Classifying Sentiment in Microblogs: Is Brevity an Advantage?

Adam Bermingham & Alan Smeaton
CLARITY: Centre for Sensor Web Technologies
School of Computing
Dublin City University
{abermingham, asmeaton}@computing.dcu.ie

ABSTRACT

Microblogs as a new textual domain offer a unique proposition for sentiment analysis. Their short document length suggests any sentiment they contain is compact and explicit. However, this short length coupled with their noisy nature can pose difficulties for standard machine learning document representations. In this work we examine the hypothesis that it is easier to classify the sentiment in these short form documents than in longer form documents. Surprisingly, we find classifying sentiment in microblogs easier than in blogs and make a number of observations pertaining to the challenge of supervised learning for sentiment analysis in microblogs.

Categories and Subject Descriptors

H.3.3 [**Information Search and Retrieval**]: Text Mining

General Terms

Algorithms, Experimentation

1. INTRODUCTION

Microblogging has become a popular method for Internet users to publish thoughts and information in real-time. Automated sentiment analysis of microblog posts is of interest to many, allowing monitoring of public sentiment towards people, products and events, as they happen.

The short length of microblog documents means they can be easily published and read on a variety of platforms and modalities. This brevity constraint has led to the use of non-standard textual artefacts such as emoticons and informal language. The resulting text is often considered "noisy".

It is reasonable to assume that the short document length introduces a succinctness to the content. The focused nature of the text and higher density of sentiment-bearing terms may benefit automated sentiment analysis techniques. On the other hand, it may also be that the shorter length and language conventions used mean there is not enough context for sentiment to be accurately detected. It is unclear which of these is true.

These issues motivate our research questions: (i) How does sentiment classification accuracy in the microblogging domain compare to that for microreviews, another short-form textual domain? How do these accuracies compare to those for their long-form counterparts? and (ii) How do different feature vector representations and classifiers affect sentiment classification accuracy for microblogs? How does this compare to the corpora explored in (i)?

CIKM'10, October 26–29, 2010, Toronto, Ontario, Canada.

2. RELATED WORK

Sentiment analysis has been successfully used to analyse and extract opinion from text in recent years [12]. Some exploratory works have been completed in the microblog domain. Diakapolous and Shamma used manual annotations to characterize the sentiment reactions to various issues in a political debate [5]. They find that sentiment is useful as a measure for identifying controversy. Jansen *et al.* studied the Word-Of-Mouth effect on Twitter using an adjective-based sentiment classifier, finding it useful for brand analytics on Twitter. Bollen *et al.* analysed sentiment on Twitter according to a six-dimensional mood representation [2] finding that sentiment on Twitter correlates with real-world values such as stock prices and coincides with cultural events. The latter two studies report positive results from using automated sentiment analysis techniques on Twitter data.

Noise in Computer-Mediated-Content has been the subject of much research. Tagliamonte and Denis studied instant messaging [13], finding that the penetration of non-standard English language and punctuation is far less than is reported in the media. In a study of classification of customer feedback, Gamon found a high level of accuracy for supervised sentiment classification despite their noisy nature [7]. One strategy to deal with noise in the domain put forward by Choudhury *et al.* is to use Hidden Markov Models to decode text into standard English [4] reporting a high level of success for SMS data. Agarwal *et al.* showed that by simulating noise in text classification, a good classifier should perform well up to about 40% noise [1] suggesting that although noise may be present in text, this may not prove to be important for supervised learning tasks. Carvalho *et al.* found that non-standard surface features such as a heavy punctuation and emoticons are key to detecting irony in user-generated content [3].

Collectively, these studies all support our assumption that new textual domains exhibit domain specific features. We also see that there is significant value in being able to model

Table 1: Microblog annotation labels and associated document counts.

Label	#Documents
Relevant, Positive	1,410
Relevant, Negative	1,040
Relevant, Neutral	2,597
Relevant, Mixed	146
Not relevant	498
Unannotatable	603
Unclear	530
Total	6,824

sentiment in these domains. To our knowledge, this is the first work to explore the challenges that the shortness of microblog documents present to feature vector representations and supervised sentiment classification.

3. METHODOLOGY

The microblog posts used in these experiments are taken from a collection of over 60 million posts which we gathered from the Twitter public data API[1] from February to May 2009. We examined the trending topics from this period and identified five recurring themes: *Entertainment*, *Products and Services*, *Sport*, *Current Affairs* and *Companies*. We selected 10 trends from each of these categories to be used as sentiment targets. By making the topic set diverse and challenging, we hope to better test the performance of our approach and build a classifier representative of a real world generic sentiment classification scenario.

