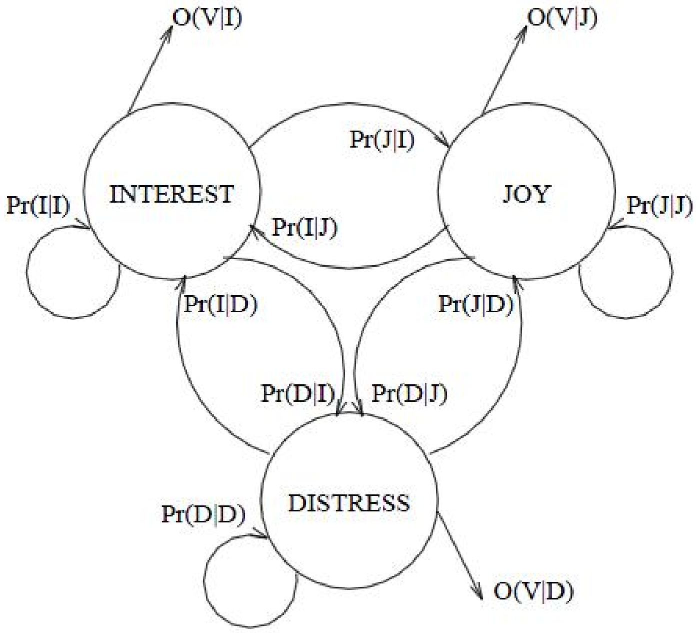


Figure 2.2 Picard's representation of emotional states (Picard, 1995)

relate to a particular emotion. With this in mind the following chapter considers how the literature presented in this chapter will contribute to the annotation of emotional expressions in text.

Chapter 3

Annotation

To give a computer the ability to recognise emotions that are expressed in text, we must first provide an annotation scheme which a system can use effectively to evoke a sense of pseudo-understanding. As seen in the previous chapter, emotions should be described and modelled in a framework which captures some sense of the way in which they act and the information that they hold. Computational systems are not sentient from creation, and it is in our nature to define how a machine performs a task in order to achieve its goal. In the case of recognising what emotions are conveyed in text, machine learning approaches can be used, some of which require empirical data in order to determine how a model should be established. To make use of the empirical learning set, the data is annotated with information relevant to the particular emotional category that it is related to. In the problem of efficiently annotating data, Sorokin & Forsyth (2008) note there are two key matters which must be resolved. First, an annotation protocol must be defined. Second, you must determine what data should be annotated. This chapter will outline the work carried out so far in data annotation for emotion recognition and sentiment analysis systems, and the effects that they have had on the machine learning processes that they have been used with.

When creating a system for sentiment analysis that relies on the empirical data to train the models of a system, there are a series of corpora and annotation schemes that have been utilised. We can group the schemes that have been implemented on the level of document granularity that has been annotated, which ranges from the document level down to the word level. Furthermore, we can also consider the category weighting assigned to the particular level of annotation as a differentiator in annotation schema.

This research will build upon the belief expressed by Strapparava & Mihalcea (2008), that the effectiveness of sentiment analysis systems can be increased by implementing a fine-grained emotion annotation scheme. Emotion is a key part of our everyday interactions, and deconstructing this complex occurrence into a naive binary scale does no justice to the progression of this field.

3.1 Granularity of Annotation

The phrase *document granularity* refers to the lexical units within a document that are annotated. This phrase is relevant to four different levels: Word, phrase, sentence and document level. It could be argued that paragraph level is also valid when annotating documents, but the attributes of document level granularity can be applied here as they similarly are just combinations of sentences.

The choice in annotation level is due to the way that emotion is expressed in text. Words alone can incite emotions, yet when combined, to form larger lexical units such as phrases and sentences, different emotions are often found to be expressed (McDonald et al., 2007). The following subsections will describe the corpora and levels of granularity that have been annotated for use in sentiment analysis.

3.1.1 Document Level

Annotating at the document level assigns a label to the general emotion of the whole document. In this instance, every word of a document is assigned the same emotional category. Pang et al. (2002) developed a corpus of film reviews from the website IMDb[1]. The reviews selected for the corpus were chosen due to the star rating that the user had associated with the film. In doing this, it was assumed that the rating was relevant to the emotion exhibited in the review. This meant that human annotation was unnecessary, and reviews could be automatically extracted from IMDb. Altogether, 1301 positive reviews (4/5 stars) and 752 negative reviews (1/2 stars) were extracted from 144 unique users. In this process, no scaling of emotional intensity was used, only flat category labels. Although this corpus was released for use in the research community in 2002, it is still used in current research (Taboada et al., 2011). Another similar English corpus is the SFU Review Corpus, created by Taboada & Grieve (2004). This again collected data using a crawler to scrape data from Epinions. This collected reviews on topics ranging from books and cars to hotels and music. It then assigned the label of positive or negative to a document based on whether a reviewer had indicated that they would recommend the product in question or not, which is more emotionally relevant than the scaling technique of Pang et al. (2002).

Corpora that are annotated at the document level are straightforward to obtain using machine extraction techniques (Pang et al., 2002). However, there are drawbacks in using corpora where each word is assigned the same emotional category. When a document has several sentences, with the possibility of each expressing different emotions, by globally labelling each sentence with a particular emotion, it is overlooking the occurrence of different emotions. If such data is used in training data for supervised machine learning algorithms, the model will be skewed, hence decreasing performance. The issue here is the choice of gathering large amounts of annotated data in a convenient way, or spending time and resources to robustly annotate a corpus. This thesis will consider ways in which the two can be combined where a corpus with primary document level annotations is used.

3.1.2 Sentence Level

To annotate a document at a level more granular than the document level, the sentence level can be used. Pang & Lee (2004) compiled a corpus of 5000 subjective sentences from film review site Rotten Tomatoes[2], and 5000 objective sentences of film plot summaries from IMDb. The subjective sentences were not annotated with emotional categories, which could be attributed to the difficulties in annotating sentence level data with correct emotional information. Riloff & Wiebe (2003) support this decision by noting that it is difficult to obtain collections of individual sentences that can easily be identified as objective or subjective. By viewing emotions as a more complex subjective entity,

[1] http://www.imdb.com

[2] http://www.rottentomatoes.com/

sentence level annotation becomes a difficult task. If we consider the assumption that lexical units require substantial context then this is an agreeable conclusion to reach.

Building upon their previous work, Pang & Lee (2005) annotate an additional corpus of film reviews at the sentence level. In this dataset, they annotate the sentences with a relative polarity on a fine-grained scale, which is their own questionable interpretation of fine-grained. This scale ranges from [0, 5], which aims to imitate the star rating that a user gives a review. Trivialising emotion to a five-point range seems to over simplify the process however. As before, the data was automatically extracted, along with the star rating. This overcomes the need for human participation at their end, by utilising the data that people have posted on the internet. This could be seen as a strength of the annotation process, as no pressure is put on the user to annotate with a strict rule base in mind. However, this freedom means that those with a vested interest could possibly annotate the data incorrectly. Another drawback of this annotation scheme, and duly a limitation for others that annotate on a scale, is the assumption that star rating is linked to the emotion of a user. The user could well have given a film five stars, but whether that was because it made them happy, or moved them emotionally (say made the user cry), denotes a big difference in emotion expressed. Therefore it would be preferable if in the process of writing a review, the intended emotion was captured in addition to the review being written, in a data object such as a meta tag. This would aid in supervised machine learning approaches to sentiment analysis, but it is often not convenient for a user to tag their emotional state alongside what they have written due to the inherently fast paced speed with which content creators often work.

