Sentimentor: Sentiment Analysis of Twitter Data

James Spencer and Gulden Uchyigit

School of Computing, Engineering and Mathematics
University of Brighton,
Brighton, BN2 4GJ
{j.spencer1,g.uchyigit}@brighton.ac.uk

Abstract. In this paper we present Sentimentor, a tool for sentiment analysis of Twitter data. Sentimentor utilises the naive Bayes Classifier to classify Tweets into positive, negative or objective sets. We present experimental evaluation of our dataset and classification results, our findings are not contridictory with existing work.

Keywords: sentiment analysis, opinion mining, classification, machine learning

1 Introduction

Social networks have revolutionised the way in which people communicate. Information available from social networks is beneficial for analysis of user opinion, for example measuring the feedback on a recently released product, looking at the response to policy change or the enjoyment of an ongoing event. Manually sifting through this data is tedious and potentially expensive.

Sentiment analysis is a relatively new area, which deals with extracting user opinion automatically. An example of a positive sentiment is, *"natural language processing is fun"* alternatively, a negative sentiment is *"it's a horrible day, i am not going outside"*. Objective texts are deemed not to be expressing any sentiment, such as news headlines, for example *"company shelves wind sector plans"*.

There are many ways in which social network data can be leveraged to give a better understanding of user opinion such problems are at the heart of natural language processing (NLP) and data mining research.

In this paper we present a tool for sentiment analysis which is able to analyse Twitter data. We show how to automatically collect a corpus for sentiment analysis and opinion mining purposes. Using the corpus we build a sentiment classifier, that is able to determine positive, negative and objective sentiments for a document.

1.1 Related Work

The increase in social media networks has made sentiment analysis a popular research area, in recent years. In Turney[4] reviews are classified by calculating the summation of polarity of the adjectives and adverbs contained within text. This study utilised movie and car reviews, where thumbs up and thumbs down ratings indicate positive and negative sentiment respectively. A discrepancy between the accuracy of the movie and car reviews was observed with the car reviews getting a higher accuracy. This was attributed to the fact that movie reviews, whilst being positive, can have a lot of adjectives and adverbs that do not fully relate to the overall enjoyment of the film and can actually be more a description of the scenes within the film itself. The PMI-IR (Pointwise Mutual Information - Informations Retrieval) algorithm was used to classify documents. This algorithm works by taking the relevant bigrams from the document then using the near function on a search engine to see how many times this bigram appears near a word that expresses strong positive or negative sentiment, a large number of matches indicates a stronger polarity.

Pang[1] consider word presence vs frequency where word presence is found to be more effective than word frequency for sentiment analysis. Word position within a given sentence can also be effective, where such information can be used to decide if a particular word has more strength at the beginning or the end of a given sentence.

Go[2] train sentiment classifier on Twitter data. This itself presents a new challenge as there is no explicit rating system such as star rating or thumbs rating like in previous work. This issue is negated through the use of Twitter's search functionality by searching for emoticons such as :) :(representing positive and negative sentiment respectively. This system is highly limited as it is restricted to binary classification and does not take into account objective texts. This work explored the use of several different classifiers across different n-grams with and without the use of POS tags. A combination of using Unigrams and Bigrams give the best results across all classifiers. The inclusion of POS tags with unigrams had a negative impact across all classifiers however this still performed better than using bigrams. Our work considers combination of Bigrams and POS tags.

Pak[3] considers objective tweets as well as those that are positive and negative sentiment. This paper discuses the method for collecting corpus data, this again is similar to [2] by using emoticons for positive and negative sets. As it is also concerned with collecting data for an objective set it looks at the tweets from a collection of news sources such as the New York Times, Financial Times etc. Pak[3]. provide a rigorous analysis of their corpus, showing sets of texts differ in terms of the POS tag distributions. Generally there is a far greater difference in the objective and subjective texts than positive and negative sets, such differences show that using POS tags can be a strong indicator of the difference between types of text. The objective and subjective comparison shows that the interjections and personal pronouns are strong indicators of subjective texts whilst common and proper nouns are indicators of objective texts. Subjective texts are often written in first or second person in past tense whilst objective texts are often written in third person. The difference between the positive and negative sets do not give a strong indication, however they are good indicators in the difference between the amount of superlative adverbs and possessive endings, both indicating positive sentiment whilst the negative set often contains more verbs in the past tense as people are often expressing disappointment.

