Sentiment Analysis of Political Tweets: Towards an Accurate Classifier

Akshat Bakliwal[1], **Jennifer Foster**[2], **Jennifer van der Puil**[3*],
Ron O'Brien[4], **Lamia Tounsi**[2] and **Mark Hughes**[5]
[1]Search and Information Extraction Lab, IIIT-Hyderabad, India
[2]NCLT/CNGL, School of Computing, Dublin City University, Ireland
[3]Department of Computer Science and Statistics, Trinity College, Ireland
[4]Quiddity, Dublin, Ireland
[5]CLARITY, School of Computing, Dublin City University, Ireland
[1]akshat.bakliwal@research.iiit.ac.in
[2,5]{jfoster,ltounsi,mhughes}@computing.dcu.ie
[3]jvanderp@tcd.ie
[4]ron@quiddity.ie

Abstract

We perform a series of 3-class sentiment classification experiments on a set of 2,624 tweets produced during the run-up to the Irish General Elections in February 2011. Even though tweets that have been labelled as sarcastic have been omitted from this set, it still represents a difficult test set and the highest accuracy we achieve is 61.6% using supervised learning and a feature set consisting of subjectivity-lexicon-based scores, Twitter-specific features and the top 1,000 most discriminative words. This is superior to various naive unsupervised approaches which use subjectivity lexicons to compute an overall sentiment score for a <tweet,political_party> pair.

1 Introduction

Supervised machine learning using minimal feature engineering has been shown to work well in binary positive/negative sentiment classification tasks on well-behaved datasets such as movie reviews (Pang et al., 2002). In this paper we describe sentiment analysis experiments in a more complicated setup: the task is three-class positive/negative/neutral classification, the sentiment being classified is not at the general document level but rather directed towards a topic, the documents are tweets, and the topic is politics, specifically the Irish General Election of February 2011.

The dataset used in the experiments contains tweets which were collected in the run up to the election and which were subsequently doubly annotated as positive, negative or neutral towards a particular political party or party leader. The annotators also marked a tweet as sarcastic if its literal sentiment was different to its actual sentiment. Before exploring the thorny issue of sentiment classification in the face of sarcasm, we simplify the problem by first trying to establish some sentiment analysis baselines for those tweets which were not deemed to be sarcastic.

We first explore a naive approach in which a subjectivity lexicon is used as the primary source of information in determining whether sentiment towards a political party or party leader is positive, negative or neutral. The best version of this method achieves an accuracy of 58.9, an absolute improvement of 4.9 points over the majority baseline (54%) in which all tweets are classified as neutral. When these lexicon scores are combined with bag-of-word features and some Twitter-specific features in a supervised machine learning setup, this accuracy increases to 61.6%.

The paper is organised as follows: related work is described in Section 2, followed by a brief discussion of the 2011 Irish General Election in Section 3, a description of the dataset in Section 4 and a description of the natural language processing tools and resources employed in Section 5. In Section 6, the unsupervised lexicon-based approach is presented and its limitations discussed. Section 7 describes the machine-learning-based experiments and Section 8 concludes and provides hints towards fu-

[*] Akshat Bakliwal and Jennifer van der Puil carried out their part of this work while employed as summer interns at the Centre for Next Generation Localisation(CNGL) in the School of Computing, DCU.

Proceedings of the Workshop on Language in Social Media (LASM 2013), pages 49–58,
Atlanta, Georgia, June 13 2013. ©2013 Association for Computational Linguistics

ture work with this new dataset.

2 Previous Work

The related work can be divided into two groups, general sentiment analysis research and research which is devoted specifically to the political domain.

2.1 General Sentiment Analysis

Research in the area of sentiment mining started with product (Turney, 2002) and movie (Pang et al., 2002) reviews. Turney (2002) used Pointwise Mutual Information (PMI) to estimate the sentiment orientation of phrases. Pang et al. (2002) employed supervised learning with various set of n-gram features, achieving an accuracy of almost 83% with unigram presence features on the task of document-level binary sentiment classification. Research on other domains and genres including blogs (Chesley, 2006) and news (Godbole et al., 2007) followed.

