Map the Debate

Understanding the web's response

Author: Joseph Root <jsr08@ic.ac.uk>
Supervisor: Francesca Toni <f.toni@ic.ac.uk>

June 2011

Imperial College London

CONTENTS

PART I | ANALYSIS

1

INTRODUCTION

1.1 Motivation

From women's rights to civil rights, the influence of public opinion on government policy has been pivotal. Within a healthy democracy the voice of the electorate should be heard and recognised by those chosen to represent them. Throughout history platforms have often been provided for public opinion to be made known, from the early public forums of Greece and Rome, to speaker's corner and the house of commons today. Providing a means for people to express their opinion enables them to both challenge and shape the direction their elected governments take. Finding ways of gathering and understanding this opinion has increasingly proven fundamental if a government wishes to be successful.

Current methods of measuring opinion are largely statistical, with methods such as polling looking at the opinion of a sample group, before using their results to make further predictions. These can be very accurate, however their small sampling rates mean that figures can often be askew. Furthermore polling is both costly and time consuming to conduct, and thus can neither be used to find opinion on breaking news or on a variety of topics. None the less, as methods for measuring public opinion have increased both in accuracy and detail, politicians and policy makers are starting to look to them not only for affirmation of their policies, but for guidance and new initiative.

As the web has become more prevalent throughout society, it is increasingly becoming a platform for discussion and opinion. The initial growth of blogging demonstrated the web's ability to serve as a forum for debate and opinion. However the technical knowledge required to start a blog, alongside the time required to write a post meant that adoption was limited. In the past two years we have seen the rise of micro-blogging (essentially 140 character blog posts) through services such as Twitter. These have seen unprecedented levels of adoption, with Twitter's 200 million users posting 25 billion micro-blog posts in 2010. Due to the simple nature of writing short posts, micro-blog discussion tends to break quickly

around news topics, and offers genuine insight into public opinion surrounding news topics.

This project hopes to utilise the growth of publicly available opinion on the web, using it as a source upon which new methods for analysing and measuring public opinion can be built. In particular the project will focus on understanding sentiment on micro-blogging services such as Twitter.

1.2 Contributions

1. Ruby implementation of a sentiment analysis engine based upon current research. The engine should be able to correctly identify sentences containing sentiment, and classify the sentiment as positive or negative. Different implementations should be tested and compared to determine which algorithms and techniques work best.
2. An algorithm for classifying sentiment as a range of emotion and feeling, rather than just a score along a scale of positive to negative. The algorithm should be implemented in Ruby and included in the sentiment analysis engine.
3. Research optimisations for current algorithms, in order to tailor the sentiment analysis engine for micro-blog posts from services such as Twitter. These optimisation should be implemented within the engine.
4. Research optimisations for current algorithms, in order to better facilitate the understanding of Politically focussed micro-blog data. These optimisation should be implemented within the engine.
5. A Ruby based Twitter module to store and classify live data from Twitter.
6. Visualisations should be designed and implemented to help better understand the data and classification results.

2

BACKGROUND

From it's early forums through to the 'social web' of today, the Internet has served as a continually expanding platform for discussion. The result has been an explosion in the amount of readily available, computer-formatted textual opinion. With this growth has come an increasing desire to computationally understand the wealth of opinion now so easily accessible. Combining elements of linguistics, natural language processing and machine learning, this field of exploration has come to be known as *opinion mining* or *sentiment analysis*. In the following chapter we will first briefly examine Twitter as a backdrop to our discussion on sentiment analysis. We will then go on to explore the general problems posed by sentiment analysis along with the common approaches and solutions taken in addressing them. In sections 2.2.2 - 2.2.4 we will discuss in detail the areas and methods of sentiment analysis which will bear relevance to this project's Twitter-based setting. Finally in section 2.4 we will explore emotion in general, particularly looking at it's scope and ways of classifying it.

2.1 Twitter

Twitter is a social-networking web-service. It enables users to post and read 140 character messages known as *tweets*. A user's *timeline* serves as a publicly viewable history of their tweets. Furthermore if someone chooses to *follow* another user, they will be notified of changes to that user's timeline. This simplicity has seen Twitter's user-base rapidly expand, with over 200 million active users today. From football transfers to revolutions Twitter has become the go-to service for spreading news quickly and efficiently.

Since its launch in 2006, certain protocols have emerged from within the Twitter community. These have been embraced by Twitter, enabling it to serve not only as an efficient platform for spreading news, but also as a rich and sophisticated medium for conversation. Notable protocols include:

Hashtags enable users to tag their tweets with any word or combination of characters they deem appropriate. Although this may seem basic at first, through common hashtags, it enables users to take part in a community-wide discussion. For example, during the recent voting reform referendum, the hashtags '*#yes2av*' and '*#no2av*' were used to form a debate on the strengths and weaknesses of the Alternative Vote.