In the annotation process, Wilson's definition of sentiment was used: *"Sentiment analysis is the task of identifying positive and negative opinions, emotions, and evaluations."* [14] Our team of annotators consisted of 9 PhD students and Postdoctoral researchers. To ensure sufficient agreement among the annotators, the annotation was preceded by a number of training iterations, consisting of group meetings, consensus annotations and one-on-one discussions. See Table 3 for a breakdown of annotations by label.

In total, 9 annotators annotated 17 documents for each of the 50 topics. 463 documents were doubly annotated for inter-annotator agreement (6.78%). For the 7 labels, the Kappa agreement was 0.65. For the 3 classes which we use for training (positive, negative and neutral) Kappa was 0.72. If we just consider the binary sentiment classes, positive and negative, this increases to 0.94. These relatively high values for kappa are consistent with our previous annotation of blogs [10].

To contrast with our microblogs corpus, we derive a corpus of blog posts from the TREC *Blogs06* corpus [8]. We use a templating approach to extract positive, negative and neutral blog post content and comments from the corpus, using the TREC relevance judgments as labels.

As much of sentiment analysis literature concerns review classification, in parallel to our experiments on the microblog and blog corpora, we also conduct our experiments on a corpus of microreviews and a corpus of reviews. In January 2010 we collected microreview documents from the microreview website, Blippr[2]. Blippr reviews bear a similarity to microblog posts in that they share the same character limit of 140 characters. Reviews on Blippr are given one of four ratings by the author, in order from most negative to most positive: *hate*, *dislike*, *like* and *love*. In our corpus we use only reviews with strongly polarised sentiment: *hate* and *love*. We have made our microreview and microblog corpora available for other researchers[3].

The reviews corpus we use as comparison is perhaps the mostly widely studied sentiment corpus, Pang and Lee's movie review corpus [11]. This corpus contains archival movie reviews from USENET. We refer to the microblog and microreview datasets as the *short-form* document corpora and the blog and movie review datasets as the *long-form* document corpora.

Our datasets are limited to exactly 1000 documents per class in line with the movie review corpus. This allows us to eliminate any underlying sentiment bias which may be present. While this is obviously a consideration for a real-world system, in our experiments we wish to examine the challenges of the classification without biasing our evaluation towards any particular class. As the sentiment distribution is different in each of the domains, this also makes accuracies comparable across datasets.

For our experiments we use Support Vector Machine (SVM) and Multinomial Naive Bayes (MNB) classifiers, giving us an accurate representation of the state-of-the-art in text classification. We use an SVM with a linear kernel and the parameter `c` set to 1. In preliminary experiments we found binary feature vectors more effective than frequency-based vectors and found no benefit from stopwording or stemming. Where possible, we replaced topics with pseudo-terms to avoid learning topic-sentiment bias. We also replace URLs and usernames with pseudo-terms to avoid confusion during tokenization and POS tagging. Each feature vector is L2 Normalized and for the the long-form corpora only features which occurred 4 or more times were used, as Pang and Lee did in their original movie review experiments. Accuracy was measured using 10 fold cross-validation and the folds were fixed for all experiments.

As a baseline for binary (positve/negative) classification we use a classifier based on a sentiment lexicon, SentiWordNet [6]. This unsupervised classifier calssifies documents using the mean sentiment scores of the synsets its words belong to. Despite their naivety, this type of classifier is often used as it does not require expensive training data.

4. RESULTS AND DISCUSSION

Unigram binary (positive/negative) classification accuracy for microblogs is 74.85% using an SVM. This is an encouraging accuracy given the diversity in the sentiment topics. As we have balanced datasets, a trivial classifier achieves 50% accuracy for binary classification. For microreviews, the accuracy is considerably higher at 82.25% using an SVM. As expected, the classifier finds it easier to distinguish between polarised reviews than to identify sentiment in arbitrary posts.

Sentiment classification of the long-form documents yields some surprising results. Blog classification accuracy is significantly lower than for microblogs. However, movie review classification is higher than for microreviews, confirming Pang and Lee's results of 87.15% for SVM with unigram

[1] `http://apiwiki.twitter.com`

[2] `http://www.blippr.com`

[3] `http://www.computing.dcu.ie/~abermingham/data/`

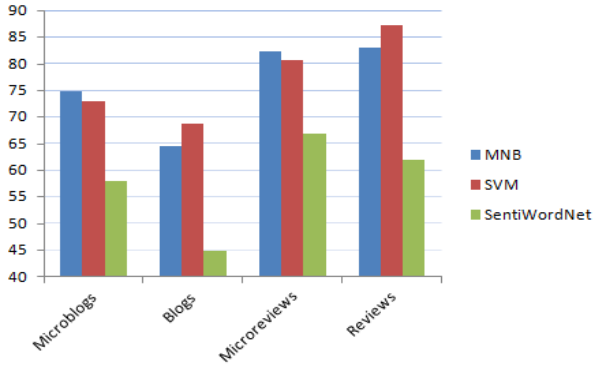

Figure 1: Accuracies for unigram features.

features. At first this may seem contradictory — surely the classifier should perform consistently across textual domains? We speculate that this behaviour is due to within-document topic drift. In the two review corpora the text of the document has a high density of sentiment information about the topic, and a low noise density. In the blogs dataset, this is not necessarily the case; the sentiment in a blog post may be an isolated reference in a subsection of the document. Although topic drift also occurs in the microblog corpus, there is less opportunity for non-relevant information to enter the feature vector and our classifier is not as adversely affected as in the blog domain.