3.1.3 Phrase Level

Implementing an annotation scheme at the phrase level enables the capture of a mixture of emotional and non-emotional data within a sentence. Wilson et al. (2005) developed a phrase level annotation scheme for use with the Multi-perspective Question Answering Corpus (Wiebe et al., 2005). In this scheme, annotators were asked to tag the polarity of subjective expressions as positive, negative, both or neutral. The tags are for expressions of emotion (*I'm happy*), evaluations (*Good idea*) and stances (*He supports the teachers*). The both tag was used for phrases displaying opposing polarities (*They were graceful in defeat*). Neutral was used where subjective expressions did not express emotion, such as suggestive phrases (*You should eat the food*). An important step in the annotation process, was asking the annotators to judge the polarity once the sentence had been fully interpreted, not as the sentence was being read. The example given by (Wiebe et al., 2005) is to stop phrases such as *They will never succeed* being tagged with a negative polarity in the context of the sentence: *They will never succeed in breaking the will of the people.* This example highlights the important role of capturing contextual information when annotating a corpus. Altogether, 15991 phrases were annotated in this corpus.

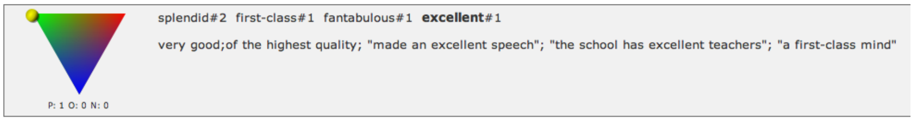
Another way to capture phrase level information is to annotate the n-grams of a document. This is a sequence of n items, and in the case of sentiment analysis, these items are words. Potts & Schwarz (2008) created the UMass Amherst Linguistics Sentiment Corpora of English, Chinese, German and Japanese reviews. These corpora consist of data from Amazon, TripAdvisor and MyPrice, and contains approximately 700,000 documents each annotated at the trigram, bigram and unigram level of granularity. Each n-gram was tagged with a score from [1, 5], which was reflective of the review score on the respective websites. These corpora suffer from the same document level annotation error, where given a positively labelled document that contains negative n-grams, the n-grams will wrongly be tagged with the incorrect sentiment. This is especially prevalent where unigrams are considered, as if this data was to be used to construct a sentimental lexicon for research purposes, the data would unfortunately not be a good representation of the true sentiment.

3.1.4 Word Level - Affective Lexicons

Annotating each word of a corpus with its emotional connotations would be both a time and resource consuming job. Due to this innovative ways of compiling corpora of words annotated with emotional information have been developed. The word level annotations can be viewed as a lexicon of affective words. These are important resources to use where background knowledge is lacking, as they give a generalisation of the affect associated with a word. These can be used for primitive word matching techniques, but this is limited due to the lack of context that can be understood from a single word.

SentiWordNet, developed by Esuli & Sebastiani (2006) is a lexicon with the task of valenced sentiment analysis in mind. All the synsets from WordNet (Fellbaum, 1998) are annotated with scores of the scale of positivity, negativity and objectivity associated with a word in the interval [0,1]. As opposed to other annotation schemes where intensity is considered (Turney, 2002), the total score for all categories must equal 1.

To develop SentiWordNet, seed sets were used as a starting point in WordNet. Seed sets are a collection of words, which in this case have emotional connotations. These are used as starting points in traversing a lexicon's synonym set, and therefore help gather words which have a similar meaning. These seed words were the same as those used by Turney & Littman (2003) in regard to their work on the General Inquirer corpus. An example of such a word is the seed word*excellent*. No humans were used directly in the annotation process, which meant that a semi-supervised classification algorithm had to be used. Q-WordNet (Agerri & Garcia-Serrano, 2010) was created in a similar fashion to SentiWordNet, and similarly was annotated with polarity data on a binary scale, but it is significantly smaller.



splendid#2 first-class#1 fantabulous#1 **excellent**#1

very good;of the highest quality; "made an excellent speech"; "the school has excellent teachers"; "a first-class mind"

Figure 3.1 SentiWordNet's annotation for the word *excellent*

3.2 Emotional Annotations

3.2.1 SemEval-2007 Headline Corpus

At the 4^{th} International Workshop on Semantic Evaluations, Strapparava & Mihalcea (2007) introduced the task of automatically annotating news headlines with both polarity and emotions. The goal of this task was to observe the underlying relationships between lexical semantics and emotions. News headlines were used as they typically aim to provoke an emotion, and try to do so with only a few words. This posed a suitable challenge, as machine learning approaches rely on a reasonable amount of input data to learn linguistic models of emotion, and typically, fine-grained annotation is notably harder than polarity based labels (Agerri & Garcia-Serrano, 2010).

Table 1. Inter-annotator agreement

Emotion	r
Anger	49.55
Disgust	44.51
Fear	63.81
Joy	59.51
Sadness	68.19
Surprise	36.07

A corpus of 1250 headlines was compiled from both news websites and newspapers. The set of emotions that each headline was annotated with was the set of six basic emotions proposed by Ekman et al. (1983) : anger, disgust, fear, joy, sadness and surprise. The scale for emotions annotations was [0, 100], where zero indicated that the emotion was not present in the headline, and 100 meant that the presence of the emotion was maximal. This enabled the annotators to mark up the headlines with a varying degree of emotional intensity, as opposed to a binary presence annotation which would not capture the varying degrees of emotional expression.

Six independent annotators were involved in the process of labelling the data. Each was instructed to annotate the headlines based on the emotions evoked by words, phrases, and overall feeling. Although these three criteria were used, no words or phrases were given specific emotional values, and were seemingly lost in the overall annotation scheme. Inter-annotator agreement was determined using the Pearson product-moment correlation coefficient measure, *r*. By computing agreement between each annotator and the average of the five other annotators, and taking the average of the outcome, this produced the agreement statistics shown in Table 1.

The results in Table 1 indicate that agreement between the annotators is surprisingly low, despite the small number of annotators involved. This is surprising, considering the results achieved by Aman & Szpakowicz (2007) in a similar experiment involving the annotation of emotions in blog posts, where the agreement scores were significantly higher. These low scores may be explained by the fact that as the text spans in news headlines utilize loaded expressions in a limited space, these can easily be misinterpreted. Another possible explanation is that the backgrounds of the annotators varied, meaning that their levels of understanding of the headlines may differ. Therefore the argument could be put forward that annotator demographic should be taken into account when labelling data.

3.2.2 Suicide Note Corpus

Recently, I was involved in a Medical NLP Challenge, organised by Informatics for Integrating Biology & the Bedside[3]. This challenge focused on emotion recognition in suicide notes. The data for the challenge was from a corpus of 1,000 notes of those who unfortunately had died due to suicide. While being a corpus containing relatively few documents, the notes had been hand annotated with 15 different emotions, at the sentence level. Table 2 summarizes how data in the

[3] https://www.i2b2.org/

training set (600 notes) had been annotated. Each sentence could be annotated with zero, one or two different emotions, which added significant complexity to this challenge.

One of the first points to consider regarding this challenge was the range of emotions that the challenge organisers believed were expressed in the corpus. This suicide note dataset was unique, with only the SemEval-2007 (Task 14) (Strapparava & Mihalcea, 2007) dataset showing similarities in annotation through the range of emotional categories that it utilised. This however was still annotated with far fewer emotional categories. Following from these numerous categories, we must consider the skew that is present in the annotations. Of the four emotional categories which were identified most frequently in the text[4], these represent 73.99% of the annotated sentence data. For a supervised machine learning algorithm, this introduces bias to the learned data model. Consequently, this decreases the performance of the machine learning method in correctly recognising and returning the sentences it believes hold some form of emotion. Ideally, one would hope for a corpus that does not introduce bias to the algorithms that will run on it, however in the case of emotional data such as a suicide note, emotions will rarely be balanced so as to provide suitable input for a system.