Mentions allow users to reference other users in their tweets. Furthermore if a user is mentioned in a tweet, Twitter will notify the mentioned user. Through this, Twitter users can take part in a direct conversations with one or more other users. For example, if we wanted to ask Stephen Fry a question, we could tweet '*what are you eating for breakfast @stephenfry?*'.

Re-tweets give users the ability to re-post other users' tweets in their own timeline. This simple feature has had a significant impact on Twitter's ability to facilitate the rapid spread of news. For example in 2009 when the US Airways flight 1549 crash landed in the Hudson river, rapid re-tweeting of an amateur photo meant the news broke on Twitter far earlier than it did within the media at large. This has continued to be true for many more notable events such as the recent North-African revolutions.

Links have always been the popular subject of tweets, however the introduction of link-shorteners has changed the way in which they are posted. In freeing up characters by shortening a URL, users now have the option to describe or comment on the link they are tweeting. This has enabled users to engage in deeper conversation on content they have viewed online, and has neatly allowed Twitter's viral nature to better merge with it's community's desire for debate.

Through Twitter's RESTful API [1], this rich resource of live news and debate will serve as the project's main data source.

2.2 Sentiment analysis

Sentiment analysis as a field, is the exploration of how we can computationally understand opinions expressed within a body of text. In order to do this, we must first define a computational structure for expressing opinions. In general [6] this is done by breaking an opinion down into four parts. Firstly we must determine

[1] RESTful API's allow developers to retrieve, modify, create and delete data by making get, post and delete HTTP requests to specified web addresses.

the opinion's focus of discussion, also known as it's *topic* [2]. This in practise can encompass anything from Government policy to mobile phone battery life. Often opinion is not necessarily that of the author, but of a referenced person or group, therefore it is important to determine the opinion's *holder*. Along with this it is also often necessary to determine the *time* at which the opinion was expressed. Finally, we hope to *classify* (or in some cases quantify) the opinion which has been passed. Leading research [10, 12] has typically focussed on discrete classification, such as deciding whether an opinion is positive, negative or neutral. A fifth *object* component is sometimes introduced for larger documents, which serves as an identifier for related topics. For example, within a phone review the majority of opinions may share the same object, in this case the phone, but focus on different topics such as battery life or call quality.

How do we computationally discover opinions and identify their parts? In general the approach can be loosely split into two components, *sentence-level classification* and *document-level classification*. Sentence-level classification determines whether a sentence expresses an opinion along with classifying that opinion if it exists. Furthermore if an opinion is found, sentence-level classification will try to determine it's topic, holder and the time at which the opinion was cast. Document-level classification goes on to collate the sentence-level results, in order to form a general description of the document's sentiment. Both these approaches draw heavily upon machine learning techniques. It is important to note here however, a core criticism of the field. Linguists such as Chomsky [8] observe that rather than truly trying to understand and define the semantics of sentiment, the field takes a heavily statistical approach. This means that rather than determining sentiment by forming a semantic conclusion, the field uses a limited linguistic foundation to predict sentiment based upon experience. Nonetheless, redefining natural language processing and sentiment analysis is not within the scope of this project, and we shall proceed with the field's successfully tried and tested approaches.

As we shall discuss in more detail in chapter 4, only sentence-level classification is relevant to this project. Furthermore, methods for determining an opinion's holder and time are unnecessary and will not be discussed here. The remainder of this section will instead focus on the three relevant topics from within sentence-level classification. Firstly we shall explore what exactly an opinionated sentence is and how we can computationally determine this. Next we will look at common approaches to classifying sentiment, before finally examining how we determine the topic of an opinion. Before this however, we shall briefly outline the concepts and methods of *supervised learning* as this shall form the core for each of our classification problems.

[2] This is more commonly referred to in literature as an opinion's *feature*, however to avoid later confusion with the machine learning term, we will use the term *topic*.

2.2.1 Supervised learning

Supervised learning is a task within machine learning which infers a function from a set of training data. This approach is well suited to classification problems, and in our case is particularly relevant to classifying opinion and determining polarity. Thus, the remainder of this section will discuss supervised learning with respect to the classification of sentences.

2.2.1.1 Defining the problem

In both problems we want to find an approximate hypothesis function h for our actual function c. Both functions will map an input sentence $s \in S$, to a discrete classification $o \in O$, where S is the set of all possible sentences and O is the set of all possible classifications, for example $O = \{positive, neutral, negative\}$, such that:

$$h \approx c : S \rightarrow O \tag{2.1}$$

In order to find our best fit hypothesis function h, we will first need to determine a set of *features* for our sentences. Within machine learning, features are the attributes which best describe and discriminate our input data when trying to classify it. For example if we are trying to learn a function to decide whether we should play tennis or not, features might include humidity and sunlight. In essence we want to identify a list of the most useful features $f_1, f_2, \ldots, f_n$ for our sentences, such that:

$$h \approx c' : \langle f_1, f_2, \ldots, f_n \rangle \rightarrow O \tag{2.2}$$

Once a set of features has been chosen we can approximate h by training it. In order to find the perfect hypothesis function for classifying subjective functions, $h = c$, we would require knowledge of every single possible sentence along with it's correct classification. Clearly we could never produce the set of all possible sentences, let alone determine every sentence's classification. Instead, we select a sample of training sentences $T \subseteq S$, and manually *label* each sentence $t \in T$ with a classification $l \in O$. This is our *training data* D, such that:

$$D = \{(t, l) : \forall t \in T \textit{ there exists a manually labelled classification } l\} \tag{2.3}$$