Our unsupervised lexicon-based classifier performs poorly across all datasets. For the blogs corpus, it is outperformed by a trivial classifier. The accuracy gap between supervised and unsupervised classification accuracy in the long-form corpora is much more pronounced. This makes intuitive sense as the probability of the polarity of a word in a document expressing sentiment towards a topic is again much higher in the short-form domains.

Of the two supervised classifiers, SVM outperforms MNB in the long-form domains, whereas the opposite is true in the short-form domains. SVMs scale better with larger vector dimensionality so this is most likely the reason for this observation; the number of unique terms in the longer documents is over three times their shorter counterparts, even when infrequent features have been excluded.

Having established a reasonable performance in sentiment classification of microblog posts, we wish to explore whether we can improve the standard bag of words feature set by adding more sophisticated features. Using sequences of terms, or n-grams, we can capture some of the information lost in the bag-of-words model. We evaluated two feature sets: (unigrams + bigrams) and (unigrams + bigrams + trigrams). We found that although an increase in classification accuracy is observed for the movie reviews, this is not the case for any of the other datasets (see Table 2). We also examined POS-based n-grams in conjuction with a unigram model and observed a decrease in accuracy across all corpora. This indicates that the syntactic patterns represented by the POS n-gram features do contain information which is more discriminative than unigrams.

The most promising results came from a POS-based stopwording approach proposed by Matsumoto *et al.* ([9]). This approach (which Matsumoto *et al.* refer to as "word subsequences") consists of an n-gram model, where terms have been stopworded based on their POS. We use the same POS list as Matsumoto. These features increase accuracy across all corpora for unigrams + POS-stopworded bigrams. This suggests that a better understanding of the linguistic context of terms is similarly advantageous in all domains.

Table 3: Most discriminative unigrams, bigrams and trigrams according to Information Gain Ratio for binary classification.

	Microblogs	Blogs	Microreviews	Reviews
1	!	witherspoon	great	bad
2	<Url>	joaquin	boring	worst
3	<Topic>	reese witherspoon	best	stupid
4	amazing	joaquin phoenix	terrible	boring
5	. .	sharon	the best	the worst
6	! !	ledger	worst	waste
7	?	heath ledger	n't	ridiculous
8	! ! !	heath	love	wasted
9	love	johnny cash	loved	awful
10	<Topic> !	palestinians	?	?

Table 4: Ternary Unigram Classification Accuracies: Positive, Negative, Neutral

	MNB	SVM	#features
Microblogs	61.3	59.5	8132
Blogs	52.13	57.6	28805

To examine the performance of individual features, we use a standard measure of discriminability, Information Gain Ratio (see Table 3). Immediately obvious is the significant role that punctuation plays in expressing sentiment in microblog posts. This suggests that these are being used specifically in microblog posts to express sentiment, perhaps as indicators for intonation. The discriminative features for both the reviews and microreviews are largely similar in nature, typically polarised adjectives. The blog classifier appears to have learned a certain amount of entity bias as many of the discriminative features are people or places. Note that none of these entities are topic terms (topic terms were removed in pre-processing), though they do appear to be entities associated with topics.

Results of our ternary classification on microblogs and blogs can be seen in Table 4. The accuracy is, as expected, significantly less than for binary classification with SVMs again outperforming MNB on the longer blog documents.

5. CONCLUSION

The results of our experiments on the whole are encouraging for the task of analysing sentiment in microblogs. We achieve an accuracy of 74.85% for binary classification for a diverse set of topics indicating we can classify microblog documents with a moderate degree of confidence. In both of our short-form corpora we find it difficult to improve performance by extending a unigram feature representation. This is contrary to the long-form corpora which respond favourably to enriched feature representations. We do however see promise in sophisticated POS-based features across

Table 2: Binary accuracy summary (figures as %)

