Sentiment Analysis Of Micro Blogs

Course: MSc Software Development (CR_KSDEV_9)
Module: COMP9028

This report represents substantially the result of my own work except where explicitly indicated in the text

Signed _____ Date_____

Abstract. Sentiment analysis refers to the computational analysis of text and the evaluation of any underlying positive or negative sentiment. In recent years it has been the focus of much research. This has been due, in part, to the increased amounts of user-generated content, such as product reviews, movie reviews, blogs, political discourse etc., available on the Internet. This increase in user-generated content has led to an exploration of viable methods for analyzing the sentiment and opinion that it may contain. From a commercial point of view applications that could extract such information would be of interest not only to commercial organizations and governments but also to consumers seeking, not just general information and product specification, but the opinion of the public in general, about products or services. With the recent surge of interest in micro blogging services, such as twitter, a new source public opinion and sentiment has emerged. Micro blogs differ significantly to more traditional user generated content and therefore present new and unique challenges to sentiment analysis. As yet, very little research has taken place with regard to sentiment analysis of micro blog posts. The goal of this paper is to present an overview of past research in the area of sentiment analysis and explore, by means of experiment, whether methods proposed in the literature can be applied to micro blogs.

Table of Contents

1 Introduction

The terms sentiment analysis and opinion mining refer to the computational analysis of text and the evaluation of any underlying positive or negative sentiment, or opinion, expressed by the writer. This is a relatively easy task for human beings but proves to be exceptionally difficult for machines to achieve.

The problem is that, while computers can readily understand formal languages such as programming languages, natural languages like English, present a far greater challenge.

A formal language is constructed from a finite set of words and symbols each with a specific meaning within the context of the language. How these words and symbols are used is defined by a very rigid set of rules called a grammar. Natural languages have grammars too, but they are more flexible and allow ambiguity E.g. flexibility is evident in: “It is nice to be important” or “To be important is nice” where both are grammatically correct. Ambiguity arises in the following examples: “I once shot an elephant in my pajamas” or “Happily they left”. Another obstacle for computational analysis of natural language is that the text is written by humans and so there is no guarantee that all grammatical rules have been followed and therefore *even sense make of this thing that is they have written hard is*.

1.1 Motivation

Natural Language Processing has been an established field of research since the 1940’s when the first attempts were made to use computers to translate text from one language to another.

One of the most commonly experienced product of this field of research is the topic categorization that occurs every time a search term is entered into an Internet search engine like Google.

Sentiment analysis is a relatively new area of research in Natural Language Processing and there has been a lot of research carried out in the area in since 2000. In their 2008 survey of research in the area of sentiment analysis, Pang and Lee (Pang & Lee 2008) note the following factors for this increased interest:

- Increased efficiency in information retrieval
- Improved and increased machine learning methods for natural language processing

- An increase in popularity in blog and review oriented sites on the internet which provide a vast amount of data for these machine learning methods to be trained on
- A realization of the intellectual challenges and the financial potential of creating Sentiment Analysis applications

Sentiment analysis has many practical applications. It could be used as a *flame*[1] detection mechanism for message boards or blog sites, where filtering could be automatically applied to posts that are deemed to be inappropriate (Turney 2002). Others expect the next evolution of search engines to include sentimental summaries of search results (Pang & Lee 2008). This could be particularly useful as; it is human nature that, when confronted with a decision, people seek the opinions of others. For example, if a person, who knows nothing about cameras, were to buy a camera, it is quite probable that, the decision about which camera to buy would be based on advice either from friends or a sales person. The Internet is increasingly becoming a source people turn to when they need information on which to base their decisions.

The Pew Research Center is a non-partisan, non-profit organization, which collects data on the trends and attitudes of American Internet users.

Table 1 shows a subset of the data collected by the Pew Research Center with regard to trends among American Internet users from 2000 to 2010.

Usage	**Date**	**%**
Get news online	March 2000	60%
	May 2010	75%
Get financial information online	March 2000	44%
	May 2010	37%
Research a product or service	March 2000	74%
	September 2007	81%
Look for news or information about politics or the upcoming campaign	March 2000	35%
	December 2009	68%
Look for information about a place to live	March 2000	27%
	August 2009	39%
Look for information about movies, books or other	March 2000	62%

[1] Flames are overly heated or antagonistic comments posted on a message board or in response to a blog entry with the deliberate intent to insult.

leisure activities	March 2002	73%
Create or work on your own online journal or blog	June 2002	3%
	January 2010	14%
Ever read someone else's online journal or blog	February 2004	17%
	December 2008	32%
Rated a product, service or person using an online rating system	May 2004	26%
	April 2009	31%
Used online social or professional networking sites like Friendster or LinkedIn	February 2006	7%
	May 2010	61%
Posted a comment or review about a product you bought or service you received	September 2007	30%
	-	-
Use twitter or other service to share updates about yourself or see updates about others	August 2008	6%
	May 2010	17%

Table 1 Internet Usage 2000 - 2010[2]

Not surprisingly we can see there has been an increase in users posting content and examining the content posted by other, known or unknown, individuals. All this content is full of opinion about products, services, travel destinations, politics, movies, books, music and so on.

Finally a program that can accurately extract sentiment and opinion would be a powerful tool for information analysis in government, commercial, and political domains where it would be of particular benefit to be able to assess the attitudes and feelings of people though social media and other online sources (Wiebe, Wilson & Cardie 2005).

1.2 Blogs and Micro Blogs

Due to the lack of online resources, early work in the area of sentiment analysis used works of fiction (Wiebe 1994) or news articles (Hatzivassiloglou & McKeown 1997) as data sources.

At the end of the 1990's and the start of the 2000's there was an increase in online user generated content such as online reviews, online journals and opinion pieces i.e. blogs. Online reviews prove to be ideal for the task of sentiment analysis as they are plentiful and usually provide a rating system, for example, in the case of movie reviews where a star rating is used to indicate a positive or negative review. Indeed

[2] http://www.pewinternet.org/Static-Pages/Trend-Data/Usage-Over-Time.aspx

the area of reviews was exhaustively used by many researchers (Pang, Lee & Vaithyanathan 2002) (Turney 2002) (Pang & Lee 2005) (Pang & Lee 2004) (Ohana 2009), with some finding that the accuracy of their experiments was greatly affected by the domain under review. For example Turney in (Turney 2002) achieves 84% accuracy when determining the sentiment of automobile reviews yet the same technique achieves 64% accuracy when determining the sentiment of movie reviews. Turney observes that, while the word unpredictable, when used in a phrase such as "unpredictable steering" has a negative connotation with regard to automobiles, the same word, when used in a phrase such as "unpredictable plot", has a positive connotation with regard to movies.

While the start of the 2000's saw the blogosphere increase dramatically in popularity, the end of the decade saw the birth and phenomenal growth of the Micro Blog. From Table 1 we can see that in less than two years, August 2008 to May 2010, the amount of American Internet users who use twitter or some other micro blog service increased from 6% to 17%. This compares to the increase of 11% in the number of users who maintained a journal or blog, between June 2002 and May 2010.

While social network sites and many others offer micro blog services, twitter.com is the front-runner when it comes to micro blogging. The numbers are really quite staggering. Twitter has an estimated 200 million users worldwide who are responsible for 65 million tweets per day[3].

The power of such a medium is well documented. In the 2008 US presidential election Barack Obama employed twitter to great effect by sending out regular alerts and announcements. As a result of this, the use of micro blogging and other social media are viewed as essential to any political campaign (Gross 2008).

While micro blogs present a rich source of data for sentiment analysis they also present new and unique challenges. The limited nature of posts (140 characters) means that there is a need for brevity which gives rise to the use of slang, acronyms and colloquialisms. An example of this is seen in the following tweet.

"*Now dhat I think about it... Steven Spielberg created E.T. which iz classic... Super 8 def finna be raw as ever.. automatic classic*"

The casual nature of micro blogs also gives rise to spelling mistakes and also means that one tweet can concern itself with multiple topics. This is seen in the following

[3] http://en.wikipedia.org/wiki/Twitter

tweet where the author expresses sentiment about both his "honeybunn" and the movie Pirates of the Caribbean.

"*My date last night with my honeybunn was goood. Pirates of the Carribean was AMAZINGGG of course!!((:*"

1.3 The Problem Of Sentiment Analysis

The field of sentiment analysis is a relatively new field and as such has no clearly defined terminology. The term "*sentiment analysis*" first surfaced in a 2001 paper on extracting market sentiment from Internet message boards. *Opinion mining* was coined later in 2003. This area of study has also been referred to as *review mining* and *appraisal extraction*.

Textual information can be either objective (facts or statements that can be substantiated) or subjective (opinions or ideas held by the individual (Wiebe, Wilson & Cardie 2005)). Although sentiment is more likely to occur in subjective text, sentiment expression is not limited to subjective text (Liu 2010). Objective text may also contain positive or negative sentiment. Take for example the tweet "*@htc, the thunderbolt still has GPS issues, freezes when deleting a long text message thread. Phone is a mess"*, which expresses negative sentiment toward the HTC Thunderbolt mobile phone. While "phone is a mess" can be considered an opinion held by the author, "has GPS issues" and "freezes when deleting a long text message thread" can be considered facts. So any textual information can contain sentiment. Text that contains sentiment or opinions is called *opinionated text*.

Consider the following, which illustrates some different aspects that the problem of sentiment analysis presents.

"On a happier note, we set up Boot Camp and ran benchmarks in Windows 7 to provide some more context to our numbers, and the (1)MacBook Pro crushed those tests as well -- the (2)VAIO Z only got a better (3)PCMarkVantage score because of its fast (4)SSD, and the (5)Envy 17 just barely pulled out better (6)graphics performance. (You can configure the (7)MBP with a 128GB (8)SSD for $100 extra, which should probably be standard over the pokey 5,400RPM (9)hard drive.) Playing a little Batman: Arkham Asylum while booted in Windows netted a smooth (10)60fps at native resolution while meandering about, with a dip to (11)55fps during fights".[4]

If we look at this segment of a review for the new MacBook Pro we can see there is quite a lot to take in. It is apparent, even from this small segment, that the reviewer

[4] http://www.engadget.com/2011/03/04/macbook-pro-review-early-2011/

has a positive opinion towards the new MacBook Pro. However for a computer to discern this is more difficult.

To start with let's identify the topic under review, the MacBook Pro, which is referred to twice in the piece (1) and (7) (MBP being MacBook Pro). There are a number of features of the MacBook Pro being examined – the system performance (3), the graphics performance (6), the hard drive (9), and the frame rate per second, fps, (10) and (11).

Two other lap tops are mentioned, the Sony VAIO Z (2) and the Hewlett Packard Envy 17 (5). The system performance (3) and the SSD (4) are two features of the VAIO Z mentioned and graphic performance (6) is a feature of the Envy 17 that is mentioned. These features are used as a comparison to the equivalent features in the MacBook Pro.

If we look at the points of this review segment we get:

- MacBook Pro system performance is not as good as the VAIO Z
- MacBook Pro graphics performance is not as good as the Envy 17
- The SSD of the MacBook Pro costs an extra $100
- The hard drive rpm 5400 which is "pokey"
- The frame rate per second ranges from 60 to 55

So how do we surmise that the sentiment here is positive toward MacBook Pro? Firstly the paragraph is about benchmark tests performed on the MacBook Pro and we are told early on that the MacBook Pro "*crushed*" these tests. This gives us an overall sense that the reviewer is very impressed and what is to follow is a continuation of this initial narrative.

The comparisons made show the MacBook to be inferior in some aspects to competing products. Domain knowledge plays a part here in determining the sentiment. The VAIO Z is the top scoring laptop for system performance because it uses a Solid State Disk, which is characterized by lower access times and latency and therefore will outperform a Hard Disk Drive. Also, the Envy 17 is top of the line when it comes to graphics performance in a laptop. Taking this into account, and in combination with the language used, we see that the MacBook Pro compares favorably to the top of the line competitors.

Domain knowledge also is important when looking at the statistics for frame rate per second and a range of 60 to 55 is a high level of performance for a laptop.

Context is also important. This is one paragraph from a piece that is a glowing review of the new MacBook Pro. With this in mind, suggestions that the 5400-rpm hard disk drive is "pokey" can be dismissed, as the rest of the review does not seem to indicate that the reviewer really finds this to be the case.
One more point we may consider is the source of the review. Did an Apple evangelist who waxes lyrical about everything Apple post the review or is this a truly objective review of the MacBook Pro? Obviously, if the case is the former, rather than the latter, what is the value of this review?
For a computer to extract such information is a difficult task and Liu (Liu 2010) identifies different approaches and techniques studied to tackle the problems associated with sentiment analysis.