Given this training data we can now determine as accurate a hypothesis function as possible for classifying *all* sentences. There are numerous, largely statistical

methods for training our hypothesis function. Each brings their own positives and negatives, and there has been extensive research [9] into which methods perform best for opinion based classification. We will discuss the most appropriate methods, features and training data for each classification problem in their respective parts.

2.2.1.2 Common approaches

2.2.2 Discovering opinion

In general opinion manifests itself either *explicitly* through *subjective* sentences and phrases, or *implicitly* through *objective* sentences and phrases. An objective sentence expresses factual information, whilst a subjective sentence expresses a mental or emotional state, such as a sentiment or belief. A subjective sentence such as, "*I love the NHS, it's bloody marvellous*", explicitly states an opinion. Similarly however, a sentence such as "*Lost my job due to recent Coalition cuts*" although objective, could also be considered an implicit opinion. This clearly poses a difficult challenge for classification, and as Mihalcea et al. [7] note, it is one which "has often proved to be more difficult than subsequent polarity classification". As observed by Liu [6] however, opinionated sentences tend to be a subset of subjective sentences. Due to this, the approaches for classifying them are similar and the terms are taken as interchangeable. This is referred to as *subjectivity classification*.

Subjectivity classification is typically achieved through a mix of supervised and unsupervised learning. In general, unsupervised learning is used to bootstrap a relatively small but accurate training set. The bootstrapped training set is then utilised to train a classifier. Numerous feature choices have been proposed for training subjectivity classifiers. We shall first examine some of the the more commonly used features, as discussed by Wiebe et al. [19]:

Adjectives tend to be strong indicators of subjectivity, often serving as descriptions or qualifications of opinion. For example the adjectives in, "*the coalition cuts are **harsh** but **necessary***", are clear indications of subjectivity. As Wiebe et al. [19] observe a simple binary feature alone, noting the appearance of one or more adjectives, results in a classification accuracy of 56%.

Adverbs modify verbs, adjectives and phrases, for example "*they **usually** get things right*". Their presence is often an indicator of subjectivity, and although not as useful as adjective presence, their inclusion as a binary feature further improves classification rates. Wiebe et al. [19] suggest a binary feature noting the presence of any adverb other than *not*.

Pronouns are substitutions for nouns, for example *it* in place of an object. They

are often minor indicators of subjectivity, and have been shown to marginally improve classification accuracy when included as a binary feature.

Adjective orientation and gradabilty tend to be further indicators of subjectivity. Essentially orientation notes whether an adjective encodes a desirable (e.g *beautiful*) or undesirable (e.g. *ugly*) state. The gradability of an adjective denotes the relative extent to which an adjective varies in strength from the norm. For example *"small"* and *"large"* have high gradability. As shown by Wiebe et al. [13], the presence of polarised, gradable adjectives is a strong measure of subjectivity and a useful feature.

Wiebe et al. [19] observed that using the first 3 techniques, coupled with cardinal numbers and a single document-level feature, resulted in classification rates of 71.2%.

But how can we identify these features within a sentence? Adjectives, adverbs and pronouns are all known as *parts of speech (POS)*. A word's POS can take on one of eight roles within a sentence: *verb*, *noun*, *pronoun*, *adjective*, *adverb*, *preposition*, *conjunction* and *interjection*. A word's part of speech is often determined by it's position within the sentence. For example "*love*" can be a noun or a verb, dependant upon the context in which it is used. Below is an example of a sentence whose words have been *tagged* with their POS:

$$\underbrace{\text{She}}_{\text{pron.}}\ \overbrace{\text{likes}}^{\text{verb}}\ \underbrace{\text{big}}_{\text{adj.}}\ \overbrace{\text{snakes}}^{\text{noun}}\ \underbrace{\text{but}}_{\text{conj.}}\ \overbrace{\text{I}}^{\text{pron.}}\ \underbrace{\text{hate}}_{\text{verb}}\ \overbrace{\text{them.}}^{\text{pron.}} \tag{2.4}$$

2.2.2.1 *Part of speech tagging*

Given a phrase or sentence, *part of speech tagging* computationally determines each word's POS. This can be done in variety of ways. Typically basic implementations use a lexicon of words with their appropriate tags, or a more advanced dictionary such as WordNet[3]. In general these implementations are naive and often simply return a list of possibilities. More intuitive techniques tend to use machine learning to recognise patterns, or are built with a set of linguistic rules. We will discuss the merits of these techniques and their implementation in more detail in chapter 4. With a fully tagged sentence it is now possible to build a feature set based upon the relevant parts of speech.