Early sentiment analysis research focused on longer documents such as movie reviews and blogs. Microtext on the other hand restricts the writer to a more concise expression of opinion. Smeaton and Bermingham (2010) tested the hypothesis that it is easier to classify sentiment in microtext as compared to longer documents. They experimented with microtext from Twitter, microreviews from *blippr*, blog posts and movie reviews and concluded that it is easier to identify sentiment from microtext. However, as they move from contextually sparse unigrams to higher n-grams, it becomes more difficult to improve the performance of microtext sentiment classification, whereas higher-order information makes it easier to perform classification of longer documents.

There has been some research on the use of positive and negative emoticons and hashtags in tweets as a proxy for sentiment labels (Go et al., 2009; Pak and Paroubek, 2010; Davidov et al., 2010; Bora, 2012). Bakliwal et al. (2012) emphasized the importance of preprocessing and proposed a set of features to extract maximum sentiment information from tweets. They used unigram and bigram features along with features which are more associated with tweets such as emoticons, hashtags, URLs, etc. and showed that combining linguistic and Twitter-specific features can boost the classification accuracy.

2.2 Political Sentiment Analysis

In recent years, there has been growing interest in mining online political sentiment in order to predict the outcome of elections. One of the most influential papers is that of Tumasjan et al. (2010) who focused on the 2009 German federal election and investigated whether Twitter can be used to predict election outcomes. Over one hundred thousand tweets dating from August 13 to September 19, 2009 containing the names of the six parties represented in the German parliament were collected. LIWC 2007 (Pennebaker et al., 2007) was then used to extract sentiment from the tweets. LIWC is a text analysis software developed to assess emotional, cognitive and structural components of text samples using a psychometrically validated internal dictionary. Tumasjan et al. concluded that the number of tweets/mentions of a party is directly proportional to the probability of winning the elections.

O'Connor et al. (2010) investigated the extent to which public opinion polls were correlated with political sentiment expressed in tweets. Using the Subjectivity Lexicon (Wilson et al., 2005), they estimate the daily sentiment scores for each entity. A tweet is defined as positive if it contains a positive word and vice versa. A sentiment score for that day is calculated as the ratio of the positive count over the negative count. They find that their sentiment scores were correlated with opinion polls on presidential job approval but less strongly with polls on electoral outcome.

Choy et al. (2011) discuss the application of online sentiment detection to predict the vote percentage for each of the candidates in the Singapore presidential election of 2011. They devise a formula to calculate the percentage vote each candidate will receive using census information on variables such as age group, sex, location, etc. They combine this with a sentiment-lexicon-based sentiment analysis engine which calculates the sentiment in each tweet and aggregates the positive and negative sentiment for each candidate. Their model was able to predict the narrow margin between the top two candidates but failed to predict the correct winner.

Wang et al. (2012) proposed a real-time sentiment analysis system for political tweets which was based on the U.S. presidential election of 2012. They collected over 36 million tweets and collected the sentiment annotations using Amazon Mechanical Turk. Using a Naive Bayes model with unigram features, their system achieved 59% accuracy on the four-category classification.

Bermingham and Smeaton (2011) are also concerned with predicting electoral outcome, in particular, the outcome of the Irish General Election of 2011 (the same election that we focused on). They analyse political sentiment in tweets by means of supervised classification with unigram features and an annotated dataset different to and larger than the one we present, achieving 65% accuracy on the task of *positive/negative/neutral* classification. They conclude that volume is a stronger indicator of election outcome than sentiment, but that sentiment still has a role to play.

Gayo-Avello (2012) calls into question the use of Twitter for election outcome prediction. Previous works which report positive results on this task using data from Twitter are surveyed and shortcomings in their methodology and/or assumptions noted. In this paper, our focus is not the (non-) predictive nature of political tweets but rather the accurate identification of any sentiment expressed in the tweets. If the accuracy of sentiment analysis of political tweets can be improved (or its limitations at least better understood) then this will likely have a positive effect on its usefulness as an alternative or complement to traditional opinion polling.