In creating this dataset for the challenge, the corpus underwent a strict annotation process. The suicide notes were first subject to a rigorous anonymisation procedure before they could be presented to the annotators. This was due to the possibility that bias could occur through subject identification. Once complete, all notes were marked up by only three independent annotators, despite the number of emotional categories that the data could be labelled with. Where differing emotions were thought to be expressed in a sentence by the annotators, the majority decision was taken as the expressed emotional annotation. If the opinions of the annotators differed greatly, and no final annotation decision could be reached, the sentence was left with no mark up. However, sentences that were not assigned emotions in the initial annotation stage were viewed as having no emotions present. Therefore, this lack of mark-up could either mean that there was disagreement in annotation, or no emotion was found. The two were not distinguished, which led to significant confusion when manually observing the data. An example of this confusion is shown in the following example. The sentence *I love you .* was annotated five times with the emotion love, as would be expected; but twice it was not annotated at all. There is a lack in the consistency of the annotations here, as is shown in the utterance of this particular example, where it could be hypothesised that love should be the prominent emotion expressed. The limiting factor here, which could lead to the lack of stability in the annotations, could be attributed to the context in which the utterance occurs. Due to this, a conclusion can be drawn that contextual information, such as a preceding or proceeding emotional expression, should be stored with the annotated data.

3.3 Discussion

This chapter has highlighted the annotation schema that have been used for sentiment analysis, and the level of granularity at which the annotations were added to a document. What can be drawn from this literature review is that the annotation of emotional information to a document is difficult when the process of emotional expression is trivialised to a blanket labelling of the various substructures within a document. This has occurred due to the relative lack of effort required in setting up a crawler to scrape the data from a resource and compile it into a labelled corpus. This overcomes the issue of requiring human annotators to label a data set, but in the process,

[4] Instructions, hopelessness, love & information.

Table 2. Annotation variation in suicide note data. N = Number of sentences annotated with emotion

Emotion	N
Instructions	820
Hopelessness	455
Love	296
Information	295
Guilt	208
Blame	107
Thankfulness	94
Anger	69
Sorrow	51
Hopefulness	47
Fear	25
Happiness/Peacefulness	25
Pride	15
Abuse	9
Forgiveness	6
TOTAL	2522

accuracy of annotation is sacrificed for ease of access. A suggestion would be to revert back to using human annotators, but instead the task should be crowdsourced using platforms such as Amazon's Mechanical Turk[5] as it has been demonstrated to be useful and resource considerate in natural language processing tasks (Snow et al., 2008).

Whilst this research must take careful consideration of the resources required to build a corpus that is useful for fine-grained sentiment analysis, a more pressing issue that effects the outcome of a learned classifier is the way in which the data has been annotated. This chapter has highlighted the numerous differences in approaches to the granularity of annotation, and one clear issue is that many corpora are only annotated on a single level. If annotation was to occur across multiple levels, then it is hypothesised that by utilising this approach in machine learning techniques, that the general performance of sentiment analysis would improve. The following chapter will observe a subset of machine learning techniques that are able to benefit from this annotated data.

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

Chapter 4

Machine Learning

Samuel (1959) defines machine learning as the *field of study that gives computers the ability to learn without being explicitly programmed.* Using this definition, machine learning can be appropriately applied to the problem of text classification, and by way of inheritance, can duly be related to sentiment analysis. It would take a substantial effort to program a computer with all possible emotional utterances. Due to this, machine learning techniques have the potential to contribute an efficient solution to the problem of sentiment analysis. Both supervised and unsupervised machine learning approaches have been applied to the challenge of sentiment analysis, and for some limited domains that exhibit little topical variation, performance has been good. However, previous approaches have viewed emotion in a naive way, and the discrete categories of positive and negative opinion have been the sole labels in the class set. The collection of emotions is larger than this initial group, and therefore a greater challenge to machine learning is posed. This chapter will assess the current techniques from both the supervised and unsupervised literature. The usage of a dependency parser will also be observed in this chapter. First however, the roots of sentiment analysis, text classification, must be discussed.

4.1 Text Classification

Text classification refers to the computational assignment of predefined categories to textual documents. For example, the sentence:

> *"Gloucester drive Bath to distraction to hog derby pride."*

is classifiable as Sport, but cannot be placed into a more granular category by a computational method without further contextual information revealing that rugby is being referenced in this extract. In addition to the topic, rugby, being the category label for this passage, the sentiment of the document can also be used to classify this text. A positive sentiment can be assigned to the entity *Gloucester*, whereas a negative one could be assigned to *Bath*. The challenge here however, if possible, is to assign an overall sentimental category to this passage. This is as sentiment is incredibly subjective, and depends upon a number of variables. For this reason sentiment analysis goes far beyond primary topic-based text classification, and the literature demonstrates that traditional text classification methods should be augmented in order to advance towards the problem of textual emotion recognition.

Nonetheless, traditional text classification has progressed greatly in terms of efficiency from its roots in rule-based classification. A general approach to this was to manually define heuristics which captured patterns in a corpus, such as keywords which would in turn identify a class. Accordingly, for large data sets, this was a laborious and time-consuming task, which was often incredibly specific to a domain for which the rules were crafted. For example, any rules created for sport would not apply to a political domain without the possibility of a detrimental effect on the classifier emerging. This drawback significantly limits rule-based classification. With the technological advancements in computational power that happen almost annually, it goes without saying that machine learning based text classification techniques have been a focus of the NLP community.

4.2 Supervised Methods

This section will observe the literature regarding the supervised machine learning approaches that have been applied to sentiment analysis. Supervised classification techniques construct a system based on the labelled empirical data that is given as input. As a result of this, a classifier is created that can model a domain with ease. This leads to adequate classification performance for a given task, particularly in topic based classification tasks such as spam filtering (Drucker et al., 1999). However, results vary in supervised sentiment analysis techniques due to the quality of training data available to the learner.

In the supervised domain, there are a number of learning algorithms. Such algorithms include the Maximum Entropy classifier (Berger et al., 1996; Nigam et al., 2000; Pang et al., 2002), Support Vector Machines (Joachims, 1998; Pang et al., 2002) and Decision Trees (Jia et al., 2009). Whilst these all have their relative merits in text categorisation, when they are considered for use in machine learning approaches to sentiment analysis, these techniques are not representative of the intuitive emotion recognition process, as demonstrated by Picard (1995), who implements the Bayesian process as a basis for emotional state generation. Therefore the following section will focus on the implementation of a Bayesian classifier as an approach to sentiment analysis.

4.2.1 Naive Bayes Classification

A Naive Bayes (NB) classifier is one of the simpler methods of automatic categorization that has been applied to text classification. Consequently, it has also been utilised in attempting to solve the problem of sentiment analysis. In this algorithmic setting, the lexical units of a corpus are labelled with a particular category or category set, and processed computationally. During this processing, each document is treated as a bag-of-words, so the document is assumed to have no internal structure, and no relationships between the words exist. Once processing has completed, a classification model is established that can be used to group unseen documents. Relative to this investigation, the labels by which documents are grouped would be emotional states that have been textually expressed. Strapparava et al. (2006) note that in discourse, each lexical unit, whether it be a word or phrase, has the ability to contribute potentially useful information regarding the emotion that is being expressed. However, it is typically a combination of these lexical units which motivates the communication and understanding of an emotional expression.

Contrary to this, a universal feature of NB classification is the conditional independence assumption. In this each individual word is assumed to be an indication of the assigned emotion. The assumption is made that the occurrence of a particular lexical unit does not affect the probability of a different lexical unit in the passage conveying a particular emotional meaning. This contrasts the argument proposed by Firth (1957), which asserts that the meaning of a word is highly contextual. In this argument he puts forward the claim that the meaning of a term is dependent on the meaning of any words that co-occur alongside it. This opposes the Bayesian independence assumption. Agreeing with the statement of Firth (1957) should render the algorithm flawed for sentiment analysis, yet this is not always the case when an NB classifier is used. Before we discuss the reasoning behind this, it will be of use to discuss how a NB classifier works, and how the independence assumption is an unavoidable part of the algorithm.