1.3.1 Sentiment And Subjectivity Classification.

This area of research treats the problem of sentiment analysis as a text classification problem. Two distinct tasks can be identified here. These tasks are sentiment classification and subjectivity classification.
Sentiment classification is done at document level. This means that an entire document is evaluated and classified as containing negative or positive sentiment toward the subject matter of the document. The main assumptions here are, that the document being analyzed is an *opinionated document*, that is, it contains *opinionated text* and that the sentiment expressed can be determined by *features* of the language used.
Subjectivity classification is done at sentence level, that is, individual sentences of a document are classified as objective or subjective. These subjective sentences are further evaluated with the aim of determining whether they express positive or negative sentiment.
It should also be noted here that there is an overlap between sentiment analysis and *subjectivity analysis*. The goal of subjectivity analysis is to determine whether text is objective or subjective. One approach to identifying subjective text is to determine the *semantic orientation* or *polarity*[5] of words contained in the text, as opinionated language is an indicator of subjectivity (Wiebe 1994) (Hatzivassiloglou & McKeown

[5] Semantic orientation or polarity refers to the sentiment a word or phrase conveys, and can be positive ("honest", "happy") or negative ("disturbing", "sad") and can also vary in strength.

1997). In this respect sentiment analysis is a subtask of subjectivity analysis (Pang & Lee 2008).

1.3.2 Feature Based Sentiment Analysis.

This aim of this work is to evaluate any sentiment expressed about features, attributes or characteristics of a target. One of the main challenges involve in this form of analysis is identifying these features, attributes or characteristics. Using the MacBook Pro review segment as an example, the features that have sentiment expressed about them are the system performance, the graphic performance, the hard disk drive etc.

1.3.3 Sentiment Analysis Of Comparative Sentences.

Comparative sentences are sentences where one object is compared to another. In the MacBook Pro example, above, there are two examples. The system performance of the MacBook Pro is compared to that of a Sony VAIO Z and the graphics performance is compared to the graphics performance of a HP Envy 17. This is a common method of expressing sentiment or opinions and it is sentiment and opinion expressed in this way that this approach attempts to identify and classify.

1.3.4 Opinion Search And Retrieval.

The tasks studied in this kind of research are, identifying and retrieving opinionated documents that are relative to a search term, classifying the sentiment and ranking the documents retrieved by some sentiment metric.

1.3.5 Opinion Spam And Utility Of Opinions.

Opinion spam refers to fake Internet content that tries to mislead users by giving positive or negative opinions to either enhance or damage the reputation of a person, product or service. Opinion utility refers the usefulness of an opinion. The purpose of this research is to measure the utility of opinions in order to negate opinion spam.

1.3.6 Past Research

Although there are many threads in the overall research area of sentiment analysis, this paper mainly focuses on subjectivity classification and sentiment classification. Research in these areas has three main goals: exploring methodologies and approaches to determining subjectivity within text, exploring methodologies and approaches to determine the positive or negative sentiment contained in text and the

construction of manual and/or machine annotated corpora and lexicons to aid further research and the development of applications.

1.3.6.1 Machine Learning

The use of Machine learning algorithms has become the most popular approach to the problem of sentiment analysis (Liu 2010). Both *supervised* (Pang & Lee 2005) (Pang & Lee 2004) (Pang, Lee & Vaithyanathan 2002)and *unsupervised* (Turney 2002) (Ohana 2009) learning algorithms have been employed with varying results. One aspect of machine learning approaches is the use of text classifiers.

A classifier is a program that "*takes inputs, which can be just about anything, and returns a classification for this input over a finite number of discrete categories*" (Carpenter 2010).

The experimental portion of this paper uses a Bayes Classifier to classify tweets. A Bayes classifier is a based on Bayes theorem

$$P(H|E) = \frac{P(E|H) \times P(H)}{P(E)}$$

Equation 1 Bayes Theorem

where

- H is a hypothesis to be tested
- E is the evidence associated with that hypothesis. (Roiger & Geatz 2003)

Or in the case of sentiment analysis of tweets

- H is the sentiment class
- E is the tweet being classified.

The term Naïve is applied to this type of classifier because of the simplistic underlying assumption that all inputs are independent of one another and of equal weight. Although this assumption of independence is rarely observed, this type of classifier still performs well in sentiment classification (Go, Bhayani & Huang 2009) (Pang, Lee & Vaithyanathan 2002) (Parikh & Movassate 2009) (Ohana 2009).

To understand how this works in practice consider the following

- In a newsagent, 90% of customers who buy playboy are men.
- 1 out of every 40 customers that go to the newsagents buys playboy.
- 40% of customers that go to the newsagents are men.

Bayes theorem can be used to calculate the probability that, when a man enters the newsagents he is heading for that top shelf, intent on making a purchase.

If m is the proposition that a man enters the newsagent and b is the proposition that the customer will buy playboy we have

$P(m|b) = 0.9$

$P(b) = \frac{1}{40} = 0.025$

$P(m) = 0.4$

$P(b|m) = \frac{P(m|b)P(b)}{P(m)} = \frac{(0.9 * 0.025)}{0.4} = 0.0526$

P(b) and P(m) are also referred to as prior probabilities.

We can see that, although the probability that a customer will buy a playboy is only 0.025, when a man enters the newsagent the probability that a playboy will be purchased more than doubles.

A text classifier needs to be trained using a predefined set of *features* that **can** occur in a document. This set of features can be denoted as $\{f_1 \dots \dots \dots f_m\}$.

N-grams are commonly used as features when classifying text. An n-gram can be described as a sequence of n tokens (Carpenter 2010). A token can be anything from a single character to a whole sentence. Frequently, however, when text is being classified the tokens used are whole words, and as such n-grams, generally refer to strings of n words. For example, "IPAD2" is a unigram, "Enda Kenny" is a bigram and "honeybunn was good" is a trigram.

The set of *features* can, however, contain anything. In some instances features can consist of manually selected keywords and in other instances features consist of machine generated keywords extracted from the data that is being categorized. Non-lexical tokens such as punctuation (Pang, Lee & Vaithyanathan 2002) or emoticons (Read 2005) can also be used as features.

As $\{f_1 \dots \dots \dots f_m\}$ defines a set of features that can occur in a document we may define $n_i(d)$ as the number of times a feature (f_i) occurs in a document (d).

This implies that a document can be represented by the following vector:

$$\vec{d} \coloneqq \left(n_1\langle d\rangle, n_2(d), \dots \dots n_m(d)\right)$$

Vector 1 Bag Of Words Framework

This vector is described as the "bag of words" or "bag of features" framework (Pang, Lee & Vaithyanathan 2002).

The bag of words framework can be used with a Naïve Bayes classifier to classify a document in the following way.

The sentiment of document d = class c^* which is defined as

$c^* = arg\ max_c\ P(c|d)$

Using Bayes Theorem, $P(c|d)$ is calculated in the following way

$$P(c|d) = \frac{P(c)(\prod_{i=1}^{m} P(f_i|c)^{n_i(d)})}{P(d)}$$

Equation 2 Naive Bayes With Bag Of Words

The Ling Pipe implementation of the Naïve Bayes classifier calculates the maximum likelihood estimations $P(c)$ and $P(f_i|c)^{n_i(d)}$using simple frequency counts. That is $P(c)$ is the number of times class c occurs in the training data divided by the number of training instances and $P(f_i|c)^{n_i(d)}$ is the number of times the feature f_i appears in in a document, d, whose class is c. $P(d)$ can be ignored as it is the same for every document and therefore does not affect the outcome (Pang, Lee & Vaithyanathan 2002).

1.3.6.1.1 Supervised Learning

A supervised learning algorithm relies on labeled training data to classify a given set of test data.

Many machine learning algorithms used in sentiment and subjectivity classification have been chosen because of previously recorded high levels of accuracy in the related task of topic classification of text.

Pang and Lee (Pang & Lee 2004) achieve up to 86.15% accuracy classifying movie reviews in a comparative analysis of text classifiers that had previously proved accurate in topic classification. Parikh and Movassate (Parikh & Movassate 2009) also achieve high levels of accuracy using supervised learning techniques to classify micro blog posts collected from twitter.

While supervised learning methods achieve over 90% accuracy in the task of topic classification they have not been as successful in the task of sentiment and subjectivity classification. This is due to the more complex nature of the task of sentiment and subjectivity classification (Pang, Lee & Vaithyanathan 2002).

1.3.6.1.2 Unsupervised Learning

In contrast to supervised learning algorithms, an unsupervised algorithm relies on no manually classified training data. In many cases statistical methods are used to cluster, or group, outcomes into classes. The use of lexicons is also considered as an

unsupervised learning method, as it does not rely on training data to classify test data (Pang & Lee 2008).

One of the most common uses of unsupervised learning algorithms is in the automated generation of sentiment lexicons (Esuli & Fabrizio 2006). This process is known as lexicon induction.

Turney (Turney 2002) presents an unsupervised approach to sentiment classification and achieves 84% accuracy when classifying automobile reviews and 64% accuracy when classifying film reviews. This approach leverages the mutual information between bigrams extracted from reviews and the seed words “excellent” and “poor”.

Ohana (Ohana 2009) also presents an unsupervised approach to sentiment classification that uses information from the SentiWordNet (Esuli & Fabrizio 2006) sentiment lexicon to train a Naïve Bayes classifier that is then used to classify movie reviews. This approach achieves 69.10 % accuracy.

Go et al. (Go, Bhayani & Huang 2009) also report high levels of accuracy using an unsupervised approach to classify tweets.

1.4 This Paper

The purpose of this paper is to present a survey of some of the different research that has taken place in the field of sentiment analysis, with particular attention paid to the task of sentiment classification. This survey is presented in section 2.

Section 3 of this paper describes the data that is used for the experimental aspect of this research project. This section outlines how the data is collected and manually classified. A brief analysis of the data is also described and discussed in this section.

The programming part of this research project focuses on the problem of sentiment classification of micro blogs. This problem is explored in the form of an experiment where a number of micro blog posts, or tweets, from the micro blogging site twitter are classified as positive or negative using one supervised learning method and one unsupervised learning method. The accuracy of both methods is measured and each result set is analyzed. The effectiveness of various preprocessing and feature selection methods is also examined. The experiments and their implementation are described in section 4.

In section 5 of this paper the results of the experiments are presented and analyzed.

Section 6 of this paper is an overall summary outlining the results and proposing avenues for further research.

2 Literature Survey

2.1 Early Research

A large amount of research in the area of sentiment analysis has been carried out since the year 2000. Prior to this watershed a there have been a number papers in the field of computation linguistics and natural language processing that have concerned themselves with tasks such as the interpretation of metaphor, interpretation of point of view, interpretation of believes expressed in text and tracking narrative (Pang & Lee 2008).

Research by Wiebe (Wiebe 1994) is frequently cited as a forerunner to sentiment analysis (Pang & Lee 2008). This paper presents an algorithm to track the *psychological point of view* in third party narratives in works of fiction and is partly based on the work of literary theorist Ann Banfield who defines psychological point of view as being closely related to subjectivity. The proposed algorithm tracks psychological point of view by identifying subjective sentences and their subjective characters. Through observation Wiebe identifies common methods used by authors to manipulate point of view. The algorithm to track psychological point of view is based on these methods.

Wiebe identifies 12 potentially subjective elements.

1. Exclamations and direct questions • *Ugh!* she [the girl] thought. How could the poor thing have married him in the first place?
2. Elements that express evaluation or judgment • adjectives such as *awful* or *poor* (ambiguous between evaluative meaning and non-evaluative one – "Poor John . . . " vs. "John is poor") • noun phrases such as *old bag* • adverbs such as *oddly* and *incredibly* • auxiliary verbs and phrases that express judgment of obligation, such as *had better*, *ought to*, *should*, and *be supposed to* • adverbs such as *scarcely* and *hardly* when used as minimizer subjuncts ("She could hardly be expected to live there.")
3. Elements that express a lack of knowledge • subordinators such as *whoever* and *whatever* when used in reference to

particular individuals
4. Sentence fragments such as • So this was a seraph. *Tall, even taller than the twins.*
5. Kinship terms such as *Dad* and *Aunt Margaret*
6. Evidentials, which qualify the information conveyed by a statement • evidentials that express (un)certainty, such as *surely* and *might* • evidentials that express (un)certainty and that one's knowledge is based partly on evidence, such as *evidently*, *seemingly*, *must have*, *appear to be*, *as if*, etc. • hedges, e.g. adverbs such as *more or less* and *sort of* when used as modifiers or adjectives • evidentials signaling that expectations have been met (of course) or have not been met (*just*, *merely*, *only*)
7. Adverbials that are conjuncts, which connect units of discourse (*first, in addition*, *after all*, *anyway*, etc.)
8. Conditional clauses
9. Comparative *like*
10. Habitual sentences (*often*, *frequently*)
11. The past perfective, but only in the main verb phrase
12. The progressive but only in the main verb phrase

Table 2 Wiebe's 12 potentially subjective elements

While this paper is more concerned with recognizing subjectivity within text, than sentiment analysis, there is particular relevance to sentiment analysis in the recognition by Wiebe that opinion-oriented language is useful in distinguishing subjective text from objective text. This is evident in Wiebe's 12 potentially subjective elements with a number of them being language elements that express opinion such as points 2 (*Elements that express evaluation or judgment*) and 6 (*Evidentials, which qualify the information conveyed by a statement*).