3 #ge11: The Irish General Election 2011

The Irish general elections were held on February 25, 2011. 165 representatives were elected across 43 constituencies for the Dáil, the main house of parliament. Eight parties nominated their candidates for election and a coalition (Fine Gael and Labour) government was formed. The parties in the outgoing coalition government, Fianna Fáil and the Greens, suffered disastrous defeats, the worst defeat of a sitting government since the foundatation of the State in 1922.

Gallagher and Marsh (2011, chapter 5) discuss the use of social media by parties, candidates and voters in the 2011 election and conclude that it had a much more important role to play in this election than in the previous one in 2007. On the role of Twitter in particular, they report that *"Twitter was less widespread among candidates [than Facebook], but it offered the most diverse source of citizen coverage during the election, and it has been integrated into several mainstream media"*. They estimated that 7% of the Irish population had a Twitter account at the time of the election.

4 Dataset

We compiled a corpus of tweets using the Twitter search API between 20th and the 25th of January 2011 (one month before the election). We selected the main political entities (the five biggest political parties – Fianna Fáil, Fine Gael, Labour, Sinn Féin and the Greens – and their leaders) and perform query-based search to collect the tweets relating to these entities. The resulting dataset contains 7,916 tweets of which 4,710 are retweets or duplicates, leaving a total of 3,206 tweets.

The tweets were annotated by two Irish annotators with a knowledge of the Irish political landscape. Disagreements between the two annotators were studied and resolved by a third annotator. The annotators were asked to identify the sentiment associated with the topic (or entity) of the tweet. Annotation was performed using the following 6 labels:

- **pos**: Tweets which carry positive sentiment towards the topic
- **neg**: Tweets which carry negative sentiment towards the topic
- **mix**: Tweets which carry both positive and negative sentiment towards the topic
- **neu**: Tweets which do not carry any sentiment towards the topic
- **nen**: Tweets which were written in languages other than English.
- **non**: Tweets which do not have any mention or relation to the topic. These represent search errors.

In addition to the above six classes, annotators were asked to flag whether a tweet was sarcastic.

The dataset which we use for the experiments described in this paper contains only those tweets

Positive Tweets	256	9.75%
Negative Tweets	950	36.22%
Neutral Tweets	1418	54.03%
Total Tweets	2624	

Table 1: Class Distribution

that have been labelled as either positive, negative or neutral, i.e. non-relevant, mixed-sentiment and non-English tweets are discarded. We also simplify our task by omitting those tweets which have been flagged as sarcastic by one or both of the annotators, leaving a set of 2,624 tweets with a class distribution as shown in Table 1.

5 Tools and Resources

In the course of our experiments, we use two different subjectivity lexicons, one part-of-speech tagger and one parser. For part-of-speech tagging we use a tagger (Gimpel et al., 2011) designed specifically for tweets. For parsing, we use the Stanford parser (Klein and Manning, 2003). To identify the sentiment polarity of a word we use:

1. **Subjectivity Lexicon (SL)** (Wilson et al., 2005): This lexicon contains 8,221 words (6,878 unique forms) of which 3,249 are adjectives, 330 are adverbs, 1,325 are verbs, 2,170 are nouns and remaining (1,147) words are marked as *anypos*. There are many words which occur with two or more different part-of-speech tags. We extend SL with 341 domain-specific words to produce an extended SL.

2. **SentiWordNet 3.0 (SWN)** (Baccianella et al., 2010): With over 100+ thousand words, SWN is far larger than SL but is likely to be noisier since it has been built semi-automatically. Each word in the lexicon is associated with both a positive and negative score, and an objective score given by (1), i.e. the positive, negative and objective score sum to 1.

$$ObjScore = 1 - PosScore - NegScore \quad (1)$$

6 Naive Lexicon-based Classification

In this section we describe a naive approach to sentiment classification which does not make use of labelled training data but rather uses the information in a sentiment lexicon to deduce the sentiment orientation towards a political party in a tweet (see Liu (2010) for an overview of this unsupervised lexicon-based approach). In Section 6.1, we present the basic method along with some variants which improve on the basic method by making use of information about part-of-speech, negation and distance from the topic. In Section 6.2, we examine some of the cases which remain misclassified by our best lexicon-based method. In Section 6.3, we discuss briefly those tweets that have been labelled as sarcastic.

