On Using Twitter to Monitor Political Sentiment and Predict Election Results

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

Abstract

The body of content available on Twitter undoubtedly contains a diverse range of political insight and commentary. But, to what extent is this representative of an electorate? Can we model political sentiment effectively enough to capture the voting intentions of a nation during an election capaign? We use the recent Irish General Election as a case study for investigating the potential to model political sentiment through mining of social media. Our approach combines sentiment analysis using supervised learning and volume-based measures. We evaluate against the conventional election polls and the final election result. We find that social analytics using both volume-based measures and sentiment analysis are predictive and we make a number of observations related to the task of monitoring public sentiment during an election campaign, including examining a variety of sample sizes, time periods as well as methods for qualitatively exploring the underlying content.

1 Introduction

For years, standard methodologies such as polls have been used by market researchers to measure the beliefs and intentions of populations of individuals. These have a number of disadvantages including the human effort involved and they can be costly and time-consuming. With the advent of social media, there is now an abundance of online information wherein people express their sentiment with respect to wide variety of topics. An open research question is how might we analyse this data to produce results that approximate what can be achieved through traditional market research. An automated solution would mean that we could "poll" a population on demand, and at low cost.

This is a challenging task however. How can we ensure that our sample has a representative distribution? How much confidence do we put in noisy signals such as sentiment analysis? The wisdom of crowds teaches us that sufficient scale should at least somewhat mitigate these problems. In this paper we review a live system we developed for the Irish General Election, 2011. Our system used a variety of techniques to provide a live real-time interface into Twitter during the election. Using the volume and sentiment data from this system we review a number of sampling approaches and methods of modelling political sentiment, replicating work of others as well as introducing novel measures. We evaluate the error with respect to polls, as well as with respect to the election result itself.

In the next section we review related research. This is followed in Section 3 by a description of our methodology. We present our results in Section 4, followed by discussion in Section 5, and we conclude in Section 6.

2 Related Work

There appears to be three research areas emerging in terms of using online sentiment to monitor real world political sentiment. First is event monitoring, where the aim is to monitor reactionary content in social media during a specified event. In the political area this would typically be a speech, or a TV debate. An example is work by Diakopoulos and Shamma who characterised the 2008 US presidential debate in terms of Twitter sentiment (Diakopoulos and Shamma, 2010). Previously Shamma et al. examined a variety of aspects of debate modelling using Twitter, beyond individual politician performance (Shamma et al., 2009). In these studies, Twitter proved to be an effective source of data for identifying important topics and associated public reaction.

A second area which has received attention is modelling continuous sentiment functions for predicting other real-world continuous values, for example to predict stock market values. Bollen et al. have focused on modeling public mood on a variety of axes to correlate with socio-economic factors (Bollen et al., 2009) and to predict the Dow Jones Industrial Average (Bollen et al., 2010). They report a number of interesting observations such as changes in tension and anxiety around important events and find a significant improvement in predicting the Dow Jones Industrial Average when incorporating sentiment. This work is echoed by preliminary work from Zhang et al. who also focus on emotive concepts, in this case “hope” and “fear”, and correlate with a number of market indicators (Zhang et al., 2010). It is noteworthy that the emphasis in these studies is on emotive sentiment (mood states, emotions), rather than polar sentiment (positivity, negativity) which is popular in other applications. O’Connor et al. also observe leading signals in Twitter sentiment, but with respect to political opinion polls (O’Connor et al., 2010). They offer the caveat, “text sentiment is volatile ... it is best used to detect long-term trends”.

A third, related area, is result forecasting. A classic example of this is predicting election results, the focus of this paper. In result forecasting, it is the final result which is used to judge the accuracy of a particular forecasting measure, rather than a continuous series. Asur and Huberman (Asur and Huberman, 2010) used Twitter volume and sentiment to predict box office takings for movies, bettering other market indicators. They find volume to be a strong predictor and sentiment to be a useful, yet weaker predictor. They also propose a general model for linear regression social media prediction which serves as a basis for our model.

More directly, related to elections is Tumasjan et al.’s work on the German federal election in 2009 (Tumasjan et al., 2010). They found that that the share of volume on Twitter accurately reflected the distribution of votes in the election between the six main parties. It is difficult to draw general conclusions from this single result however. A focus of our study is to replicate and extend these experiments with respect to the Irish General Election. Noteworthy also is an earlier study which mined content from a political prediction website and in identifying author-party valence, trained classifiers with lexical features to identify “predictive sentiment” with promising results for predicting Canadian district elections (Kim and Hovy, 2007).

