Journal of Universal Computer Science, vol. 23, no. 3 (2017), 280-303
submitted: 15/9/16, accepted: 1/3/17, appeared: 28/3/17 © *J.UCS*

On Predicting Election Results using Twitter and Linked Open Data: The Case of the UK 2010 Election

Evangelos Kalampokis
(University of Macedonia, Thessaloniki, Greece
ekal@uom.gr)

Areti Karamanou
(University of Macedonia, Thessaloniki, Greece
akarm@uom.gr)

Efthimios Tambouris
(University of Macedonia, Thessaloniki, Greece
tambouris@uom.gr)

Konstantinos Tarabanis
(University of Macedonia, Thessaloniki, Greece
kat@uom.gr)

Abstract: The analysis of Social Media data enables eliciting public behaviour and opinion. In this context, a number of studies have recently explored Social Media's capability to predict the outcome of real-world phenomena. The results of these studies are controversial with elections being the most disputable phenomenon. The objective of this paper is to present a case of predicting the results of the UK 2010 through Twitter. In particular, we study to what extend it is possible to use Twitter data to accurately predict the percentage of votes of the three most prominent political parties namely the Conservative Party, Liberal Democrats, and the Labour Party. The approach we follow capitalises on (a) a theoretical Social Media data analysis framework for predictions and (b) Linked Open Data to enrich Twitter data. We extensively discuss each step of the framework to emphasise on the details that could affect the prediction accuracy. We anticipate that this paper will contribute to the ongoing discussion of understanding to what extend and under which circumstances election results are predictable through Social Media.
Key Words: Twitter, Social Media, Predictive Analytics, Linked Data, Election
Category: H.3.3, I.2.2, I.2.7, L.3.2

1 Introduction

In the last years, Social Media have grown in popularity with millions of users producing tremendous amounts of data in various forms such as text messages, tags and multimedia content. Twitter is a popular Social Media platform with 313 million monthly active users that publish more than 500 million posts per day (as of June 2016). It enables the distribution of small-scale messages (also called tweets) that can be collected through its APIs. Interestingly, Twitter has also become a popular tool that enables researchers, businesses and governments to elicit public behaviour and opinion regarding numerous topics and phenomena.

In this context, a number of studies have recently explored Twitter's capability to predict the outcome of phenomena such as elections [Tumasjan et al., 2010, Livne et al., 2011], stock market [Bollen et al., 2011], Oscar awards [Bothos et al., 2010], box office [Asur and Huberman, 2010], and consumers' behavior [Kalampokis et al., 2016] or the occurrence of phenomena such as pandemics [Chunara et al., 2012, Ritterman et al., 2009]. These studies analyzed various Twitter related features such as sentiment of posts, and suggested that these could predict the outcome of a phenomenon under study. However, later studies challenged the positive initial findings with elections being the main topic of dispute [Jungherr et al., 2012, Gayo-Avello, 2011, Metaxas et al., 2011]. In this context, it was argued that the main cause of the contradicting results is that, Social Media are treated as a black box i.e. *"it may give you the right answer, even though you may not know why"* [Metaxas et al., 2011].

A recent systematic analysis of studies aiming at predicting real-world phenomena through Social Media data revealed that these studies can be decomposed into a small number of steps and different approaches can be followed in each step [Kalampokis et al., 2013]. The approaches that are followed in each step impact the accuracy of the prediction. In addition, the analysis revealed that published studies aiming to predict elections outcome have the poorest results among all application areas (such as box office revenues, stock market, epidemics) because of the approaches followed. More importantly, this study provides a flexible mechanism enabling the investigation of different approaches in the different steps.

In this paper, we present the case of predicting the results of UK 2010 election using Twitter data. Towards this end, we adopt the Social Media analysis framework for predictions [Kalampokis et al., 2013] and we employ the Linked Data paradigm to semantically enrich tweets by reusing structured objective data that is freely available on the Linked Data Web. The term Linked Data refers to *"data published on the Web in such a way that it is machine-readable, its meaning is explicitly defined, it is linked to other external datasets, and can in turn be linked to from external datasets"* [Bizer et al., 2009, p. 2]. The aim of our paper is neither to prove the ability of Social Media to predict real world phenomena nor to evaluate the accuracy of the framework or the employed technologies. We anticipate that this paper will contribute towards understanding to what extend and under which circumstances it is possible to accurately predict election results using Twitter data.

The structure of the paper is as follows: Section 2 presents the background knowledge as regards Twitter and Linked Data paradigm. Section 3 describes the framework for predicting elections that we follow while section 4 presents the application of the framework to the UK elections 2010 case study. Finally, section 5 discusses our approach and the results of this study and section 6 draws conclusions.

2 Background

In this section we introduce the two main topics of interest of the article, namely information dissemination in Twitter and the Linked Data paradigm. In the latter we also describe Named Entity Recognition (NER) as a way to identify named entities in tweets, which will be the joint point to link tweets to Linked Open Data.

2.1 Twitter

Twitter is one of the most popular Social Media. While most tweets are conversation and chatter, many users share information and spread news through Twitter as well [Java et al., 2007, Naaman et al., 2010]. The number of tweets that are posted online is closely related to real world phenomena, as important events seem to trigger an increased number of tweets [Hughes and Palen, 2009]. It is interestingly estimated that the majority of the trending topics and 85% of all tweets are related to news [Kwak et al., 2010]. In Twitter realm traditional media organisations seem to play a significant role since they are by far the most active users. However, only about 15% of tweets received by ordinary users are received directly from the media [Wu et al., 2011].

Two important features of Twitter are hashtags and the '@' symbol. A hashtag is a word or phrase preceded by the # symbol and denotes some aspects of a tweet such as its topic or its intended audience. Hashtags allow users to group posts together and also facilitates the finding of tweets with the same hashtag, topic or content. The @ symbol is mainly used to address a tweet to a particular user and thus enables the identification of conversations in Twitter. Finally, retweets are also important in Twitter because they enable the forwarding of a tweet by posting it again, sometimes with additional comments [Kwak et al., 2010].

2.2 Linked Data

Linked Data has been introduced as a promising paradigm for opening up data because it facilitates the integration of datasets across the Web [Bizer et al., 2009]. The specification of the Linked Data principles [Berners-Lee, 2006] resulted in the emergence of the Web of Linked Data, which currently comprises more than 1000 datasets in various domains [Schmachtenberg et al., 2014].

Linked data is based on Semantic Web philosophy and technologies but in contrast to the full-fledged Semantic Web vision, it is mainly about publishing structured data in RDF using URIs rather than focusing on the ontological level or inferencing [Hausenblas, 2009].

Linked Data following a RESTful approach requires the identification of resources with URI references that can be dereferenced over the HTTP protocol into RDF data that describes the identified resource. In addition, Linked Data include the creation of typed links between URI references, so that one can discover more data.

