On voting intentions inference from Twitter content: a case study on UK 2010 General Election

Vasileios Lampos

Intelligent Systems Laboratory

Computer Science Department

University of Bristol

bill.lampos@bristol.ac.uk

Abstract

This is a report, where preliminary work regarding the topic of voting intention inference from Social Media – such as Twitter – is presented. Our case study is the UK 2010 General Election and we are focusing on predicting the percentages of voting intention polls (conducted by YouGov) for the three major political parties – Conservatives, Labours and Liberal Democrats – during a 5-month period before the election date (May 6, 2010). We form three methodologies for extracting positive or negative sentiment from tweets, which build on each other, and then propose two supervised models for turning sentiment into voting intention percentages. Interestingly, when the content of tweets is enriched by attaching synonymous words, a significant improvement on inference performance is achieved reaching a mean absolute error of $4.34\% \pm 2.13\%$; in that case, the predictions are also shown to be statistically significant. The presented methods should be considered as work-in-progress; limitations and suggestions for future work appear in the final section of this script.

1. INTRODUCTION

Information generated by users and published on the Social Web has enabled a new wave of experimentation and research in the past few years. There exist several works showing that textual 'observations' coming from Social Media – such as Twitter – could, for example, be used to nowcast events emerging in the real world [Lampos and Cristianini 2011], understand general emotional figures in a population [Golder and Macy 2011; Lansdall-Welfare et al. 2012] and improve branding strategies [Jansen et al. 2009].

In this report, we present a preliminary method for extracting voting intention figures from Twitter. The case study used to verify our findings is the 2010 General Election in the UK.[1] In the recent past, a few papers have been published on this topic [Lui et al. 2011; Tumasjan et al. 2010; Gayo-Avello et al. 2011; Metaxas et al. 2011; O'Connor et al. 2010], offering preliminary solutions or discussing the limitations that several approaches might have; we refer to and discuss them at the final section of this report.

We consider only the three major parties in the UK; namely the Conservative (**CON**), the Labour (**LAB**) and the Liberal Democrat Party (**LIBDEM** or **LIB**). Overall, we are using three techniques for extracting positive and negative sentiment from tweets and then two different methods to map this sentiment to voting intention percentages.

Ground truth is acquired by YouGov's Published Results[2] and consists of 68 voting intention polls dated from January to May 2010. Polls usually refer to a pair of days (in our data set only 3 of the them are

---

[1]United Kingdom General Election 2010, `http://en.wikipedia.org/wiki/United_Kingdom_general_election,_2010`.

[2]YouGov Archive, `http://labs.yougov.co.uk/publicopinion/archive/`.

1-day polls) and indicate an expectation for the voting intention percentage per political party. As we move closer to the election day (6th of May), they become more dense; there is a new poll published every day. Tweets are drawn from the same period of time, *i.e.* January to May; their total number is greater than 50 million, but not all of them are used as it will become apparent in the following sections. After applying some filtering to keep tweets regarding politics, we end up with 300,000 tweets, *i.e.* approximately 100,000 tweets per political party.

2.   EXTRACTING POSITIVE AND NEGATIVE SENTIMENT FROM TWEETS

The common characteristic in all three approaches is that at first we retrieve a set of tweets regarding each political party by using a handmade set of keywords (see Appendix A). The keywords are different per party, not many (approx. 60 per party) and as simple as possible; we use the names of the top politicians in each party, the name of the party and so on. The search is case sensitive when the keyword is an 1-gram and case insensitive when it is an n-gram, with $n > 1$. In the latter case, we are not looking for an exact match either; we are just searching for tweets that contain all 1-grams of the target n-gram. Character '#' in front of an 1-gram denotes that we are searching for a Twitter topic. Since those keywords are based mainly on names or entities, one could argue that they could also be created in an automatic manner or extracted from a repository and hence, the human involvement could become insignificant.

The three approaches build on each other; each one is a more elaborate version of its precedent. In the first approach, we are using a stemmed version of SentiWordNet 3.0 to extract positive and negative sentiment from tweets without taking into consideration the different parts of speech (verb, noun, adjective and so on). SentiWordNet is the result of automatically annotating all WordNet synsets according to their degrees of positivity, negativity and neutrality [Baccianella et al. 2010; Esuli and Sebastiani 2006]. Stemming is performed by applying Porter's algorithm [Porter 1980] and Part-Of-Speech (**POS**) tagging is 'skipped' by computing the average positive and negative sentiment weight for each stem over all possible POS that it might appear in. Stems with equal positive and negative sentiment are not considered. The positive and negative sentiment scores of a tweet are retrieved by computing the sum of positive and negative sentiment weights of the words it contains. It is noted that there might exist tweets with no words listed in SentiWordNet – those tweets have zero positive and negative sentiment and therefore are ignored. The acronym **SnPOS** (Sentiment no POS tagging) is used to denote this approach. The motivation behind the removal of POS tagging is the assumption that Twitter language might not follow the norms of formal scripts and therefore POS taggers – trained on more 'mainstream' types of text – might create inaccurate results. However, as we will see in practice in the following sections, this is not always the case, probably because tweets that refer to politics have a much better structure than the more casual ones.