6.1 Method and Results

Our baseline lexicon-based approach is as follows: we look up each word in our sentiment lexicon and sum up the scores to corresponding scalars. The results are shown in Table 2. Note that the most likely estimated class prediction is neutral with a probability of .5403 (1418/2624).

6.1.1 Which Subjectivity Lexicon?

The first column shows the results that we obtain when the lexicon we use is our extended version of the SL lexicon. The results in the second column are those that result from using SWN. In the third column, we combine the two lexicons. We define a combination pattern of Extended-SL and SWN in which we prioritize Extended-SL because it is manually checked and some domain-specific words are added. For the words which were missing from Extended-SL (SWN), we assign them the polarity of SWN (Extended-SL). Table 3 explains exactly how the scores from the two lexicons are combined. Although SWN slightly outperforms Extended-SL for the baseline lexicon-based approach (first row of Table 2), it is outperformed by Extended-SL and the combinaton of the two lexicons for all the variants. We can conclude from the full set of results in Table 2 that SWN is less useful than Extended-SL or the combination of SWN and Extended-SL.

6.1.2 Filtering by Part-of-Speech

The results in the first row of Table 2 represent our baseline experiment in which each word in the tweet is looked up in the sentiment lexicon and its sentiment score added to a running total. We achieve a classification accuracy of 52.44% with the

Method	**Extended-SL**		**SWN**		**Combined**	
3-Class Classification (Pos vs Neg vs Neu)	Correct	Accuracy	Correct	Accuracy	Correct	Accuracy
Baseline	1376	52.44%	1379	52.55%	1288	49.09%
Baseline + Adj	1457	55.53%	1449	55.22%	1445	55.07%
Baseline + Adj + S	1480	56.40%	1459	55.60%	1481	56.44%
Baseline + Adj + S + Neg	1495	56.97%	1462	55.72%	1496	57.01%
Baseline + Adj + S + Neg + Phrases	1511	57.58%	1479	56.36%	1509	57.51%
Baseline + Adj + S + Neg + Phrases + Than	1533	58.42%	1502	57.24%	1533	58.42%
Distance Based Scoring: Baseline + Adj + S + Neg + Phrases + Than	1545	58.88%	1506	57.39%	1547	**58.96%**
Sarcastic Tweets	87/344	25.29%	81/344	23.55%	87/344	25.29%

Table 2: 3-class classification using the naive lexicon-based approach. The majority baseline is 54.03%.

Extended-SL polarity	**SWN Polarity**	**Combination Polarity**
-1	-1	-2
-1	0	-1
-1	1	-1
0	-1	-0.5
0	0	0
0	1	0.5
1	-1	1
1	0	1
1	1	2

Table 3: Combination Scheme of extended-SL and SWN. Here 0 represents either a neutral word or a word missing from the lexicon.

Extended-SL lexicon. We speculate that this low accuracy is occurring because too many words that appear in the sentiment lexicon are included in the overall sentiment score without actually contributing to the sentiment towards the topic. To refine our approach one step further, we use part-of-speech information and consider only adjectives for the classification of tweets since adjectives are strong indicators of sentiment (Hatzivassiloglou and Wiebe, 2000). We achieve an accuracy improvement of approximately three absolute points, and this improvement holds true for both sentiment lexicons. This supports our hypothesis that we are using irrelevant information for classification in the baseline system.

Our next improvement (third row of Table 2) comes from mapping all inflected forms to their stems (using the Porter stemmer). Examples of inflected forms that are reduced to their stems are *delighted* or *delightful*. Using stemming with adjectives over the baseline, we achieve an accuracy of 56.40% with Extended-SL.