Twitter data has been recently considered as a source of data that if published as Linked Data and connected to other data sets would provide value to the users. It was suggested that Linked Data could alleviate the information overload problem of Twitter by replacing keyword and hashtag based search by SPARQL queries, which would specify RDF graph patterns as constraints and thus users could select a subset of data that matches their needs [Mendes et al., 2010]. Moreover, Rowe & Stankovic [Rowe and Stankovic, 2012] proposed an approach to automatically align tweets with events, while Abel et al. [Abel et al., 2011] proposed a semantically enriched faceted search method for Twitter, which is based on Linked Data paradigm.

Named Entity Recognition (NER) is the process of identifying named entities in text and classifying them into predefined categories such as person, organisation and location [Nadeau and Sekine, 2007]. NER is very important in Linked Data representation of tweets as it enables the identification of named entities included in the actual text and thus it enriches the actual representation of the tweet. NER methods can be classified into two main categories: Rule Based Methods and Machine Learning Methods. The former includes a set of rules that implements a specific grammar to identify and classify named entities while the later is divided into three categories, namely supervised learning (SL), semi-supervised learning (SSL) and unsupervised learning (UL). In SL the classifier for NER is trained using a training data set that is annotated with named entities while in UL there is no annotated data.

3 The Framework for Predicting Elections

In this section we describe in detail the framework we followed in order to predict the results of the UK 2010 elections through the analysis of Twitter data. Specifically, our approach is based on the Social Media (SM) data analysis framework for predictions [Kalampokis et al., 2013], which identified the phases, stages, steps of relevant studies, and the different approaches that can be followed in each step. The framework comprises two phases namely *Data Conditioning* phase and *Predictive Analysis* phase. The first one focuses on transforming raw SM data to high quality data that is structured based on some predictor variables. The second phase aims at creating and evaluating a predictive model for the prediction of phenomena outcomes based on new data observations.

Each of these phases is divided into a sequence of stages. The first phase comprises the *Collection and filtering of raw data* stage, which deals with the collection of raw data from various sources and its further filtering that removes irrelevant data, and the *Computation of Predictor Variables* stage, which deals with analysis of the raw data resulting from the previous stage in order to compute the values of predictor variables. The second phase comprises the *Creation of Predictive Model* stage, in which the actual model is created based on statistical or machine learning methods, and the *Evaluation of the Predictive Performance* stage, in which the predictive accuracy is evaluated against the actual outcome.

Moreover, each stage of the framework can be decomposed into a small number of steps and different approaches can be followed in each step. According to the analysis of Kalampokis et al. [Kalampokis et al., 2013] the accuracy of the prediction depends on the approach that is followed in each step of the framework, as well as on the social media type and the application area. In Table 1 the stages and steps of the framework are presented.

Table 1: The stages and steps of the Social Media Analysis Framework

Stages	Steps
Collection & Filtering of Raw Data	Determination of time window Identification of location Identification of user profile characteristics Selection of search terms
Computation of Predictor Variables	Selection of predictor variables Measurement of predictor variables Computation of predictor variables
Creation of Predictive Model	Selection of predictive method Selection and use of non-SM predictor variables Identification of data for evaluation of prediction
Evaluation of the Predictive Performance	Selection of the evaluation method Specification of the prediction baseline

4 Applying the framework to the UK elections 2010 case study

Prime Minister Gordon Brown announced the election in April 6, 2010. The election took place on May 6, 2010 in 650 parliament constituencies across the U.K. A total of 49 political parties participated in the election, with three of them being the most prominent i.e. the Conservative party led by David Cameron, the Liberal Democrat party led by Nick Clegg and the Labour party led by Gordon Brown. Smaller parties that finally received more than 1% of the total votes include the UK Independence Party (UKIP), the British National Party (BNP), the Scottish National Party (SNP), the Green Party and the Pirate Party.

In this section we apply the framework presented in the previous section to the case study of the UK Elections 2010. Specifically, this section presents in detail the four stages of the framework and the approaches followed in the respective steps. In each stage, we also describe relevant research endeavours aiming at predicting elections results through Twitter data.

4.1 Collection and Filtering of Raw Data

The *Collection and filtering of raw data* stage deals with raw SM data collection from various sources and filtering of data in order to determine those relevant. This stage is important for the identification of both the complete and correct set of SM data that is related to a phenomenon under exploration. This step requires to determine the following characteristics:

- The time window of the phenomenon. Time window specifies the duration of the collection activity as well as its relation to the characteristic period of the phenomenon. The time window of our phenomenon starts with the announcement of the elections and finishes with the election day. In practice, we started collecting tweets two days after the announcement of the elections (8/4/2010) until the day of the elections (i.e. 5/5/2010).

- The location of the phenomenon. The accurate extraction of the location of the collected data is significant for some phenomena such as during the occurrence of natural phenomena. In our research however the location of the collected tweets is not of primary importance because we assume that the location of all tweets that talk about the UK elections come from the UK. As a result, we discard this step from our study.

- The user profile characteristics. The identification of user profile characteristics in tweets is out of the scope of this paper. As a result, we discard this step.

- The search terms. In this study we selected search terms to i) collect tweets relevant to the 2010 UK elections, and ii) filter the collected tweets in order to classify them to the UK political party they refer to. In the literature, the different approaches used for the identification of the complete and correct search terms can fall into two broad categories: manual approaches where researchers set the search terms (e.g. [Polgreen et al., 2008]) and dynamic approaches where search terms are derived through a computational process (e.g. [Ginsberg et al., 2009]). However, the selection of the correct set of search terms can be challenging in application areas such as elections or macroeconomics which involve multiple and interrelated real-world entities such as political parties and politicians or complex concepts such as consumer confidence or inflation rate. Although dynamic approaches provide better prediction results, all existing studies aiming at predicting election results through Twitter use manual approaches, usually by selecting a number of keywords, hash tags or user mentions [Kalampokis et al., 2013]. In this paper, we manually selected hash tags to collect Twitter data. However, we used a dynamic search term selection employing the Linked Data paradigm to filter the collected Twitter data. The next two sub-sections describe the approaches used and the results of the i) collection and ii) the filtering of tweets.

4.1.1 Collection of Twitter Data

In this study, we used hashtags as search terms to collect relevant Twitter data. Specifically, we used the Twitter API and the #ge2010, #ukelection, #election2010 and #ge10 hashtags and collected a total of 84.375 unique tweets and 25.241 re-tweets. A list with all tweed ids can be found online[1]. We did not remove re-tweets because we consider them as a way of opinion indication. The retrieved JSON files were stored in an instance of a MongoDB NoSQL database so as to easily access and exploit them.

Figure 1 depicts the distribution of the total number of tweets regarding all political parties over the whole time period. The figure shows that the majority of the tweets were published during the six last days prior to the election with highest peak the last day before the elections.

Moreover, Figure 2 displays the distribution of tweets by different authors across the 30-days period before the elections. The X-axis shows the number of tweets in the log scale, while the Y-axis represents the corresponding frequency of authors in the log scale. The figure shows that the distribution is very close to a Zipfian distribution as a few authors are producing a large number of tweets.