	Microblogs		Blogs		Microreviews		Movies	
Feature Set	MNB	SVM	MNB	SVM	MNB	SVM	MNB	SVM
Unigram	**74.85**	72.95	64.6	68.75	**82.25**	80.8	82.95	87.15
Unigram+Bigram	74.35	72.95	64.6	68.45	82.15	**81.4**	85.25	**87.9**
Unigram+Bigram+Trigram	73.7	72.8	64.6	68.5	81.95	80.85	84.8	**87.9**
Unigram+POS n-gram (n=1)	73.25	71.6	64.7	68.45	80.8	79.5	82.4	86.95
Unigram+POS n-gram (n=1,2)	70.25	70.05	62.6	66.25	80.8	79.5	81.8	84.95
Unigram+POS n-gram (n=1,2,3)	68.8	69.7	62.45	64.6	74.7	76.9	79.95	82
Unigram+POS-stopworded Bigram	74.15	73.25	64.5	**69**	82.5	81.05	85.35	87.5
Unigram+POS-stopworded Bigram+Trigram	74.4	**73.45**	**64.85**	68.7	82.15	80.6	**85.5**	87.8

all datasets and speculate that engineering features based on deeper linguistic representations such as dependencies and parse trees may work for microblogs as they have been shown to do for movie reviews. In analysing discriminative features, we found that a significant role is played by punctuation. As a future direction for this work we hope to explore this notion with a view to incorporating it into the feature engineering process. It is surprising to see that this is not a pattern seen in our microreviews corpus indicating that this is not an artefact of the brevity of the platforms.

We conclude that although the shortness of the documents has a bearing on which feature sets and classifier will provide optimum performance, the sparsity of information in the documents does not hamper our ability to classify them. On the contrary, we find classifying these short documents a much easier task than their longer counterparts, blogs. Also, the "noisy" artefacts of the microblog domain such as informal punctuation turn out to be a benefit to the classifiers. These results provide a compelling argument to encourage the community to focus on microblogs in future sentiment analysis research.

Acknowledgments

This work is supported by Science Foundation Ireland under grant 07/CE/I1147

6. REFERENCES

[1] S. Agarwal, S. Godbole, D. Punjani, and S. Roy. How much noise is too much: A study in automatic text classification. In *ICDM*, pages 3–12, 2007.

[2] J. Bollen, A. Pepe, and H. Mao. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*, abs/0911.1583, 2009.

[3] P. Carvalho, L. Sarmento, M. J. Silva, and E. de Oliveira. Clues for detecting irony in user-generated contents: oh...!! it's "so easy" ;-). In *TSA '09: Proceeding of the 1st international CIKM workshop on Topic-sentiment analysis for mass opinion*, pages 53–56, New York, NY, USA, 2009. ACM.

[4] M. Choudhury, R. Saraf, V. Jain, A. Mukherjee, S. Sarkar, and A. Basu. Investigation and modeling of the structure of texting language. *IJDAR*, 10(3-4):157–174, 2007.

[5] N. A. Diakopoulos and D. A. Shamma. Characterizing debate performance via aggregated Twitter sentiment. In *Conference on Human Factors in Computing Systems (CHI 2010)*, 2010.

[6] A. Esuli and F. Sebastiani. Sentiwordnet: A publicly available lexical resource for opinion mining. In *In Proceedings of the 5th Conference on Language Resources and Evaluation (LREC-06)*, pages 417–422, 2006.

[7] M. Gamon. Sentiment classification on customer feedback data: noisy data, large feature vectors, and the role of linguistic analysis. In *COLING '04: Proceedings of the 20th international conference on Computational Linguistics*, page 841, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[8] C. MacDonald and I. Ounis. The TREC Blogs06 collection : Creating and analysing a blog test collection. Technical report, University of Glasgow, Department of Computing Science, 2006.

[9] S. Matsumoto, H. Takamura, and M. Okumura. Sentiment classification using word sub-sequences and dependency sub-trees. In *Proceedings of PAKDD'05, the 9th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining*, 2005.

[10] N. O'Hare, M. Davy, A. Bermingham, P. Ferguson, P. Sheridan, C. Gurrin, and A. F. Smeaton. Topic-dependent sentiment analysis of financial blogs. In *In: TSA 2009 - 1st International CIKM Workshop on Topic-Sentiment Analysis for Mass Opinion Measurement*, Hong Kong, China, 6 Nov 2009.

[11] B. Pang and L. Lee. A sentimental education: sentiment analysis using subjectivity summarization based on minimum cuts. In *ACL '04: Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, page 271, Morristown, NJ, USA, 2004. Association for Computational Linguistics.

[12] B. Pang and L. Lee. Opinion mining and sentiment analysis. *Foundation and Trends in Information Retrieval*, 2(1-2):1–135, 2008.

[13] S. A. Tagliamonte and D. Denis. LINGUISTIC RUIN? LOL! INSTANT MESSAGING AND TEEN LANGUAGE. *American Speech*, 83(1):3–34, 2008.

[14] T. Wilson, J. Wiebe, and P. Hoffmann. Recognizing contextual polarity in phrase-level sentiment analysis. *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 347–354, 2005.