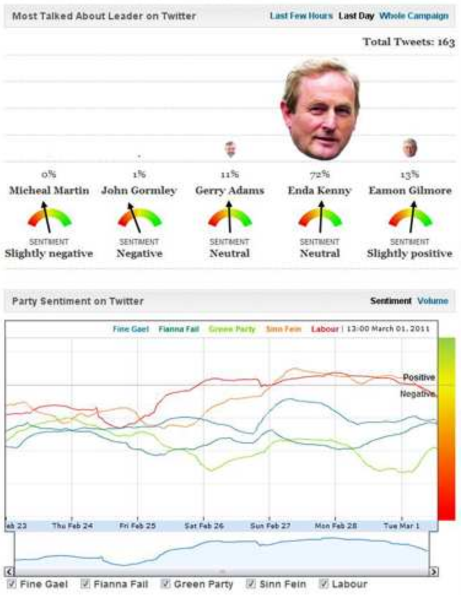


Figure 1: A screen shot of the sentiment portion of the #GE11 Real-time Twitter Tracker

The concerns around using Twitter as the basis for a prediction mechanism have been voiced by Gayo-Avello et al. who state, “we argue that one should not be accepting predictions about events using social media data as a black box.” (Gayo-Avello et al., 2011) They cite the two primary caveats with using social media to inform predictive models as selection bias (inability to determine a representative sample) and potential for deliberate influence of results (through gaming and spamming for example). This is echoed by (Jungherr et al., 2011) who argue that methods of prediction using social media analytics are frequently contingent on somewhat arbitrary experimental variables.

Thus we see that predictive systems which utilise social media are both promising and challenging. The contention of our research is that the development of techniques for political public sentiment monitoring and election prediction is a promising direction requires more research work before we fully understand the limitations and capabilities of such an approach.

3 Methodology

The system we developed to evaluate our research idea was completed in collaboration with an industrial partner, an online news company[1]. The purpose of the "#GE11 Twitter Tracker" was to allow users, and our partner's journalists, to tap into the content on Twitter pertaining to the election, through an accessible dashboard-style interface. To that end, the "Twitter Tracker" featured a number of abstractive and extractive summarization approaches as well as a visualisation of volume and sentiment over time (see Figure 1).

The Irish General Election took place on 25th February, 2011. Between the 8th of February and the 25th we collected 32,578 tweets relevant to the five main parties: Fianna Fáil (FF), the Green Party, Labour, Fine Gael (FG) and Sinn Féin (SF). We identified relevant tweets by searching for the party names and their abbreviations, along with the election hashtag, `#ge11`. For the purposes of the analysis presented here, we do not consider the independent candidates or the minority parties[2]. Tweets reporting poll results were also filtered out.

3.1 Election Polls

The standard measure of error in predictive forecasting is Mean Absolute Error (MAE), defined as the average of the errors in each forecast:

$$MAE = \frac{1}{n}\sum_{i=1}^{n} |e_i| \qquad (1)$$

where n is the number of forecasts (in our case 5) and e_i is the difference in actual result and predicted result for the i^{th} forecast. MAE measures the degree to which a set of predicted values deviate from the actual values. We use MAE to compare Twitter-based predictions with polls as well as with the results of the election. To provide a reference point for our analysis, we use nine polls which were commissioned during the election. These polls guarantee accuracy to within a margin of 3% and in comparison to the final election results, had an average MAE of 1.61% with respect to the five main parties. There have been varying reports for Twitter-based predictions in the literature where the observed error can vary from very low (1.65%) (Tumasjan et al., 2010) to much higher (17.1% using volume, 7.6% using sentiment) (Gayo-Avello et al., 2011).