[3]Wordnet is a detailed dictionary with additional levels of detail describing the semantic interlinking between words. It will be used throughout this project and shall be discussed in more detail in chapter 4.

2.2.2.2 *Use of supervised techniques*

As noted in our discussion of features, some adjectives are more useful in classifying subjectivity than others. Determining these adjectives, and in this case their polarity, would prove tedious if carried out by hand. Instead, Wiebe [14] suggests a supervised approach using a set of seed words and a large corpora of text. The corpora is examined for conjunctions, such as "*handsome* and *smart*", and disambiguations such as "*smart but cruel*". When a seed word is found within either scenario, it's fellow word's polarity can be inferred. For conjunctions, if one of the words is known as positive, then the unknown word is likely to be positive also. The converse holds for disambiguations, where the unknown word is inferred to be the opposite of the known word. This technique enables the rapid building of a polarised adjective lexicon. It is particularly useful in domains which assign their own meaning to adjectives, for example *sick* is often a positive adjective within youth culture.

Building a training set significant enough for accurate subjectivity classification can often be time consuming. Liu [6] and Akkaya et al. [18] describe a supervised method for bootstrapping an initial training set. A high precision, low recall rule based classifier, as originally proposed by Wiebe and Riloff [15], is used to build a small training set from a large corpora. The classifier does this by identifying strong and weak subjective clues within a sentence. If there are two or more strong subjective clues the sentence is classified as subjective. In order to determine objectivity, the sentences on either side are taken into account. If between them neither contain more than one strong and two weak clues, along with no strong clues in the analysed sentence, the sentence is considered objective. If the conditions for subjectivity and objectivity are not met, the classifier leaves the sentence unclassified. The use of supervised methods such as this and the lexicon builder described above are typical within the field. They provide simple and efficient ways of optimising the overall training process.

2.2.2.3 *Present research and issues*

Recent literature has also explored numerous improvements to the classic algorithm as described above. One such improvement of notable effect is *subjectivity word sense disambiguation (SWSD)*, originally presented by Wiebe et al. [17], and further refined by Akkaya et al. [18]. SWSD tries to reduce the misclassification of objective words, and thus possibly the sentence, as subjective. These false hits often occur as a result of assuming that if a word exists within a subjective lexicon, it is being used in subjective sense. For example, *pain* is often used subjectively, however within, *early symptoms include body pain*, *pain* is used in an objective sense.

SWSD attempts to eliminate this source of error. A subjective lexicon of words is built, and for each of it's words, a classifier is trained. Given a potentially subjective word within a sentence, the classifier will label the word's sense as objective or subjective. The classifier is trained using a corpora of sentences whose subjective words have been labelled as either subjective or objective. The classifier is then used to ensure that all subjective words are used in their subjective sense. Using SWSD within subjectivity classification, Wiebe et al. [18] noted a 24% reduction in error against a classifier using the regular subjectivity lexicon when looking for subjective words.

Subjectivity classification is a well researched field, however current methods do pose problems. As is typically the case within supervised learning, the classifier's ability is significantly influenced by how representative it's training set is of the input domain. Subjectivity classification, along with many other natural language approaches, is often extremely sensitive to the type of content with which it has been trained. This means that if one wants to build a subjectivity classifier for political speeches, the training corpora should be built from similar content, not for example from movie reviews. No fixed approach has been developed for this, and it is an issue we shall have to contend with during our implementation in chapter 4.

2.2.3 Classifying opinion

An opinionated sentence can express a diverse range of sentiment, and classifying this can prove difficult. Sentiment can be classified in numerous ways, for example "*I liked the tone of his speech, however I am uncertain of the proposals within it*", could be interpreted in any number of ways. At a phrase level, we might consider the first part to express some form of delight, while the latter expresses distrust. Of course, to a certain extent these are subjective, and more detailed emotional labels shall be discussed in section 2.4. A more broad classification might classify the first part as positive and the second part as negative. Developing methods for labelling a sentence's polarity has served as a focus for much of the research into classifying opinion. This field is referred to as *sentiment classification*.

But how do we determine sentiment? At first this may seem simple. For example "*I **love** the EU*" would typically be classified as positive, whilst "*I **hate** the EU's decentralisation of power*" would be negative. Clearly *love* and *hate* are strong indicators of polarity. Basic methods for classifying sentiment simply check whether any of the words within the sentence exist within a pre-defined polarity lexicon, and classify accordingly. If we explore increasingly complex phrases however, the problem becomes far less simple than simply identifying polarising words. Understanding the scope of negation can present challenges. For example the negative

in "*not nice*", simply negates its neighbour, whilst in "*no one thinks that its good*", the ensuing negation spans the phrase. In certain scenarios negation words can even strengthen polarity, such as "*not only good but amazing*". Issues of word sense, similar to those discussed in section 2.2.2, present further problems. For example "the National *Trust* may waste money" conveys an opinion which expresses the polar opposite of trust. The domain of the sentiment being expressed can also effect polarity. "*Go read the book*" may be considered positive within a book review, however for a film it is generally see as negative.