The multinomial Naive Bayes classifier takes multiple occurrences of a word into account, while still maintaining the independence assumption. The training of such a classifier is one of the quicker classifiers to train, despite taking into account multiple occurrences of a word. The following definition is adapted from Manning et al. (2009). First, we must consider the probability of a document d being labelled as expressing emotion e:

$$P(e|d) \propto P(e) \prod_{1 \leq k \leq n_d} P(w_k|e) \tag{4.1}$$

where $P(w_k|e)$ is the conditional probability of a word w_k expressing emotion e. $P(e)$ is the prior probability of an emotion being expressed in a document, dependent on the training set. It is estimated as follows:

$$P(e) = \frac{N_e}{N} \tag{4.2}$$

where N_e is the number of documents labelled with emotion e, and N is the total number of training documents given in the document set. This is adequate where a single emotional label is assigned to a document, but where there are a range of labels that could be assigned to a document, we must either deconstruct the document into further parts, with one label per part, or adapt this formula to consider this occurrence. Next, the classifier estimates the conditional probability that a word w has been labelled with emotion e:

$$P(w|e) = \frac{W_{ew} + 1}{W_e + |V|} \tag{4.3}$$

where W_{ew} is defined as the total occurrences of a word w in training documents that express emotion e, W_e is the total occurrences of all words that express emotion e, and V is the vocabulary of the document set. The plus one is used for smoothing, as a uniform prior for each w.

The goal with a supervised classifier is to assign the best class or group of classes to an unseen document. An instance of a Naive Bayes classifier is no different, and calculates this in the following way:

$$e_{best} = \arg\max_{e \in E} P(e|d) = \arg\max P(e) \prod_{1 \leq k \leq n_d} P(w_k|e) \tag{4.4}$$

With this we take the maximum value from the product of all previous conditional probabilities of a document's words, and assign the predicted emotional label to the document which is undergoing classification. The difficulty arises when attempting to assign multiple labels to a document. If there are multiple emotions that could label a document, and the $\arg\max$ for each is within a threshold, we must attempt to recognise at what threshold we allow multiple labels to be assigned to a document in the domain of sentiment analysis.

Pang et al. (2002) use the Naive Bayes classifier in their study of the sentiment analysis of film reviews. In this they employ both the multinomial model which has just been defined, and the multi-variate Bernoulli model. This differs from the multinomial model by replacing word counts with a binary presence representation for a word. The word value in the vector will be 1 if the word is present in the given document, and 0 if not. Pang et al. (2002) report that binary presence returned better results than using the word frequency approach of the multinomial model. They suggest that this indicates a significant gap between text classification and sentiment analysis. This contradicts previous results reported by McCallum & Nigam (1998) on the use of these two derivations of the NB classifier. Pang et al. (2002) do not elaborate on the reasons for this difference, however such a shift in classifier performance suggests that previous approaches are no longer feasible, and must be adapted for the task of sentiment analysis.

Although the Naive Bayes algorithm has performed well in the above experiments on consumer-based data sets, it has not performed well under different circumstances. Strapparava & Mihalcea (2008) attempted to use a Naive Bayes classifier to classify news headlines into a set of six emotions. The results however were somewhat underwhelming, with an average precision of 12.04% and a recall of 18.01% across all categories. Similarly, in recent experiments I carried out for the Computational Medicine Suicide Note challenge, the results were above the system baseline, but did not compete in standard with the winner of the competition. The best system that we developed was a Naive Bayes based classifier, optimized with a dependency parser. However, in this challenge there were fifteen emotional categories into which documents could be classified, as opposed to the six that were used in the Semeval-2007 task.

This poor performance could be attributed to a number of common supervised machine learning issues: The relevance of the training data, the dimensionality of the input data and the amount of training data provided. Usually, if insufficient training data is provided, the classifier is unable to cope with unseen features, which would explain a drop in the test results. In this case however, the corpus of emotional blog posts that was used as training data consisted of 8,761 documents. While not being an enormous corpus, something of this size should supply sufficient training examples. If we consider the dimensionality of the data as a issue in the weakness of the Naive Bayes approach, the data used in the current experiment has only six possible dimensions, which were the set of basic emotions defined by Ekman (1992). While not significantly greater than a binary sentiment analysis problem, this should not be an issue, as the training data, if sufficient, should be able to cope with this by giving satisfactory evidence to the classifier. This leaves the relevance of the data as the possible cause of the problem, which is agreeable due to the difference of domain between the test set, and the training data set. The training data came from blog posts on the LiveJournal website, whereas headlines from professional news outlets were provided as the test set. As news headlines tend to be written by a trained journalist, in contrast to the amateur diary-like entries of a blog post, this difference can pose enough of a linguistic difference to significantly affect the output values of the classifier. Therefore, this issue of domain must either be overcome, or ignored if we are to successfully experiment with supervised machine learning methods. The alternative to the supervised approach is the unsupervised approach, which fortunately does not suffer from the issue of domain dependence.

4.3 Unsupervised Methods

Unsupervised approaches to machine learning differ significantly from their supervised relatives. Unsupervised algorithms do not require labelled input data to find patterns in a corpus, which makes them impenetrable to the biases and annotator mistakes of empirical knowledge. Instead of relying on a labelled training corpus, these systems use statistical inferences alone to learn from the data. In turn, these unsupervised methods will group together items that exhibit a distinct similarity. The methods can return the lexical items that showed a high similarity, and therefore give an insightful view of a corpus.

In the unsupervised domain, there are a number of classifiers which have been implemented for machine learning problems, such as document clustering (Steinbach et al., 2000). Examples of algorithms applied to these problems are the k-means (Li & Wu, 2010) and k-nearest neighbour classifiers (Tan & Zhang, 2008). However, these classifiers only operate at the word level, and do not go beyond this in grouping documents by other features such as expressed emotion. Latent Semantic Analysis (LSA) on the other hand works on the semantic level, and groups documents that are semantically similar, not just documents that have similar features. Therefore, this section will focus on LSA, in particular observing the way that it has been applied to both coarse and fine-grained sentiment analysis, and emotion recognition systems.

4.3.1 Latent Semantic Analysis

Landauer (2006) expresses the opinion that in order to determine the message of a document, a function of the meaning of all the words and their context should be perceived. Nevertheless, words are often polysemous, so disambiguating the intended meaning is difficult. In addition, many terms in a passage can refer to the same idea. This is known as synonymy. A solution to these two issues of lexical ambiguity proposed by Deerwester et al. (1990) is called Latent Semantic Analysis (LSA). This looks at the underlying semantic structures of a corpus, and highlights lexical items that are used in similar contexts, which consequently could have similar meanings. It does this by taking a term-document matrix containing term frequency counts and decomposing this into a resulting matrix of singular values. This process is known as singular value decomposition (SVD). Each individual document can be viewed as a vector of term frequencies, and when decomposed, these high-dimensional vectors are mapped into a low dimensional representation in a latent semantic space. Then, similar documents or words within documents are discovered by estimating the cosine of the angles between their relative vectors.

By discovering these latent semantic structures, LSA should be an agreeable approach to take for emotion detection. It has been applied in a number of ways to the problem of sentiment analysis. Strapparava & Mihalcea (2008), make use of LSA in three different ways in their experiments. First, they carry out what they call single word LSA. This determines the similarity between a document and an emotion. It is implemented by performing LSA on the vector representation of the document, and a vector containing the relevant emotional term. Second, they calculate the similarity using a vector which contains the synonym set of the emotional word. Finally, they use a vector of all possible words relating to an emotional concept in the vector. These words are taken from the emotive lexicon, WordNet Affect. The results from the experiments indicate that the final technique using all possible emotion words was better than the previous two approaches, and consistently outperforming the other techniques in terms of recall. However, this was at the sacrifice of precision, which was extremely low. The work does not elaborate on the reason for the low precision, but this could be attributed to the fact that this method was taking to general an approach to classification, and over-fitting the data, by returning a number of false-positives. Deerwester et al. (1990) attribute a low precision to the high presence of polysemic words. Without seeing the data, it would be wrong to assume only polysemic words caused a low precision value.