This is also evident in the association Wiebe makes between subjective sentences and what are referred to as the "*private states*" of characters. Private states are "*states of an experiencer holding an attitude, optionally toward an object*" (Wiebe 1994), in other words opinions, speculations evaluations and emotions (Bruce & Wiebe 1999).

Identifying subjective indicators in text is a feature of much of the literature presented the end of the 1990's. Many of these studies focus on the use of adjectives in particular. In a study by Bruce and Wiebe (Bruce & Wiebe 1999), the results of a manual classification project were statistically analyzed. In this project four judges manually classified the sentences of 14 articles, chosen at random from the Wall street Journal, as being either subjective or objective. The purpose of the statistical analysis was to determine if there is any correlation between individual *features,* or aspects of the sentences being classified, and the subjective class. The result of the statistic analysis demonstrates a strong correlation between subjectivity and the presence of an adjective in a sentence. This supports previous studies where a correlation was indicated. Analysis of the type of adjectives used also finds that *Dynamic Adjectives* show a higher correlation than *Stative Adjectives*[6]. The correlation is established using the G^2 log-likelihood ratio test.

Bruce and Wiebe's findings are the bases for a further study by Hatzivassiloglou and Wiebe (Hatzivassiloglou & Wiebe 2000). This premise of this paper is that certain types of adjectives are more likely to indicate subjectivity in a text than others. This paper proposes sentences that contain *gradable adjectives* used in conjunction with *gradable adverbs*[7] are more likely to indicate the presence of subjectivity. In addition to gradable adjectives the paper also uses the semantic orientation of adjectives as an indicator of subjectivity in sentences (i.e. if a gradable adjective in a sentence has a negative orientation, is the likely hood that the sentence is subjective increased?).

2.2 Supervised Learning

One early attempt to use a supervised learning algorithm to determine semantic orientation of adjectives and create a sentiment lexicon, through the process of

[6] Stative Adjectives denote the static characteristics of objects. Examples of stative adjectives are tall, short, pretty. Dynamic Adjectives on the other hand denote the dynamic properties of objects. All dynamic adjectives can be used in the imperative while stative adjectives cannot. Examples of dynamic adjectives are patient, careful. The following examples illustrate the difference in use between stative and dynamic adjectives

- The lady is being **careful** /**patient** (Dynamic)
- The lady is being **pretty** / **tall** (Stative)
- Be **careful**/ **patient**! (Dynamic)
- Be **pretty** / **tall**! (Stative)

[7] Gradable adjectives are adjectives that describe a property of an object that can be measured in degrees such as size or anger. These adjectives can be used in comparative or superlative forms or with gradable adverbs such as very or extremely. Examples of gradable adjectives used with gradable adverbs are

- The man is **very tall**.
- The man is **extremely angry**.

lexicon induction, is described in (Hatzivassiloglou & McKeown 1997). The authors use the conjunctions "*and*", *"but*" and *"or*" to indirectly determine the *semantic orientation* or *polarity* of the adjectives they conjoin. Hatzivassiloglou and McKeown hypothesize that "*and*" and "*or*" connect adjectives of the same semantic orientation and "*but*" connects adjectives with different semantic orientation. For example "brutal and corrupt" is a correct use of the conjunction "and" but, "brutal but corrupt" is an incorrect use of the conjunction "but".

A four-step process is used to determine the semantic orientation of all adjectives in a corpus.

From a 21 million-word corpus, created from articles from the Wall Street Journal, all conjunctions of adjectives are extracted. The most frequently occurring (20 times or more) adjectives, that have a positive or negative semantic orientation, are then manually classified. The manually classified adjectives are used to validate the hypothesis and used as training data for text classifiers. This manual classification process is verified by comparing the training set with the results of a manual classification exercise carried out by 4 other annotators.

This training set is then used to validate the hypothesis that the semantic orientation of adjectives is constrained by conjunctions. This is achieved by extracting all conjoins from the corpus where both adjectives are present in the training data and examining whether the semantic orientations of the adjectives are the same in the case of "and" and "or" and different in the case of "but". The findings are quite interesting with the results showing that the semantic orientation of conjoined adjectives is the same for 81.73% of "and" conjunctions and 77.05% of "or" conjunctions. Semantic orientation differs for 69.16 % of "but" conjoins.

The results also show that, in the case of all conjunctions, 77.84% have the same semantic orientation. This level of accuracy is used as a baseline against which the machine learning method of classification can be compared.

The next step in the process is to use a log linear regression model to predict whether the semantic orientation of all extracted adjectives is the same or different. This model is weighted with characteristics of the conjunctions observed in the training set. The model plots a graph where each node is an adjective and each conjunction indicates whether the semantic orientation is the same or different.

The third step uses a clustering algorithm that creates two subsets of adjectives where conjunctions between the two subsets mainly indicates adjectives with differing

semantic orientation and conjunctions within each set mainly indicates adjectives with the same semantic orientation.

The final step is to classify the distinct subsets that are generated from the clustering method. As adjectives with a positive semantic orientation are statistically more frequently used than adjectives with a negative semantic orientation, the adjectives in the largest subset are classified as positive.

Steps 2 to 4 are repeated using different subsets of the training data. The accuracy in classifying semantic orientation ranges between 78.08% and 92.37% depending on the subset of the training data used. In all cases the larger of the two subsets is verified as being positive.

Another supervised learning approach to the problem of sentiment classification is taken in Pang et al. (Pang, Lee & Vaithyanathan 2002). Pang et al. do not try to use lexical features that are indicative of sentiment polarity like, for example, Bruce and Wiebe (Bruce & Wiebe 1999) or Hatzivassiloglou and Wiebe (Hatzivassiloglou & Wiebe 2000). Instead they use a number of machine-extracted features of the text that is being classified. The point of doing this is to avoid some of the problems with domain that Turney encountered whereby the accuracy of his approach varies between different domains.

In this supervised learning experiment Pang et al. hypothesis that sentiment classification of documents is similar to the topic classification of documents and therefore, text classifiers that have proved to be successful in the task of topic classification should also be effective in the task of sentiment classification. The goal of the paper is to determine the polarity of a number of movie reviews i.e. “Thumbs up” or “Thumbs down”. The accuracy of three text classifiers, Naïve Bayes, Maximum Entropy and Support Vector Machines, are evaluated. These particular classifiers are chosen as they achieve over 90% accuracy when used for topic classification of documents.

Pang et al. use 1400 movie reviews (700 positive, 700 negative) as the data for this research. The choice of movie reviews is because they offer convenience. They are convenient because they are subjective in nature, have a positive or negative rating and are therefore easily manually classified and also they are readily available on the Internet.

As a baseline for this experiment Pang et al. use a list of positive and negative key words selected from the data by two human subjects. Using these key words, the data

is analyzed and for each review the cumulative totals of the occurrences of positive and negative key words is calculated. If the total of positive words exceeds the total of negative keywords the review is deemed to be positive and vice versa. This base line study achieves 58% accuracy using the key words chosen by human subject 1 and 64% accuracy using the keywords chosen by human subject 2. In a further attempt to create a baseline the authors use automated methods to extract the most frequently occurring tokens[8] from the data. From this list they choose seven positive and seven negative tokens and use these as key words to analyze the data. This resulted in 69% accuracy. An interesting finding of this experiment is that the human subjects only choose adjectives while the statistical method chooses some less obvious tokens such as "?" and "!". This indicates that, despite human intuition and understanding of a piece of text, by using statistics, machines can automatically select more accurate key words or features for the task of sentiment analysis.

The data for the experiment is prepared by removing the HTML tags from the reviews. The effect of negation is also considered and all words between a negation word, such as "not" or "wouldn't" are prepended with the string "NOT_".

In an experiment, validated using 3 fold cross validation, the accuracy of three classifiers are compared. The following features are used

- Unigrams
- Unigrams and bigrams
- Bigrams
- Unigrams annotated with Part Of Speech Tags.
- Adjectives
- The most frequently occurring unigrams. These are all unigrams that occur at least four times in the data.
- Unigrams tagged with the additional information of the position they occur in the text. This final feature is based on an intuition. Pang et al. find that in the domain of movie reviews, the documents being analyzed usually start with an overall statement of the sentiment toward the movie being review, proceed with a discussion of the plot and end with a summary of the author's overall view of the film

The results of these experiments are shown in table

[8] Punctuation is included

Feature	Naïve Bayes	Maximum Entropy	Support Vector Machines
Unigrams	81.0	80.4	82.9
Unigrams + Bigrams	80.6	80.8	82.7
Bigrams	77.3	77.4	77.1
Unigrams + POS	81.5	80.4	81.9
Adjectives	77.0	77.7	75.1
Top Unigrams	80.3	81.0	81.4
Unigrams + position	81.0	80.1	81.6

Table 3 Results for Pang et al.

In general Support vector machines achieve the highest accuracy but not significantly. Naïve Bayes is found to perform the worst. Also all results for features that contain a unigram component are higher than those that do not. This seems to contradict other studies, which find that higher order n-grams such as bigrams outperform unigrams (Pak & Paroubek 2010).

This paper is significant for a number of reasons. Firstly the fact that the machine learning approach produces higher levels of accuracy than the human base line study shows that the approach is viable. The results also show that the accuracy of the text classifiers is lower for sentiment classification than the accuracy of the same classifiers when used for topic-based classification. This shows that topic-based classification differs from sentiment classification. While topic-based classification relies solely on the recognition of keywords, sentiment classification is subtler and relies more on context. Another complexity in sentiment classification that Pang et al. find, is that in the domain of movie reviews there is a tendency for reviewers to "thwart expectations" by building up the imagined potential of a movie and contrasting it against the perceived reality. Movie reviewers also consider a number of different elements such as the performance of an actor in a movie. This is problematic when the quality of the performance of an actor is contradictory to the quality of the movie.

The accuracy of this approach is improved significantly by Pang and Lee in (Pang & Lee 2004). This improvement is achieved by a refinement to the preprocessing applied to the movie reviews that are being classified. Pang and Lee find that the removal of objective elements not only decreases the amount of data that the

classifiers have to process but also increases the accuracy of the classification process with accuracies improving to 86.4% for the Naïve Bayes classifier and 86.15% for Support Vector Machines classifier.

2.3 Unsupervised Learning

An interesting unsupervised learning approach is taken by Turney (Turney 2002). Like Pang et al. (Pang, Lee & Vaithyanathan 2002) the aim of the experiment undertaken in this paper is to assign a rating of "thumbs up" or "thumbs down" to a selection of reviews posted by users of the web site Epinion.com.

The first part of Turney's approach is to extract bigrams that indicate subjectivity from a review. Turney then calculates a numeric value for the semantic orientation of these phrases where a positive value indicates positive sentiment and a negative value indicates negative sentiment. The sentiment of the overall document is then calculated by summing these values.

Turney chooses bigrams as indicators of subjectivity because, although earlier work shows that adjectives and adverbs are strong indicators of subjectivity (Hatzivassiloglou & McKeown 1997) (Bruce & Wiebe 1999) (Hatzivassiloglou & Wiebe 2000), in isolation they do not provide context. For example, the adjective "unpredictable" may have a negative orientation in an automobile review when used in a phrase such as "unpredictable steering", but it could have a positive orientation in a movie review, in a phrase such as "unpredictable plot". In the two examples above "steering" and "plot" add context.

This first step of the process uses a Part of Speech tagger to tag each word in the reviews with Part Of Speech information. Bigrams that exhibit the Part Of Speech patterns, shown in table 4, are then extracted from the reviews.