3.2 Predictive Measures

It is reasonable to assume that the percentage of votes that a party receives is related to the volume of related content in social media. Larger parties will have more members, more candidates and will attract more attention during the election campaign. Smaller parties likewise will have a much smaller presence. However, is this enough to reflect a popularity at a particular point in time, or in a given campaign? Is measuring volume susceptible to disproportionate influence from say a few prominent news stories or deliberate gaming or spamming? We define our volume-based measure as the proportional share of party mentions in a set of tweets for a given time period:

$$SoV(x) = \frac{|Rel(x)|}{\sum_{i=1}^{n} |Rel(i)|} \qquad (2)$$

where $SoV(x)$ is the share of volume for a given party x in a system of n parties and $|Rel(i)|$ is the number of tweets relevant to party i. This formula has the advantage that the score for the parties are proportions summing to 1 and are easily compared with poll percentages. The sets of documents we use are:

- *Time-based*: Most recent 24 hours, 3 days, 7 days
- *Sample size-based*: Most recent 1000, 2000, 5000 or 10000 tweets
- *Cumulative*: All of the tweets from 8th February 2011 to relevant time
- *Manual*: Manually labelled tweets from pre-8th February 2011

When we draw comparison with a poll from a given date, we assume that tweets up until midnight the night before the date of the poll may be used. The volume of party mentions was approximately consistent in the approach to the election, meaning the *cumulative* volume function over time is linear and monotonically increasing.

3.3 Sentiment Analysis

Our previous research has shown that supervised learning provides more accurate sentiment analysis than can be provided by unsupervised methods such as using sentiment lexicons (Bermingham and Smeaton, 2010). We therefore decided to use classifiers specifically trained on data for this

[1] `http://www.thejournal.ie`

[2] There is a difficulty with the minority parties and independent candidates for this election in that many of the official parties were more commonly referred to by their party alliance. This made relevance difficult to determine and such an exercise is outside the scope of this work

	Positive	Negative	Neutral	Mixed	Total
Week 1	255	1,248	1,218	47	2,768
Week 2	629	1,289	2,411	106	4,435
Total	884	2,537	3,629	153	7,203

Table 1: Annotation counts

election. On two days, a week apart before the 8th of February 2011, we trained nine annotators to annotate sentiment in tweets related to parties and candidates for the election. The tweets in each annotation session were taken from different time periods in order to develop as diverse a training corpus as possible.

We provided the annotators with detailed guidelines and examples of sentiment. Prior to commencing anntoation, annotators answered a short set of sample annotations. We then provided the gold standard for these annotations (determined by the authors) and each answer was discussed in a group session. We instructed annotators not to consider reporting of positive or negative fact as sentiment but that sentiment be one of emotion, opinion, evaluation or speculation towards the target topic. Our annotation categories consisted of three sentiment classes (positive, negative, mixed), one non-sentiment class (neutral) and the 3 other classes (unannotatable, non-relevant, unclear). This is in line with the definition of sentiment proposed in (Wilson et al., 2005).

We disregard unannotatable, non-relevant and unclear annotations. A small subset (3.5%) of the documents were doubly-annotated. The inter-annotator agreement for the four relevant classes is 0.478 according to Krippendorff's Alpha, a standard measure of inter-annotator agreement for many annotators (Hayes and Krippendorff, 2007). We then remove duplicate and contradictory annotations leaving 7,203 document-topic pairs (see Table 1). Approximately half of the annotations contained sentiment of some kind.

The low level of positive sentiment we observe is striking, representing just 12% of the document-topic pairs. During this election, Ireland was in a period of economic crisis and negative political sentiment dominated the media and public mood. This presents a difficulty for supervised learning. With few training examples, it is difficult for the learner to identify minority classes. To mitigate this effect, when choosing our machine learning algorithm we optimise for F-measure which balances precision and recall across the classes. We disregard the mixed annotations as they are few in number and ambiguous in nature.

		Recall			
classifier	accuracy	pos	neg	neu	F-score
trivial	50.19	0	0	1	0.335
MNB	62.94	0.007	0.561	0.832	0.584
ADA-MNB	65.09	0.334	0.689	0.7	0.645
SVM	64.82	0.201	0.634	0.768	0.631
ADA-SVM	64.28	0.362	0.623	0.726	0.638

Table 2: Accuracy for 3-class sentiment classification

Our feature vector consists of unigrams which occur in two or more documents in the training set. The tokenizer we use (Laboreiro et al., 2010) is optimised for user-generated content so all sociolinguistic features such as emoticons (":-)") and unconventional punctuation ("!!!!") are preserved. These features are often used to add tone to text and thus likely to contain sentiment information. We remove all topic terms, usernames and URLs to prevent any bias being learned towards these.