Furthermore, Figure 3 shows how the number of tweets per unique author changes over the 30 days for the three most famous political parties. The figure

[1] `https://tinyurl.com/zxooew7`

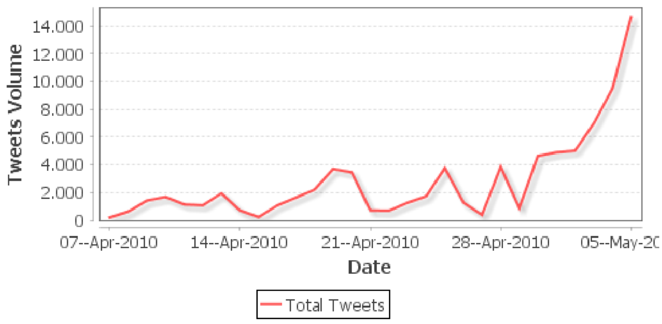

Figure 1: Tweet volume over 30 days before the elections

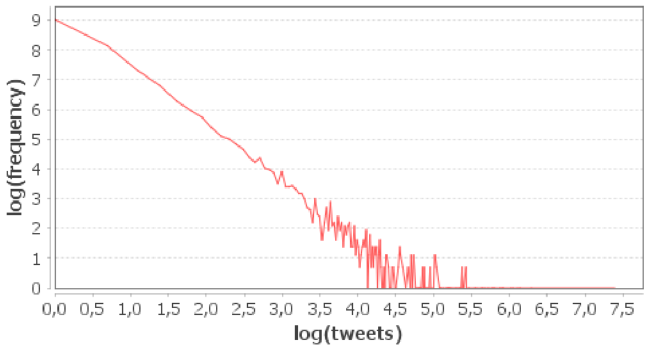

Figure 2: Log distribution of authors and tweets

indicates that the ratio ranges between 1 and 1.8 tweets per day during the period of interest.

4.1.2 Filtering of Twitter Data

In this paper we follow a dynamic search term selection approach to filter the collected Twitter data and classify them to political parties. Our approach comprises two steps: i) the extraction of named entities from the text of tweets by employing NER, and ii) the creation of Linked Data including the establishment of links to DBpedia. As an example, figure 4 depicts four tweets that were published prior to the elections. Each of these includes a different named entity that is related to the elections, namely Cameron, George Osborn, Tory and Conservatives. In Twitter's realm these entities and thus the tweets that include them do not have any connection. However, DBpedia includes linked data descriptions of the same entities having also links among them. In particular, DBpedia suggests

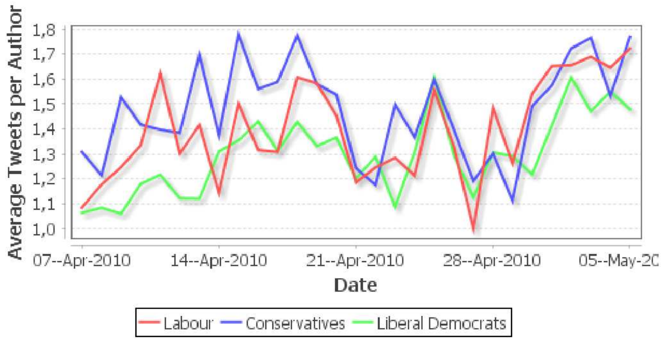

Figure 3: Average number of tweets per unique author for the three parties

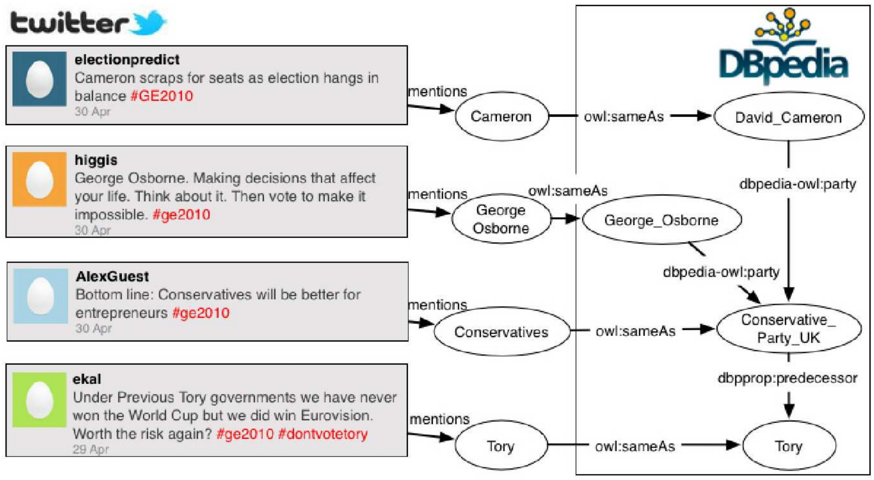

Figure 4: Linking entities extracted from Tweets to DBpedia

that David Cameron and George Osborne belong to the Conservative party and that the Tory party is predecessor of the Conservative party. The integration of the data from DBpedia with the data from Twitter could enable an analyst to understand that the four tweets of figure 4 actually refer to the Conservative party.

We used a Conditional Random Field (CRF) sequence classifier provided by Stanford NER in order to identify named entities in the tweets. More specifically, we created the gold standard of our data set by randomly sampling 2000 tweets and manually annotated them with 3 different Named Entity Types: Person, Organisation and Location. Thereafter, the classifier was trained on 1000 manually annotated tweets (training data set) from the gold standard and then tested on the remaining 1000 tweets (test data set). The CRF classifier achieved 85.89 F1 score in the data set. The F1 score specifies the accuracy of NER as a harmonic mean of Recall and Precision. Recall is the ratio of the total number of correct entities identified to the total number of entities while Precision the ratio of the total number of correct entities identified to the total number of entities identified.

Thereafter, the extracted entities along with tweet's metadata were transformed into RDF, which was stored into a Virtuoso store. URI aliases were identified between the Twitter data set and DBpedia and owl:sameAs links were established. The aim of this interlinking was to perform (a) entity disambiguation i.e. to specify the extracted entities that refer to the same real world entity and (b) data enrichment by enabling the use of existing linked data on the Web. In order to perform the interlinking we searched for candidate linking entities in DBpedia using DBpedia Lookup Service. More particularly, we used the Keyword Search API of DBpedia Lookup Service that employs keyword(s) to search for related DBpedia resources. For example, the query with QueryClass=Person and QueryString=Cameron searches DBpedia for URIs relative to the keyword Cameron. The QueryClass restricts the searching of this query to only resources that are instances of the Person DBpedia class. The specific query returns five results: "David Cameron", "James Cameron", "Cameron Diaz", "Thomas Fairfax, 3rd Lord Fairfax of Cameron" and "Cameron Crowe".