Pak[3] use multi nominal naive Bayes classifier to compare unigrams, bigrams and trigrams they conclude that bigrams give the best coverage in terms of context and expression of sentiment. Pak[3] also compare the usage of using negation attachment to words although this process may be considered unorthodox it does improve the classification process by 2% on average. Pak[3] also consider use of two different methods for reducing the influence of words which occurrence is ambiguous between sets, entropy and salience, out of these two salience was found to work better however the use of these methods can introduce ambiguity into the system meaning that the classification process may fail depending on the filter value selected. The simplification of their calculation for classification is potentially dangerous as this assumes there is equal word distribution across sets, having run this test on our data set we have found that the negative set contained over 4% more words than the positive set, showing clear bias in the classification process. The method used for reporting accuracy, is through the process of plotting accuracy against decision. This essentially allows the system to cherry pick data and claim high accuracy across a small subsection of the testing data whilst ignoring the rest.

2 Sentimator: Sentiment Analysis Tool

Sentimator[1] is a web based tool which uses naive Bayes Classifier to classify live Twitter data based on positivity, negativity and objectivity. Sentimentor has an interface which enables the user to analyse the word distributions(see Figures 1 and 2. Sentimentor presents classification results in a easy to understand pictorial format (see Figure 3). Other functionalities of Sentimentor include: the text type details(see Figure 4); The analysis of the twitter message (see Figure 5); search (see Figure 6).

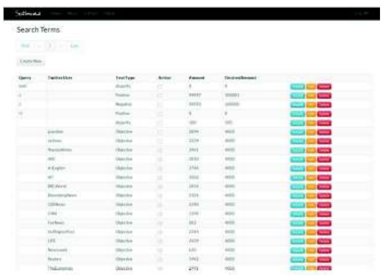

Fig. 1. Screenshot of the search term index

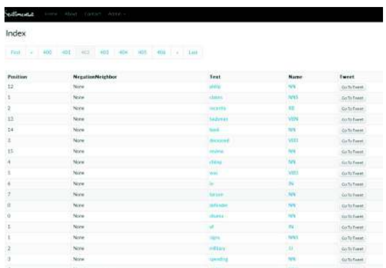

Fig. 2. Screenshot of the word position index

[1] http://sentimentor.co.uk

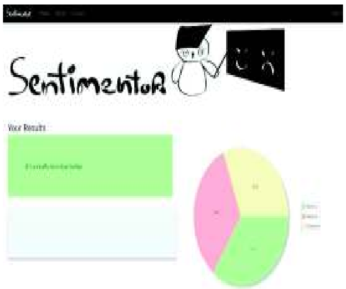

Fig. 3. Sentiment analysis of a piece of text

Fig. 4. Screenshot of the text type details

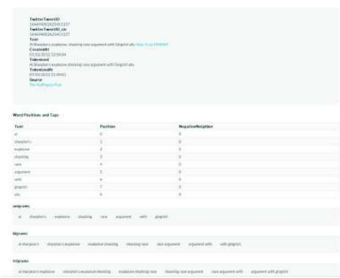

Fig. 5. Screenshot of the tweet details page

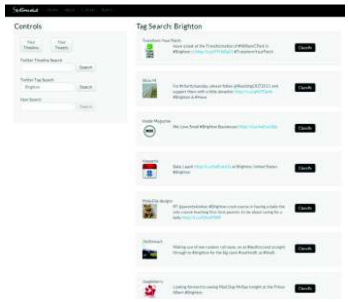

Fig. 6. Screenshot of the twitter search functionality

3 Data Collection and Preprocessing

Twitter API was used for the data extraction process. Negative, positive and objective texts were collected by following the same procedures as in([2] and [3]). Tokenization process from [2] and [3] was followed for the data preprocessing task. The steps followed included the removal of any urls and usernames (usernames follow the @symbol) and removal any characters that repeat more than twice turning a phrase such as OOMMMGGG to OOMMGG, which is applied by a regular expression. Table 1, shows an example of the tokenization process. Finally, the stopset words were removed from the data. The stopset is the set of words such as “a”, “and”, “an”, “the”, these are words that do not have any meaning on their own. The second phase is associated with determining the POS tag for each word. The OpenNLP library was used for POS tagging and the extraction of unigrams and bigrams. An example of bigrams extracted from our dataset is shown in Tables 2, 3 and 4.