The second approach is based on the same basic principles as SnPOS, only this time POS tagging is applied. When one stem is mapped to a particular POS more than once (this can happen due to stemming), the average positive and negative sentiment weights are assigned to it. POS tagging of tweets is carried out by using Stanford POS Tagger, which is a Java implementation of the log-linear POS taggers described in [Toutanova and Manning 2000; Toutanova et al. 2003]. By summing over the sentiment weights of a tweet's terms, we retrieve its sentiment score. This method is denoted as **SPOS** (Sentiment with POS tagging).

Finally, we extend SPOS, by incorporating the core word senses of WordNet [Miller 1995], a semi-automatically compiled list of 5,000 terms.[3] WordNet gives a set of synonyms for each term in this list and we use those synonyms to extend the content of a tweet. That is, if a tweet contains a word listed in WordNet's core terms, the tweet is extended by attaching the synonyms of this word to it. Again, the sentiment score of a tweet is computed by summing over the sentiment weights of its POS tagged terms. This method is identified by the acronym **SPOSW** (Sentiment with POS tagging and WordNet's core terms).

---

[3]Available at `http://wordnet.princeton.edu/wordnet/download/standoff/`.

The main motivation behind extending the content of a tweet with synonyms is the fact that the short length of a tweet might reduce expressiveness and therefore, adding more words could enhance its semantic orientation. Furthermore, SentiWordNet does not include all English words and by attaching synonyms, we achieve to compute pseudo-sentiment for a greater number of tweets.

3. INFERRING VOTING INTENTIONS

By applying a method from the ones described in the previous section, we retrieve the positive and negative sentiment scores for a set of t tweets, which in turn has been originally extracted by using the set of keywords for a political party. The next task is to turn those scores into a percentage which will represent the voting intention for this particular party. In this section, we describe two methods that address this task.

Optionally, one could first remove tweets with almost equal positive and negative sentiments. The semantic orientation of those tweets is unclear and therefore might not always help. Later on, when we present our experimental findings, for every experimental setting we test, we are also replicating the same experiment by introducing a threshold δ, which removes the top 20,000 (this is equal to approx. 6.67% of the entire number of tweets) most 'semantically unclear' tweets. Learning the optimal value of δ from data is possible, but also dependent to the existence of a larger sample.

For the remaining $m \leq t$ tweets we compute their mean positive and negative sentiment scores, say μ_{pos_m} and μ_{neg_m} respectively. The sentiment score assigned to this set of tweets is derived by simply subtracting those two quantities

$$\mathrm{Senti}(m) = \mu_{\mathrm{pos}_m} - \mu_{\mathrm{neg}_m}. \tag{1}$$

Suppose that we have computed the sentiment scores for a set of n time instances (a time instance is equal to one or two days based on the target voting intention poll) for all three parties, $\mathrm{Senti}_{\mathrm{CON}}^{(n)}$, $\mathrm{Senti}_{\mathrm{LAB}}^{(n)}$ and $\mathrm{Senti}_{\mathrm{LIB}}^{(n)}$. To calibrate the relation between sentiment scores and voting intention percentages, we regress each one of these vectors with the corresponding ground truth using OLS and compute the three weights w_{CON}, w_{LAB} and w_{LIB}. For example,

$$w_{\mathrm{CON}} = \arg\min_{w} \sum_{i}^{n} \left(\mathrm{Poll}_{\mathrm{CON}}^{(i)} - \mathrm{Senti}_{\mathrm{CON}}^{(i)} w \right)^2, \tag{2}$$

where $\mathrm{Poll}_{\mathrm{CON}}^{(i)}$ denotes the official voting intention poll percentage for time instance i. We do not introduce bias terms in OLS regression, as during our experimental process it became evident that they receive large values, reducing significantly the freedom of our model and causing overfitting.

In this case study, we are only considering the percentages of the three major parties. After the inference process, the results are normalised in order to represent valid percentages (having a sum equal to 1); in the special (and also rare) case, where inferences are lower than 0, negative results are thresholded to 0 and normalisation does not take place, unless the nonzero percentages sum up to a value greater than 1. Official voting intentions, however, take into consideration the existence of other political 'forces' as well as people's desire not to vote and so on. To create an equivalent and comparable representation, we have normalised the official voting intention percentages as well – based on the fact that the sum of voting intention percentages for the three major parties is on average equal to approx. 90%, this normalisation does not change the picture much (and we can always normalise the results back to a 90% level, if required).