In our case, DBpedia lookup used the entities identified after the NER process. However, from a total of 6392 distinct entities found in our tweets, DBpedia lookup identified only 3712 matches (58.07%). The unmatched entities include non-real world entities such as "tory_letter" and entities wrongly annotated by our NER classifier. For example, the classifier annotated William Hague who is a politician as a Location. Thereafter this entity was searched using QueryClass Place in DBpedia lookup, hence returning no results.

In the case where more than one resource was returned by DBpedia lookup, the challenge was to select the most appropriate one. To this end, we first created a "bag of words" by retrieving and tokenising the rdfs:label of the resources that are connected with the dcterms:subject property to the returned by DBpedia Lookup resources when searching for the eight most popular UK political parties (i.e. using as keywords "UK_Independence_Party", "Conservative_Party_(UK)", "Labour_Party_(UK)", "British_National_Party", "Liberal_Democrats", "Green_Party_(UK)", "Pirate_Party_UK", and "Scottish_National_Party"). We then created a "bag-of-words" for each of the returning results of each NER entity lookup by retrieving and tokenising again the rdfs:label of the resources that are connected with the dcterms:subject property to the results. We thereafter computed the TF-IDF vectors for each of the bag of words and compared the TF-IDF vector of each result to the TF-IDF vector of the domain-related bag-of-words using cosine similarity as a measure. The best cosine similarity indicated the most appropriate resource among the resulting resources.

The total number of distinct entities for each named entity type in the produced dataset are 36 Organisations (e.g. dbpedia:Liberal_Democrats, dbpedia:Sky_News, dbpedia:BBC), 337 Persons (e.g. dbpedia:Gordon_Brown, dbpedia:David_Cameron, dbpedia:Nick_Griffin) and 260 Locations (e.g. dbpedia:Bradford, dbpedia:Edinburgh, dbpedia:Islington).

Finally, in order to classify the tweets into the different political parties we queried the resulted RDF data in order to identify tweets that have entities that are linked to the DBpedia representation of a political party or to a resource that is connected through the dbpedia-owl:leader, dbpprop:leader, dbpedia-owl:party, dbpprop:party or dbpedia-owl:otherParty properties to a political party.

Table 2 shows the number of tweets that were classified to the three most popular political parties, namely Labour, Conservatives, and Liberal Democrats.

Table 2: The number of tweets for the major parties after the classification

	Labour	Conservatives	Liberal Democrats
Volume of tweets	13751	16992	9851

4.2 Computation of Predictor Variables

This stage analyses the data collected and filtered in the previous stage in order to compute the values of the predictor variables.

In general, most of the predictor variables are being measured at successive time instants separated by uniform time intervals. A careful selection of time intervals is significant at this stage in order to facilitate the comparison of predictor variables with the actual outcome. We first measured our predictor variables using daily time intervals. As a result, a number of time series were created to depict the daily measurements of our predictor variables during the whole month before the elections.

The variables that are typically used to create predictive models using Social Media data are related to (a) the volume of Social Media posts (e.g. number of tweets, number of reviews etc.), (b) the sentiment expressed through the data, and (c) profile characteristics of online users (e.g. Facebook friends).

Some application areas, such as natural phenomena, do not involve the expression of a subjective opinion that may involve sentiment but only the statement of an objective fact e.g. “it is raining”. In elections, however, the expression and thus the computation of the expressed sentiment is crucial. Despite that, the majority of the existing research endeavours use volume-related variables in their predictive models, with only three of them [Metaxas et al., 2011, He et al., 2012, Gayo-Avello, 2011] using sentiment-related variables.

In this paper we exploit both volume and sentiment related variables in order to create the predictive model. We selected two predictor variables for our research:

- Relative frequency (RF) of tweets. This variable is related to the volume of Social Media data. In particular, we define RF of a political party as the tweet volume per day. For example, if in a particular day a total of 100 tweets that refer UK political parties have been posted and 20 of them regard the Labour party, then the RF of the Labour party at this day will be 0.2 or 20%.

- Positive-negative ratio (PNR). This variable is related to the sentiment of Social Media data. PNR is used to quantify the sentiment expressed in tweets about a political party. In particular, for each political party, we define PNR on the day t as the ratio of positive over negative tweets published that day for this political party:

$$PNR_t = \frac{count_t(pos)}{count_t(neg)} \tag{1}$$

  PNR is more than one when the number of positive tweets for a party is larger than the number of the negative tweets. The next subsection describes in detail our approach in computing sentiment in tweets.

We used two alternative ways to measure both predictor variables: i) using moving average over a window of previous days and ii) without using moving average of the previous days. In the first case, moving average was used to measure the value of the predictor variable for a day using the average of the previous k days. A moving average over a window of the past k days for variable x is calculated as:

$$MA_t = 1/k(x_{t-k+1} + x_{t-k+2} + \ldots + x_t) \tag{2}$$

We used three different values for k i.e. k=2, k=3 and k=4 to measure the predictor variables.

4.2.1 Computation of Volume

The computation of volume-related variables is straightforward through the count of posts that satisfy some criteria. In our work, we computed the volume

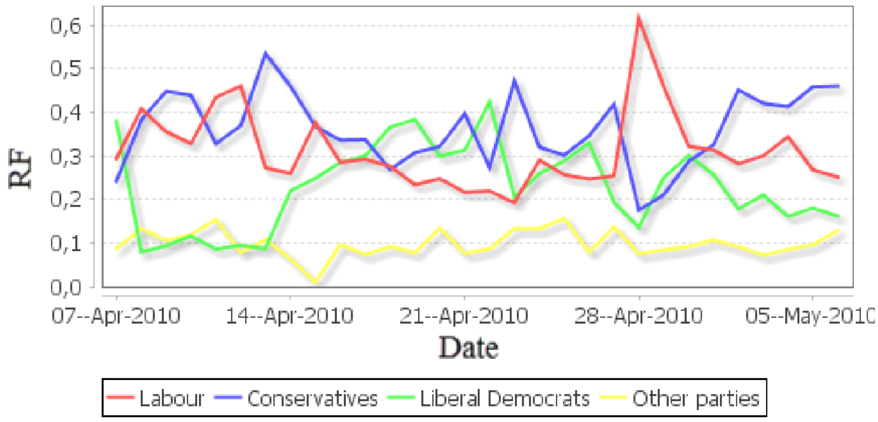

Figure 5: RF of political parties

of tweets for each political party by counting the number of tweets that were classified to the political party in the filtering process (Section 4.1.2). Figure 5 presents the RF of all political parties.

4.2.2 Computation of Sentiment

The computation of sentiment expressed in SM is significant because it may provide poor results, which impact the prediction accuracy.

In general, the approaches used for sentiment computation fall into two categories i.e. lexicon-based, where sentiment in a text is determined by the occurrence of keywords included in a pre-defined lexicon, and machine learning, where sentiment is computed by language model classifiers. The use of machine learning approaches in sentiment analysis supports the creation of more accurate predictive models than the ones created with lexicon-based sentiment analysis [Kalampokis et al., 2013].