Table 1. Example of the Tokenization Process

	Before	After
1	I wanna go to @AvrilLavigne 's concert stadium merdeka sooooooooo badly :(love you avril! Xo	in I wanna go to s concert in stadium merdeka soo badly love you avril Xo
2	@chuckcomeau 1:45am!!! OMG I WAS SLEPT AT 11:00pm WOOOOOOW I WANT A SKATE :)	am OMG I WAS SLEPT AT pm WOOW I WANT A SKATE
3	British adventurer Felicity Aston becomes first woman to ski across Antarctica	British adventurer Felicity Aston becomes comes first woman to ski across Antarctica alone

Table 2. Positive Bigram Counts

Bigram	Count
i love	2899
valentines day	2797
happy valentines	2191
thank you	2141
love you	2133
follow back	1516
d rt	1491
think i'm	1410
follow me	1342
if you	1263

Table 3. Negative Bigram Counts

Bigram	Count
i miss	3292
i have	2440
i don't	2041
i was	1922
i want	1881
but i	1813
i know	1760
miss you	1681
want to	1609
i can't	1595

Table 4. Objective Bigram Counts

Bigram	Count
to be	916
front page	574
new york	524
if you	506
in today's	496
out of	430
will be	426
mitt romney	418
us your	397
more than	395

3.1 Evaluation of the data set

An original corpus of Twitter data was collected and compared with the corpus presented in Pak[3]. The percentage distribution of POS tags across sets is shown in Figure 7.

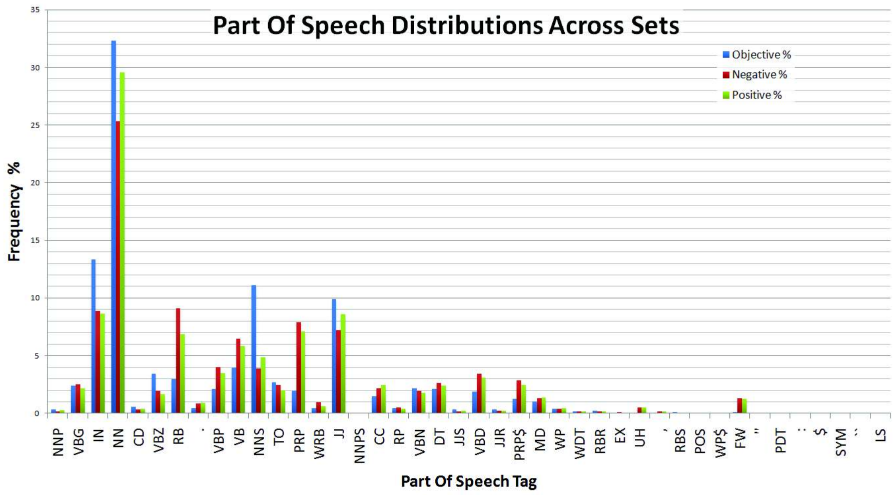

Fig. 7. The percentage distribution of POS tags across the sets

Overall singular noun (NN) is the most common POS tag, occurring 29.08% across the whole corpus. Preposition or conjunction (IN) occur 10.28% of the time with it being clear that there is a significant difference between the occurrence in all sets. To better understand the differences between sets we have calculated the percentage difference between the percentage distribution of each POS tag. This has been done for the difference between the objective and subjective sets and between the positive and negative sets, this is displayed in Figure 8 and Figure 9 respectively. Figure 8 shows a significant difference in the amount of interjections (UH) and personal pronouns (PRP, PRP$) favouring the subjective set as reported by Pak[3] . The common nouns and proper nouns are a strong indicator of the subjective set by looking at common noun plural (NNS) nouns proper singular (NNP) and noun common singular (NN). According to Pak[3] we expect writers of subjective text to be talking in the first or second person, we can partially confirm this by looking at the difference of verb present tense not third person singular (VBP) and verb past tense (VPD) however verb, present participle (VBG) contradicts this as it prevails in the objective set. This could have happened because the selected news outlets might have more comment on news than original reporting or this could be a difference in the POS tagger, however this is of little concern because the difference is relatively negligible. Likewise we can expect objective texts to be in third person the results for