Voting intention inferences are made on unseen data. Again, we have a set of k time instances, which do not overlap with the n time instances used for learning the calibration weights, and we first compute the sentiment scores for each party per time instance. Then, those sentiment scores are multiplied by the corresponding calibration weight; for example the inferred score for the Conservative Party will be equal to:

$$\mathrm{CON}^{(k)} = \mathrm{Senti}_{\mathrm{CON}}^{(k)} \times w_{\mathrm{CON}}. \tag{3}$$

After computing the triplet of inferences, $\mathrm{CON}^{(k)}$, $\mathrm{LAB}^{(k)}$ and $\mathrm{LIB}^{(k)}$, we normalise them so that their sum is equal to 1, for example:

$$\mathrm{CON}^{(k)}_{\mathrm{norm}} = \frac{\mathrm{CON}^{(k)}}{\mathrm{CON}^{(k)} + \mathrm{LAB}^{(k)} + \mathrm{LIB}^{(k)}}, \tag{4}$$

represents the k normalised inferences of the voting intention for the Conservative Party. We denote the method described above as Mean Thresholded Sentiment (**MTS**).

The second method is identical to MTS, but uses a different function to compute $\mathrm{Senti}(m)$ (Equation 1). After retrieving the m tweets that satisfy a threshold δ, we count how many of them have a positive sentiment score that is higher than the negative and vice versa. The overall sentiment score of those tweets is then computed by

$$\mathrm{Senti}(m) = \frac{\#\{\mathrm{pos} > \mathrm{neg}\} - \#\{\mathrm{neg} > \mathrm{pos}\}}{m}, \tag{5}$$

where $\#\{\mathrm{pos} > \mathrm{neg}\}$ is the number of tweets with a positive sentiment greater than the negative one and $\#\{\mathrm{neg} > \mathrm{pos}\}$ the number of tweets with a negative sentiment greater than the positive. We refer to this method as Dominant Sentiment Class (**DSC**).

4.   EXPERIMENTAL PROCESS AND RESULTS

We measure the performance of our methods using two loss functions. Primarily, we use the Mean Absolute Error (**MAE**) (and its standard deviation) between the inferred and the target voting intention percentages. This metric allows for an easier interpretation because it can be read as a percentage, *i.e.* MAE has the same units as the inferred values. In addition, aiming to assess how good is the ranking of the political parties based on the inferred voting intention percentages, we are measuring a Mean Ranking Error (**MRE**). For each triplet of voting intention percentages, the ranking error is defined as the sum over all three parties of the distance between the correct and the inferred ranking. For example, if there exists one incorrect ranking of size 1 (where size measures the difference between a correct and an inferred position), then – since we are dealing with 3 variables only – either there exists one more incorrect ranking of size 1 and therefore the triplet's total error is equal to 2, or two more ranking errors of size 1 and 2, which will make the triplet's total error equal to 4. Obviously, a triplet with correct rankings has an error equal to 0 and the maximum error ($\mathrm{max(RE)}$) per triplet is equal to 4. MRE ranges in $[0, 1]$ and is computed by:

$$\mathrm{MRE}(n) = \frac{1}{n * \mathrm{max(RE)}} \sum_{i=1}^{n} \mathrm{RE}(i). \tag{6}$$

In total, by combining the three methods for extracting positive and negative sentiment from tweets (SnPOS, SPOS and SPOSW) and the two methods for converting sentiment scores to voting intentions (MTS and DSC), we come up with six experimental setups. We run those experiments for $\delta = 0$, *i.e.* for all tweets with a sentiment score, but also for the value of δ which removes the top 20,000 most 'semantically unclear' tweets.

We retrieve the first performance figures by performing leave-one-out cross validation on the aforementioned experimental setups. The performance results for MTS and DSC are presented in Tables I and II respectively. Under both methods, we see that thresholding tends – in most occasions – to improve the inference performance in terms of MAE and MRE. SnPOS performs significantly better than SPOS, which also fails to rank voting intentions properly under MTS. Using DCS results in a better performance on average, but does not deliver the overall best performance, which is derived by SPOSW under MTS using a thresholded data set. Figure 1 depicts those best performing inferences against the corresponding ground

Table I. : MAE ± MAE's standard deviation and MRE for Mean Thresholded Sentiment (**MTS**) by performing leave-one-out cross validation. δ denotes the minimum distance between the positive and negative sentiment score for a tweet in order to be considered.