Unsatisfied with the recall from either Support Vector Machines (SVM) or Multinomial Naive Bayes (MNB) classifiers, we evaluated a boosting approach which, through iterative learning, up-weights training examples from minority classes, thus improving recall for these classes. We used Freund and Schapire's Adaboost M1 method with 10 training iterations as implemented in the Weka toolkit[3] (Freund and Schapire, 1996). Following from this, we use an Adaboost MNB classifier which achieves 65.09% classification accuracy in 10-fold cross-validation for 3 classes (see Table 2).

3.4 Incorporating Sentiment

It is difficult to say how best to incorporate sentiment. On the one hand, sentiment distribution in the tweets relevant to a single party is indicative of the sentiment towards that party. For example, if the majority of the mentions of a party contain negative sentiment, it is reasonable to assume that people are in general negatively disposed towards that party. However, this only considers a party in isolation. If this negative majority holds true for *all* parties, how do we differentiate public opinion towards them? In a closed system like an election, relative sentiment between the parties perhaps has as much of an influence.

To address the above issues, we use two novel measures of sentiment in this study. For inter-

[3] http://www.cs.waikato.ac.nz/ml/weka/

party sentiment, we modify our volume-based measure, SoV, to represent the share of positive volume, SoV_p, and share of negative volume, SoV_n:

$$SoV_p(x) = \frac{|Pos(x)|}{\sum_{i=1}^{n} |Pos(i)|} \tag{3}$$

$$SoV_n(x) = \frac{|Neg(x)|}{\sum_{i=1}^{n} |Neg(i)|} \tag{4}$$

For intra-party sentiment, we use a log ratio of sentiment:

$$Sent(x) = \log_{10} \frac{|Pos(x)| + 1}{|Neg(x)| + 1} \tag{5}$$

This gives a single value for representing how positive or negative a set of documents are for a given topic. Values for $Sent(x)$ are positive when there are more positive than negative documents, and negative when there are more negative than positive for a given party. 1 is added to the positive and negative volumes to prevent a division by zero. The inter-party share of sentiment is a proportional distribution and thus prediction error can be easily measured with MAE. Also, as it is non-parametric it can be applied without any tuning.

We fit a regression to our inter-party and intra-party measures, trained on poll data. This takes the form:

$$\begin{aligned} y(x) = \quad & \beta_v SoV(x) + \beta_p SoV_p(x) + \beta_n SoV_n(x) \\ & + \beta_s Sent(x) + \varepsilon \end{aligned}$$

This builds on the general model for sentiment proposed in (Asur and Huberman, 2010). The purpose of fitting this regression is threefold. Firstly, we wish to identify which measures are the most predictive and confirm our assumption that both sentiment and proportion of volume have predictive qualities. Secondly, we want to compare the predictive capabilities of our two sentiment measures. Lastly, we want to identify under optimum conditions how a Twitter-model for political sentiment could predict our election results.

For many applications there is little to be gained from measuring sentiment without being able to explain the observed values. We conclude our study with a suggestion for how such sentiment data may be used to explore Twitter data *qualitatively* during an election.

4 Results

Comparing our non-parametric inter-topic measures with the election result, our lowest error comes for when we use all available data, with volume performing marginally better (5.58%) than the share of positive volume. Interestingly, in many of the other data sets, the share of positive volume outperforms the share of volume in terms of result prediction. Unsurprisingly, share of negative volume performs worst in all cases. Also interesting is the fact that among the worst-performing is the more accurate manually labelled data. Perhaps this is due to the gap in time between when the documents were labelled and the election. See Table 3 for the MAE for result prediction for our data sets.

	MAE		
Dataset	SoV	SoV_p	SoV_n
cumulative	0.0558	0.0576	0.0658
1 day	0.0841	0.0574	0.1248
3 days	0.0920	0.0805	0.1203
7 days	0.0790	0.0718	0.0982
last 1000	0.0805	0.0857	0.1088
last 2000	0.0795	0.0663	0.1335
last 5000	0.0723	0.0701	0.1066
last 10000	0.0926	0.0808	0.1206
manually labelled	0.0968	0.1037	0.1128

Table 3: Mean absolute error for non-parametric measures compared to election result