At it's heart sentiment classification poses a significant linguistic challenge, and the approaches vary as a result of it. They can be broadly split into two approaches however, supervised and unsupervised. Unsupervised methods propose that sentiment can be understood by analysing it's linguistic form. By understanding the rules which allow sentiment to be expressed, we should be able to both identify and understand it within a sentence. Supervised methods suggest that the complexities of language make unsupervised methods too specific and difficult to identify. Instead it hopes to make use of machine learning's supervised techniques in order to better classify sentiment. We will explore and contrasts these two methods. In particular, we will focus upon the unsupervised approach put forward by Turney [12], and the supervised approach proposed by Pang et al. [10].

2.2.3.1 Unsupervised sentiment classification

Turney [12] suggests a two part approach to supervised sentiment classification. As discussed when exploring subjectivity in section 2.2.2, adjectives tend to be a significant grammatical structure through which sentiment is expressed. Thus, Turney proposes extracting phrases containing adjectives and whose structure indicates an expression of sentiment. Given a sentence, we tag its parts of speech, before extracting any two-word phrases whose structure can be found within the following linguistic patterns:

Table 2.1: Extraction patterns for identifying opinionated two-word phrases

Rule	First word	Second word	Third word (*not extracted*)
1.	JJ	NN, NNS	anything
2.	RB, RBR, RBS	JJ	not NN, not NNS
3.	JJ	JJ	not NN, not NNS
4.	NN, NNS	JJ	not NN, not NNS
5.	RB, RBR, RBS	VB, VBD, VBN, VBG	anything

Once these phrases have been identified, we can then determine their sentiment's polarity. This is done by first selecting two words commonly associated with strong

positive and negative sentiment. Turney suggests *excellent* and *poor* as the benchmark words for positive and negative polarity. This is largely due to their prevalent use within reviews as descriptions for high and low ratings. In order to calculate a phrases sentiment, we attempt to measure the association between it and benchmark's words. Co-occurence between two words is calculated using their *Pointwise Mutual Information (PMI)*, defined as:

$$\text{PMI}(word_1, word_2) = \log_2 \left(\frac{\text{p}(word_1 \ \& \ word_2)}{\text{p}(word_1)\,\text{p}(word_2)} \right) \tag{2.5}$$

Where $\text{p}(word_1, word_2)$ is the probability that $word_1$ and $word_2$ co-occur within a corpora, and $\text{p}(word)$ is the probability that $word$ occurs. Now that we have definition for PMI, we can define the *semantic orientation (SO)* of a $phrase$ as:

$$\text{SO}(phrase) = \text{PMI}(phrase, \text{"}excellent\text{"}) - \text{PMI}(phrase, \text{"}poor\text{"}) \tag{2.6}$$

The resulting semantic orientation is a measure of a phrase's sentiment. An SO larger than 0 denotes positive polarity, while an SO less than zero indicates negative polarity. Thus, a sentence's overall polarity is simply the average of it's phrases' SO. This approach to supervised sentiment classification has proven effective across a variety of review domains. Turney reports an impressive 80% when classifying bank reviews and an even better 84% accuracy for automobile reviews. He does note however, that movie reviews present a challenge for his supervised approach, reporting an accuracy of 65.83% within the movie domain. Nonetheless, across domains Turney reports classification rates of 74.39%, demonstrating the strong potential which lies within unsupervised methodologies.

2.2.3.2 Supervised sentiment classification

Shortly after Turney published his paper on supervised approaches [12], Pang et al. put forward a counter paper. This addressed the potential of supervised learning within the same domain of internet reviews as Turney's original paper. At it's core, Pang et al. address the issue that often sentiment can be expressed in very subtle ways. For example, "*How could anyone sit through this movie?*" does not express negative opinion in any readily apparent way. Essentially the proposition put forward by Pang et al. is that the nuanced structures through which we express opinion are too vast and varied. They cannot simply be whittled down into a simple set of rules, and rather, we should look to experience to guide our classification.