	δ	CON	LAB	LIBDEM	All Parties	MRE
SnPOS	0	12.24 ± 10.66	10.66 ± 10.39	10.1 ± 11.54	11 ± 8.31	0.4559
SnPOS	0.025	12.17 ± 10.63	10.54 ± 10.47	10.11 ± 11.44	10.94 ± 8.27	0.4779
SPOS	0	36.06 ± 7.09	17.86 ± 11.29	15.12 ± 11.65	23.02 ± 5.23	0.9412
SPOS	0.0072	35.99 ± 7.17	17.88 ± 11.26	15.06 ± 11.65	22.98 ± 5.3	0.9412
SPOSW	0	4.22 ± 2.78	3.75 ± 2.86	5.1 ± 3.76	4.36 ± 2.13	0.2059
SPOSW	0.0238	4.17 ± 2.88	3.76 ± 2.82	5.07 ± 3.7	4.34 ± 2.13	0.1912

Table II. : MAE ± MAE's standard deviation and MRE for Dominant Class Sentiment (**DCS**) by performing leave-one-out cross validation.

	δ	CON	LAB	LIBDEM	All Parties	MRE
SnPOS	0	9.68 ± 8.82	9.26 ± 8.76	9.34 ± 9.23	9.42 ± 6.66	0.4338
SnPOS	0.025	8.23 ± 6.83	8.66 ± 8.34	9.28 ± 9.12	8.72 ± 6.39	0.375
SPOS	0	11.61 ± 11.99	7.29 ± 5.27	7.9 ± 6.06	8.93 ± 6.21	0.5074
SPOS	0.0072	11.62 ± 11.89	7.52 ± 5.34	8.09 ± 6.33	9.08 ± 6.46	0.5074
SPOSW	0	5.5 ± 4.12	4.19 ± 3.35	5.52 ± 4.82	5.07 ± 3.04	0.2647
SPOSW	0.0238	5.47 ± 3.9	3.95 ± 3.34	5.49 ± 4.49	4.97 ± 2.88	0.2279

Table III. : MAE ± MAE's standard deviation and MRE for Mean Thresholded Sentiment (**MTS**). δ denotes the minimum distance between the positive and negative sentiment score for a tweet in order to be considered.

	δ	CON	LAB	LIBDEM	All Parties	MRE	p-value
SnPOS	0	12.97 ± 10.89	10.56 ± 10.05	6.66 ± 7.08	10.06 ± 9.72	0.4038	0.222
SnPOS	0.025	12.98 ± 10.92	10.51 ± 10.03	6.7 ± 7.19	10.06 ± 9.74	0.4038	0.259
SPOS	0	35.03 ± 6.96	21.31 ± 9.4	12.71 ± 7.16	23.02 ± 12.11	0.9231	0.531
SPOS	0.0072	35.14 ± 6.82	21.33 ± 9.41	12.77 ± 7.13	23.08 ± 12.1	0.9231	0.527
SPOSW	0	4.44 ± 3.18	2.66 ± 1.85	3.74 ± 2.86	3.61 ± 2.76	0.1731	0.019
SPOSW	0.0238	4.44 ± 3.31	2.65 ± 1.8	3.84 ± 2.71	3.64 ± 2.75	0.1346	0.006

Table IV. : MAE ± MAE's standard deviation and MRE for Dominant Class Sentiment (**DCS**).