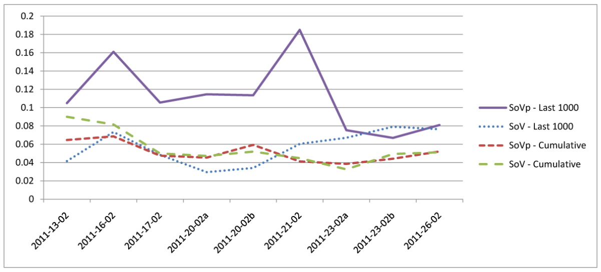


Figure 2: Mean absolute error for *cumulative* and *last 1000* sample data for SoV and SoV_p

To understand better how each of these predictive measures is performing, we look closer at two of our datasets: *cumulative* and *last 1000*. We choose *cumulative* as it performs best out of all our datasets and we choose *last 1000* as this sample size is easy to reproduce and a number frequently used in polling for sufficient sample size. This also allows us to compare a cumulative data set with a fixed volume dataset.

In Figure 2 we can see that broadly the error for the cumulative datasets improves compared to each successive poll over time. The performance of the positive share of volume and overall volume are strongly positively correlated. For the most recent 1000 document samples however, we see

Features	MAE cumulative	MAE last 1000
s	0.0996	0.1029
n	0.071	0.0661
n,s	0.0448	0.0645
p	0.0471	0.066
p,s	0.04	0.064
p,n	0.04	0.0594
p,n,s	0.0388	0.0608
v	0.0551	0.0573
v,s	0.0403	0.0547
v,n	0.0434	0.0533
v,n,s	0.0377	0.0502
v,p	0.0466	0.0538
v,p,s	0.0399	0.0542
v,p,n	0.0383	0.0486
v,p,n,s	0.0367	0.0486

Table 4: Error for regressions, trained and tested on poll data $v = SoV, p = SoV_p, n = SoV_n, s = Sent$

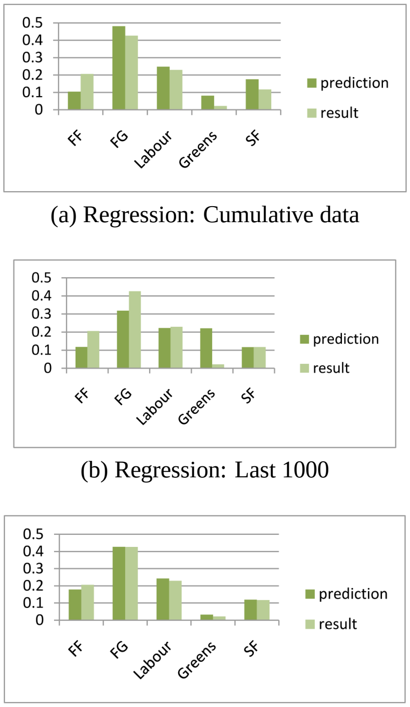
	MAE
Regression (cumulative)	0.0585
Regression (last 1000)	0.0804
Exit poll	0.0108

Table 6: Error for regressions, trained on poll data and official exit poll, compared to election results

the error for share of positive volume vary wildly, likely due to the low volume of tweets classified as positive. This does appear to lessen as the election draws nearer, eventually reaching the same level as the overall share of volume in the recent 1000 documents. After the initial polls however, the cumulative scores give a much lower error.

Using intra-party sentiment in Figure 3 we see that in the weeks before the election, it is difficult to discern any salient pattern. The party sentiment values all seem to be relatively close, with an average sentiment score of 0.75, approximately equal to a ratio of 1 positive document for every 6 negative documents. In the days before polling day however we observe a divergence of sentiment which continues through polling day and beyond, showing overall positive sentiment for Labour and Sinn Féin, both of whom won a record number of seats in the parliament. This trend continues for a few days after the election but by a week later has returned to values similar to those observed earlier in the campaign.

Looking at the results of the regressions which were fitted to the poll results, we see a low error, particularly for the cumulative data which has an MAE of 3.67%. In Table 5 we can see how the regression has weighted each of the factors.



(a) Regression: Cumulative data

(b) Regression: Last 1000

(c) Exit poll

Figure 4: Exit poll, election results and election predictions for regression trained on poll data using all features

For both datasets, the regression has placed a high weight on share of volume. Intuitively, the share of positive volume receives a positive weight and the share of negative volume receives a negative weight. Each of the sentiment scores are weighted higher for the cumulative data. In Table 4, we can see that adding in more features improves the regression accuracy but taking just two features (for example SoV_p and SoV_n) we can approach similar accuracy. In terms of the final election results, the cumulative regression outperforms the 1000 sample regression significantly with an MAE of 5.85% (see Table 6 and Figure 4).