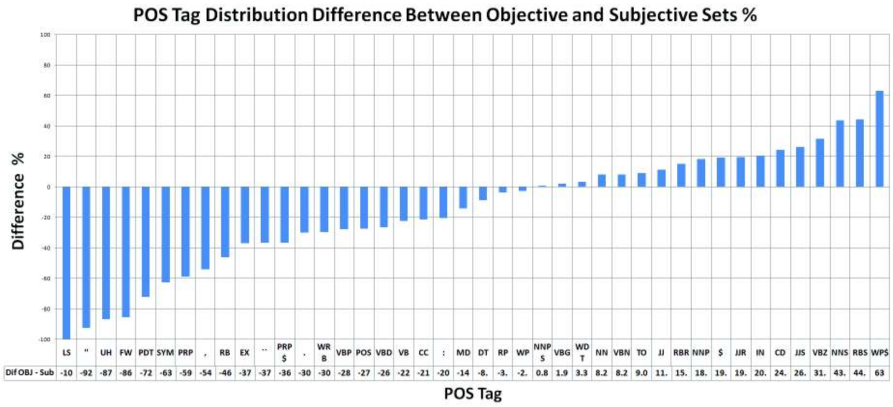

Fig. 8. Graph showing the percentage difference of POS tags frequencies between objective and subjective texts

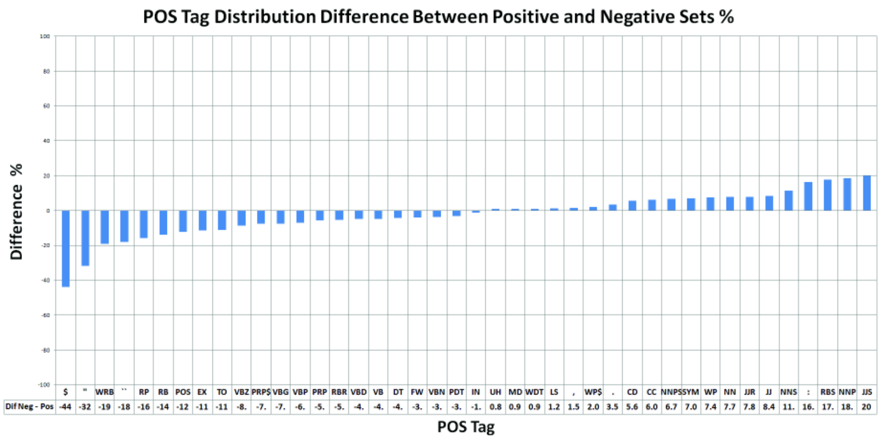

Fig. 9. Graph showing the percentage difference of POS tag frequencies between positive and negative sets

verb present tense 3rd person singular (VBZ) can confirm this. In our dataset Superlative adverbs (RBS) have a very Strong weighting for objective text this is contrary to Pak[3] , where this is not significant [16]. The List item marker (LS) has a -100% difference as this doesn't occur in the objective set this tag isn't present in Pak [3] data. The symbols that have not been removed by the tokenizer are a potential source of error as these represent significant difference between sets. The POS tagger used has the ability to detect foreign words (FW) which have a Strong indication of the text being subjective, the reasoning for this is because news outlets would only be expected to use correctly structured English, standard user tweets may contain a mix of languages despite the fact that the Twitter search was limited to English tweets. Now looking at Figure 9 we can see that these two sets are a lot closer in terms of POS difference, which is expected as both sets are subjective. The strongest indicators for negative sentiment is Currency ($) and quotation marks while an individual is highly likely to express their fiscal issues in a negative sentiment but as there are only 19 occurrences of currency in the system this is not a good indicator of what set the text belongs to, also the inclusion of quotation marks here is likely going to introduce error into the system. Wh-adverb - negative (WRB) , particle (RB, RP) genitive marker (POS) are all strong indicators on negative sentiment, however Pak[3]. state that (POS) may be an indicator of positive sentiment, the results we have collected contradict this. Superlative adverb (RBS) , proper noun singular (NNP) , adjective superlative (JJS) , Noun (NNS) common plural are all indicating strong positive sentiment. The appearance of (RBS) confirms that this is a good indicator of a positive sentiment.

3.2 Classification

The naive Bayes classifier was used for classification this decision is primarily based on findings by Pak and Go [[2],[3]], that the naive Bayes classifier show good performance results.

$$P(C|m) = P(C) \prod_{i=1}^{n} P(f_i|C)$$

where C is the class positive, negative or objective sets, m is the twitter message and f is a feature. In our experiments the features are POS tags, unigrams or bigrams.