As with any supervised problem, the learning experience is largely guided by our choice of features. Before we examine these, it is important to introduce the

concept of *n-grams* and how they work as features. For example, if we use unigram feature set, there is a feature for every possible word. A feature set this large is unnecessary however, as the only words which will be important in classification are those we encounter in training. Thus we build a feature set from the words we encounter when training. If our training set only contained "*I love the NHS*", we would have the following feature set for classification $\langle f_I, f_{love}, f_{the}, f_{NHS} \rangle$. Alternatively if we used bigrams (2 word phrases), we would have a feature set $\langle f_{(I,love)}, f_{(love,the)}, f_{(the,NHS)} \rangle$. But what values do we assign to these features when given a sentence to classify? Pang et al. experiment with two options:

1. *Term presence* denotes whether the n-gram phrase that a feature represents occurs within our sentence. For example, using the unigram and bigram feature sets above, and given a sentence "*I hate the NHS*", we would have the following feature sets:

$$\langle f_I, f_{love}, f_{the}, f_{NHS} \rangle = \langle true, false, true, true \rangle$$
$$\langle f_{(I,love)}, f_{(love,the)}, f_{(the,NHS)} \rangle = \langle false, false, true \rangle$$

2. *Term frequency* denotes how frequently each feature's n-gram phrase occurs within our sentence. For example, using the unigram and bigram feature sets above, and given a sentence "*I hate the NHS, but I love my GP*", we would have the following feature sets:

$$\langle f_I, f_{love}, f_{the}, f_{NHS} \rangle = \langle 2, 1, 1, 1 \rangle$$
$$\langle f_{(I,love)}, f_{(love,the)}, f_{(the,NHS)} \rangle = \langle 1, 0, 1 \rangle$$

Pang et al. also experiment with appending POS tags to the end of each word, thus distinguishing between their possible uses. In order to handle negation, any words between a negative word such as *not* and the next punctuation mark are tagged with a NOT. For example "I do not like the NHS" would result in a feature set $\langle f_I, f_{do}, f_{not}, f_{NOT-like}, f_{NOT-the}, f_{NOT-NHS} \rangle$.

The different feature sets were tested within the movie review domain. The presence feature set for unigrams performs strongest in their experiments with an accuracy of 82.9%. The combination of unigrams and bigrams sees a marginal drop in accuracy to 82.7%. Interestingly POS tags also have a slight negative effect on accuracy, seeing it drop to 81.9% when coupled with a unigram presence feature set. In domains where the expression of sentiment is subtle, supervised approached have a clear benefit over their unsupervised counterparts. However, supervised learning requires one to build a training set, which can often prove time consuming. Furthermore it's understanding of sentiment is based upon experience, thus it could never really explain why it reached it's decision. Deciding which approach is better is difficult, and we shall explore this is more detail in section 5.

2.2.3.3 Present research and issues

Recent research has focussed on how combinations of supervised and unsupervised learning can be used to improve classification rates. Essentially these improvements have hoped to introduce greater linguistic detail into the supervised approach described by Pang et al.. In the following section we shall provide a general overview of two improved methodologies put forward by Wilson et al. [16] and Benamara et al. [2]. We shall explore these approaches in greater detail in section 5.

Although Wilson et al. [16] acknowledge the need for elements of supervised learning, they observe that the sentence-level approach put forward by Pang et al. is to general. Instead they propose that to truly understand sentiment, we must approach it at a phrase level. The main motivation behind this is the common misclassification of *clue* words as polar, when the sense in which they are being used means they are in fact neutral. This problem is of particular relevance to the supervised approach discussed above. The method put forward by Pang et al. essentially creates a lexicon of polar words during training and later uses them as clue's for classifying polarity. As mentioned in our introduction to opinion classification, this can lead to words being taken out of context to classify neutral statements as polar. Wilson et al. propose a two step solution to this. The first step identifies all clue phrases within a sentence, before using a supervised approach to classify each one as polar or neutral. The polarity of each polar phrase is then disambiguated to give it an overall classification of either *positive*, *negative* or *both*. Not only does this approach provide a more rigorous framework for sentiment classification, unlike the methods put forward above it also acknowledges the potential neutrality of phrases within a sentence.

Alongside the influential research into phrase-level sentiment by Wilson et al., other prominent research has focussed on measuring sentiment strength. Benamara et al. [2] highlight the important role of adverbs as measures of opinion. These adverbs are known as *adverbs of degree*. Within this subset of adverbs, five clear classifications can be noted:

1. *Affirmation* adverbs such as *certainly* and *absolutely* strengthen adjectives.
2. *Doubt* adverbs such as *possibly* and *seemingly* weaken adjectives.
3. *Strong intensifying* adverbs such as *exceedingly* and *extremely* strengthen adjectives.
4. *Weak intensifying* adverbs such as *barely* and *scarcely* weaken adjectives.
5. *Negation/minimising* adverbs such as *hardly* and *rarely* invert or neutralise adjectives.

Using a lexicon containing adverbs of degree and their appropriate classification, all unary and binary adverb adjective combinations are found. A unary combination has the form $\langle adverb \rangle \langle adjective \rangle$, whilst a binary combination has the form $\langle adverb_i, adverb_j \rangle \langle adjective \rangle$. Each adjective in the matching phrases has it's polarity strength adjusted according to the classification of the adverbs which proceed it. For unary combinations the score is a product of the adjective and adverb strengths. For binary combinations, the strength of $\langle adverb_j \rangle \langle adjective \rangle$ is calculated first as if it were a unary combination, before calculating the strength combination of the resulting score and $adverb_i$. Benamara et al. report results almost on par with human strength classification, highlighting the proposed method as not only viable but effective.