In order to explore the content according to sentiment we define *Sentiment TF-IDF*. In this measure, we consider the entire set of documents to be the tweets relevant to a topic and thus the document frequency for a term is the number of relevant documents in which a term appears. To calculate the term frequencies for a topic-sentiment class we then concatenate all documents of that class into a single document and calculate word frequencies. Doing this for the positive and negative classes for each party provides us with the ranked terms list in Table 7. These terms may be thought of as those terms that most characterise

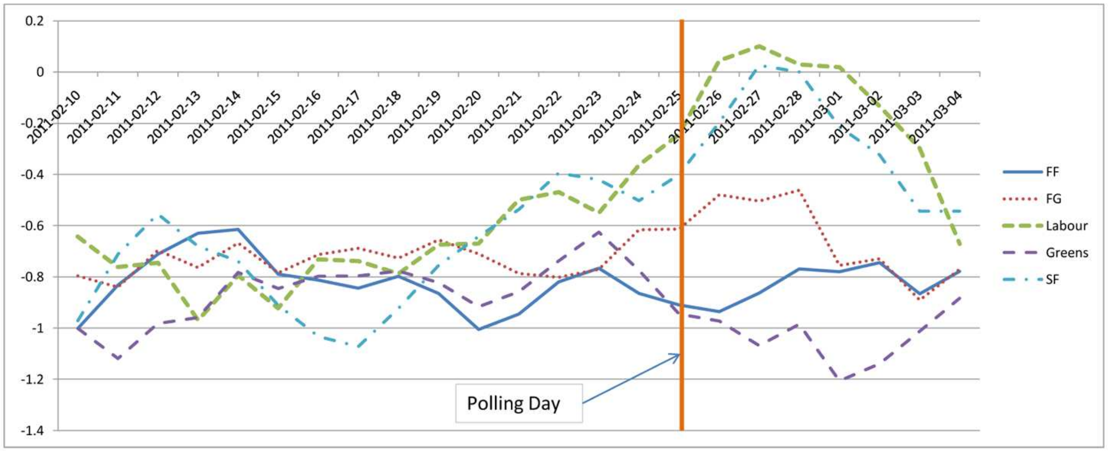


Figure 3: Daily sentiment: each data point is average of the daily $Sent$ score for a party over the previous three days

	SoV	SoV_p	SoV_n	$Sent$	ε	MAE	Correlation Coefficient
Cumulative	1.3444	0.6516	-1.0019	0.2193	0.1801	0.0367	0.9524
Last 1000	1.3339	0.2125	-0.6708	-0.0075	0.0196	0.0486	0.896

Table 5: Linear regression coefficients, error and correlation coefficient for regression fitted to poll data

the sentiment-bearing documents for that party.

5 Discussion

Overall, the best non-parametric method for predicting the result of the first preference votes in the election is the share of volume of tweets that a given party received in total over the time period we study. This is followed closely by the share of positive volume for the same time period which, despite considering only a fraction of the documents considered by share of volume, approaches the same error. Either overall share of volume or share of positive volume performs best for each dataset. As expected, negative share of voice consistently performs worst, though in some cases rivals the other measures. This is likely due to a correlation with the overall share of volume.

The error compared with the individual polls is telling as we see a downward trend for the cumulative data as more data is available. This pattern does not appear in the *last 1000* sample volume data so this is likely linked to quantity of data rather than temporal proximity to election day. The share of positive volume for the *last 1000* sample is much more erratic than we observe for the cumulative data suggesting that 1000 is perhaps too small to rely on metrics derived from subsets of the data.

Perhaps the most intriguing results is the sentiment pattern over the course of the election. In Figure 3 we see that there is a dramatic change in sentiment towards the parties for the days after polling day but that this sentiment shift had already begun before polling day. This period, from a few days before the election to approximately a week afterwards, is a period where public sentiment appears to have settled at a range of values for the parties. Outside of this time period it is difficult to separate the parties in terms of sentiment. Perhaps this is a case of Twitter users being more honest and considered with the vote and results imminent, rather than simply reactionary. The fact that this sentiment appears to be leading makes for an interesting avenue to pursue in future studies.