6.1.3 Negation

"*Negation is a very common linguistic construction that affects polarity and, therefore, needs to be taken into consideration in sentiment analysis*" (Councill et al., 2010). We perform negation handling in tweets using two different approaches. In the first approach, we first identify negation words and reverse the polarity of sentiment-bearing words within a window of three words. In the second approach, we try to resolve the scope of the negation using syntactic parsing. The Stanford dependency scheme (de Marneffe and Manning, 2008) has a special relation (`neg`) to indicate negation. We reverse the sentiment polarity of a word marked via the `neg` relation as being in the scope of a negation. Using the first approach, we see an improvement of 0.6% in the classification accuracy with the Extended-SL lexicon. Using the second approach, we see an improvement of 0.5%. Since there appears to be very little difference between the two approaches to negation-handling and in order to reduce the computational burden of running the Stanford parser each time to obtain the dependencies, we continue further experiments with the first method only. Using baseline + stemming + adjectives + neg we achieve an accuracy of 56.97% with the Extended-SL lexicon.

6.1.4 Domain-specific idioms

In the context of political tweets we see many sentiment-bearing idioms and fixed expressions, e.g. *god save us*, *X for Taoiseach*[1], *wolf in sheep's clothing*, etc. In our study, we had a total of 89 phrases. When we directly account for these phrases, we achieve an accuracy of 57.58% (an absolute improvement of 0.6 points over the last step).

6.1.5 Comparative Expressions

Another form of expressing an opinion towards an entity is by comparing the entity with some other entity. For example consider the tweet:

Fast Food sounds like a better vote than Fianna Fail. (2)

In this tweet, an indirect negative sentiment is expressed towards the political party *Fianna Fáil*. In order to take into account such constructions, the following procedure is applied: we divide the tweet into two parts, left and right. The left part contains the text which comes before the *than* and the right part contains the text which comes after *than*, e.g.
Tweet: 'X is better than Y'
Left: 'X is better'
Right: 'Y'.

We then use the following strategy to calculate the polarity of the tweet oriented towards the entity:

$$\begin{aligned} S\_left \;&=\; sentiment\; score\; of\; Left. \\ S\_right \;&=\; sentiment\; score\; of\; Right. \\ Ent\_pos\_left \;&=\; if\; entity\; is\; left\; of \\ &\quad `than', \; then\; 1,\; otherwise\; -1. \\ Ent\_pos\_right \;&=\; if\; entity\; is\; right\; of \\ &\quad `than', \; then\; 1,\; otherwise\; -1. \\ S(tweet) = Ent\_pos\_left &* S\_left\; + \\ &Ent\_pos\_right * S\_right. \end{aligned} \tag{3}$$

So in (2) above the entity, *Fianna Fáil*, is to the right of *than* meaning that its *Ent_pos_right* value is 1 and its *Ent_pos_left* value is -1. This has the effect of flipping the polarity of the positive word *better*. By including the "than" comparison, we see an improvement of absolute 0.8% (third last row of Table 2).

6.1.6 Distance Scoring

To emphasize the topic-oriented nature of our sentiment classification, we also define a distance-based scoring function where we define the overall score of the tweet as given in (4). Here $dis(word)$ is defined as number of words between the topic (i.e. the political entity) and the sentiment word.

$$S(tweet) = \sum_{i=1}^{n} S(word_i)/dis(word_i). \tag{4}$$

The addition of the distance information further enhanced our system accuracy by 0.45%, taking it to 58.88% (second last row of Table 2). Our highest overall accuracy (58.96) is achieved in this setting using the combined lexicon.

It should be noted that this lexicon-based approach is overfitting to our dataset since the list of domain-specific phrases and the form of the comparative constructions have been obtained from the dataset itself. This means that we are making a strong assumption about the representativeness of this dataset and accuracy on a held-out test set is likely to be lower.

6.2 Error Analysis

In this section we discuss pitfalls of the naive lexicon-based approach with the help of some examples (see Table 4). Consider the first example from

[1] The term *Taoiseach* refers to the Irish Prime Minister.

the table, *@username and u believe people in fianna fail . What are you a numbskull or a journalist ?* In this tweet, we see that negative sentiment is imparted by the question part of the tweet, but actually there are no sentiment adjectives. The word *numbskull* is contributing to the sentiment but is tagged as a noun and not as an adjective. This tweet is tagged as negative by our annotators and as neutral by our lexicon-based classifier.