First Word	Second Word	Third Word (Not Extracted)
JJ	NN or NNS	Anything
RB, RBR or RBS	JJ	Not NN or NNS
JJ	JJ	Not NN or NNS
NN or NNS	JJ	Not NN or NNS
RB, RBR or RBS	VBN or VBG	

Table 4 Patterns of POS tags for bigram extraction

POS Tag	Meaning

JJ	Adjective
NN	Noun, singular or mass
NNS	Noun, plural
RB	Adverb
RBS	Adverb, superlative
RBR	Adverb, comparative
VBN	Verb, past participle
VBG	Verb, gerund or present participle

Table 5 Penn Treebank tag set[9]

The semantic orientation of each extracted phrase is then calculated using Turney PMI-IR (Point wise Mutual Information[10] and Information Retrieval) algorithm.
The PMI part of this algorithm is based on the work of Church & Patrick (Church & Patrick 1990) and is a metric that measures the strength of relationships between words in a corpus. For the purpose of this experiment Turney measures the relationship between each extracted bigram, and the positive seed word "excellent" and negative seed word "poor".
The seed words "excellent" and "poor" were chosen in relation to the star rating system on the Epinion site where 1 star represents "poor" and 5 stars represents "excellent".
The algorithm to calculate Point wise Mutual Information is defined as

$$PMI(word_1, word_2) = log_2\left(\frac{p(word_1 \& word_2)}{p(word_1)p(word_2)}\right)$$

Equation 3 PMI

where

- $word_1$ is the extracted bigram
- $word_2$ is the seed word
- $p(word_1 \& word_2)$ is the probability that $word_1$and $word_2$ co occur

[9] http://www.ling.upenn.edu/courses/Fall_2003/ling001/penn_treebank_pos.html
[10] "*In our application, word probabilities P(x) and P(y) are estimated by counting the number of observations of x and y in a corpus, f (x) and f(y), and normalizing by N, the size of the corpus. (Our examples use a number of different corpora with different sizes: 15 million words for the 1987 AP corpus, 36 million words for the 1988 AP corpus, and 8.6 million tokens for the tagged corpus.) Joint probabilities, P(x,y), are estimated by counting the number of times that x is followed by y in a window of w words, fw (x,y), and normalizing by N.*" (Church & Patrick 1990)

- $p(word_1)p(word_2)$ is the probability that $word_1$and $word_2$ co occur, if the words are statistically independent
- the ratio is the measure of statistic dependence between the two words.
- the log of this ratio is the amount of information provided about one word when we observe the other.

The Semantic Orientation (SO) of a phrase is calculated using the following algorithm

$$SO(phrase) = PMI(phrase, "excellent") - PMI(phrase, "poor")$$

Equation 4 Semantic Orientation

The "IR" part of "PMI-IR" stands for information retrieval. This is because PMI is calculated by querying the Internet using a search engine. The IR part of Turney's approach makes this approach unique as the Internet itself is used as a corpus[11].

The Alta Vista search engine is used to execute queries. Alta Vista is chosen because one of the advanced search options is the NEAR operator. This operator limits results returned, to documents where the queried words that are within 10 words of each other. For example a query of "low fees" NEAR "excellent" would only return documents where an occurrence of the words "low fees" is not separated from the word "excellent" by more than 10 words. The "NOT" operator is also used to exclude documents from the Epinion web site so that the source material is not included in the search results.

Alta Vista is used in the following way

- A query is run using the extracted feature.
- A query is run using a seed word.
- A query is run for the extracted feature NEAR the seed word.
- The values $p(word_1)$, $p(word_2)$ and $p(word_1 \& word_2)$ from Equation 3, are then calculated using the results from all three search results.

Through manipulation of Equation 3 and Equation 4 we get Equation 5, which is PMI-IR.

$$SO(phrase) = \log_2 \left(\frac{hits(phrase\ NEAR\ "excellent")\ hits("poor")}{hits(phrase\ NEAR\ "poor")\ hits("excellent")} \right)$$

Equation 5 PMI-IR

where

- $hits(query)$ is the number hits returned for the given query.

[11] At the time of writing (2002) Alta Vista indexed approximately 350 million web pages

Using this method Turney achieves 84% accuracy when analyzing automobile reviews but only achieves 64% accuracy when analyzing movie reviews. This highlights the role that domain can play in sentiment analysis.

Other unsupervised approaches to sentiment classification take to approach of using automatically generated sentiment lexicons to train classifiers (Pang & Lee 2008). The method proposed by Ohana in (Ohana 2009) differs slightly from this as the author explores the possibility of using the freely available sentiment lexicon SentiWordNet (Esuli & Fabrizio 2006), an opinion lexicon where each synonym set (synset) of WordNet (Miller 1995), a thesaurus like lexicon resource, is associated with three numerical scores; positive, negative and objective. Ohana uses SentiWordNet to establish the semantic orientation of unigrams. These unigrams are then used to train a Naïve Bayes classifier.

Using the same movie review data as Pang et al. (Pang, Lee & Vaithyanathan 2002) Ohana extracts all unigrams the data. For each unigram a positive, negative and neutral value is calculated using SentiWordNet. If the value is positive the unigram is used as positive training data, if the value is negative the unigram is used as negative training data and if the value is neutral (or zero) the unigram is used as neutral training data.

The best accuracy achieved with this method, validated using 3 fold cross validation, is 69.10%. A number of factors are considered for this low (comparatively speaking) level of accuracy.

Colloquialisms and slang are not represented in the SentiWordNet lexicon the presence of such words in the testing data directly affects the overall accuracy. This is also true for text in the reviews such as film titles and actors names.

Like Pang et al. (Pang, Lee & Vaithyanathan 2002) the structure of the reviews also affects the accuracy of Ohana's experiment with Ohana also finding that using information relating to the position of a feature within the document increases accuracy.

Another reason for a low level of accuracy is inaccuracies in the Part Of Speech tagger. Miss-tagged words lead to inaccurate SentiWordNet values, as the information contained in SentiWordNet is dependent, in part, on Part Of Speech information.

Finally the problem of "thwarted expectations", experienced by Pang et al. (Pang, Lee & Vaithyanathan 2002), also effects the accuracy recorded by Ohana.

The comparative analysis preformed by Pang et al. in (Pang, Lee & Vaithyanathan 2002) is emulated by Go et al. in (Go, Bhayani & Huang 2009) with the exception that Go et al. use data collected from the micro blogging site twitter.

This work falls into a category of machine learning that is can be described as distant, or semi-, supervised learning as, instead of using manually classified tweets as training data, Go et al. use tweets that contain the emoticons ":)" and ":(" where ":)" indicates positive sentiment and ":(" indicates negative sentiment. Although sentiment expressed by emoticons sometimes may not relate to the specific query, as in the following tweet that expresses negative sentiment with regard to the topic "Barrack Obama" "*@wiskey1249 At least the GOP realizes that Obama is capable of trying every behind-the-back move he can get away with. That's good! :)*", and also in this tweet that expresses positive sentiment with regard to the topic "IPAD2" "*waiting and waitng for my ipad2 :(its soooo boring ! but illl be welll happpy when it comes !!!*", tweets that contain emoticons are likely to contain opinionated text (Pak & Paroubek 2010). This is approach is similar to the approach taken by Read in (Read 2005).

Using this method Go et al. collect a data set of 1.6 million tweets. From this data they extract 359 tweets as test data with the remainder being used as training data. The testing data is manually classified as positive or negative. All data is pre processed in the following way

- Any tweets that contain conflicting emoticons are removed from the data set i.e. tweets containing both ":)" and ":(".
- Emoticons are removed from the data. This is a necessary step as because their frequency within data causes the Maximum Entropy and Support Vector Machine classifiers to weight them too heavily. This is not an issue for the Naïve Bayes classifier.
- Retweets are removed.
- Tweets with the emoticon ":P" are removed. The reason for this is a bug in the Twitter API that also returns tweets containing the emoticon ":P" when the query string contains the emoticon ":(".
- All tweets that are repeated are removed from the dataset.

The effect of each individual step involved in preprocessing the data is not evaluated.

As a base line study for this experiment Go et al. use a list of positive and negative keywords that are used by the twitter sentiment analysis site Twittratr[12] to train their classifiers. This approach achieves 65.2 % accuracy.

Unlike Pang et al. in (Pang, Lee & Vaithyanathan 2002) there is no cross validation employed in the evaluation of the three classifiers.

The following features are used during the evaluation

- Unigrams
- Bigrams
- Unigrams and bigrams
- Unigrams appended with Part Of Speech tags

The results for this evaluation are shown in table 6.

Feature	Naïve Bayes	Maximum Entropy	Support Vector Machines
Unigrams	81.3	80.5	82.2
Bigram	81.6	79.1	78.8
Unigram + Bigram	82.7	83.0	81.6
Unigram + POS	79.9	79.9	81.9

Table 6 Results for Go et al.

The levels of accuracy outperform the base line study and are also comparable to the results achieved by Pang et al. in (Pang, Lee & Vaithyanathan 2002) but are lower to the results of Pang and Lee in (Pang & Lee 2004). Also, similar to Pang et al., the results indicate that the Support Vector Machines classifier performs the best and Naïve Bayes is the least accurate although the difference is not hugely significant. Additionally Go et al. do not experience the same problem with domain dependence that Turney experiences in (Turney 2002) even though the training data and test data are drawn from a number different topics in a number of different domains.

One of the conclusions that Go et al. draw is that twitter can effectively be used as a corpus for the sentiment classification of micro blogs. This is conclusion is examined by Pak and Paroubek (Pak & Paroubek 2010). In this study Pak and Paroubek collect 300,000 tweets from twitter that are evenly divided into three classed positive, negative and objective. The distinction between positive and negative is achieved by

[12] http://twitrratr.com

using emoticons while the objective tweets are collected from the twitter accounts of newspapers.

The data is then analyzed in two ways.

The first analysis performed is to establish if the distribution of word frequencies conforms to Zipf's law, which states,

"in any corpus of natural language utterances the frequency of any word is inversely proportional to its rank in the frequency table. Thus the most frequent word will occur approximately twice as often as the second most frequent word, three times as often as the third most frequent word, etc."[13].

The finding was that the collected data conformed to Zipf's law.

The second part of the analysis is a study of the distribution of Part Of Speech tags with regard to the objective tweets and the subjective tweets (all positive and negative tweets) and the distribution of part of speech tags with regard to the positive tweets and the negative tweets. Equation 6 is used to calculate the pair wise comparison of Part Of Speech tag distributions

$$P_{1,2}^{T} = \frac{N_1^T - N_2^T}{N_1^T + N_2^T}$$

Equation 6 Pair wise comparison of tag distributions

where

- N_1^T is the number of occurrences of tag T in set 1
- N_2^T is the number of occurrences of tag T in set 2

This kind of analysis enables the authors to observe some of the lexical features of the collected tweets. Many of the findings of the analysis of objective and subjective are in line with Wiebe's 12 potential subjective elements. Some examples of the lexical features revealed by this analysis are

- Using the set of objective and subjective tweets, Part Of Speech tags are not evenly distributed, which means that Part Of Speech could be used as an indicator of subjectivity.
- There is a tendency for objective tweets to contain more proper nouns while subjective tweets contain more personal pronouns.
- Positive tweets show a greater use of superlative adverbs such as "most" or "best".

[13] http://en.wikipedia.org/wiki/Zipf's_law

- Negative tweets show a greater use of verbs in the past tense such as “missed” or “bored”. This is because the authors of negative tweets often express feelings about personal loss or disappointment.

The next step in the process for Pak and Paroubek is to pre process the data. The data is preprocessed in the following way

- All URL links, i.e. all substrings starting with “http://”, are removed.
- All twitter user names, i.e. all substrings starting with “@”, are removed.
- All occurrences of the string “RT”, which stands for retweet, are removed.
- All emoticons are removed.
- The stop words “a”, “an” and “the” are removed.
- All negations such as “no” and “not” are concatenated to the end of the word that immediately precedes them. For example “do not” becomes “donot”.

A Naïve Bayes classifier, which uses n-grams and Part Of Speech information, observed during the analysis of the data, as features, is then used to classify the data. One particular finding of the classification process is that the accuracy when using bigrams is better than unigrams and trigrams. The authors suggest that bigrams strike a balance between unigrams that provide coverage of the words used and sentiment expression patterns that are captured in trigrams.