We achieved an MAE of 3.67% using our regressions compared to the poll results, although naturally this was overfitted, since the regressions had originally been fitted to the poll data. For that reason the error is much higher when we test with the actual result at 5.85%. It is noteworthy that this is in fact slightly worse that the best performing non-parametric measure. In both cases, the error is significantly higher than that achieved by the tradition polls.

Both the intra-party and inter-party sentiment measures appear to improve upon volume-based measures and the weights the regression assigns to them reflects this. However it is difficult to conclude that intra-party sentiment is important when inter-party sentiment is considered. In a closed system, the actual distribution of sentiment in content relevant to a given party may only matter relative to that for the other parties. Considering the regression results, it seems that capturing the share of positive volume and the share of negative volume is sufficient, particularly where a large amount of data is available. With all features for cumulative data, the coefficient for intra-party sentiment score is assigned a weight of just -0.0075 suggesting that this factor is effectively ignored by the regression.

Examining the errors, we see that our methods have particular trouble forecasting the result for the Green Party (too high) and Fianna Fáil (too low). In the former case, we suspect this is due to the selection bias in sampling Twitter. Green party members, and their supporters, tend to be more tech-savvy and have a disproportionately large presence in social media. In the latter case we speculate that although Fianna Fáil attracted low volume and plenty of negativity, they are however traditionally the largest Irish party and thus enjoyed a degree of brand loyalty.

In opinion measurement and social media analytics it is limiting to simply measure without providing means to explain measurements. Using *Sentiment TF-IDF* we can identify terms that provide a path to qualitatively exploring the dataset. We suggest using *Sentiment TF-IDF* to identify terms which can be used to identify important, sentiment-bearing documents. Doing this we were able to use the words in Table 7 to determine that people were discussing Fine Gael negatively with respect to *planting* a member of the *audience* in a popular current affairs television show. We also saw a negative reaction to the Green Party's proposal for a *citizens' assembly*. This shows that there may be further value in terms of qualitative analysis which Twitter may offer during an election.

6 Conclusion

Overall, we conclude that Twitter does appear to display a predictive quality which is marginally augmented by the inclusion of sentiment analysis. We derive two different methods for monitoring topic sentiment, intra-party and inter-party. Fitting our features to a regression we observe that volume is the single biggest predictive variable followed by inter-party sentiment. Given sufficient data, intra-party sentiment appears to be less valuable as a predictive measure. Our speculation is that the relative success of the inter-party sentiment is due to the closed nature of the system.

Our approach however has demonstrated an error which is not competitive with the traditional polling methods. A next step is to conduct a failure analysis to discern whether there is a further aspect of the content that we may able to model, or a bias we may be able to correct for which can reduce this error. We also observe a dramatic sentiment shift in the two days before polling day which hint at the election outcome. It is perhaps a deeper analysis of the sentiment distribution during this period which will produce the most beneficial application of sentiment analysis in the context of an election campaign.

There are perhaps two reasons that volume is an altogether stronger indicator than sentiment. The first is that volume may simply be a reasonable indicator of popularity in a population of people, and in this case, voting intention. The other is that sentiment in comparison is reactive and it is difficult to discriminate between sentiment which reflects the inner preferences of people, and that which is reflecting an immediate response to a given news story or event. We do see cases where sentiment is necessary. For example, the Green Party in this election had a relatively high volume, but a closer look at the content reveals that this was because people were commenting on low levels of support, an aspect not adequately captured by our sentiment analysis.

At this stage it is unclear whether confining ourselves to sentiment and volume data will allow us to approach levels of acceptable accuracy for reliable measurement. Improvement in sentiment analysis techniques and increased availability of data will likely increase performance, however the research community must address the issues of representativeness and potential for adversarial activity before these methods can be used in a credible way.

Acknowledgments

This work is supported by Science Foundation Ireland under grant 07/CE/I1147. The authors would like to thank the reviewers for their helpful feedback and comments.