	δ	CON	LAB	LIBDEM	All Parties	MRE	p-value
SnPOS	0	10.56 ± 6.75	9.82 ± 9.18	7.69 ± 9.88	9.36 ± 8.68	0.4038	0.467
SnPOS	0.025	9.66 ± 6.89	9.46 ± 9.05	7.25 ± 9.04	8.79 ± 8.35	0.3654	0.529
SPOS	0	10.63 ± 8.94	8.09 ± 6.37	6.12 ± 5.12	8.28 ± 7.14	0.3846	0.238
SPOS	0.0072	10.51 ± 9.14	7.95 ± 6.18	6.08 ± 5.5	8.18 ± 7.26	0.4038	0.149
SPOSW	0	4.51 ± 3.45	2.87 ± 2.06	3.53 ± 3.29	3.64 ± 3.04	0.1154	0
SPOSW	0.0238	4.49 ± 3.49	2.46 ± 1.81	3.51 ± 3.14	3.49 ± 2.98	0.0962	0
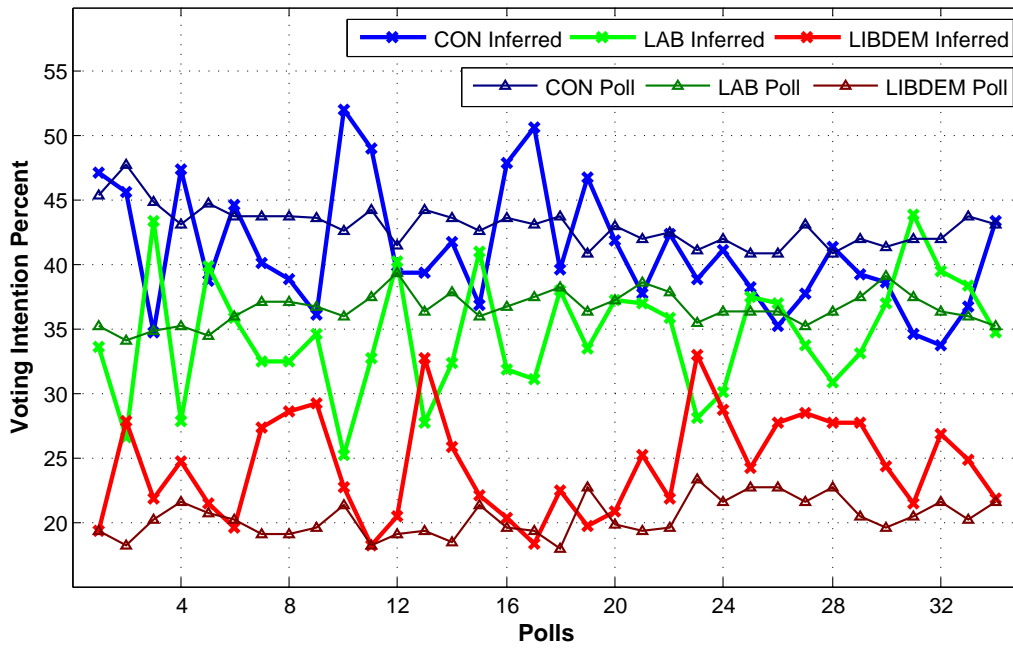
(a) First 34 time instances
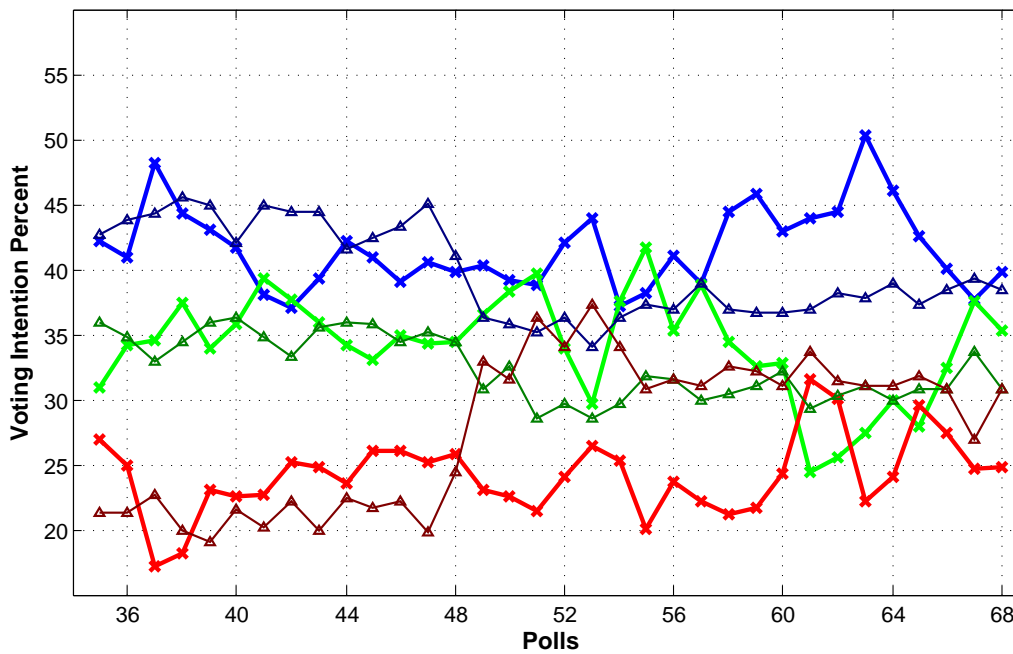
(b) Last 34 time instances