Out of the eight elections studies that employed sentiment related variables in their predictive models, two followed a lexicon-based approach [Metaxas et al., 2011, Burnap et al., 2016], five followed a machine learning approach [He et al., 2012, Bermingham and Smeaton, 2011, Franch, 2013, Ceron et al., 2014, Ceron et al., 2013], and one used both [Gayo-Avello, 2011]. In this paper, we follow a machine-learning algorithm to determine the sentiment expressed in tweets.

We used sentiment analysis to compute our PNR variable. Our sentiment analysis is based on the computation of sentiment by a machine learning language model classifier from the LingPipe package[2]. In particular, we employed

[2] http://alias-i.com/lingpipe

the DynamicLMClassifier and we performed a k-fold cross-validation to classify the tweets as Positive or Negative. The specific language model classifier accepts training events of categorised character sequences. Training is based on a multivariate estimator for the category distribution and dynamic language models for the per-category character sequence estimators. Specifically, during a k-fold cross-validation a data set is divided into evenly sized k folds and then k iterations of classifications are performed. Each iteration uses one of the folds of data (a different fold is selected for each iteration) as testing data and the remaining k-1 folds as training data.

In order to train the classifier we need a training data set, which separates positive tweets from negative tweets. Usually the training data set is obtained by manually annotating data but in our case we used a set of positive and negative hashtags to identify the relevant categories of tweets. 4053 tweets include negative (e.g. #dontvotetory, #labourout, #libdemfail) and 4274 positive hashtags (e.g. #torywin, #votelabour, #imvotinglibdem). However, our training data set contains 70% of these tweets because tweets that mentioned more than one political party were ignored in order to avoid ambiguity and increase accuracy. We also ignored all re-tweets and, furthermore, preprocessed these tweets in order to:

- Remove user mention entities and URL entities
- Remove phenomenon related hashtags e.g. #ukelection
- Remove all stop-words
- Replace party and candidate names with $< PP >$ and $< CA >$ respectively

We trained the classifier using an n-gram model with n=6. The classifier was trained to predict two classes i.e. positive and negative. Our cross-validation sentiment analysis experiment resulted in a model with an (average) accuracy of 83.67%.

Thereafter, we applied our sentiment analysis model to all tweets that mention only one political party in order to predict their sentiment. We used a threshold of 85% for the conditional probability of the predicted sentiment in order to keep only high confidence predictions for the sentiment of tweets. The result was 21.487 (out of the 24.265) tweets mentioning only one political party and showing sentiment for it: 10.074 of them showing positive and 11.413 showing negative sentiment respectively.

Figure 6 presents the PNR of all political parties using moving average with k=4.

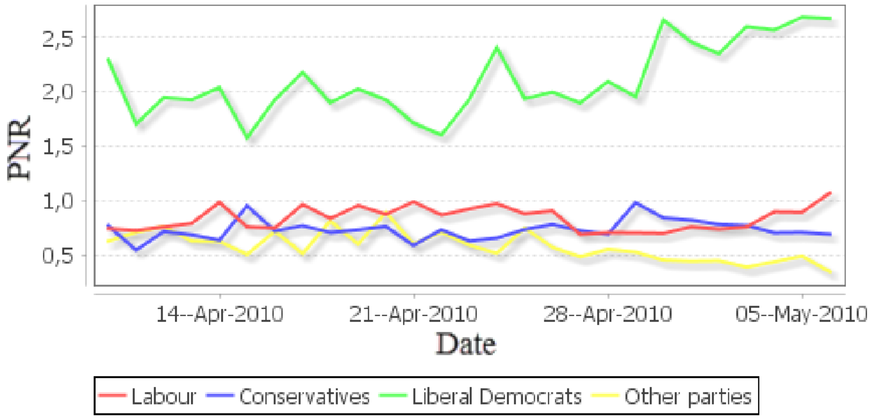

Figure 6: PNR of political parties

4.3 Creation of Predictive Model

This stage creates the actual predictive models. We aim to develop one predictive model per political party. Towards this end we create and compare multiple predictive models for all the combinations of moving average with different k_s and political parties. Specifically, for each political party, we create three models using the moving average technique (k=2, k=3 and k=4) and a model without using the moving average technique, and then compare the models to select the most accurate. We also use historical data from YouGov to train our predictive models. Precisely we create our models using data from the polls of YouGov for the first twenty days of the month before the elections (i.e. 8/4 - 26/4). The data of YouGov regarding the rest nine days (i.e. 27/4 - 5/5) will be used in the next stage of our approach, for the evaluation of the predictive models (see next section).

The predictive method that we use is regression analysis, one of the most common methods in literature [Kalampokis et al., 2013]. The equation for the creation of our regression models is:

$$y_t = b_0 + b_1 x_{1t} + b_2 x_{2t} + \epsilon \tag{3}$$

where x_{1t} and x_{2t} are the independent variables, y_t is the dependent variable and b_0 and b_1 are computed using least squares analysis. ϵ is the error of the regression. In our case, y_t is YouGov's poll for a political party on day t, and x_{1t} and x_{2t} are RF and PNR respectively on day t.

Table 3 presents a summary for each predictive model we created. Each summary includes the p-value and the t-value for each predictor variable and also the R^2 of the model.

Table 3: Summaries of the predictive models for the major political parties

		k=4		k=3		k=2		k=1	
		RF	PNR	RF	PNR	RF	PNR	RF	PNR
	P-value	0.348	0.128	0.255	0.273	0.337	0.732	0.097	0.68
Labour	T-value	0.972	-1.62	1.181	-1.138	0.99	-0.348	1.755	0.419
	R^2	0.018		0.374		0.16		0.02	
	P-value	1.01e-05	0.295	**1.13e-06**	0.749	8.41e-06	0.233	0.0004	0.02
Cons	T-value	6.7	-1.087	7.826	0.326	6.424	1.24	4.38	1.99
	R^2	0.763		0.8		0.72		0.023	
	P-value	7.03e-07	0.94	**5.04e-08**	0.893	4.26e-06	0.95	0.004	0.577
Libdem	T-value	8.467	-0.072	9.994	0.137	6.8	-0.064	3.278	-0.569
	R^2	0.842		0.873		0.743		0.047	

From the table we observe that:

- the p-value of the PNR variable in all regression models is higher than 0.05. This suggests that there is no statistical relation between PNR and the dependent variable.
- the p-value of the RF variable in all regression models created from Labour's data is higher than 0.05. This also suggests that there is no statistical relation between RF and the dependent variable in the Labours' models and, thus, we cannot create a statistically significant model.
- the p-values of the RF variable of Conservatives' and Liberal Democrats' regression models are quite low. This suggests that there is a statistical relation between RF and the dependent variable for the Conservatives and Liberal Democrats political parties.
- R^2 is high in the models of Conservatives and Liberal Democrats. This suggests that these models fit better to the data.
- Regression models that use moving average with k=3 have better results than the rest of the models.