	The Green Party		Fianna Fáil		Fine Gael		Sinn Féin		Labour	
	pos	neg	pos	neg	pos	neg	pos	neg	pos	neg
1	vote	happen	lesson	vinb	vote	vote	election	vinb	vote	tv3ld
2	mid	election	unparalleled	tv3ld	team	vinb	luck	plan	north	vinb
3	flyer	happening	failure	bad	children	ha	fought	point	cllr	tv3news
4	dublin	made	interesting	vote	bucket	voting	candidates	vote	dublin	vote
5	west	citizens	wake	tv3news	bearable	don	seat	job	central	baby
6	rx	assembly	east	country	day	gay	vote	creation	prefs	lost
7	oireachtas	proposal	record	voting	giving	tv3news	constituency	banks	donegal	bunch
8	candidate	hard	flyer	anglo	picture	twitter	hard	playing	fair	eating
9	welcomes	final	education	door	bebo	facebook	rain	blinder	running	communists
10	preference	obliterated	smacking	telling	yellow	planted	biased	disgraceful	great	opportunity
11	urban	poor	election	things	hope	audience	helping	don	good	major
12	man	week	b4	screwed	red	answering	campaign	racist	today	posters
13	guidelines	idea	seats	anarchist	plan	twolicy	poised	vincent	west	back
14	achieved	ireland	brilliant	day	roses	script	tonight	money	2nd	won
15	statutory	hoax	ad	friend	equality	priceless	today	tonight	govt	advising

Table 7: The most positive and negative terms for each party according to *Sentiment TF-IDF*

References

Sitaram Asur and Bernardo A. Huberman. 2010. Predicting the future with social media. *Computing Research Repository*.

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*, CIKM '10, pages 1833–1836, New York, NY, USA. ACM.

Johan Bollen, Alberto Pepe, and Huina Mao. 2009. Modeling public mood and emotion: Twitter sentiment and socio-economic phenomena. *CoRR*, abs/0911.1583.

Johan Bollen, Huina Mao, and Xiao-Jun Zeng. 2010. Twitter mood predicts the stock market. *CoRR*, abs/1010.3003.

Nicholas A. Diakopoulos and David A. Shamma. 2010. Characterizing debate performance via aggregated Twitter sentiment. In *Conference on Human Factors in Computing Systems (CHI 2010)*.

Yoav Freund and Robert E. Schapire. 1996. Experiments with a New Boosting Algorithm. In *International Conference on Machine Learning*, pages 148–156.

Daniel Gayo-Avello, Panagiotis T. Metaxas, and Eni Mustafaraj. 2011. Limits of electoral predictions using twitter. In *Proceedings of the International Conference on Weblogs and Social Media (ICWSM) 2011, July 17-21, 2011*.

A. F. Hayes and K. Krippendorff. 2007. Answering the call for a standard reliability measure for coding data. In *Communication Methods and Measures*.

Andreas Jungherr, Pascal Jrgens, and Harald Schoen. 2011. Why the Pirate Party Won the German Election of 2009 or The Trouble With Predictions: A Response to Tumasjan, A., Sprenger, T. O., Sander, P. G., & Welpe, I. M. Predicting Elections With Twitter: What 140 Characters Reveal About Political Sentiment. *Social Science Computer Review*.

Soo-Min Kim and Eduard Hovy. 2007. Crystal: Analyzing Predictive Opinions on the Web. In *Proceedings of the Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)*.

Gustavo Laboreiro, Luís Sarmento, Jorge Teixeira, and Eugénio Oliveira. 2010. Tokenizing microblogging messages using a text classification approach. In *AND '10: Proceedings of The Fourth Workshop on Analytics for Noisy Unstructured Text Data*, New York, NY, USA, October. ACM.

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 AAAI Conference on Weblogs and Social Media*.

David A. Shamma, Lyndon Kennedy, and Elizabeth F. Churchill. 2009. Tweet the debates: understanding community annotation of uncollected sources. In *Proceedings of the first SIGMM workshop on Social media*, WSM '09, pages 3–10, New York, NY, USA. ACM.

Andranik Tumasjan, Timm Sprenger, Philipp Sandner, and Isabell Welpe. 2010. Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. In *International AAAI Conference on Weblogs and Social Media 2010*.

T. Wilson, J. Wiebe, and P. Hoffmann. 2005. Recognizing Contextual Polarity in Phrase-Level Sentiment Analysis. *Proceedings of the 2005 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 347–354.

Xue Zhang, Hauke Fuehres, and Peter A Gloor. 2010. Predicting Stock Market Indicators Through Twitter I hope it is not as bad as I fear. In *Collaborative Innovations Networks Conference (COINs)*.