2.4 Other Literature

This importance of the concept of private states, introduced by Wiebe in (Wiebe 1994), is further explored by Wiebe et al, in (Wiebe, Wilson & Cardie 2005). The main goal of this paper is to produce a corpus that is annotated with detailed information about opinions and emotions that could be used to aid the development of Natural Language Processing applications that use such information. Using this corpus these applications could not only identify the opinion orientation of a snippet of text, they could also determine the strength of that opinion. This goal is achieved by the manual annotation of phrases within a large corpus that express private states. Wiebe et al also introduce a low-level XML-like annotation scheme.

While a lot of work in sentiment analysis had been done in domain of reviews this work focuses on news articles. The nature of the way news is reported means that there can be multiple opinions expressed in one sentence. The corpus itself was collected between June 2001 and May 2002 and consists of 565 documents from 187 different newspapers.

Each private state, that is identified, is annotated with what is described as a *private state frame*. A *private state frame* contains attributes such as the source, subject, type of attitude expressed, the intensity and the significance of what is expressed in the private state. The annotations are done at word and phrase level as opposed to document or sentence level annotations. The context of the word or phrase is also considered by the annotator, which means the resulting annotations are fine grained. Three kinds of private states are annotated

- Explicit mentions of private states such as fear, love, hate, etc. In the example *"The U.S.* ***fears*** *a spill-over," said Xirao-Nima* fears is the explicit mention of a private state.
- Speech events expressing a private state. A speech event is any speaking or writing events. An example of this is *"The report is full of absurdities," Xirao-Nima* ***said*** where "said" denotes a speech event.
- Expressive subjective elements. These are elements in speech events where a private state is implicitly expressed. In the example *"The report* ***is full of absurdities****," Xirao-Nima said*, "full of absurdities" is an expressive subjective element.

These three kinds of private states give rise to three private state frame types

- Expressive subjective element frames
- Objective speech event frames
- Direct subjective frames

When annotating, expressive subjective private state frames can be nested in objective speech event private state frames.

Three annotators performed the annotation manually and the findings are validated in an inter annotator agreement study. This study is divided into three parts

1. Agreement between annotators with regard to expressive subjective element frames
2. Agreement between annotators with regard to direct subjective and objective speech event frames (explicit frames).
3. Agreement between annotators with regard to the distinction between direct subjective and objective speech event frames.

Cohen's K-coefficient or Kappa is the de-facto standard statistical measure of inter-rater, or inter-annotator, agreement. The first two parts of the study cannot be

measured using K however as there is no guarantee that the annotators identify the same set of expressions. As a result of this, a measure of the intersection between annotators is calculated using the agr metric

$$agr(a||b) = \frac{A\ matching\ B}{A}$$

Equation 7 Agr Metric

where

- A is the total set of text anchors annotated by annotator a
- B is the total set of text anchors annotated by annotator b

The following example illustrates the criteria for A matching B. In the sentence "*We applauded this move because it was not only just, but it made us begin to feel that we, as Arabs, were an integral part of Israeli society"* annotator a annotates "not only" and "integral" while annotator b annotates "because", "not only just", "but" and "integral part". In this instance there is a match between "not only" and "not only just" and also a match between "integral" and "integral part".

The findings of parts 1 and 2 of the inter annotator agreement study is an average pair wise score of 0.72 with regard to expressive subjective element frames and an average pair wise score of 0.82 with regard to direct subjective and objective speech event frames. This indicates that the annotators find it easier to identify direct subjective and objective speech frames.

The third part of the inter annotator agreement study can be measured K as the set of expressions being evaluated are the same.

$$\kappa = \frac{P(A) - P(E)}{1 - P(E)}$$

Equation 8 Cohen's k-coefficient

where

- P(A) is the calculated percentage agreement
- P(E) is the calculated percentage of chance agreement

The average pair wise score for K is 0.81which shows a high level of agreement between annotators[14]. This shows that manual annotation is possible but also that such classification is subjective.

[14] See Appendix B

3 The Data

The data for this experiment was collected from twitter between 22/05/2011 and 24/07/2011. In that time 50,244 tweets were downloaded. 10 different queries in the categories sport, electronic products, movies and politics were used. The queries chosen are shown in table 7. These queries were chosen as at the time of harvesting they were in the media spotlight and the hope was that they would be emotive.

Query	**# of Tweets**	**Category**
Manchester United	6204	Sport
Leinster[15]	3868	Sport
IPAD2	5810	Electronic Product
Queen Elizabeth	4269	Politics
Enda Kenny	2353	Politics
Pirates Of The Carribean[16]	4847	Movies
Kindle	6484	Electronic Product
HTC Thunderbolt	3822	Electronic Product
Super 8	5620	Movies
Obama	6967	Politics

Table 7 Queries

As Parikh & Movassate in (Parikh & Movassate 2009) find that on average 5 to 10 out of every 200 tweets contain sentiment there is a need to eliminate tweets that could be reasonably deemed useless to this experiment. This was done using the following selection process:

- Any tweet containing a URL, identified by the sub string "http://", was considered spam and removed.
- Any tweet starting with the substring "RT", which signifies a retweet[17], were removed to avoid duplicate data.
- In an attempt to only be left with subjective tweets to manually classify, the remaining tweets were Part Of Speech tagged using the Stanford Part Of Speech Tagger and classified as subjective or objective using Turney's Part Of

[15] This should have been Leinster Rugby as after the hype from the Heineken Cup died down Leinster proved to be a very ambiguous search.
[16] This highlights the frequent spelling mistakes in tweets by the number of tweets returned for this query, although to be honest it was a not a deliberate spelling mistake on my part.
[17] A retweet is a tweet that is reposted by another user

Speech patterns (Turney 2002) from table 4[18]. All objective tweets are discarded.

This process eliminated some 42,237 tweets leaving 9,541 tweets to be manually classified. The breakdown of the remaining tweets is shown in table 8

Query	**# of Tweets**	**Category**
Manchester United	634	Sport
Leinster	1061	Sport
IPAD2	852	Electronic Product
Queen Elizabeth	725	Politics
Enda Kenny	581	Politics
Pirates Of The Carribean	1739	Movies
Kindle	820	Electronic Product
HTC Thunderbolt	487	Electronic Product
Super 8	1683	Movies
Obama	959	Politics

Table 8 Subjective Tweets

The remaining tweets were then manually classified as positive or negative. The target here was to create a dataset of 600 positive tweets and 600 negative tweets. Although some preprocessing was done manual classification was still a difficult task as many tweets still had to be removed as they were unusable.

Some tweets did not relate to the specific topic. This occurred frequently for the topic "Super 8" as twitter returned tweets containing the word "super" and the number "8". For example "*#Mizzou softball won 8-0 to move onto super regionals Chelsea Thomas is a #beast throwing a no hitter*" or "*#Prediction: Dublin-Kilkenny will contest every hurling final this year. Walsh Cup, NHL, Leinster, only All-Ireland to go. #gaa #1500thTweet*" which does not refer to Leinster Rugby.

Other tweets were removed because I considered them to be neutral, such as "*Do you know that Obama will fly in his own steps ahead of AF1 that he will use to get off the plane? #obamavisit*". While relevant to the query "Obama" there is no sentiment expressed.

Some tweets were strange. "*@damienmulley Enda Kenny issues statement denying he was found wandering through Dail last night wearing the Queens dress. #nottherapture*".

[18] This classification was not verified and only used as a rough guide to help reduce the workload involved in the manual classification process.

Although all tweets containing the substring "http://" were considered as spam and removed, there were still tweets that could also be considered as spam. "*#FollowFriday @TheNJFGroup - #iPad2 giveaway when we hit 3000 followers - already given 3 away plus 5 @Jusslondon candles RT AND FOLLOW NOW*".

Also there were tweets that just made no sense "*GRABE PIRATES OF THE CARRIBEAN I CAN'T EVEN*".

Positive tweets occurred more frequently than negative tweets and this means after the manual classification task was completed there were 1,165 positive tweets and 601 negative tweets. Table 9 shows how these tweets were distributed by query.

Query	**# Positive**	**# Negative**
Manchester United	86	15
Leinster	86	15
IPAD2	111	26
Queen Elizabeth	46	14
Enda Kenny	168	101
Pirates Of The Carribean	283	77
Kindle	77	29
HTC Thunderbolt	78	87
Super 8	178	54
Obama	52	183

Table 9 Breakdown of manually classified tweets.

600 negative and 600 positive tweets were used during experiments.

3.1 Data Analysis

To establish whether or not the collected data can be considered as a valid corpus the frequency distribution of the all the words[19] in the data is calculated. Figure 1 shows the frequency distribution plotted on a graph. This distribution shows that the data is consistent with Zipf's law, which means the data shows the proper characteristics of a corpus. One interesting feature of the data is that the first two most commonly used words in the English language "the" and "of" are also the first two most commonly used words in the collected tweets.

[19] All strings greater than length 1

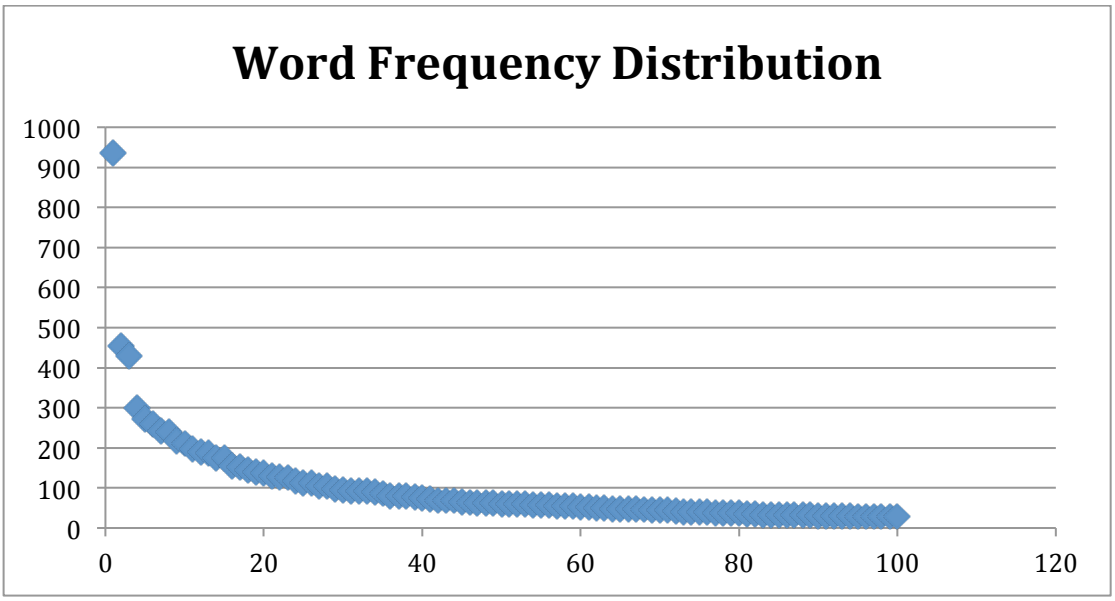


Figure 1 Word Frequency Distribution[20]

Next the distribution of Part Of Speech tags for all negative and positive tweets is examined. The equation to calculate P^T (equation 6) is used.

In both sets there are outliers that are exclusive to category, WH-pronoun possessive[21] (/WP$) in the case of positive tweets and list item marker[22] (/LS) in the case of negative. These tags can be discounted as having no meaning due to the frequency with which they occur (each occurs only once in their respective datasets).

From this data we see that positive tweets show a number of characteristics such as use of verbs in the past tense (/VBD), use of cardinals or numerics (/CD), use of symbols (/SYM), adverbs (/RB), interjections (/UH) and pre-determiners (/PDT).

The use of verbs in the past tense is contrary to the findings in (Pak & Paroubek 2010) although it must be noted that the dataset used by Pak & Paroubek is larger than the dataset used in this research paper.

Also of interest here is the (/FW) tag, which denotes a foreign word, but in the case of this data is applied to slang, misspellings and the personal pronoun "I" when it is used in lower case. This implies that the authors of positive tweets are more likely to use slang, misspell words or use "I" in the incorrect case and suggests that positive tweets are less considered and more casual that negative tweets. This is also evident by the higher frequencies of Part Of Speech tags overall in negative tweets.

[20] Top 100 words
[21] "Whose"
[22] A., B., C., D., First, Second, Third, Three, Two, a, b, etc.