For the above reasons, we decided to use only the models for Conservatives and Liberal Democrats for our predictions using RF as the only independent variable and using the moving average method with k=3.

The following equation represents the model we used for the Conservatives party:

$$y = 0.3856RF + 0.1791 \tag{4}$$

Moreover, the following equation represents the model we used for the Liberal Democrats party:

$$y = 0.5973RF + 0.1191 \tag{5}$$

4.4 Evaluation of the Predictive Performance

The evaluation of the predictive performance is very important as it provides the actual result of the study as a whole.

The evaluation of the predictive performance is often not performed using predictive measures and out-of-sample assessment. It is indicative that half of the studies that were reviewed by [Kalampokis et al., 2013] do not use predictive analytics to draw conclusions on the predictive performance of Social Media. When it comes to elections only four studies [Franch, 2013, DiGrazia et al., 2013, Ceron et al., 2014, Ceron et al., 2013] make use of predictive analytics.

In this paper we employ predictive analytics as an evaluation method. Out-of-sample data from the YouGov opinion polling for the last 9 days before the elections (i.e. 27/4 - 5/5) are employed to predict the elections results.

We employed mean absolute error as a statistical metric for the assessment of our evaluation.

Our two regression models are used to predict the percentage of votes of the Liberal Democrats and the Conservatives' party. Table 4 shows the predicted percentages for the Liberal Democrats and Conservatives party respectively for the 9 days that were used to test the model. We can see that, for the Liberal Democrats, the predicted vote percentage using our independent variable has an average prediction error of 0.024 (or 2.4% of votes) when comparing with the forecast that YouGov made. In the same way, for the party of Conservatives, the predicted vote percentage using our independent variable has an average prediction error of 0.027 (or 2.7% of votes) when comparing with the forecast that YouGov made.

Interestingly however, both our regression models predicted more accurately the percentage of the Liberal Democrats and the Conservatives at the day of the elections. More specifically having in mind that Liberal Democrats gained the 23% of the total UK votes in the elections, our model forecasted 23.9% remarkably better than YouGovs 28% forecast. In the same way, having in mind that Conservatives gained the 36.1% of the total UK votes in the elections, our model forecasted 36.8% which is also better than YouGovs 35% forecast.

Table 4: Prediction of the UK elections' results for the Liberal Democrats and Conservatives party. SE stands for (absolute) Standard Error.

	Liberal Democrats				Conservatives			
Date	YouGov	RF	Predic.	SE	YouGov	RF	Predic.	SE
27/4	0.28	0.267	0.279	0.001	0.33	0.465	0.358	0.028
28/4	0.31	0.226	0.254	0.056	0.34	0.396	0.332	0.008
29/4	0.28	0.187	0.231	0.049	0.34	0.323	0.304	0.036
30/5	0.28	0.23	0.257	0.023	0.34	0.238	0.271	0.069
1/5	0.28	0.282	0.288	0.008	0.35	0.3	0.295	0.055
2/5	0.29	0.266	0.278	0.012	0.34	0.392	0.33	0.010
3/5	0.28	0.233	0.258	0.022	0.35	0.445	0.351	0.001
4/5	0.24	0.199	0.238	0.002	0.35	0.483	0.365	0.015
5/5	0.28	0.2	0.239	0.041	0.35	0.49	0.368	0.018
			AVG:	0.024			AVG:	0.027

5 Discussion

In this paper we presented a case for predicting election results using Twitter data. This case falls into the broader category of studies that exploit Social Media data to predict the outcome of real-world phenomena. The aim of our paper is not to prove the ability of SM to predict these phenomena but to enable understanding to what extend and under which circumstances the prediction is possible. Even in the case of weather forecasting we cannot say that weather is predictable or not because weather forecasts are typically quite reliable looking a few days ahead, but they become increasingly less accurate three, four and five days out [Tetlock and Gardner, 2016]. In this context, we do not claim that our work can be established as a general framework for predicting election results. Introducing such a general framework would require a lot of further work to be done such as evaluating the framework using similar frameworks or other case studies. In this section, however, we discuss our work vis-a-vis other research works in the area in order to gain insight on the challenges related to the prediction of election results through Social Media data analysis.

A number of relevant efforts in literature aim to predict political phenomena using SM [Tumasjan et al., 2010, Gayo-Avello, 2011, He et al., 2012, Jin et al., 2010, Jungherr et al., 2012, Lui et al., 2011, Metaxas et al., 2011, Skoric et al., 2012, Burnap et al., 2016, Bermingham and Smeaton, 2011, Franch, 2013, Di-Grazia et al., 2013, Ceron et al., 2014, Caldarelli et al., 2014, Ceron et al., 2013]. The novelty of the presented case in relation to the literature is two fold. The first one regards the approach that is followed in studies aiming at predicting election results. Our study is the first one that is compatible with all suggestions made by a theoretical framework for social media data analysis [Kalampokis et al., 2013] with regards to the data analysis steps followed and the approaches adopted in each step. In particular our study utilises a dynamic approach to identify keywords for data collection and filtering, employs both volume and sentiment related variables, computes the sentiment variable through a machine-learning algorithm, uses time-series to create the predictive model, and evaluates the predictive performance using predictive metrics and out-of-sample data. The second novelty of our study is related to the utilisation of the Linked Data paradigm to semantically enrich tweets and reuse objective data that is freely available on the Web in order to support the understanding of the tweets. Although this technique has been already proposed in the literature, our study is the first one that adopts it in a case of exploring the predictive power of Social Media.

Two other studies in the literature aim also at predicting the UK election of 2010 through Social Media data [Franch, 2013, He et al., 2012]. Franch [Franch, 2013] used data from Facebook, Twitter, Google and YouTube and computed numerous variables referring to both volume and sentiment. He also created a linear regression prediction model using poll data from YouGov. As regards the vote percentage of Liberal Democrats, the author found the higher prediction accuracy (average of three bivariate models 23.8%) when used Facebook related variables, while he did not present results for Twitter data. As in our case, the result was close to the real outcome (23%) but it presented a big error in relation to the YouGov poll (28%) that was used to train the model. However, the author does not provide information regarding the collection and filtering stage and the computation of the sentiment variables. On the other hand, He et al. [He et al., 2012] collected data from Twitter and applied a lexicon-based approach as well as unsupervised and supervised machine learning algorithms to compute the public sentiment in UK expressed through the tweets for the three major parties. Although, they did not express the variables as time series and they did not evaluate the predictive performance using predictive analytics, they challenged the power of Twitter to predict the results of the UK elections of 2010.