The paper does not detail a similarity threshold value that indicates the presence of emotion, which is surprising as this information is vital to the outcome of the algorithms. If one was to set the threshold too low, then recall would be high, as lower cosine similarity scores would be allowed to filter through. However, this would be at the cost of the precision of the algorithms. If the threshold were to be raised, it would be fair to predict that the precision would rise, but the recall would fall. This aspect is not highlighted in this proposal, yet time permitting, this is a research path which could be considered.

4.4 Dependency Parsing

While not being a supervised or unsupervised machine learning technique, it seemed fitting to include a short review of the literature regarding dependency parsers and how they fitted in with the machine learning methodology. As Manning et al. (2009) note, a dependency parser is used to understand the grammatical relationships in a sentence. It is used in place of a phrase structure tree, due to its intuitive nature.

The Stanford Dependency parser (Manning et al., 2009) calculates dependencies within a given sentence as a triple of the form *relation(governor, dependent)*. The relation is a grammatical relation, such as a nominal subject, which is a noun phrase that acts as the subject of a clause. The governor in this case could be a verb, but it is not restricted to this, and the dependent is often a noun. An example of such a dependency would be *nsubj(usurped, Gaddafi)*.

A dependency parser is of use to sentiment analysis, as often in machine learning, feature selection is required in order to optimize the classification technique being performed. Frequent words or proper nouns are often seen as noise, and can detract from the performance of a classifier. Therefore, certain dependency based features, such as the aforementioned nominal subject relation, or the direct object of a verb phrase can be selected as the representative features of a document, and classifiers can train solely on this data.

In the CMC Suicide Note challenge, a slightly different approach to using the dependency parser was used. Instead of relying solely on specific relations, dependants and governors of different parts of speech were focused on. The best performing system used the NB classifier that had been trained using only verbs, adverbs and adjectives that appeared in the governor position of all relations that were discovered. These parts of speech were used under the assumption that the governor of a dependency contained more sentimental information than the dependent, which was shown to be a correct assumption in a simple experiment, but further investigation would be needed to statistically confirm this occurrence.

4.5 Summary

This chapter has looked at the influence of supervised and unsupervised machine learning techniques on sentiment analysis. The Naive Bayes classifier, and Latent Semantic Analysis have been the focus of the machine learning techniques, and these will form the basis for the experiments of this research. The methodological approach to the experiments, along with their evaluation, will be outlined in the following chapter.

Chapter 5

Methodology & Evaluation

A review of machine learning techniques alongside an outline of relevant models of emotion has been given in previous chapters. The proposed work will build upon this literature, and therefore this chapter will describe the methodology and evaluation that will be used in this work. The intention of this research is to apply existing models of emotion to sentiment analysis, and investigate the effects that these paradigms have on machine learning techniques and their universal reliability. This will expand current techniques into a finer-grained domain, and will examine the suitability of machine learning approaches to the detection of verbal expressions of emotion.

In order to proceed with this research, a series of experiments will be devised with the aim of producing a sentiment analysis technique, or group of techniques, that detect the expressed emotion in a corpus in ways that go beyond current state of the art method. A high level objective of this research is to produce a framework for processing corpora through which future research on fine-grained sentiment analysis that uses machine learning techniques can be based. In order to achieve an infrastructure supportive of the needs of this research, the following pieces of work will be carried out.

5.1 Pilot Study

In the initial stages of experimentation, a pilot study will be implemented with the aim of assessing an appropriate emotional model with which to annotate a corpus, and the level at which annotation must occur. The pilot study is timetabled in the following chapter, and will proceed as follows:

- Manually extend the annotation scheme of the SemEval-2007 gold-standard training corpus, which contains 1,250 documents, according to the dyadic model of emotion defined by Plutchik (1997).
  - Maintain the valence assumption outlined by Ortony et al. (1988) through the expansion of polarity-based categories into the dyadic categories of Plutchik (1997). This should be done at varying levels of annotation granularity, as noted in Chapter 3, in order for a range of supervised machine learning experiments to be carried out.
  - Annotate the corpus with the direct and indirect expressive nature of words outlined by Strapparava et al. (2006). Expand this to also annotate the corpus at varying levels of document granularity.

    - Analyse the corpus for concordance and collocation data, taking note of the annotations and whether these are representative of the expressed emotion given the context within which they appear. Stop words will be removed from the concordance list, and the 20 most frequent words from the list will be taken and analysed in their relative collocative settings in order to glean insight on their usage and emotional expressiveness.

Whilst this is taking place a framework to support the corpus examination and mark-up will be developed in order to be used as a basis for future work specifically focusing on the annotation of corpora for sentiment analysis. Current off the shelf frameworks for text processing problems such as *LingPipe*[1] will be used as a basis for developing software. However, these were not created with the problem of sentiment analysis in mind, so will have to be tailored to the needs of this research through use of their respective APIs.

Following from the annotation of a corpus, pilot studies will take place that focus on variations of the Naive Bayes classifier that have been adapted for sentiment analysis.

- Adapt the Naive Bayes variations used by Pang et al. (2002) with unigram, bigram and trigram features alongside dependency structures and apply to the previously annotated corpus. Train on the 250 document test set which has been manually re-annotated, and test on the remaining 1,000 documents, to keep experimentation consistent with the SemEval 2007 task.
    - Experiment with the relations of the Stanford Dependency parser, in particular testing on a variety of different combinations of the dependency structures in order to determine the most effective combination for sentiment analysis.
    - Ensure careful observation of the governor and dependent is maintained in order to verify any statistical patterns that may occur.
    - Further to the basic training which was consistent with the SemEval-2007 methods, train the classifier and corroborate results using a ten-fold validation technique to ensure robustness of the classification performance (Kohavi, 1995).
    - Experiment with threshold values of classifier parameters to find an optimal value when the classifiers are used for multi-category classification. Following this, observations must be made which study how the relaxing of this threshold affects the classification outcome. Ensure that the value is significant through statistical confirmation in the ten-fold cross validation technique of the classifier input data.

Following this it will be of interest to compare the results of running variations of the LSA unsupervised machine learning algorithm over the same data to compare the categories returned with the Naive Bayes classifier in order to determine if there are similarities between the results that the two may return. If it is the case that similar categorisations are returned, or specific patterns are highlighted for particular emotional classes, then this could lead to interesting theoretical questions regarding both the learning techniques and the nature of emotions.

- Run variations of LSA over corpora in order to establish the rationale of annotation and the NB results. The variations will incorporate those described by Strapparava & Mihalcea (2008), which made changes to the contents of the comparison vector.

[1] http://alias-i.com/lingpipe/index.html

Once this pilot study is completed, results will be analysed and evaluated in order to determine the quality of the annotation schemes used, the emotional models their relevancy, and the effectiveness of the machine learning techniques. In light of this, if changes need to be made, they will be factored into the experiments of this research. Experimentation will then take place but at a larger scale, and the results will be analysed appropriately.

5.2 Evaluation

In order to evaluate this work, it is worth restating the research questions and hypotheses set out in the introduction of this thesis proposal:

RQ1 Which model of emotion is best suited to sentiment analysis?

(a) Are the emotions expressed in text suited to an ontology?

RQ2 How should documents be annotated with emotional information?

(a) What spans of text should be annotated?

(b) How will structural information be captured?

(c) How will the different forms of expression be captured?

RQ3 Which machine learning techniques are most suited to the task of textual emotion recognition?

Hypothesis 1 - (RQ1) Emotions can be structured as a tree, with valenced categories acting as the root node, and fine-grained emotional categories at the leaves.

Hypothesis 2 - (RQ2) Expressed emotion is not a sum of its parts, and therefore documents should be annotated across various levels to capture this expression.

Hypothesis 3 - (RQ3) Supervised machine learning techniques in combination with a dependency structure are most suited to sentiment analysis.

In order for these research questions to be successfully evaluated, they will have to be validated through annotator agreement studies and statistical results.