Although significant improvements have been made within the field, sentiment classification is still far from perfect. Many of it's problems have been reduced in size, however they have not been eradicated. One could argue that this is largely due to the statistical nature of supervised learning, and clearly the field still has a lot to learn from linguistics. Most importantly to this project however, is the fact that little research has explored beyond the confines of polarity and into the realm of emotion. We shall explore the field's limited approach to the classification of emotion in more detail later, in section 2.4.

2.2.4 Topic extraction

2.3 Sentiment on Twitter

With the recent and rapid growth of Twitter has come an interest in understanding the sentiment expressed on it. Although at it's heart an issue of sentiment analysis, Twitter's constraints and protocols pose new and different issues for current approaches. Literature is still limited, and solutions to the problems within it are varied. In this section we will focus on some of the more prominent approaches. In particular we will outline the framework proposed by Barbosa and Feng [1], whilst looking at some of the innovative improvements and observations put forward by Go et al. [5] and Bermingham et al. [3].

Barbosa and Feng [1] propose a two stage sentiment analysis framework. Firstly the subjectivity of a tweet is determined, and if subjective, the tweet's sentiment is then classified. This framework bares many similarities to sentence-level sentiment classification, however the approach within each stage is in many ways very different. Particular emphasis is placed upon the need for strong subjectivity detection. There is a lot of *noise* on Twitter through adverts and spam accounts, thus it is important to filter this out if we ever hope to obtain an accurate overview. A

typical noisy tweet might be:

> *Get a FREE $500 Starbucks Gift Card >> Special Online Offer ...: Starbucks is celebrating its first forty years ... http://bit.ly/iepyV5*

Barbosa and Fang propose some previously unconsidered features for helping distinguish noise from subjective tweets. As evident from analysing the tweet above, Barbosa and Fang note that *link presence* and *uppercase letter frequency* serve as particularly useful subjectivity clues. A novel approach is taken to training the subjectivity classifier using existing online Twitter sentiment classifiers. Subjective tweets are scraped from three such sites, and any tweet appearing as subjective in all three is added to the training data. They report that although this can lead to slight bias, it serves as an effective bootstrapping method.

Barbosa and Fang take an entirely supervised approach to polarity classification using many of the features discussed in section 2.2.3. Uppercase letter frequency again proves particularly useful, along with a feature for *good emoticons*. An emoticon is a text-character face expressing an emotion, for example happy is commonly represented as `:)` while sad is `:(`. Barbosa and Fang note significant improvements both in subjectivity and sentiment classification when using tweet-based features, as opposed to the typical approaches described in sections 2.2.2 and 2.2.3. Using unigrams alone for sentiment classification, Barbosa and Fang report an error rate of 44.5%, whilst the introduction of Twitter based features reduces this to 25.1%. Although far from perfect, the improvements are notable, and suggest that a better understanding of the intricacies of Twitter could lead to further improvements.

Interestingly recent work by Bermingham and Smeaton [3] suggests that further linguistic detail when building a feature set in fact harms classification rates. Rather than using POS tagging or larger n-grams, they note that features such as link presence and punctuation mark frequency serve as far better discriminators for subjectivity and polarity. They report accuracy rates of 74.85%, which are strikingly similar to those achieved by Barbosa and Fang.

Building a training set for Twitter can prove difficult due to the need for large data sets. There is an extraordinary diversity of structure, language and grammatical approach on Twitter, thus a large training set is necessary if we hope to be able to accurately classify it's broad range of opinion. Further to the innovative approach taken by Barbosa and Fang, Go et al. [5] suggest a further innovative technique for quickly building a large data set. By searching Twitter for all tweets containing positive and negative emoticons, Go et al. were able to quickly assemble list of polarised opinion. This method for building a training set proved remarkably successful, and simply using unigrams as features, they report an accuracy rate of 82.2%.

Although literature regarding sentiment analysis on Twitter is limited, there have been significant advances in accuracy. Interestingly, as Bermingham and Smeaton observe, detailed linguistic features seem to be of little benefit when classifying subjectivity and polarity. However, as noted by Go et al., the size of the training set seems to have a marked effect on classification accuracy. Clearly Twitter poses many new challenges for sentiment analysis, and although progress has been made, more in depth research is needed before the best approaches can be truly identified.

2.4 Emotion

Defining emotion has been a problem that has puzzled philosophers and thinkers as far back as Cicero and Descartes. Although there is no unifying theory, or completely accepted classification, in general many have agreed there to be two broad types of emotion, the *basic emotions* and *complex emotions*. Basic emotions are biologically innate within all humans, whilst complex emotions are culturally specific amalgamations of our basic one. Deciding upon both what constitutes are basic and complex emotions however has been the cause of significant debate. Perhaps the two most prominent classifications of recent times are those put forward by Ekman [4] and Plutchik [11].