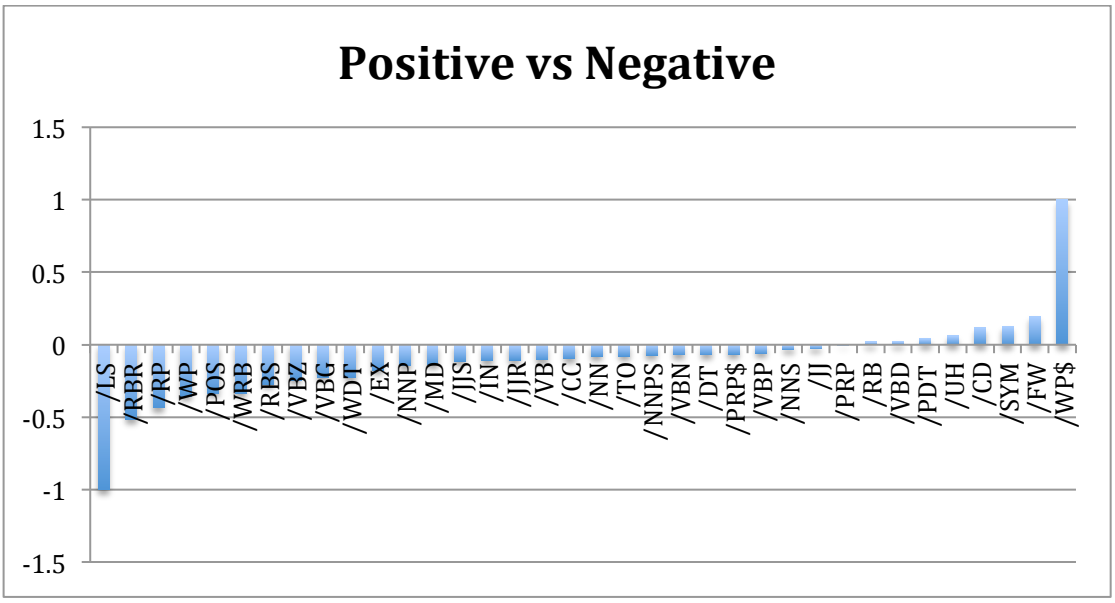


Figure 2 P^T values for positive vs. negative

This analysis does, however, offer little information that could be used to train a classifier as the distribution is too heavily biased to the negative class.

4 Experiment

Figure 1 outlines the process flow of this experiment.

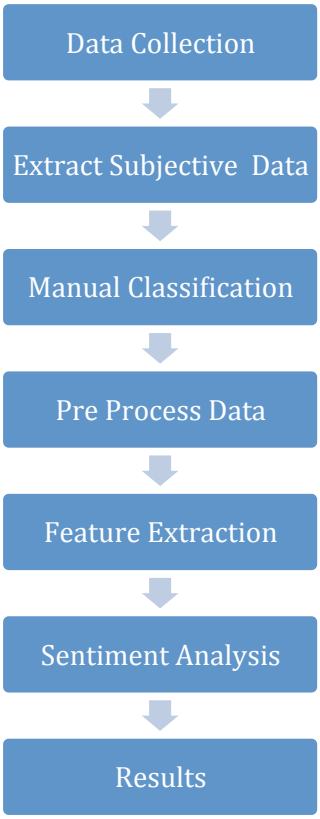


Figure 3 Experiment Process Flow

The goal of the experiment is determine if previous approaches to sentiment classification text can be applied to micro blog posts. To establish this the accuracy of one supervised learning method of sentiment classification and the one unsupervised learning method of sentiment classification are evaluated.

4.1 Supervised Learning

In this supervised learning approach to sentiment classification the problem is treated as a binary text classification problem using two classes, positive and negative. To classify the tweets a text classifier is used. The text classifier employed in this experiment is the implementation of the Naïve Bayes classifier provided in the Ling Pipe Natural Language Processing library. Although other classifiers mentioned in the literature, such as Support Vector Machines, are also implemented in Ling Pipe, Naïve Bayes is chosen because it is relatively easy to configure, features more frequently in the literature and high levels of accuracy (Pang & Lee 2004) (Go, Bhayani & Huang 2009)have been achieved when used for the task of sentiment classification.

For the purpose of this paper the classification is validated using 10-fold cross validation.

In a 10-fold cross validation process the manually classified data is divided into 10 equal parts or folds. The data is then classified in an iterative process. During each iteration nine folds are used to train the text classifier and one fold is used as testing data. This means that the classifier evaluates every fold once. The accuracy for each iteration is then averaged to give an overall accuracy.

Using 10 fold cross validation also allows a kappa value to be calculated.

As a base line study the data is classified using a Naïve Bayes Classifier trained with unigram features extracted from the same training data as Go et al[23]. this approach achieves an accuracy of 63%.

4.2 Unsupervised Learning Approach

The unsupervised learning approach to sentiment classification uses the SentiWordNet (Esuli & Fabrizio 2006) sentiment lexicon to assign a numeric value to each word in each tweet. The score assigned to each word is the semantic orientation

[23] 1.6 million tweets

of the word. This value will be positive in the case of a positive semantic orientation and negative in the case of a negative semantic orientation.
Tweets are then classified as positive or negative in the following way

- For each tweet an overall value is calculated by summing the numeric values for semantic orientation
- If the summed value is positive the tweet is classified as positive
- If the summed value is negative the tweet is classified as negative
- If the summed value is zero the tweet is classified as negative if it contains more words with negative semantic orientation than positive semantic orientation and vice versa.

To allow for the effect of negation, the positive and negative scores extracted from SentiWordNet for negated words are reversed, by multiplying each by -1.
If a word is immediately preceded by an intensifier, such as “very” or “really”, the positive and negative values are doubled.
As a base line study for this experiment the approach taken by Ohana in (Ohana 2009) is emulated. In this approach a Naïve Bayes classifier is trained with a set of unigram features whose class is determined by their semantic orientation, which is established using SentiWordNet. This approach achieves 57.7% accuracy.

4.3 Experiment Setup

The experiment is set up in the following way:

- Tweets are collected from twitter.
- A crude mechanism is used to extract tweets that show indications of being subjective to reduce the amount of tweets that need to be manually classified. At this point Retweets[24] and spam[25] are also removed.
- The tweets are manually classified as positive or negative and therefore the numbers of positive and negative tweets are known. This is essential so that the accuracy is can be measured and also for training the classifier in the supervised learning experiments.
- A number of different preprocessing methods are applied to the tweets before feature extraction and classification. The purpose of preprocessing is to clean

[24] A retweet is a re post of a message by another user. Any post starting with the letters “RT” is considered a retweet.
[25] All tweets containing the substring ”http://” are considered spam.

and normalize the data. These methods can be turned on or off so their effects on the overall accuracy of the process can be measured. The methods employed for these experiments are

- Spelling correction. Using a spell checker any words that are misspelled are replaced by the spell checker's top ranked suggestion.
- Stemming. The process of stemming attempts to reduce a word to its canonical form. For example "seems" becomes "seem" or "attacking" becomes "attack". This process does not, however, guarantee that the resulting form is an actual word. For example "catholic" becomes "cathol" and "very" becomes "veri".
- Removing any sentences from individual tweets that do not contain the query string. For example "*Super 8 was stupid btw. I didn't like it*." becomes "*Super 8 was stupid btw*".
- Removing the query string from individual tweets and replacing it with the string <TOPIC>. For example "*@Darius_wafflez haha! I want the HTC Thunderbolt. It looks like a cool phone*." becomes "*@Darius_wafflez haha! I want the <TOPIC>. It looks like a cool phone*."
- Removing stop list words. Stop list words are commonly used words that by themselves convey little or no information such as "and', "the" and "by"[26].
- Removing punctuation marks and other noisy data from tweets such as "/", ")", ".", ",", ":", ";" etc.
- Removing user names from tweets. User names in twitter can be identified by the symbol "@" followed immediately by a string.
- Removing hash tags from tweets. Hash tags are used to group threads together and add context to a tweets. They are identified by the symbol "#" immediately followed by a string.
- Negation. The string "***NOT_"*** is prepended to all words that occur between negation words such as "not" or "wouldn't" and a non-alphanumeric character. For example "*Seen pirates of the carribean saturday .....yeah not so good. The third still the best*" becomes "*Seen pirates of the carribean saturday .....yeah not* ***NOT_****so* ***NOT_****good. The third still the best*" and

[26] See Appendix A for the full list.

"*Enda Kenny not doing too bad on the Late Late #latelate"* becomes "*Enda Kenny not **NOT**_doing **NOT**_too **NOT**_bad **NOT**_on **NOT**_the **NOT**_Late **NOT**_Late #latelate*".

- The classification of tweets is done using the bag of words framework. A number of different features are used. These are
    - All adjectives. All adjectives are extracted from the manually classified data to train the classifier.
    - Turney's Bigrams. All bigrams that match the patterns suggested by Turney in table 4 are extracted and used to train the classifier.
    - Unigrams. All individual words are extracted and used to train the classifier. For example the unigram features of "*Enda Kenny not doing too bad on the Late Late #latelate*" would be "Enda", "Kenny", "not", "doing", "too" and so on.
    - Bigrams. All word pairs are extracted and used to train the classifier. For example the bigram features of "*Enda Kenny not doing too bad on the Late Late #latelate*" would be "Enda Kenny", "Kenny not", "not doing", "doing too", "too bad" and so on.
    - Trigrams. All word trios are extracted and used to train the classifier. For example the trigram features of "*Enda Kenny not doing too bad on the Late Late #latelate*" would be "Enda Kenny not", "Kenny not doing", "not doing too", "doing too bad", "too bad on" and so on.
    - Hash tags. Hash tags are used to group threads together and add context to a tweets. They are identified by the symbol "#" immediately followed by a string.
- Finally the tweets are classified as either positive or negative and the results are recorded. The classification is done using ling pipe's implementation of the Naïve Bayes classifier in the case of the supervised learning portion of this experiment and using a classifier based on SentiWordNet in the case of the unsupervised portion of this experiment.

4.4 Implementation

Three separate applications were constructed for the purpose of carrying out this experiment. These applications were programmed in the java programming language and use a number of open source libraries. These libraries are Twitter, the ling pipe

natural language processing tool kit, the Stanford University part of speech tagger (Toutanova et al. 2003), and the Suggested spell checking java library. In addition to these libraries the unsupervised learning portion of the experiment required the WordNet (Miller 1995) and SentiWordNet (Esuli & Fabrizio 2006) lexicons.
Data is stored in a MySQL database.

4.4.1 JTwitter

JTwitter[27] is a java library produced by Winterwell Associates that leverages the public Twitter API[28]. This library is used to connect to Twitter and submit queries. This library also allows the user to specify extra query parameters such as language.

4.4.2 Ling Pipe

Ling Pipe is an extensive java library developed for the use in performing many Natural Language Processing and Computational Linguistic tasks. For the purpose of this paper a number of classes provided by Ling Pipe are used.

4.4.3 Stanford University Part Of Speech Tagger

The Stanford University Part Of Speech Tagger is a java library that reads text and assigns part of speech tags, such as noun, verb, punctuation mark etc., to each word or symbol contained in the text. This library is versatile can use different language model files that contain taggers that have been trained using different languages and source material. The language model used for this paper is the "bidirectional-distsim-wsj-0-18.tagger". This model is constructed for the English language using articles from the Wall Street Journal and has been found to have a high level of accuracy (97.28% accuracy tagging Wall Street Journal articles, 90.46 % accuracy tagging unknown texts). This library uses the Penn Treebank Part Of Speech tag set.

4.4.4 Suggester Spell Checking Library

Suggester spell checker is a spell-checking library that not only identifies misspelling but also offers spelling suggestions that could be possible replacements for the misspelled word.

[27] http://www.winterwell.com/software/jtwitter.php
[28] Application Programming Interface.

4.4.5 Tweet Collection Tool

This application is used to gather the initial data for the experiment and also perform some preprocessing tasks to help reduce the workload needed for the manual classification task. No user interface is needed for this application, as no user intervention is required.

This application uses JTwitter to connect to and query twitter. JTwitter connects to twitter using the OAuth (Open Authorization) protocol. This the security protocol used by twitter to allow applications to use the Twitter API.

The Twitter API allows 350 requests per hour for applications using OAuth authentication. This means that for each query this application can only send $\frac{350}{\#\ of\ queries}$ requests per hour to twitter in compliance with this limit. Another restriction imposed by the twitter API is that a maximum 100 tweets may be returned per request.

To help eliminate duplicate data being retuned the last twitter id for each query is added as a parameter to each request. Also only tweets in the English language are requested, although the twitter still returns a small number of tweets in foreign languages.