Fig. 1: Leave-one-out cross validation by applying SPOSW and MTS thresholded (in two parts for a better visualisation).
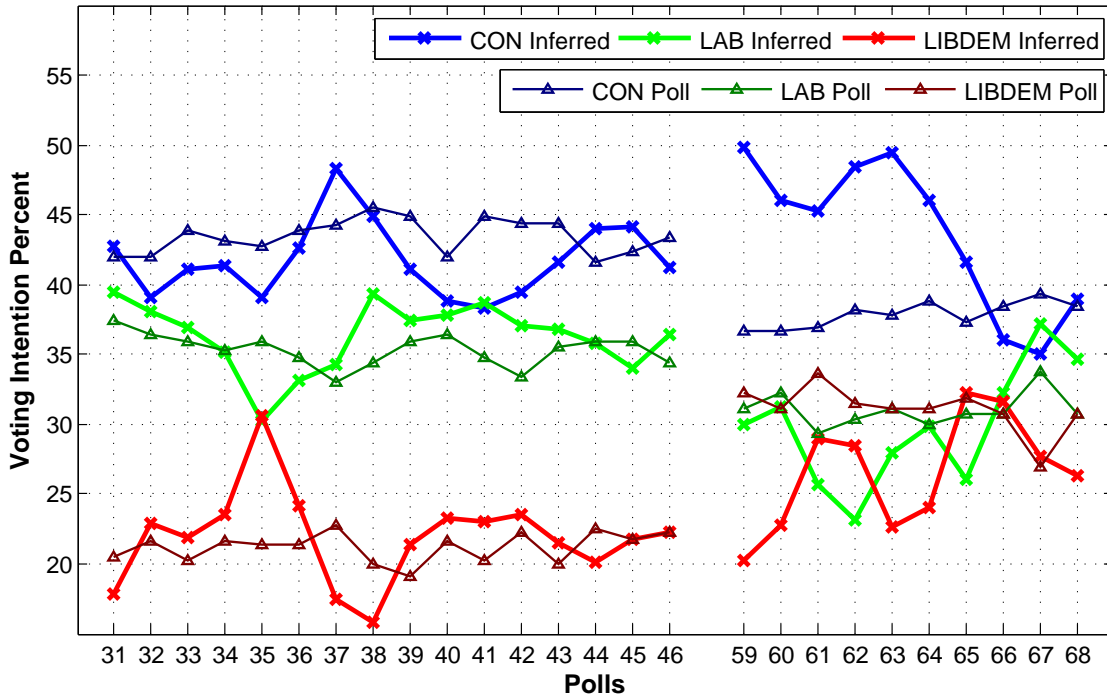
Fig. 2: SPOSW under thresholded DCS.

truth; in this case, MAE across all parties is equal to 4.34 with a standard deviation of 2.13 and MRE is 0.1912.

Since our main data set is of small size and based on the fact that the oscillations in the actual voting intention signals are not major (something that might affect the training and testing processes), we also perform a more focused testing. From the 68 polls, we use 1–30 and 47–58 for training and we perform testing on the remaining 26 ones (31–46 and 59–68). The number of tweets that we retrieve using our search terms is increasing as we get nearer to the election day. This is why the training set has been sliced into two parts, one 'way' before the election day (05/01 to 24/03) and another much closer (13/04 to 25/04). In this experiment, we additionally retrieve a statistical significance indication for our inferences. To do that, we randomly permute the outcomes of MTS or DSC and come up with a randomised training and test set; we repeat this process for 1,000 times and count how many times a model that is based on randomly permuted data delivers a better inference performance than the actual one – this fraction gives the p-value. We consider that a p-value lower than 0.05 indicates statistical significance.

Inference results for MTS and DCS are presented in Tables III and IV respectively. Again, it becomes apparent that combining SentiWordNet with POS tagging and extending tweets with WordNet's core senses (SPOSW) gives out the best inference performance as well as that DCS performs on average better than MTS. SnPOS and SPOS not only show a fairly bad performance in terms of MAE and MRE, but also do not deliver statistically significant results. On the contrary, SPOSW's inferences are shown to be statistical significant; its best performance is now achieved under DCS by performing thresholding reaching an MAE of 3.49 ± 2.98 and an MRE of 0.0962 (see Figure 2).

5. RELATED WORK AND FURTHER DISCUSSION OF THE RESULTS

The topic of voting intention or electoral result inference from the content of Twitter is quite new in the scientific literature. Tumasjan *et al.* have published a paper, which after providing proof that Twitter is a platform where political discussions are indeed conducted, proposes a method for predicting the result of the German elections in 2009 [Tumasjan et al. 2010]. Their method uses the Linguistic Inquiry and Word Count (LIWC2007) for semantic analysis [Pennebaker et al. 2007]; this tool produces a set of 12 dimensions – not just positive or negative sentiment – and therefore Tumasjan *et al.* introduced an averaging technique to acquire 1-dimensional representations. Their method like ours uses sets of keywords (party and politician names) to select politically oriented tweets and then proposes a model for matching the Twitter traffic per party to the final result of the elections. Another paper [O'Connor et al. 2010] presents a different method for tracking voting intention polls based on the ratio of positive versus negative sentiment per tweet; their learning approach has some similarities to ours (it is based on linear regression), but only deals with bivariate problems, *i.e.* polls with two outcomes. Nevertheless, a paper published after indicated that the methods in [Tumasjan et al. 2010] and [O'Connor et al. 2010] were problem specific and not very generic, as their application was proven to not do better than chance in predicting the result of the US congressional elections in 2009 [Gayo-Avello et al. 2011]. Another paper showed that the popular tool of Google Trends has a limited capacity in predicting election results [Lui et al. 2011].

Moreover, an interesting paper conducted a further analysis of the aforementioned methodologies and proposed a triplet of necessary standards that any theory aiming to provide a consistent prediction of elections using content from the Social Media should follow [Metaxas et al. 2011]. The authors recommend that firstly the prediction theory should be formed as a well defined algorithm, secondly the analysis should be aware of the different characteristics arising in the Social Web and thirdly that the method must include experimental justification on why it works.