Consider the second example from Table 4, *@username LOL . A guy called to our house tonight selling GAA tickets . His first words were : I'm not from Fianna Fail .* This is misclassified because there are no sentiment bearing words according to the sentiment lexicon. The last tweet in the table represents another example of the same problem. Note however that the emoticon :/ in the last tweet and the web acronym LOL in the second tweet are providing hints which our system is not making use of.

In the third example from Table 4, *@username Such scary words .' Sinn Fein could top the poll ' in certain constituencies . I feel sick at the thought of it . '* In this example, we have three sentiment bearing words: *scary*, *top* and *sick*. Two of the three words are negative and one word is positive. The word *scary* is stemmed incorrectly as *scari* which means that it is out of the scope of our lexicons. If we just count the number of sentiment words remaining, then this tweet is labelled as neutral but actually is negative with respect to the party *Sinn Féin*. We proposed the use of distance as a measure of relatedness to the topic and we observed a minor improvement in classification accuracy. However, for this example, the distance-based approach does not work. The word *top* is just two words away from the topic and thus contributes the maximum, resulting in the whole tweet being misclassified as positive.

6.3 Sarcastic Tweets

"*Political discouse is plagued with humor, double entendres, and sarcasm; this makes determining political preference of users hard and inferring voting intention even harder.*"(Gayo-Avello, 2012)

As part of the annotation process, annotators were asked to indicate whether they thought a tweet exhibited sarcasm. Some examples of tweets that were annotated as sarcastic are shown in Table 5.

We made the decision to omit these tweets from the main sentiment classification experiments under the assumption that they constituted a special case which would be better handled by a different classifier. This decision is vindicated by the results in the last row of Table 2 which show what happens when we apply our best classifier (*Distance-based Scoring: Baseline+Adj+S+Neg+Phrases+Than*) to the sarcastic tweets – only a quarter of them are correctly classified. Even with a very large and highly domain-tuned lexicon, the lexicon-based approach on its own will struggle to be of use for cases such as these, but the situation might be improved were the lexicon to be used in conjunction with possible sarcasm indicators such as exclamation marks.

7 Supervised Machine Learning

Although our dataset is small, we investigate whether we can improve over the lexicon-based approach by using supervised machine learning. As our learning algorithm, we employ support vector machines in a 5-fold cross validation setup. The tool we use is SVMLight (Joachims, 1999).

We explore two sets of features. The first are the tried-and-tested unigram presence features which have been used extensively not only in sentiment analysis but in other text classification tasks. As we have only 2,624 training samples, we performed feature selection by ranking the features using the Chi-squared metric.

The second feature set consists of 25 features which are inspired by the work on lexicon-based classification described in the previous section. These are the counts of positive, negative, objective words according to each of the three lexicons and the corresponding sentiment scores for the overall tweets. In total there are 19 such features. We also employ six Twitter-related presence features: positive emoticons, negative emoticons, URLs, positive hashtags, negative hashtags and neutral hashtags. For further reference we call this second set of features our "hand-crafted" features.

The results are shown in Table 6. We can see that using the hand-crafted features alone barely improves over the majority baseline of 54.03 but it does improve over our baseline lexicon-based approach (see first row of Table 2). Encouragingly, we see some benefit from using these features in conjunc-