As the task of manual classification is a time consuming process this application employs a crude mechanism to discard tweets that do not show indications of being subjective. This is done using the Stanford University Part Of Speech tagger and the bigram patterns Turney uses in (Turney 2002). These patterns are described in table 4. I describe this process as crude as there is no verification as to whether this process correctly identifies subjective tweets or not.

In addition to removing objective tweets to help reduce the overhead in the manual classification task, tweets that can be considered spam or retweets are also eliminated. This is done using the following rules of thumb

- If a tweet contains the substring "http://" it is considered spam
- If a tweet starts with the substring "RT" it is considered a retweet.

All tweets that remain after this process are stored in a MySQL database.

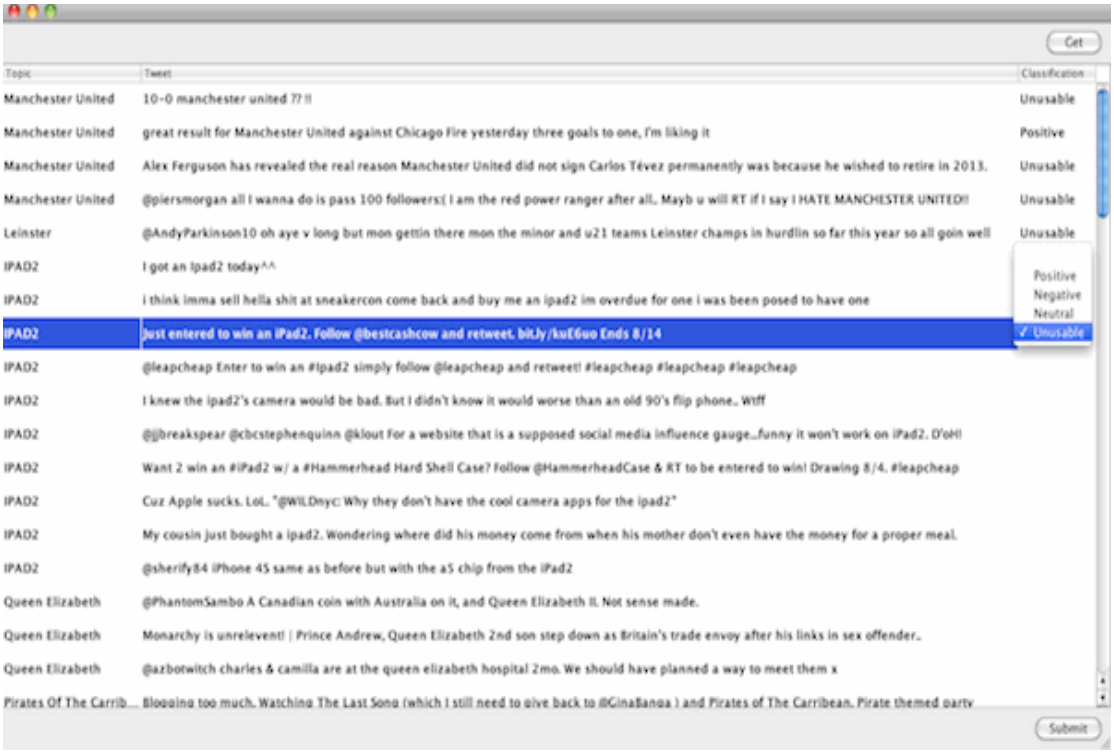
4.4.6 Manual Classification Tool

This application is used to display tweets to the user and allow the user to manually classify individual tweets as positive, negative, neutral or unusable. The manually classified tweets are then stored in the MySQL database.

The process of manual classification is time intensive and by no means as easy as it first seems. The user has to process a large amount of data, approximately 10,000 tweets and therefore the user interface for this application needs to be a simple as possible for the following reasons

1. The user interface should not distract the user from the task at hand as it is important to remain focused.
2. Tweets need to be easily read.
3. Accuracy is important so the process of selecting and classifying tweets needs to be easily achieved.

Screenshot 1 shows the user interface for this application.



Screenshot 1 Manual Classification Tool User Interface

The application retrieves a maximum of 100 tweets at a time to be classified. These tweets are displayed in tabular form to the user. Only the query string and tweet are displayed to the user. The third column in the table that is displayed to the user is for the user to enter the classification of each tweet. This is achieved by selecting positive, negative, neutral or unusable from a drop down selection list. The classified tweets can be submitted to the MySQL database by pressing the submit button.

4.4.7 Sentiment Analysis Tool

This application is used to perform preprocessing tasks, feature extraction and sentiment classification. No user interface is needed for this application as it performs a number of predefined experiments that require no user interaction.

A framework is developed to facilitate the creation of experiments that combine different a number of different prepossessing tasks, features and classifiers. In this framework preprocessing tasks, feature extraction tasks and classifiers are treated as separate components.

Using this framework each experiment is comprised of an array of preprocessing tasks, an array of feature extraction tasks and a text classifier.

This application can perform the following pre processing tasks

- Remove punctuation from tweets
- Remove usernames from tweets
- Remove Stop List words from tweets
- Normalize the case of tweets by converting all tweets to lower case
- Replace the query string with <TOPIC>
- Perform spelling corrections
- Perform Stemming
- Indicate Negation
- Part Of Speech tag all tweets
- Remove all sentences that do not contain the query string.

This application can extract the following features from tweets

- Unigrams
- Bigrams
- Trigrams
- Hash tags
- Adjectives
- Turney's Bigrams

The application uses the following classifiers

- Naïve Bayes Classifier. This classifier uses the Ling Pipe implementation of the Naive Bayes classifier to classify the data.
- SentiWordNet Classifier. This classifier is constructed from scratch and uses the SentiWordNet lexicon resource to calculate a value for the semantic

orientation of unigram features extracted from each individual tweet. The sentiment classification of each tweet is then calculated by summing the values of the semantic orientation of each word contained in the tweet. In the case of a tie break a tweet is classified based on the number of positively oriented or negatively oriented words it contains.

The supervised learning experiments that the application executes are described in table 10 in the Section 5.1 of this paper.

The supervised learning experiments that the application executes are described in table 12 in the Section 5.2 of this paper.

5 Results

5.1 Supervised Learning Approach

Table 10 shows the results for the supervised learning experiments.

	Accuracy	Kappa	Positive	Negative
Baseline using approach outlined in Go et al. 2009	0.63	0.26	430	326
Experiment 1: *Features*: • Adjectives	0.61	0.203	282	440
Experiment 2: *Features*: • Bigrams matching patterns using by Turney	0.559	0.118	105	495
Experiment 3: *Features*: • Unigrams	0.664	0.228	527	246
Experiment 4: *Features*: • Bigrams	0.636	0.273	442	322
Experiment 5: *Features*: • Trigrams	0.604	0.208	258	467
Experiment 6: *Features*: • Unigrams and Bigrams	0.741	0.483	483	406
Experiment 7: *Features*: • Unigrams, Bigrams and Trigrams	0.735	0.47	477	405
Experiment 8: *Preprocessing*:	0.753	0.506	507	397

• Normalize Case • Remove Punctuation • Spelling Correction *Features*: • Unigrams and Bigrams				
Experiment 9: *Preprocessing*: • Normalize Case • Remove Punctuation • Spelling Correction • Remove Stop List Words *Features*: • Unigrams and Bigrams	0.761	0.523	503	411
Experiment 10: *Preprocessing*: • Normalize Case • Remove Punctuation • Spelling Correction • Remove Stop List Words • Remove all sentences that do not contain the query string Features: • Unigrams and Bigrams	0.745	0.49	491	403
Experiment 11: *Preprocessing*: • Normalize Case • Remove Punctuation • Spelling Correction • Remove Stop List Words • Negation *Features*: • Unigrams and Bigrams	0.761	0.523	503	411
Experiment 12: *Preprocessing*: • Normalize Case • Remove Punctuation • Spelling Correction • Remove Stop List Words • Stemming *Features*: • Unigrams and Bigrams	0.768	0.536	508	414
Experiment 13: *Preprocessing*: • Normalize Case • Remove Punctuation • Spelling Correction • Remove Stop List Words • Stemming	0.772	0.524	509	418

• Negation *Features*: • Unigrams and Bigrams • Hash tags				
Experiment 14: *Preprocessing*: • Normalize Case • Remove Punctuation • Spelling Correction • Remove Stop List Words • Stemming *Features*: • Unigrams and Bigrams • Hash tags	0.777	0.554	508	425
Experiment 15: *Preprocessing*: • Normalize Case • Remove Punctuation • Spelling Correction • Remove Stop List Words • Stemming • Remove User Names *Features*: • Unigrams and Bigrams • Hash tags	0.773	0.546	510	418
Experiment 16: *Preprocessing*: • Normalize Case • Remove Punctuation • Spelling Correction • Remove Stop List Words • Stemming • Remove User Names • Remove Query String *Features*: • Unigrams and Bigrams • Hash tags	0.77	0.541	497	428

Table 10 Supervised Learning Results

The results of this experiment are promising with highest accuracy of 77.7%, which is significantly better than the accuracy of the base line. Like in Pang et al. (Pang, Lee & Vaithyanathan 2002) the presence of unigram features is a significant factor, with the accuracy recorded for unigrams on their own better than that of bigrams and trigrams. The combination of unigrams and bigrams significantly increases accuracy. This suggests that while unigrams provides baseline coverage of the positive and negative

classes, bigrams provide extra contextual information that is important for sentiment classification (Pak & Paroubek 2010).

One interesting feature that increases accuracy is hash tags. The aim of using a hash tag in a tweet is to group threads of tweets and also add context to tweets. It would appear from these results that this simple mechanism is successful in terms of adding context. The inclusion of this feature is based on an intuition.

With the exception of the inclusion of hash tags, the performance of the features used are pretty much in line with the literature. The preprocessing steps applied, however, show potential for further investigation and improvement.

The preprocessing steps that perform well all act to make the data more uniform. These steps are

- Case Normalization
- Removal Of Punctuation
- Spelling Correction
- Stemming
- Removal of Stop words.

This shows that, in terms of sentiment classification, tweets do contain noisy data that needs to be either removed or made more uniform in order to accurately classify the data.

While these preprocessing steps perform well, other preprocessing steps do not.

Remove sentences not containing the query string. The aim of removing any sentences that do not contain the query string is to negate the effects of tweets containing mixed sentiment toward a number of topics. The implementation for this experiment is crude as it does not consider continued narrative between sentences and thus sentiment information is lost. A more effective method might identify the target of any of a narrative and eliminate narrative that is not relevant. While such a task would be difficult to achieve, Wiebe's paper on tracking physiological point of view (Wiebe 1994) may provide a starting point.

Remove usernames. Removing user names also decreases accuracy. This is surprising for two reasons. Firstly there is not an over abundance of usernames contained in the data (approx. 493 user names in 21052 words). Secondly, the expectation is that, based on previous research, which discards such data (Pak & Paroubek 2010) (Go, Bhayani & Huang 2009), usernames are noisy data and their

inclusion should decrease the accuracy of the data being classified. The fact that removing usernames decreases accuracy leads to the conclusion that certain users are associated with certain sentiments. This may be due to the relatively small data set and if a larger data set, constructed using a larger pool of queries, were used, the effect usernames has on overall accuracy may be diluted.

Remove the query string. The fact that removing the query string from all tweets has a negative effect on accuracy is surprising at first. But considering the size of the data set and the distribution of query results between topics, it is inevitable that query strings will become associated with categories. Again a larger dataset may dilute the effect that query strings have on the overall accuracy.

Negation. Negation is interesting as, when used with hash tag features, the accuracy is slightly decreased. The most likely cause of this is a hash tag containing a word that is negated. Again a larger set of data may reduce such anomalies.

Kappa is the de-facto standard metric for inter rater agreement. In this case the Kappa value measures the agreement of the classifiers over the iterations of the cross validation process. The highest Kappa value is 0.554. By the strictest interpretation of Kappa no conclusions can be drawn from the results of these experiments. However the evaluation of Kappa values is open to interpretation (Di Eugenio 2000) and using Reitveld and Van Hout's[29] scale the value of 0.554 for Kappa indicates moderate agreement.

Finally, to examine the effects of dataset size the most accurate experiment is re run with incrementing amounts of data. The results are shown in table 11. With the exception of the second iteration, the level of accuracy increases from one iteration to the next. This indicates that further investigation with a larger data set could yield more positive results.

# positive tweets	# negative tweets	Accuracy
100	100	0.675
200	200	0.765
300	300	0.74
400	400	0.748
500	500	0.762
600	600	0.777

[29] See Appendix B

Table 11 Incremental results

5.2 Unsupervised Learning Approach

Table 12 shows the results for the unsupervised learning experiments.