The first issue that may introduce uncertainty in the approach is the Named Entity Recognition (NER) process. As described in the approach we achieved 85.89 F1 score in the identification and classification of the named entities that were included in the collected tweets. Although this score is quite high we should note that we used a supervised classifier, which requires the creation of a manually annotated dataset of entities and tweets. If it is to widely adopt this approach unsupervised classifiers have to be used and evaluated as well. Moreover, the identification of representations in DBpedia of the named entities that were recognised in the tweets may also introduce uncertainty in our approach. In this process we used DBpedia lookup, although other tools can be also used, e.g. Silk Framework [3], which is a very popular tool for discovering relationships between data items within different Linked Data sources. It is indicative that in our case, out of 6392 distinct named entities identified from the NER process only 3712 were matched with one or more DBpedia resources (i.e. 58% of the entities). We should also note that we have performed the same process using Silk but the results were less accurate. Moreover, in the case where more than one resource was returned by DBpedia lookup, we introduced an extra step that involves manual effort. In particular, we created a case-specific bag of words using the representations in DBpedia of the eight most popular UK political parties and we computed the cosine similarity between the identified resources and this case-specific bag of words. Although the domain-specific bag of words can be also created in other problem areas or cases as well, it may introduce uncertainty and produce poor results. For example, this process was able to identify relationships between "Cameron" entity and dbpedia:David_Cameron, "Cleggmania" entity and dbpedia:Nick_Clegg, "Lib_Dems" entity and dbpedia:Liberal_Democrats, and "UK" entity and dbpedia:United_Kingdom. We should however note that there where cases where entities were connected to a relative but not identical DBpedia entity. For example, our mechanism connected Tony_Blair entity with dbpedia:Premiership_of_Tony_Blair entity and not dbpedia:Tony_Blair because these two DBpedia resources produced almost similar cosine similarity with the entity "Tony-Blair", with dbpedia:Premiership_of_Tony_Blair presenting better results. Finally, there were cases were the selected lookup result was wrong. We hence used a cosine similarity threshold in order to avoid establishing those irrelevant connections.

Literature suggests that the use of machine learning approaches in sentiment analysis facilitates the creation of more accurate predictive models than the ones created with lexicon-based sentiment analysis [Kalampokis et al., 2013]. In our study, although we included sentiment as an independent variable in our initial regression models, it proved that this variable is not statistically related to the dependent variable. However, although we used machine learning to compute sentiment, the training dataset was manually annotated using a set of positive and negative hashtags. This fact does not allow us to generalise our results regarding the strength of sentiment as a predictor variable.

We should also note that our case presents another limitation that is related to the volume of the collected tweets. We collected 84,375 tweets using four relevant to the election hash tags, while for example He et al [He et al., 2012] collected 919,662 tweets using more hash tags. Using more hashtags would be useful for collecting more relevant tweets.

Another limitation of this work is the assumption we made regarding the

[3] http://silkframework.org/

location of the tweets. In particular, we assumed that all tweets have posted from the UK. This assumption could make sense in 2010 when tweets were collected. However today, especially when it comes to international events that take place in a specific country, a large number of relevant tweets are usually published from users outside this country.

In general, a predictive model enables accurate prediction of phenomenon outcomes based on a new set of observations. In election related cases, however, it is difficult to evaluate a model with a new set of election results because one has to wait four or five years until the next election to take place. In this paper, we assume that YouGov opinion polling represents an alternative expression of political preference, similar to voting. We hence created our model using data from YouGov and evaluate the model based on both the prediction of YouGov and the actual election results. We should note, however, that, in the case of using the YouGov poll for the evaluation, the YouGov prediction of the last day before the elections differ from the actual voting results with Liberal Democrats presenting a difference of 5% and Conservatives a difference of 1% in the voting percentage. This fact should be carefully analysed when interpreting the accuracy of such a predictive model.

6 Conclusion

During the last years a number of studies exploited Twitter data to predict the outcome of real-world phenomena such as elections, stock market, oscar awards and product sales. The results of these studies are controversial with elections being the main phenomenon of dispute. The black box approach that were followed in some case has been criticised as the main cause of the contradicting results, while some later works identified the steps and the approaches that affect the prediction accuracy.

In this paper we aim at analysing Twitter data to predict the outcome of the UK's 2010 general election. Towards this end, we employ an approach that is based on (a) a generic theoretical Social Media data analysis framework for predictions to define the steps of the analysis and the approaches followed in each step, and (b) Linked Open Data to semantically enrich tweets and thus support their interpretation.

The results of the analysis suggest that the vote percentage of the two out of three major parties can be accurately predicted through the volume of tweets. The limitations of our method include the accuracy of both the Named Entity Recognition and linking operations, the small volume of collected tweets, and the use of opinion polling data for creating the model.

As in all efforts that aim to predict real-world phenomena, the real value of this case study comes from understanding to what extend and under which circumstances we can predict the elections. We anticipate that both the approach we follow and the results will contribute to the ongoing research in the area and enable researchers to design new studies and propose future research directions.

References

[Abel et al., 2011] Abel, F., Celik, I., Houben, G.-J., Siehndel, P. (2011). Leveraging the semantics of tweets for adaptive faceted search on twitter. In Aroyo, L., Welty, C., Alani, H., Taylor, J., Bernstein, A., Kagal, L., Noy, N., Blomqvist, E., editors, *The Semantic Web – ISWC 2011*, volume 7031 of *Lecture Notes in Computer Science*, pages 1–17. Springer Berlin Heidelberg.

[Asur and Huberman, 2010] Asur, S. Huberman, B. A. (2010). Predicting the future with social media. In *Proceedings of the 2010 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology - Volume 01*, WI-IAT '10, pages 492–499, Washington, DC, USA. IEEE Computer Society.

[Bermingham and Smeaton, 2011] Bermingham, A. Smeaton, A. F. (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), Asian Federation of Natural Language Processing*, page 210.

[Berners-Lee, 2006] Berners-Lee, T. (2006). Design issues: Linked data.

[Bizer et al., 2009] Bizer, C., Heath, T., Berners-Lee, T. (2009). Linked data - the story so far. *International Journal on Semantic Web and Information Systems*, 5(3):1–22.

[Bollen et al., 2011] Bollen, J., Mao, H., Zeng, X. (2011). Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1 – 8.

[Bothos et al., 2010] Bothos, E., Apostolou, D., Mentzas, G. (2010). Using social media to predict future events with agent-based markets. *IEEE Intelligent Systems*, 25(6):50–58.

[Burnap et al., 2016] Burnap, P., Gibson, R., Sloan, L., Southern, R., Williams, M. (2016). 140 characters to victory?: Using twitter to predict the {UK} 2015 general election. *Electoral Studies*, 41:230 – 233.

[Caldarelli et al., 2014] Caldarelli, G., Chessa, A., Pammolli, F., Pompa, G., Puliga, M., Riccaboni, M., Riotta, G. (2014). A multi-level geographical study of italian political elections from twitter data. *PloS one*, 9:e95809.

[Ceron et al., 2013] Ceron, A., Curini, L., Iacus, S. (2013). To what extent sentiment analysis of twitter is able to forecast electoral results? evidence from france, italy and the united states. In *ECPR General Conference*, pages 5–8.