Based on those propositions, we tried to formalise the steps in our methodology, we provided statistical significance figures for our results and finally by forming the three different schemes (SnPOS, SPOS and SPOSW) for extracting sentiment from tweets, we tried to encounter some of the special characteristics of Twitter's textual stream. Particularly in SPOSW, where we enriched tweets by adding synonymous words in order to enhance their semantic interpretation, we observed a significant improvement on inference performance.

Similarly to the aforementioned works, the first two sentiment extraction methods that we tried (SnPOS and SPOS) showed poor performance and were not statistically significant. On average, modelling sentiment with DCS performed better than MTS. This is quite interesting as – by definition – DCS compresses more information than MTS; one might argue that this compression results in a much clearer signal, *i.e.* removes some noisy observations, especially when applied on extended tweets. The contribution of thresholding in our experiments is ambiguous; sometimes it improves the inference performance and in most occasions reduces the MRE. As we mentioned in the beginning, the optimal value as well as the contribution of thresholding could be investigated further.

We remind the reader that in this report the presented methods for modelling voting intention polls are preliminary. Several aspects are a matter of future research. Primarily, our methods should be applied to other data sets as well, to come up with experimental proof about their capability to generalise. Another important factor is the choice of keywords that are used to select tweets relevant to the task at hand. We argued that those keywords can be selected from topic-related repositories; still, the influence of each keyword should be modelled. Ideally, an automatic mechanism or algorithm should be compiled in order not only to select an optimal subset of keywords, but also to quantify the contribution of each term. Moreover, possible biases introduced by the use of different sentiment analysis tools should also be considered; tools

that incorporate special emotional expressions used in instant messaging and Twitter such as ';-)' or ':D' might achieve a better performance.

A general conclusion that could be extracted or reconfirmed from the presented work is that Social Media do encapsulate content related to political opinion; extending tweets with synonymous terms probably assists in the amplification of this signal. However, it is very important to contemplate that the validity of voting intention polls is questionable. Are those polls a good representation of what is actually happening? Usually, polls from different agencies tend to have significant differences in their percentages and quite often those differences exceed, for example, the lowest total MAE derived in our experimental process (3.49%). Therefore, the formation of a consistently good ground truth is another issue that approaches like this one must resolve.

Acknowledgements

REFERENCES

BACCIANELLA, S., ESULI, A., AND SEBASTIANI, F. 2010. SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining. In *Proceedings of the 7th Conference on International Language Resources and Evaluation (LREC)*. ELRA, 2200–2204.

ESULI, A. AND SEBASTIANI, F. 2006. SentiWordNet: A publicly available lexical resource for opinion mining. In *Proceeding of the 5th Conference on Language Resources and Evaluation (LREC)*. 417–422.

GAYO-AVELLO, D., METAXAS, P. T., AND MUSTAFARAJ, E. 2011. Limits of Electoral Predictions using Twitter. In *Proceedings of the Fifth International AAAI Conference on Weblogs and Social Media (ICWSM)*. 490–493.

GOLDER, S. A. AND MACY, M. W. 2011. Diurnal and Seasonal Mood Vary with Work, Sleep, and Daylength Across Diverse Cultures. *Science 333,* 6051, 1878–1881.

JANSEN, B. J., ZHANG, M., SOBEL, K., AND CHOWDURY, A. 2009. Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology 60,* 11, 2169–2188.

LAMPOS, V. AND CRISTIANINI, N. 2011. Nowcasting Events from the Social Web with Statistical Learning. *ACM Transactions on Intelligent Systems and Technology (TIST) 3,* 4.

LANSDALL-WELFARE, T., LAMPOS, V., AND CRISTIANINI, N. 2012. Effects of the Recession on Public Mood in the UK. In *Mining Social Network Dynamics (MSND) session on Social Media Applications in News and Entertainment (SMANE) at WWW '12*. ACM, 1221–1226.

LUI, C., METAXAS, P. T., AND MUSTAFARAJ, E. 2011. On the predictability of the U.S. elections through search volume activity. In *Proceedings of the IADIS International Conference on e-Society*.

METAXAS, P. T., MUSTAFARAJ, E., AND GAYO-AVELLO, D. 2011. How (Not) To Predict Elections. In *IEEE 3rd International Conference on Social Computing (SocialCom)*. IEEE Press, 165 – 171.

MILLER, G. A. 1995. WordNet: a lexical database for English. *Communications of the ACM 38,* 11, 39–41.

O'CONNOR, B., BALASUBRAMANYAN, R., ROUTLEDGE, B. R., AND SMITH, N. A. 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*. Number May. AAAI Press, 122–129.