4 Results

We have tested our classifier against a training set which contains 216 manually tagged tweets. We have provided the test results for unigrams and bigrams both with and without the use of POS tags these results are detailed in Tables 5,6,7, 8 and 9. Table 9 details the accuracy of each of the previously mentioned tests. The test with the highest accuracy is the one using bigrams without POS tags with an accuracy of 52.31% and the lowest is Unigrams without POS tags at 46.76%. Accuracy would be far higher if we were to carry out these tests using binary classification and it should be stated that this is one of the further complexities of using microblogging data as appose to using reviews as these are not expected to be objective. The use of bigrams has show an increase in performance with or without the use of POS tags. This also reduces the amount of false positives in the objectivity classifier however there is also notable increase in false positives by the positive classifier, the negative classifier does not seem to be effected much by this. Overall the use of POS tags has had a negative effect on the accuracy of the calssification proccess, this is caused by the Ambiguity of POS tag occurances across sets this is most likely also the case because we using the summation of POS tags in a given phrase and not looking for binary occurance as disscused in [1]. It may potentially benifit the classifcation proccess to give less wheight to the POS tags or to experiment with diffrent n-grams of POS tags. We have confirmed previous works finding to be correct in there conclusion that bigrams give better results than unigrams. The overall performance of the system is satisfactory, however we would still like to further improve this as outlined in our future work section.

Table 5. Results for Unigrams

Sentiment	Number of Samples	Correctly Identified	False Positives
Positive	108	37	9
Negative	75	45	45
Objective	33	19	61

Table 6. Results for Unigrams and POS Tags

Sentiment	Number of Samples	Correctly Identified	False Positives
Positive	108	39	10
Negative	75	45	42
Objective	33	18	62

Table 7. Results for Bigrams

Sentiment	Number of Samples	Correctly Identified	False Positives
Positive	108	47	16
Negative	75	47	44
Objective	33	19	43

Table 8. Results for Bigrams and POS Tags

Sentiment	Number of Samples	Correctly Identified	False Positives
Positive	108	46	19
Negative	75	45	43
Objective	33	17	46

Table 9. Results compared

Test	Correctly Identified	False Positives
Unigrams	101	46.75%
Unigrams POS	109	47.2%
Bigrams	113	52.31%
Bigrams POS	108	50%

5 Conclusion and Future Work

In this paper we have presented a way in machine learning techniques can be applied to large sets of data to establish membership, in this case positivity, negativity and objectivity. We have looked at common process in NLP that can help us derive the meaning or context of a given phrase. We have demonstrated how to collect an original corpus for sentiment classification and the refinement that is needed with such data. We have applied a naive Bayes classifier to this set conduct sentiment analysis and have found this process to be successful. On analysis of our results we have confirmed that bigrams offer better performance when conducting the classification process supporting Pak[3] results. We has also confirmed some of Pak[3] findings when looking at the differences between the objectivity and subjectivity set, the same can't be be said for the positive and negative sets which prove to be far more ambiguous. We have discovered that collecting data across a short amount of time may be a potential source of error when determining sentiment, this is due to the fact that opinions can shift over time as can the meaning of words. The classification process itself has been successful with and accuracy of 52.31% however it is felt that this could be further improved, this is outlined in future work. One of our future works is to experiment with different classifiers on our dataset. We also intend on developing an application which carries our textual analysis on video games servers analysing what a player is expressing and adjusting the game enviroment accordingly.

References

1. Bo Pang, L.L.: Opinion mining and sentiment analysis. Foundations and Trends in Information Retrieval January Volume 2 Issue 1-2, 1–94 (2008), `http://www.cs.`

cornell.edu/home/llee/omsa/omsa.pdf
2. Go, A., Bhayani, R., Huang, L.: Twitter sentiment classification using distant supervision. Processing 150(12), 1–6 (2009), http://www.stanford.edu/~alecmgo/papers/TwitterDistantSupervision09.pdf
3. Pak, A., Paroubek, P.: Twitter as a corpus for sentiment analysis and opinion mining. In: Chair), N.C.C., Choukri, K., Maegaard, B., Mariani, J., Odijk, J., Piperidis, S., Rosner, M., Tapias, D. (eds.) Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10). European Language Resources Association (ELRA), Valletta, Malta (may 2010)
4. Turney, P.D.: Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews. In: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. pp. 417–424. ACL '02, Association for Computational Linguistics, Stroudsburg, PA, USA (2002), http://dx.doi.org/10.3115/1073083.1073153