[Ceron et al., 2014] Ceron, A., Curini, L., Iacus, S. M., Porro, G. (2014). Every tweet counts? how sentiment analysis of social media can improve our knowledge of citizens political preferences with an application to italy and france. *New Media & Society*, 16(2):340–358.

[Chunara et al., 2012] Chunara, R., Andrews, J. R., Brownstein, J. S. (2012). Social and news media enable estimation of epidemiological patterns early in the 2010 haitian cholera outbreak. *The American Journal of Tropical Medicine and Hygiene*, 86(1):39–45.

[DiGrazia et al., 2013] DiGrazia, J., McKelvey, K., Bollen, J., Rojas, F. (2013). More tweets, more votes: Social media as a quantitative indicator of political behavior. *PloS one*, 8(11):e79449.

[Franch, 2013] Franch, F. (2013). (wisdom of the crowds)2: 2010 uk election prediction with social media. *Journal of Information Technology & Politics*, 10(1):57–71.

[Gayo-Avello, 2011] Gayo-Avello, D. (2011). Don't turn social media into another 'literary digest' poll. *Commun. ACM*, 54(10):121–128.

[Ginsberg et al., 2009] Ginsberg, J., Mohebbi, M. H., Patel, R. S., Brammer, L., Smolinski, M. S., Brilliant, L. (2009). Detecting influenza epidemics using search engine query data. *Nature*, 457(7232):1012–1014.

[Hausenblas, 2009] Hausenblas, M. (2009). Exploiting linked data to build web applications. *IEEE Internet Computing*, 13(4):68–73.

[He et al., 2012] He, Y., Saif, H., Wei, Z., Wong, K.-F. (2012). Quantising opinions for political tweets analysis. In Association, E. L. R., editor, *Eight International Conference on Language Resources and Evaluation*, pages 3901–3906.

[Hughes and Palen, 2009] Hughes, A. L. Palen, L. (2009). Twitter adoption and use in mass convergence and emergency events. *International Journal of Emergency Management*, 6(3):248–260.

[Java et al., 2007] Java, A., Song, X., Finin, T., Tseng, B. (2007). Why we twitter: Understanding microblogging usage and communities. In *Proceedings of the 9th WebKDD and 1st SNA-KDD 2007 Workshop on Web Mining and Social Network Analysis*, WebKDD/SNA-KDD '07, pages 56–65, New York, NY, USA. ACM.

[Jin et al., 2010] Jin, X., Gallagher, A., Cao, L., Luo, J., Han, J. (2010). The wisdom of social multimedia: Using flickr for prediction and forecast. In *Proceedings of the International Conference on Multimedia*, MM '10, pages 1235–1244, New York, NY, USA. ACM.

[Jungherr et al., 2012] Jungherr, A., Jürgens, P., Schoen, H. (2012). 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*, 30(2):229–234.

[Kalampokis et al., 2016] Kalampokis, E., Karamanou, A., Tambouris, E., Tarabanis, K. (2016). Applying brand equity theory to understand consumer opinion in social media. *Journal of Universal Computer Science*, 22(5):709–734. `http://www.jucs.org/jucs_22_5/applying_brand_equity_theory`.

[Kalampokis et al., 2013] Kalampokis, E., Tambouris, E., Tarabanis, K. (2013). Understanding the predictive power of social media. *Internet Research*, 23(5):544–559.

[Kwak et al., 2010] Kwak, H., Lee, C., Park, H., Moon, S. (2010). What is twitter, a social network or a news media? In *Proceedings of the 19th International Conference on World Wide Web*, WWW '10, pages 591–600, New York, NY, USA. ACM.

[Livne et al., 2011] Livne, A., Simmons, M. P., Adar, E., Adamic, L. A. (2011). The party is over here: Structure and content in the 2010 election. In *ICWSM*.

[Lui et al., 2011] Lui, C., Metaxas, P. T., Mustafaraj, E. (2011). On the predictability of the us elections through search volume activity. In *IADIS International Conference e-Society*, pages 165–172.

[Mendes et al., 2010] Mendes, P., Passant, A., Kapanipathi, P., Sheth, A. (2010). Linked open social signals. In *Web Intelligence and Intelligent Agent Technology (WI-IAT), 2010 IEEE/WIC/ACM International Conference on*, volume 1, pages 224–231.

[Metaxas et al., 2011] Metaxas, P., Mustafaraj, E., Gayo-Avello, D. (2011). How (not) to predict elections. In *Privacy, security, risk and trust (passat), 2011 ieee third international conference on and 2011 ieee third international conference on social computing (socialcom)*, pages 165 –171.

[Naaman et al., 2010] Naaman, M., Boase, J., Lai, C.-H. (2010). Is it really about me?: Message content in social awareness streams. In *Proceedings of the 2010 ACM Conference on Computer Supported Cooperative Work*, CSCW '10, pages 189–192, New York, NY, USA. ACM.

[Nadeau and Sekine, 2007] Nadeau, D. Sekine, S. (2007). A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26.

[Polgreen et al., 2008] Polgreen, P. M., Chen, Y., Pennock, D. M., Nelson, F. D., Weinstein, R. A. (2008). Using internet searches for influenza surveillance. *Clinical Infectious Diseases*, 47(11):1443–1448.

[Ritterman et al., 2009] Ritterman, J., Osborne, M., Klein, E. (2009). Using prediction markets and twitter to predict a swine flu pandemic. In *Proceedings of the 1st International Workshop on Mining Social Media*, pages 9–17.

[Rowe and Stankovic, 2012] Rowe, M. Stankovic, M. (2012). Aligning tweets with events: Automation via semantics. *Semantic Web*, 3(2):115–130.

[Schmachtenberg et al., 2014] Schmachtenberg, M., Bizer, C., Paulheim, H. (2014). Adoption of the linked data best practices in different topical domains. In Mika, P., Tudorache, T., Bernstein, A., Welty, C., Knoblock, C., Vrandečić, D., Groth, P., Noy, N., Janowicz, K., Goble, C., editors, *The Semantic Web – ISWC 2014*, volume 8796 of *Lecture Notes in Computer Science*, pages 245–260. Springer International Publishing.

[Skoric et al., 2012] Skoric, M., Poor, N., Achananuparp, P., Lim, E.-P., Jiang, J. (2012). Tweets and votes: A study of the 2011 singapore general election. In *System Science (HICSS), 2012 45th Hawaii International Conference on*, pages 2583–2591.

[Tetlock and Gardner, 2016] Tetlock, P. Gardner, D. (2016). *Superforecasting: The art and science of prediction.* Random House.

[Tumasjan et al., 2010] Tumasjan, A., Sprenger, T., Sandner, P., Welpe, I. (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*, pages 178–185.

[Wu et al., 2011] Wu, S., Hofman, J. M., Mason, W. A., Watts, D. J. (2011). Who says what to whom on twitter. In *Proceedings of the 20th International Conference on World Wide Web*, WWW '11, pages 705–714, New York, NY, USA. ACM.