$Hypothesis 1$ and $RQ1$, this will be evaluated through an agreement study with annotators. By asking annotators to re-annotate a corpus that contains sentimental mark-up, a new tag set is created that enables comparison to occur. The differing tag sets will be compared using Cohen's kappa value (Cohen, 1960). Kappa values that are greater than 0.6 are seen as a figure for substantial agreement (Landis & Koch, 1977). Nonetheless, in order for what is viewed as good reliability to occur, and therefore contribute significantly to $RQ1$, a kappa score that is greater than 0.8 will be required. This will denote whether the model of emotion used is suited to sentiment analysis. If kappa values closer to zero are obtained however, then this raises interesting questions regarding the model used. Through further investigation significant differences will be considered between the two annotation sets, looking in particular at the features on which these difference of annotation occurred, alongside annotator demographic information.

To evaluate $RQ1_{(a)}$, and furthermore evaluate $Hypothesis 1$ the results of the experimentation between the extended annotation set and the basic annotation set will be taken into account. The values of precision, recall, accuracy and f1 scores for all emotional categories will be observed and compared appropriately. If it is found that for the extended set, which employs an ontological approach to emotion achieves higher scores than the basic set, then the question can confirm that emotions expressed in text are suited to an ontology. However, if they do not achieve better results than the basic set, or the results are mixed, then this will revoke the claim, and reasoning for this outcome will be investigated, and will nonetheless contribute to knowledge.

An inter-annotator agreement study will again be carried out for the evaluation of $Hypothesis2$ and $RQ2$. By providing an experiment where annotators are able to tag the spans of text the believe express emotion at varying levels of document granularity, an inter-annotator agreement study can be implemented. This will be evaluated using kappa values, with a score of 0.8 denoting a good reliability score, and therefore providing a solution to the problem of how documents should be annotated with emotional information. If scores are significantly lower than this threshold for all the types of annotation tested, then it will become apparent that these methods are not suited to sentiment analysis, and through analysis and observation, new approaches must be sought.

Finally, $Hypothesis3$ and $RQ3$ will be evaluated by analysing the results of the machine learning experiments between the supervised and the unsupervised techniques. This will be achieved by holding back a test set of the data to experiment on, so no algorithm has the opportunity to train or learn from these documents. This training set will be set at different proportions of the whole document set in order to validate the results. Similar to the evaluation of $RQ1_{(a)}$, the evaluation will depend on statistical metrics in order to determine which set of algorithms is preferable for sentiment analysis.

5.3 Conclusion

This chapter has outlined the methodological approach that will be taken to this research, and has described how the work will be evaluated in order to assess it. A pilot study has been defined, which will act as the basis for this research. Following from this the evaluation has reiterated the research questions and hypotheses, and given clear criteria by which this work can be assessed. The next chapter will detail how this methodology will fit into the allotted time for research.

Chapter 6

Proposed Timetable

Due to the variable nature of research, the following timetable is only an outline, and therefore is flexible to change. Any changes will be reported in further RSMG reports, and the timetable will be amended duly.

- *November - December 2011*:
  - Extend the annotation scheme of the SemEval-2007 gold-standard corpus with Plutchik's model.
  - Prototype sentiment analysis corpora evaluation software.
- *January - March 2012*:
  - Develop machine learning classifiers.
  - Experiment with classifiers on the extended-annotation corpus.
- *April 2012*:
  - Evaluate results from experiments.
  - Write a paper given the results, to submit to either the 24^{th} International Conference on Computational Linguistics (COLING 2012) or the Recent Advances in Natural Language Processing (RANLP 2012).
  - Write RSMG4 report.
- *May - August 2012*:
  - Amend classification techniques in light of results.
  - Finalize framework for sentiment analysis software, integrating corpus analysis with machine learning techniques.

The following gantt chart displays this timetable in a graphical form:

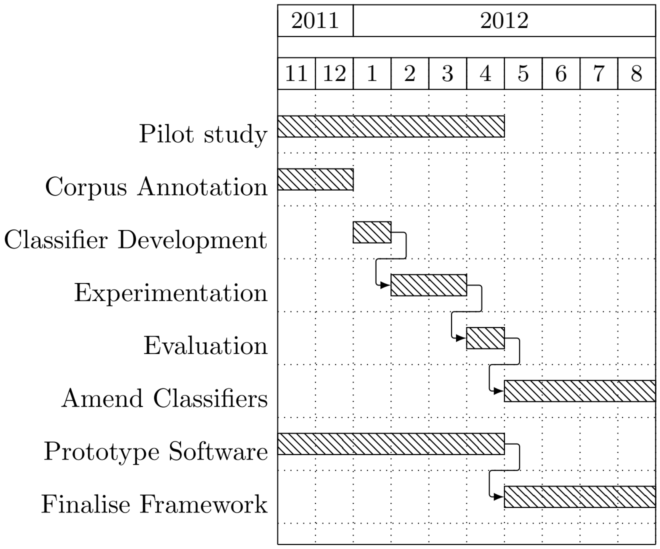


Chart 1: Ten month Gantt Chart

Beyond this, time scales will be dependent on the outcome of the pilot study, so further experimentation, data collection and data analysis will rely on the conclusions drawn from this. However, work can still be planned for the coming years, and the research is scheduled as follows:

- *August 2012 - January 2013*:
  - Collect emotional data from online review sites for own corpus.
  - The corpus should be annotated manually by five human annotators according to annotation schema proposed from the pilot study.
  - Write paper for the International Joint Conference on Artificial Intelligence (IJCAI 2013) on implementing emotional models in a machine learning environment, based on results from previous experimentation.
- *January - October 2013*:
  - Experiment with machine learning methods on own corpus.
  - Analyse results from experiments.
  - Submit paper to ACL & IJCAI
- *October 2013 - March 2014*:
  - Write-up thesis.
  - Submit paper to ACL, IJCAI & Computational Linguistics
  - Submit thesis.
- *March - April 2014*:
  - Contingency time.