2.4.1 Current defenitions

After years of work within the field, and having observed the Fore tribesmen of Papua New Guinea, Ekman's 1969 paper presented what he believed to be the six core emotions. These were *anger*, *disgust*, *fear*, *happiness*, *sadness*, *suprise*. Within his work he notes that the Fore tribesmen could identify these emotions when presented with photos of faces expressing them, regardless of their cultural origin.

In 1980, Robert Plutchik [11] presented his research into human emotion. Within it, he uses five of the emotions put forward by Ekman, whilst introducing three new emotions. Plutchik expands these emotions further, referring to them as *dimensions*, within which different emotions can express varying degrees of their dimension. Furthermore, each of the eight emotion definitions also has a polar opposite definition within the list:

Taking this original list of eight, Plutchik also proposes eight complex emotions, resulting from combinations of the original eight. These are:

The level of detail within Plutchik's research provides a wider scope of definition

Figure 2.1: Robert Plutchik's eight basic emotions *(* proposed by Ekman)*

Basic Emotion	Polar Emotion	Degrees (strong to weak)		
Joy	Sadness	Ecstasy	Joy	Serenity
Trust	Disgust	Admiration	Trust	Acceptance
Fear*	Anger	Terror	Fear	Apprehension
Surprise	Anticipation	Amazement	Surprise	Distraction
Sadness*	Joy	Grief	Sadness	Pensiveness
Disgust*	Trust	Loathing	Disgust	Boredom
Anger*	Fear	Rage	Anger	Annoyance
Anticipation*	Surprise	Vigilance	Anticipation	Interest

Figure 2.2: Robert Plutchik's eight complex emotions

Combined basic emotions	Complex Emotion	Polar Emotion
Anticipation *and* Joy	Optimism	Disappointment
Joy *and* Trust	Love	Remorse
Trust *and* Fear	Submission	Contempt
Fear *and* Surprise	Awe	Aggressiveness
Surprise *and* Sadness	Disappointment	Optimism
Sadness *and* Disgust	Remorse	Love
Disgust *and* Anger	Contempt	Submission
Anger *and* Anticipation	Aggressiveness	Awe

than that put forward by Ekman. For this reason it shall serve as the classification system we attempt to computationally replicate. Furthermore, Plutchik's proposal introduces concepts of polarity and strength to emotion. Both these concepts bare strong similarities to research within sentiment classification 2.2.3, and we will explore the benefits of this similarity in chapter 5.

2.4.2 Computational classification

PART II | IMPLEMENTATION

PART III | EVALUTATION

BIBLIOGRAPHY

[1] L Barbosa. Robust Sentiment Detection on Twitter from Biased and Noisy Data. *research.att.com*.

[2] F Benamara and C Cesarano. Sentiment analysis: Adjectives and adverbs are better than adjectives alone. *Proceedings of the ...*, 2007.

[3] A Bermingham. Classifying sentiment in microblogs: is brevity an advantage? *Proceedings of the 19th ACM ...*, 2010.

[4] P Ekman. The repertoire of nonverbal behavior: Categories, origins, usage, and coding. *Semiotica*, 1969.

[5] A Go and R Bhayani. Twitter sentiment classification using distant supervision. *CS224N Project Report*, 2009.

[6] B Liu. Sentiment analysis and subjectivity. *Handbook of Natural Language Processing,*, 2010.

[7] R Mihalcea and C Banea. Learning multilingual subjective language via cross-lingual projections. ... *MEETING-ASSOCIATION FOR ...*, 2007.

[8] Peter Norvig. On chomsky and the two cultures of statistical learning, May 2011.

[9] B Pang. A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. 2004.

[10] B Pang and L Lee. Thumbs up?: sentiment classification using machine learning techniques. ... *of the ACL-02 conference on ...*, 2002.

[11] R Plutchik. The nature of emotions. *American Scientist*, 2001.

[12] P Turney. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. *Proceedings of the 40th Annual Meeting on ...*, 2002.

[13] J Wiebe. Effects of adjective orientation and gradability on sentence subjectivity. *Proceedings of the 18th conference ...*, 2000.

[14] J Wiebe. Learning subjective adjectives from corpora. *Proceedings of the National Conference on Artificial ...*, 2000.

[15] J Wiebe. Learning extraction patterns for subjective expressions. In *Proceedings of the 2003 conference on Empirical ...*, 2003.

[16] J Wiebe. Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the conference on ...*, 2005.

[17] J Wiebe. Word sense and subjectivity. ... *of the 21st International Conference on ...*, 2006.

[18] J Wiebe. Subjectivity word sense disambiguation. *Proceedings of the 2009 ...*, 2009.

[19] Janyce M Wiebe, Rebecca F Bruce, and Thomas P O'Hara. Development and Use of a Gold-Standard Data Set for Subjectivity Classifications. In *the 37th annual meeting of the Association for Computational Linguistics*, pages 246--253, Morristown, NJ, USA, 1999. Association for Computational Linguistics.