PENNEBAKER, J. W., CHUNG, C. K., IRELAND, M., GONZALES, A., AND BOOTH, R. J. 2007. The Development and Psychometric Properties of LIWC2007. Tech. rep., The University of Texas at Austin & University of Auckland, New Zealand.

PORTER, M. F. 1980. An algorithm for suffix stripping. *Program 14,* 3, 130–137.

TOUTANOVA, K., KLEIN, D., MANNING, C. D., AND SINGER, Y. 2003. *Feature-rich part-of-speech tagging with a cyclic dependency network*. Vol. 1. Association for Computational Linguistics.

TOUTANOVA, K. AND MANNING, C. D. 2000. Enriching the knowledge sources used in a maximum entropy part-of-speech tagger. *Proceedings of the Joint SIGDAT conference on Empirical methods in natural language processing and very large corpora held in conjunction with the 38th Annual Meeting of the Association for Computational Linguistics 13*, 63–70.

Tumasjan, A., Sprenger, T. O., Sandner, P. G., and Welpe, I. M. 2010. Predicting elections with Twitter: What 140 characters reveal about political sentiment. In *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media.* AAAI, 178–185.

A. APPENDIX: SEARCH TERMS

Terms used to select tweets related to the major political parties in the UK. Character '␣' denotes that an empty space and '#' denotes a Twitter topic. Tables V, VI and VII hold the terms for the Conservative (52 terms), Labour (57 terms) and Liberal Democrat party (62 terms) respectively.

Table V. : Terms for the Conservative Party (**CON**) – 52 terms

#conservative	CONSERVATIVE	Eric Pickles	Margaret Thatcher
#Tories	Conservative	Francis Maude	Michael Gove
#tories	Conservative Party	George Osborne	Oliver Letwin
#Tory	Crispin Blunt	Gerald Howarth	Owen Paterson
#tory	David Cameron	Grant Shapps	Patrick McLoughlin
␣TORIES	David Davis	Hugo Swire	Philip Hammond
␣tories	David Mundell	Iain Smith	President Cameron
Alan Duncan	David Willetts	Jeremy Hunt	Theresa May
Andrew Lansley	debate Cameron	John Redwood	Tories
Andrew Mitchell	Dominic Grieve	Ken Clarke	vote Cameron
Boris Johnson	Duncan Smith	Lady Warsi	vote conservative
Caroline Spelman	Edward Leigh	Leader Cameron	vote Tory
Cheryl Gillan	Edward Vaizey	Liam Fox	William Hague

Table VI. : Terms for the Labour Party (**LAB**) – 57 terms

#labour	debate Brown	LABOUR	Rosie Winterton
␣Ed ␣Balls	Diane Abbott	Labour	Sadiq Khan
Alan Johnson	Douglas Alexander	Labour Party	Shaun Woodward
Alastair Campbell	Ed Miliband	Leader Brown	Stella Creasy
Andy Burnham	Gordon Brown	Liam Byrne	Tessa Jowell
Angela Eagle	Government Brown	Luciana Berger	Tom Harris
Ann McKechin	Harriet Harman	Maria Eagle	Tom Watson
Ben Bradshaw	Hilary Benn	Mary Creagh	Tony Blair
Caroline Flint	Ivan Lewis	Meg Hillier	Tony Lloyd
Charles Clarke	Jim Murphy	MILIBAND	vote Brown
Chris Bryant	John Denham	Miliband	vote labour
Chuka Umunna	John Healey	Peter Hain	Yvette Cooper
David Blunkett	John Prescott	Peter Mandelson	
David Lammy	Jon Trickett	President Brown	
David Miliband	Kerry McCarthy	Prime Minister Brown	

Table VII. : Terms for the Liberal Democrat Party (**LIB**) – 62 terms

#Clegg	Fiona Hall	LIBERAL	Norman Backer
#clegg	Greg Mulholland	Liberal	Norman Lamb
#liberal	Jenny Willott	Liberal Democrat	Paul Burstow
_Ed Davey	Jeremy Browne	Lord McNally	President Clegg
_Jo Swinson	John Hemming	Lord Shutt	Sarah Teather
Alistair Carmichael	John Pugh	Lorely Burt	Simon Hughes
Andrew Stunell	John Thurso	Lynne Featherstone	Steve Webb
Charles Kennedy	Julian Huppert	MP Clegg	Tessa Munt
Chris Huhne	Leader Clegg	Mark Hunter	Tim Farron
Dan Rogerson	Lembit Opik	Martin Horwood	Tom McNally
Danny Alexander	Lib_ _Dem	Michael Crockart	Vince Cable
David Heath	LIBDEM	Mike Crockart	Vincent Cable
David Shutt	LibDem	Mike Hancock	vote Clegg
debate Clegg	Libdem	Nicholas Clegg	vote liberal
Duncan Hames	libDem	Nick Clegg	
Edward Davey	libdem	Nick Harvey	