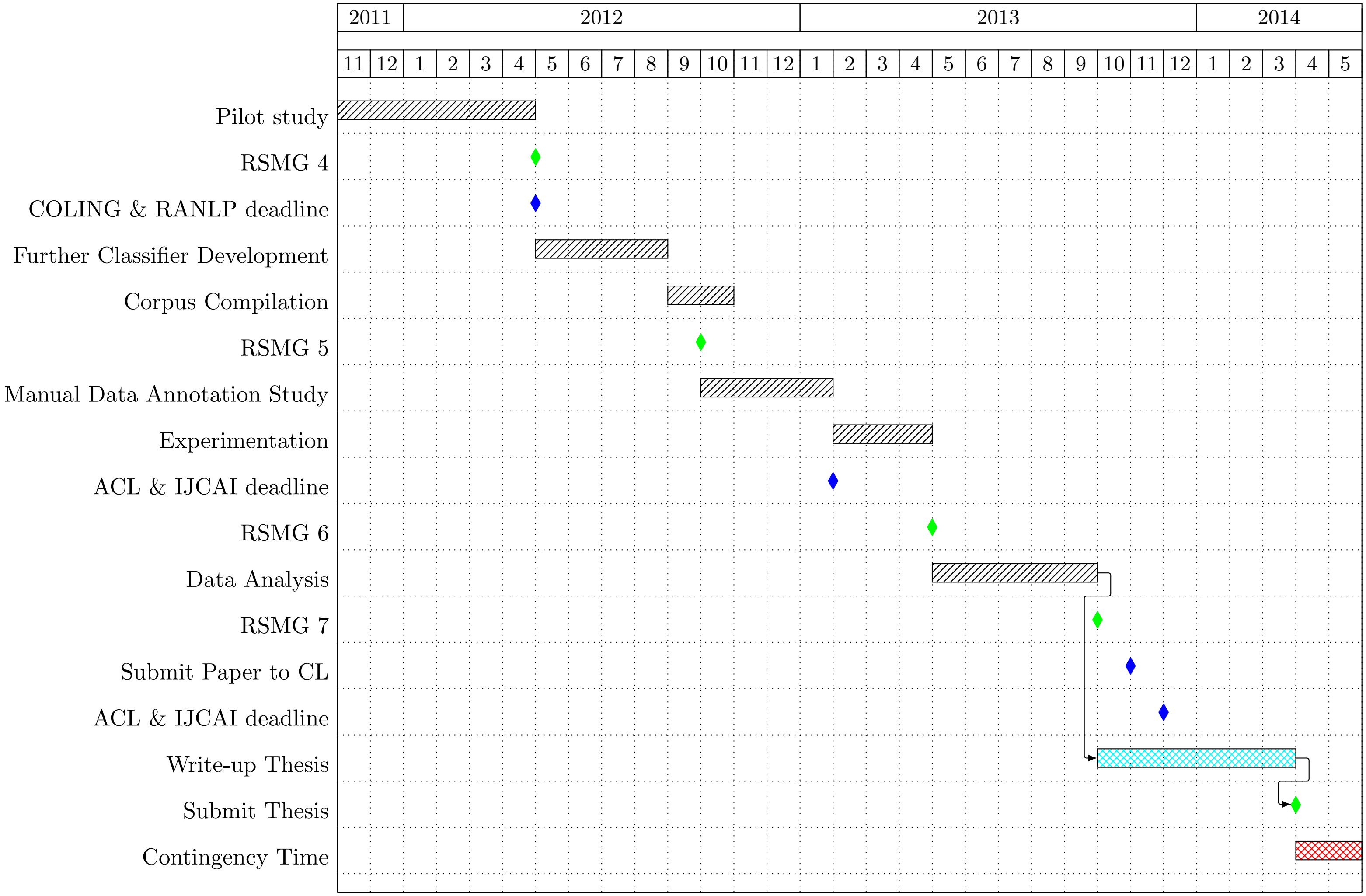


Chart 2: Complete timetable for future work

Appendices

Appendix A

Table of basic emotions

References	Emotions	Basis for inclusion
Arnold (1960)	Anger, aversion, courage, dejection, desire, despair, fear, hate, hope	Relation to action tendencies
Ekman et al. (1983)	Anger, disgust, fear, joy, sadness, surprise	Universal facial expressions
Frijda (personal communication, September 8, 1986)	Desire, happiness, interest, surprise, wonder, sorrow	Forms of action readiness
Gray (1982)	Rage and terror, anxiety, joy	Hardwired
Izard (1971)	Anger, contempt, disgust, distress, fear, guilt, interest, joy, shame, surprise	Hardwired
James (1884)	Fear, grief, love, rage	Bodily involvement
McDougall (1926)	Anger, disgust, elation, fear, subjection, tender-emotion, wonder	Relation to instincts
Mowrer (1960)	Pain, pleasure	Unlearned emotional states
Oatley & Johnson-Laird (1987)	Anger, disgust, anxiety, happiness, sadness	Do not require propositional content
Panksepp (1982)	Expectancy, fear, rage, panic	Hardwired
Plutchik (1980b)	Acceptance, anger, anticipation, disgust, joy, fear, sadness, surprise	Relation to adaptive biological processes
Tomkins (1984)	Anger, interest, contempt, disgust, distress, fear, joy, shame, surprise	Density of neural firing
Watson (1930)	Fear, love, rage	Hardwired
Weiner & Graham (1984)	Happiness, sadness	Attribution independent

References

Agerri, R. & Garcia-Serrano, A. (2010), "Q-WordNet: Extracting Polarity from WordNet Senses", *in Proceedings of the* 7th *Conference on International Language Resources and Evaluation (LREC'10)*, Valletta, Malta: European Language Resources Association (ELRA).

Aman, S. & Szpakowicz, S. (2007), "Identifying Expressions of Emotion in Text", *in* Matousek, V. & Mautner, P. (Eds.), *Text, Speech and Dialogue*, Springer Berlin / Heidelberg, volume 4629 of *Lecture Notes in Computer Science*, pp. 196–205.

Arnold, M. B. (1960), *Emotion and personality.*, Columbia University Press.

Berger, A. L., Pietra, S. A. D. & Pietra, V. J. D. (1996), "A Maximum Entropy approach to Natural Language Processing", *Computational Linguistics* 22(1), pp. 39–71.

Blitzer, J., Dredze, M. & Pereira, F. (2007), "Biographies, Bollywood, Boom-boxes and Blenders: Domain Adaptation for Sentiment Classifications", *in Proceedings of the* 45th *Annual Meeting of the Association of Computational Linguistics*, Prague: ACL, pp. 440–447.

Cohen, J. (1960), "A Coefficient of Agreement for Nominal Scales", *Educational and Psychological Measurement* 20(1), pp. 37–46.

Cowie, R., Douglas-Cowie, E., Tsapatsoulis, N., Votsis, G., Kollias, S., Fellenz, W. & Taylor, J. (2001), "Emotion Recognition in Human-Computer Interaction", *Signal Processing Magazine, IEEE* 18(1), pp. 32 –80.

Dave, K., Lawrence, S. & Pennock, D. M. (2003), "Mining the peanut gallery: opinion extraction and semantic classification of product reviews", *in Proceedings of the* 12th *International Conference on World Wide Web*, New York: ACM, pp. 519–528.

Deerwester, S. C., Dumais, S. T., Landauer, T. K., Furnas, G. W. & Harshman, R. A. (1990), "Indexing by Latent Semantic Analysis", *Journal of the American Society of Information Science* 41(6), pp. 391–407.

Dellaert, F., Polzin, T. & Waibel, A. (1996), "Recognizing Emotion in Speech", *in Proceedings of the* 4th *International Conference on Spoken Language Processing*, Philadelphia, PA: ICSLP, pp. 1970–1973.

Drews, M. (2007), "Visualization of Robert Plutchik's Psychoevolutionary Theory Of Basic Emotions", University of Applied Sciences Potsdam, Germany.

Drucker, H., Wu, D. & Vapnik, V. N. (1999), "Support Vector Machines for Spam Categorization", *IEEE Transactions on Neural Networks* 10(5), pp. 1048–1054.

Ekman, P. (1992), "An Argument for Basic Emotions", *Cognition & Emotion* 6(3), pp. 169–200.

Ekman, P., Levenson, R. & Friesen, W. (1983), "Autonomic Nervous System Activity Distinguishes among Emotions", *Science* 221(4616), pp. 1208–1210.

Esuli, A. & Sebastiani, F. (2006), "SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining", *in Proceedings of the* 5*th* *Conference on Language Resources and Evaluation*, Genoa: ELRA, pp. 417–422.

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

Firth, J. (1957), *Papers in linguistics: 1934-1951*, London: Oxford University Press.

Gray, J. A. (1982), *The neuropsychology of anxiety: An enquiry into the functions of the septo-hippocampal system.*, Clarendon Press.

Izard (1971), *The face of emotion*, New York: Plenum Press.

James, W. (1884), "What is an emotion?", *Mind* 9(34), pp. 188–205.

Jia, L., Yu, C. & Meng, W. (2009), "The effect of negation on sentiment analysis and retrieval effectiveness", *in Proceeding of the 18th ACM conference on Information and Knowledge Management*, CIKM '09, New York, USA: ACM, pp. 1827–1830.

Joachims, T. (1998), "Text categorization with support vector machines: learning with many relevant features", *in* Nédellec, C. & Rouveirol, C. (Eds.), *Proceedings of ECML-98,* 10*th* *European Conference on Machine Learning*, Heidelberg et al.: Springer, pp. 137–142.

Kleinginna, P. R. & Kleinginna, A. M. (1981), "A Categorized List of Emotion Definitions, with Suggestions for a Consensual Definitions", *Motivation and Emotion* 5(4), pp. 345–379.

Kohavi, R. (1995), "A study of cross-validation and bootstrap for accuracy estimation and model selection", *in Proceedings of the 14th international joint conference on Artificial intelligence - Volume 2*, San Francisco, CA, USA: Morgan Kaufmann Publishers Inc., pp. 1137–1143.