Tweet	Topic	Manual Polarity	Calculated Polarity	Reason for misclassification
@username and u believe people in fianna fail . What are you a numbskull or a journalist ?	Fianna Fáil	neg	neu	Focus only on adjectives
@username LOL . A guy called to our house tonight selling GAA tickets . His first words were : I'm not from Fianna Fail .	Fianna Fáil	neg	neu	No sentiment words
@username Such scary words .' Sinn Fein could top the poll ' in certain constituencies . I feel sick at the thought of it .	Sinn Féin	neg	pos	Stemming and word distance order
@username more RTE censorship . Why are they so afraid to let Sinn Fein put their position across . Certainly couldn't be worse than ff	Sinn Féin	pos	neg	contribution of *afraid*
Based on this programme the winners will be Sinn Fein & Gilmore for not being there #rtefl	Sinn Féin	pos	neu	Focus only on adjectives
#thefrontline pearce Doherty is a spoofer ! Vote sinn fein and we loose more jobs	Sinn Féin	neg	pos	Focus only on adjectives & contribution of phrase *Vote X*
@username Tread carefully Conor . BNP endorsing Sinn Fin etc . etc .	Sinn Féin	neg	neu	No sentiment words
@username ah dude . You made me go to the fine gael web site ! :/	Fine Gael	neg	neu	No sentiment words

Table 4: Misclassification Examples

Feature Set	# Features	Accuracy
# samples = 2624		SVM Light
Hand-crafted	25	54.76
Unigram	7418	55.22
	Top 1000	58.92
	Top 100	56.86
Unigram + Hand-crafted	7444	54.73
	Top 1000	**61.62**
	Top 100	59.53

Table 6: Results of 3-Class Classification using Supervised Machine Learning

tion with the unigram features. Our best overall result of 61.62% is achieved by using the Top 1000 unigram features together with these hand-crafted features. This result seems to suggest that, even with only a few thousand training instances, employing supervised machine learning is still worthwhile.

8 Conclusion

We have introduced a new dataset of political tweets which will be made available for use by other researchers. Each tweet in this set has been annotated for sentiment towards a political entity, as well as for the presence of sarcasm. Omitting the sarcastic tweets from our experiments, we show that we can classify a tweet as being positive, negative or neutral towards a particular political party or party leader with an accuracy of almost 59% using a simple approach based on lexicon lookup. This improves over the majority baseline by almost 5 absolute percentage points but as the classifier uses information from the test set itself, the result is likely to be lower on a held-out test set. The accuracy increases slightly when the lexicon-based information is encoded as features and employed together with bag-of-word features in a supervised machine learning setup.

Future work involves carrying out further exper-

Sarcastic Tweets
Ah bless Brian Cowen's little cotton socks! He's staying on as leader of FF because its better for the country. How selfless!
So now Brian Cowen is now Minister for foreign affairs and Taoiseach? Thats exactly what he needs more responsibilities http://bbc.in/hJI0hb
Mary Harney is going. Surprise surprise! Brian Cowen is going to be extremely busy with all these portfolios to administer. Super hero!
Now in its darkest hour Fianna Fail needs. . . Ivor!
Labour and Fine Gael have brought the election forward by 16 days Crisis over Ireland is SAVED!! #vinb
@username Maybe one of those nice Sinn Fein issue boiler suits? #rtefl
I WILL vote for Fine Gael if they pledge to dress James O'Reilly as a leprechaun and send him to the White House for Paddy's Day.

Table 5: Examples of tweets which have been flagged as sarcastic

iments on those tweets that have been annotated as sarcastic, exploring the use of syntactic dependency paths in the computation of distance between a word and the topic, examining the role of training set class bias on the supervised machine learning results and exploring the use of distant supervision to obtain more training data for this domain.

Acknowledgements

Thanks to Emmet O Briain, Lesley Ni Bhriain and the anonymous reviewers for their helpful comments. This research has been supported by Enterprise Ireland (CFTD/2007/229) and by Science Foundation Ireland (Grant 07/CE/ I1142) as part of the CNGL (www.cngl.ie) at the School of Computing, DCU.

References

Stefano Baccianella, Andrea Esuli, and Fabrizio Sebastiani. 2010. Sentiwordnet 3.0: An enhanced lexical resource for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Akshat Bakliwal, Piyush Arora, Senthil Madhappan, Nikhil Kapre, Mukesh Singh, and Vasudeva Varma. 2012. Mining sentiments from tweets. In *Proceedings of the WASSA'12 in conjunction with ACL'12*.

Adam Bermingham and Alan F. Smeaton. 2010. Classifying sentiment in microblogs: is brevity an advantage? In *Proceedings of the 19th ACM international conference on Information and Knowledge Management*.