Preprocessing and Feature	% Accuracy	Positive	Negative
Baseline using approach outlined in Ohana 2009	57.7	348	306
Experiment 1: *Features*: • Adjectives	58.25	487	212
Experiment 2: *Features*: • Bigrams matching patterns using by Turney	57.0	457	227
Experiment 3: *Features*: • Unigrams	61.41	426	311
Experiment 4: *Preprocessing*: • Spelling Correction *Features*: • Unigrams	61.66	400	340
Experiment 5: *Preprocessing*: • Spelling Correction • Remove User Names *Features*: • Unigrams	61.83	398	344
Experiment 6: *Preprocessing*: • Spelling Correction • Remove User Names • Remove Query String *Features*: • Unigrams	63.5	420	342
Experiment 7: *Preprocessing*: • Spelling Correction • Remove User Names • Remove Query String • Remove Hash Tags *Features*: • Unigrams	63.41	419	342
Experiment 8: *Preprocessing*: • Spelling Correction • Remove User Names	64	420	348

• Remove Query String • Remove Hash Tags • Negation *Features*: • Unigrams			
Experiment 8: *Preprocessing*: • Spelling Correction • Remove User Names • Remove Query String • Remove Hash Tags • Negation • Remove all sentences that do not contain the query string *Features*: • Unigrams	62.5	459	292

Table 12 Results for unsupervised learning approach

The results for the unsupervised learning approach are significantly lower than the results for the supervised learning approach with best accuracy of 64 %. While this out performs the base line, which emulates the approach taken by Ohana in (Ohana 2009) it is significantly less accurate than the results achieved in the literature.

Like the supervised approach one factor for this low accuracy could be tweets addressing multiple topics and expressing conflicting sentiment.

Another factor in the low level of accuracy could be attributed to words, such as slang, colloquialisms and miss spellings, in the data not being present in the SentiWordNet database. Spelling correction slightly improves accuracy and a translation routine to normalize slang and colloquialisms could also be effective. Such a translator could employ an online dictionary such as netlingo.com.

Removing hash tags slightly decreases accuracy. This is surprising as the majority of hash tags are not proper English words, for example "*#BreakingNews*" or "*#isfeidirlinn*". This indicates however that there are enough hash tags that contain full English words to affect the accuracy of the classifier. This however is not likely to cause a problem in a larger data set, as the overall effect of removing hash tags is the misclassification of one tweet.

Removing the query string has a positive effect on accuracy. This is due to words like "Caribbean" or "Super" having a positive semantic orientation in the SentiWordNet database.

Unlike the supervised learning experiments negation is effective in increasing the overall accuracy. This finding is in line with the literature.

Removing sentences that do not contain the query string, while greatly increasing the accuracy of predicting the sentiment of positive tweets, also reduces the accuracy in predicting the sentiment of negative tweets. This suggests a tenancy for negative tweets to be expressed using a continuing narrative, for example "*How many pirates of the carribean are they going make. Enough is enough. Jeez. Overkill.com*". Again like the supervised learning approach, a method of identifying the target of sentiment may prove fruitful.

As stemming does not guarantee that the returned string is a valid word it is not appropriate to use for this approach to sentiment classification. Lemmatization is a similar process to stemming. When a word is lemmatized it is reduced to its base form. Unlike stemming the word returned is a valid word. This process may lead to increased accuracy.

6 Conclusions

The terms sentiment analysis and opinion mining refer to the computational analysis of text and the evaluation of any underlying positive or negative sentiment expressed by the writer. The area of sentiment analysis is the subject of much research over the past number of years. The core of this research has focused mainly in the domain of more traditional online user generated content such as movie reviews, product reviews etc. In recent years there has been a huge increase in the usage of social media sites with the micro blogging site twitter, in particular, experiencing phenomenal growth. Social media and in particular micro blogging represent a wealth of user-generated opinion that, if extracted and evaluated accurately, could be a valuable source of information for governments, politicians, marketers and consumers alike. The aims of this paper is to present a literature survey of different research carried out in the area of sentiment analysis and evaluate whether this research can be applied to the domain of micro blogging.

The literature suggests that machine-learning algorithms achieve the best success with regard to the sentiment classification of text with both supervised and unsupervised approaches proposed. In view of this, the experimental potion of this paper takes the form of one experiment using a supervised learning approach to sentiment

classification and one experiment using an unsupervised learning approach to sentiment classification. Both experiments use a dataset comprised of manually classified micro blog posts collected from twitter.

Analysis of the data set suggests that data collected from twitter could be used as a valid corpus as the distribution of word frequencies conforms to Zipf's law. This supports previous findings by Pak and Paroubek (Pak & Paroubek 2010). Analysis also indicates that the use of slang and colloquialism is more frequent in tweets that express positive sentiment. This may indicate that tweets that express positive sentiment are of a more causal nature and less considered than tweets that express negative sentiment.

The unsupervised learning experiment centers around the use of the sentiment lexicon SentiWordNet. In this experiment a numerical value is calculated for the words in each tweet. This value that represents the semantic orientation, positive or negative, of the words. The overall sentiment of each tweet is the sum of these values. The best accuracy achieved using this method is 64%. While this outperforms the baseline accuracy, which is calculated using the approach proposed by Ohana (Ohana 2009), the accuracy lower than the accuracies achieved in the literature. One of the main factors for this can be attributed to words in tweets, such as slang and colloquialisms, not being present in the SentiWordNet database. Also, although tweets are restricted to 140 characters, a single tweet can address multiple topics and different sentiments can be expressed toward these topics.

The results of the supervised learning experiment are far more promising with best accuracy of 77.7%. While this accuracy is lower than accuracies from previous studies, 82.2% for Go et al. (Go, Bhayani & Huang 2009) and 86.4 % for Pang and Lee (Pang & Lee 2004), the baseline, which is based the approach proposed by Go et al., is significantly exceeded. The features tested that achieve best accuracy all contain unigram components, with unigrams used in combination with bigrams and hash tags achieving best accuracy. This is due to unigrams providing coverage for both class and bigrams and hash tags providing context. The preprocessing tasks that perform well, such as stemming or spelling correction, are tasks that make the data more uniform in some way. Some of the tasks that do not perform well, such as removal of the query string or removal of usernames, suggest that the data set used may be too small.

The kappa values for the supervised learning experiments are low and suggest only moderate agreement.

Based on these findings there is much room for further study, especially with respect to the unsupervised learning approach. One area of improvement could be the use of online slang dictionaries such as netlingo.com to "translate" slang words. The aim of this is to increase the amount of valid words and therefore improve the results returned from SentiWordNet. Lemmatization, whereby each word is reduced to its base form, may also help in this respect.

In terms of the data, an increase in the amount of data used may lead to better results with regards to supervised learning approaches. Also the approach taken in this research project to the manual classification of tweets could be improved by using multiple annotators. The manual classification process could then be studied in the form of an inter annotator agreement study. The results of such a study may show patterns of agreement between annotators that could be used to train a classifier. Also data classified in such a way would be more reliable.

Finally both approaches could benefit from a more holistic process to identify relevant information within a tweet. By this I mean identifying the subject, about which sentiment information is being extracted, and discarding elements of the tweet that do not relate to this subject or express sentiment.

7 Bibliography

Bruce, RF & Wiebe, JM 1999, 'Recognising Subjectivity: A case study of Manual Tagging', *Natural Language Engineering*, vol 1, no. 6.

Carpenter, B 2010, *Text Analysis with Ling Pipe*, Ling Pipe Publishing.

Church, KW & Patrick, H 1990, 'Word association norms, mutual information, and lexicography', *Computational Linguistics*, vol 16, no. 1, pp. 22-29.

Di Eugenio, B 2000, 'On the usage of Kappa to evaluate agreement on coding tasks', *Proceedings of the Second International Conference on Language Resources and Evaluation*.

Esuli, A & Fabrizio, S 2006, 'SENTIWORDNET: A Publicly Available Lexical Resource for Opinion Mining', *Proceedings from International Conference on Language Resources and Evaluation (LREC)*.

Go, A, Bhayani, R & Huang, L 2009, *Twitter Sentiment Classification using Distant Supervision*, viewed 17 July 2011, <http://www.stanford.edu/~alecmgo/papers/TwitterDistantSupervision09.pdf>.

Gross, G 2008, *Obama Transforms Web-based Politics*, viewed 29 March 2011, <http://www.pcworld.com/businesscenter/article/155917/obama_transforms_webbased_politics.html>.

Hatzivassiloglou, V & McKeown, KR 1997, 'Predicting the semantic orientation of adjectives.', *35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics*, Madrid.

Hatzivassiloglou, V & Wiebe, J 2000, 'Effects of adjective orientation and gradability on sentence subjectivity', *Proceedings of the International Conference on Computational Linguistics (COLING)*.

Liu, B 2010, 'Sentiment Analysis and Subjectivity', in N Indurkhya, FJ Damerau (eds.), *Handbook of Natural Language Processing*, 2nd edn, Chapman and Hall/CRC.

Miller, GA 1995, 'WordNet: A Lexical Database for English', *Communications of the ACM*, vol 38, no. 11, pp. 39-41.

Ohana, B 2009, *Opinion mining with the SentWordNet lexical resource*, viewed 20 July 2011, <http://arrow.dit.ie/cgi/viewcontent.cgi?article=1019&context=scschcomdis&sei-redir=1#search=%22Opinion%20mining%20SentWordNet%20lexical%20resource%22>.

Pak, A & Paroubek, P 2010, 'Twitter as a Corpus for Sentiment Analysis and Opinion Mining', *Proceedings of LREC 2010*.

Pang, B & Lee, L 2004, 'A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts', *Proceedings of the Association for Computational Linguistics (ACL)*.

Pang, B & Lee, L 2005, 'Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales', *Proceedings of the Association for Computational Linguistics (ACL)*.

Pang, B & Lee, L 2008, *Opinion mining and sentiment analysis*, now Publishing Inc, Hanover, MA.

Pang, B, Lee, L & Vaithyanathan, S 2002, 'Thumbs up? Sentiment classification using machine learning techniques', *Conference on Empirical Methods in Natural Language Processing*.

Parikh, R & Movassate, M 2009, *Sentiment Analysis of User-Generated Twitter Updates using Various Classification Techniques*, viewed 17 July 2011, <http://nlp.stanford.edu/courses/cs224n/2009/fp/19.pdf>.

Read, J 2005, 'Using emoticons to reduce dependency in machine learning techniques for sentiment classi¯cation', *Proceedings of ACL-05, 43nd Meeting of the Association for Computational Linguistics. Association for Computational Linguistics*.

Roiger, RJ & Geatz, MW 2003, *Data Mining A tutorial-based primer*, Addison Wesley.

Takamura, H, Inui, T & Okumura, M 2006, 'Latent variable models for semantic orientations of phrases', *European Chapter of the Association for Computational Linguistics (EACL)*.

Toutanova, K, Klein, D, Manning, C & Singer, Y 2003, 'Feature-Rich Part-of-Speech Tagging with a Cyclic Dependency Network', *Proceedings of HLT-NAACL*.

Turney, P 2002, 'Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews.', *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia.

Wiebe, JM 1994, 'Tracking point of view in narrative', *Computational Linguistics*, vol 20, no. 2, pp. 233-287.

Wiebe, JM, Wilson, T & Cardie, C 2005, 'Annotating Expressions of Opinions and Emotions in Language', *Language Resources and Evaluation*, vol 39, no. 2/3, pp. 164-210.

Appendix A: Ling Pipe English Stop List Words

a, be, had, it, only, she, was, about, because, has, its, of, some, we, after, been, have, last, on, such, were, all, but, he, more, one, than, when, also, by, her, most, or, that, which, an, can, his, mr, other, the, who, any, co, if, mrs, out, their, will, and, corp, in, ms, over, there, with, are, could, inc, mz, s, they, would, as, for, into, no, so, this, up, at, from, is, not, says, to

Appendix B: Interpretations of Kappa Values

Kappa is constrained to the interval [0,1] where 0 means agreement between annotators, or raters, is no more than chance and 1 is perfect agreement.

Krippendorff's scale is the strictest interpretation of Kappa and interprets Kappa in the following way

$K < 0.67$ is discounted

$0.67 < K < 0.80$ allows tentative conclusions

$K \geq 0.80$ allows definite conclusions

Reitveld and Van Hout propose another interpretation of K where

$0.41 < K < 0.60$ indicates moderate agreement

$0.61 < K < 0.80$ indicates substantial agreement

(Di Eugenio 2000)