Landauer, T. K. (2006), *Latent Semantic Analysis*, John Wiley & Sons, Ltd.

Landis, J. & Koch, G. (1977), "Measurement of observer agreement for categorical data", *Biometrics* 33(1), pp. 159–174.

Lewis, D. D. (1998), "Naive (Bayes) at Forty: The Independence Assumption in Information Retrieval", *in Proceedings of ECML-98*, Chemnitz, Germany: Springer Verlag, pp. 4–15.

Li, N. & Wu, D. D. (2010), "Using text mining and sentiment analysis for online forums hotspot detection and forecast", *Decision Support Systems* 48(2), pp. 354 – 368.

Manning, C. D., Raghavan, P. & Schrütze, H. (2009), *Introduction to Information Retrieval*, Cambridge: Cambridge University Press.

McCallum, A. & Nigam, K. (1998), “A Comparison of Event Models for Naive Bayes Text Classifications”, *in Proceedings of the AAAI-98 Workshop on Learning for Text Categorization*, Wisconsin, USA: AAAI, pp. 41–48.

McDonald, R. T., Hannan, K., Neylon, T., Wells, M. & Reynar, J. C. (2007), “Structured Models for Fine-to-Coarse Sentiment Analysis”, *in Proceedings of the* 45^{th} *Annual Meeting of the Association for Computational Linguistics*, Prague: ACL, pp. 432–439.

McDougall, W. (1926), *An introduction to social psychology*, Boston: Luce.

Mowrer, O. H. (1960), *Learning Theory and Behavior*, New York: Wiley.

Murray, I. & Arnott, J. (1996), “Synthesizing emotions in speech: is it time to get excited?”, *in Proceedings of the* 4^{th} *International Conference on Spoken Language Processing*, Philadelphia, PA: ICSLP, pp. 1816 –1819.

Nigam, K., Mccallum, A. K., Thrun, S. & Mitchell, T. (2000), “Text Classification from Labeled and Unlabeled Documents using EM”, *Machine Learning* 39(2/3), pp. 103–134.

Oatley, K. & Johnson-Laird, P. N. (1987), “Towards a cognitive theory of emotions”, *Cognition & Emotion* , pp. 29–50.

Ortony, A., Clore, G. L. & Collins, A. (1988), *The Cognitive Structure of Emotions*, Cambridge: Cambridge University Press.

Ortony, A. & Turner, T. J. (1990), “What’s Basic About Basic Emotions?”, *Psychology Review* 97(3), pp. 315–331.

Pang, B. & Lee, L. (2004), “A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts”, *in Proceedings of the* 42^{nd} *Annual Meeting of the Association for Computational Linguistics*, Barcelona: ACL, pp. 271–278.

Pang, B. & Lee, L. (2005), “Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales”, *in Proceedings of the* 43^{rd} *Annual Meeting of the Association for Computational Linguistics*, Michigan: ACL, pp. 115–124.

Pang, B., Lee, L. & Vaithyanathan, S. (2002), “Thumbs Up? Sentiment Classification using Machine Learning Techniques”, *in Proceedings of the ACL-02 conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA: ACL, pp. 79–86.

Panksepp, J. (1982), “Toward a general psychobiological theory of emotions.”, *The Behavioral and Brain Sciences* 5, pp. 407–467.

Picard, R. (1995), “Affective Computing”, Technical Report TR 321, Massachusetts Institute of Technology.

Plutchik, R. (1980a), *Emotion: A Psychoevolutionary Synthesis*, New York: Harper & Row.

Plutchik, R. (1980b), “A general psychoevolutionary theory of emotion”, *Emotion: Theory, research, and experience.” Vol. 1. Theories of emotion* .

Plutchik, R. (1997), *Cicumplex Models of Personality and Emotions*, Washington: APA.

Potts, C. & Schwarz, F. (2008), “Exclamatives and Heightened Emotion: Extracting Pragmatic Generalizations from Large Corpora”, Ms., UMass Amherst.

Riloff, E. & Wiebe, J. (2003), “Learning Extraction Patterns for Subjective Expressions”, *in Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, Stroudsburg, PA: ACL, pp. 105–112.

Russell, J. A. (1994), “Is there universal recognition of emotion from facial expression? A review of the cross-cultural studies”, *Psychological Bulletin* 115, pp. 102–141.

Samuel, A. L. (1959), “Some Studies in Machine Learning Using the Game of Checkers”, *IBM Journal of Research and Development* 3(3), pp. 210 –229.

Snow, R., O’Connor, B., Jurafsky, D. & Ng, A. Y. (2008), “Cheap and fast—but is it good?: evaluating non-expert annotations for natural language tasks”, *in Proceedings of the Conference on Empirical Methods in Natural Language Processing*, EMNLP ’08, Stroudsburg, PA: ACL, pp. 254–263.

Sorokin, A. & Forsyth, D. (2008), “Utility Data Annotation with Amazon Mechanical Turk”, *in IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, Anchorage, Alaska: IEEE, pp. 1 –8.

Steinbach, M., Karypis, G. & Kumar, V. (2000), “A Comparison of Document Clustering Techniques”, Technical Report 00-034, University of Minnesota.

Strapparava, C. & Mihalcea, R. (2007), “SemEval-2007 Task 14: Affective Text”, *in Proceedings of the* 4*th* *International Workshop on Semantic Evaluations*, SemEval ’07, Stroudsburg, PA: ACL, pp. 70–74.

Strapparava, C. & Mihalcea, R. (2008), “Learning to Identify Emotions in Text”, *in Proceedings of the 2008 ACM Symposium on Applied Computing*, SAC ’08, New York: ACM, pp. 1556–1560.

Strapparava, C., Valitutti, A. & Stock, O. (2006), “The Affective Weight of Lexicon”, *in Proceedings of the* 5*th* *International Conference on Language Resources and Evaluation*, Genoa: ELRA, pp. 423–426.

Taboada, M., Brooke, J., Tofiloski, M., Voll, K. & Stede, M. (2011), “Lexicon-Based Methods for Sentiment Analysis”, *Computational Linguistics* 37(2), pp. 267–307.

Taboada, M. & Grieve, J. (2004), “Analyzing Appraisal Automatically”, *in Proceedings of the AAAI Spring Symposium on Exploring Attitude and Affect in Text: Theories and Applications*, Palo Alto, California: AAAI, pp. 158–161.

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

Tomkins, S. S. (1984), “Affect theory”, *in Approaches to Emotion*, Hillsdale, NJ: Erlbaum, pp. 163–195.

Turney, P. (2002), “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews”, *in Proceedings of the* 40*th* *Annual Meeting of the Association for Computational Linguistics*, Philadelphia: ACL, pp. 417–424.

Turney, P. & Littman, M. (2003), "Measuring Praise and Criticism: Inference of Semantic Orientation from Association", *ACM Transactions on Information Systems* 21, pp. 315–346.

Watson, J. B. (1930), *Behaviorism*, Chicago: University of Chicago Press.

Weiner, B. & Graham, S. (1984), "An attributional approach to emotional development", *in* Izard, C., Kagan, J. & Zajonc, R. (Eds.), *Emotion, Cognition and Behaviour*, Cambridge: Cambridge University Press, pp. 167–191.

Whitelaw, C., Garg, N. & Argamon, S. (2005), "Using Appraisal Groups for Sentiment Analysis", *in Proceedings of the* 14^{th}*ACM International Conference on Information and Knowledge Management*, New York: ACM, pp. 625–631.

Wiebe, J., Wilson, T. & Cardie, C. (2005), "Annotating Expressions of Opinions and Emotions in Language", *in Language Resources and Evaluation*, volume 39, pp. 165–210.

Wilson, T., Wiebe, J. & Hoffmann, P. (2005), "Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis", *in Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Stroudsburg, PA: ACL, pp. 347–354.