Adam Bermingham and Alan Smeaton. 2011. On using Twitter to monitor political sentiment and predict election results. In *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*.

Nibir Nayan Bora. 2012. Summarizing public opinions in tweets. In *Journal Proceedings of CICLing 2012*.

Paula Chesley. 2006. Using verbs and adjectives to automatically classify blog sentiment. In *Proceedings of AAAI-CAAW-06, the Spring Symposia on Computational Approaches*.

Murphy Choy, Michelle L. F. Cheong, Ma Nang Laik, and Koo Ping Shung. 2011. A sentiment analysis of Singapore Presidential Election 2011 using Twitter data with census correction. *CoRR*, abs/1108.5520.

Isaac G. Councill, Ryan McDonald, and Leonid Velikovich. 2010. What's great and what's not: learning to classify the scope of negation for improved sentiment analysis. In *Proceedings of the Workshop on Negation and Speculation in Natural Language Processing*, NeSp-NLP '10.

Dmitry Davidov, Oren Tsur, and Ari Rappoport. 2010. Enhanced sentiment learning using Twitter hashtags and smileys. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*.

Marie-Catherine de Marneffe and Christopher D. Manning. 2008. The stanford typed dependencies representation. In *Proceedings of the COLING Workshop on Cross-Framework and Cross-Domain Parser Evaluation*.

Michael Gallagher and Michael Marsh. 2011. *How Ireland Voted 2011: The Full Story of Ireland's Earthquake Election*. Palgrave Macmillan.

Daniel Gayo-Avello. 2012. "I wanted to predict elections with Twitter and all I got was this lousy paper".

A balanced survey on election prediction using Twitter data. *CoRR*, abs/1204.6441.

Kevin Gimpel, Nathan Schneider, Brendan O'Connor, Dipanjan Das, Daniel Mills, Jacob Eisenstein, Michael Heilman, Dani Yogatama, Jeffrey Flanigan, and Noah A. Smith. 2011. Part-of-speech tagging for Twitter: annotation, features, and experiments. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*.

Alec Go, Richa Bhayani, and Lei Huang. 2009. Twitter sentiment classification using distant supervision. In *CS224N Project Report, Stanford University*.

Namrata Godbole, Manjunath Srinivasaiah, and Steven Skiena. 2007. Large-scale sentiment analysis for news and blogs. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Vasileios Hatzivassiloglou and Janyce M. Wiebe. 2000. Effects of adjective orientation and gradability on sentence subjectivity. In *Proceedings of COLING*.

Thorsten Joachims. 1999. Advances in kernel methods. chapter Making large-scale support vector machine learning practical, pages 169–184. MIT Press, Cambridge, MA, USA.

Dan Klein and Christopher D. Manning. 2003. Accurate unlexicalized parsing. In *Proceedings of the 41st Annual Meeting on Association for Computational Linguistics - Volume 1*.

Bing Liu. 2010. Handbook of natural language processing. chapter Sentiment Analysis and Subjectivity. Chapman and Hall.

Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. 2010. From tweets to polls: Linking text sentiment to public opinion time series. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM)*.

Alexander Pak and Patrick Paroubek. 2010. Twitter as a corpus for sentiment analysis and opinion mining. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC'10)*.

Bo Pang, Lillian Lee, and Shivakumar Vaithyanathan. 2002. Thumbs up?: sentiment classification using machine learning techniques. In *Proceedings of the conference on Empirical Methods in Natural Language Processing - Volume 10*.

James W. Pennebaker, Cindy K. Chung, Molly Ireland, Amy Gonzales, and Roger J. Booth. 2007. The development and psychometric properties of liwc2007. Technical report, Austin,Texas.

Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the International Conference on Weblogs and Social Media*.

Peter D. Turney. 2002. 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*.

Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. 2012. A system for real-time Twitter sentiment analysis of 2012 U.S. presidential election cycle. In *ACL (System Demonstrations)*.

Theresa Wilson, Janyce Wiebe, and Paul Hoffmann. 2005. Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*.