Predicting Elections from Social Networks Based on Sub-event Detection and Sentiment Analysis

Sayan Unankard[1], Xue Li[1], Mohamed Sharaf[1], Jiang Zhong[2], and Xueming Li[2]

[1] School of Information Technology and Electrical Engineering,
The University of Queensland, Brisbane QLD 4072, Australia
[2] Key Laboratory of Dependable Service Computing in Cyber Physical Society
Ministry of Education, Chongqing 400044, China
{uqsunank,m.sharaf}@uq.edu.au, xueli@itee.uq.edu.au,
{zhongjiang,lixuemin}@cqu.edu.cn

Abstract. Social networks are widely used by all kinds of people to express their opinions. Predicting election outcomes is now becoming a compelling research issue. People express themselves spontaneously with respect to the social events in their social networks. Real time prediction on ongoing election events can provide feedback and trend analysis for politicians and news analysts to make informed decisions. This paper proposes an approach to predicting election results by incorporating sub-event detection and sentiment analysis in social networks to analyse as well as visualise political preferences revealed by those social network users. Extensive experiments are conducted to evaluate the performance of our approach based on a real-world *Twitter* dataset. Our experiments show that the proposed approach can effectively predict the election results over the given baselines.

Keywords: election prediction, event detection, sentiment analysis, micro-blogs.

1 Introduction

Micro-blog services such as *Twitter* generate a large amount of messages carrying event information and users' opinions over a wide range of topics. The events discussed on social networks can be associated with topics, locations, and time periods. The events can be in a variety, such as celebrities or political affairs, local social events, accidents, protests, or natural disasters. Messages are posted by users after they have experienced or witnessed the events happening in the real world and they want to share their experiences immediately. For a long-running event like a nation-wide election which usually has fixed start and end times, users may want to monitor sub-events (i.e., hierarchically nested events that break down an event into more refined parts) such as the debate or campaign-launch speech. Alternatively, policy-makers may want to know the feeling of users during the course of an election. The new research in computer science, sociology and political science shows that data extracted from social media platforms yield

B. Benatallah et al. (Eds.): WISE 2014, Part II, LNCS 8787, pp. 1–16, 2014.

accurate measurements of public opinion. It turns out that what people say on *Twitter* is a very good indicator of how they would vote in an election [1,2,3,4].

Existing studies have focused on counting of preferences or sentiment analysis on a party or candidate. They neglect the fact that the voters' attitudes and opinions of people may be different depending on specific political topics and in different geographic areas. Moreover, the same voters participating in different discussions may have different political preferences. In this paper, we are interested in predicting the result of elections from micro-blog data by incorporating sub-event detection and sentiment analysis to detect their political preferences and predict the election results at a state as well as a national level.

The main contributions of this paper are as follows. (1) We present an approach to forecast the vote of a sample user based on the analysis of his/her micro-blog messages and count the votes of users to predict the election results. (2) Sub-event detection and sentiment analysis are incorporated to predict the vote of users as different level of sub-events user engaged in the discussions will affect the prediction results. We evaluate our proposed approach with a real-world *Twitter* data posted by Australia-based users during the 2013 Australian federal election.

The rest of the paper is organized as follows. First, we describe the related work in Section 2. Second, the proposed approach is presented in Section 3. Third, we present the experimental setup and results in Section 4. Finally, the conclusions are given in Section 5.

2 Related Work

2.1 Election Prediction on Social Networks

Twitter is a micro-blog service that has been attracting growing attention from researchers in Data Mining and Information Retrieval. Recently, extensive research has been done on social networks in election prediction [1,2,3,4].

O'Connor et al. in [1] presented the feasibility of using *Twitter* data as a substitute and supplement for traditional polls. Subjectivity lexicon is used to determine opinion scores (i.e., positive and negative scores) for each message in the dataset. Then, the authors computed a sentiment score. Consumer confidence and political opinion are analysed and found to be correlated with sentiment word frequencies in *Twitter* data. However, they do not describe any prediction method. Tumasjan et al. in [2] examined whether *Twitter* can be seen as a valid real-time indicator of political sentiment. The authors also found that the mere number of messages reflects the election result and comes close to traditional election polls. Sang et al. in [3] analysed *Twitter* data regarding the 2011 Dutch Senate elections. The authors presented that improving the quality of the document collection and performing sentiment analysis can improve performance of the prediction. However, the authors need to manually annotate political messages to compute sentiment weight and only the first message of every user is taken into account. In addition, the method relies on polling data to correct for demographic bias. Makazhanov et al. in [4] proposed political preference prediction models based on a variety of contextual and behavioural features. The authors extract all interactions of the candidates, group them on a per-party basis, and build a feature vector for each group. Both a decision tree-based J48 and Logistic regression classifiers are utilized for each party. However, this method needs labelling of training examples for each user. The labelling of training set based on a set of users whose political preferences are known based on the explicit statements (e.g., *"I voted XXX today!"*) made on the Election Day or soon after. Moreover, it does not predict the election outcomes.

However, there are several works presented the problems on election prediction using *Twitter* data. Jungherr et al. in [5] presented that a lack of well-grounded rules for data collection and the choice of parties and the correct period in particular can cause the problems. Metaxas et al. in [6] concluded that *Twitter* data is only slightly better than chance when predicting elections. However, the authors described three necessary standards for predicting elections using *Twitter* data: (1) it should be a clearly defined algorithm, (2) it should take into account the demographic differences between *Twitter* and the actual population, and (3) black-box methods should be avoided. Gayo-Avello has criticized several flaws in [7]. For example, there is not a commonly accepted way of counting votes in *Twitter.* Sentiment analysis is applied as a black-box and demographics are neglected. Nevertheless, the author has outlined some of the research lines for future works in this topic. For example, researches need to clearly define which are a vote and the ground truth; sentiment analysis is a core task and researches should acknowledge demographic bias.

2.2 Sub-Event Detection from Social Networks

There are a few research works on search and retrieval of relevant information from social networks [8,9]. Abel et al. in [8] introduced *Twitcident*, a framework for filtering, searching and analysing information about real-world incidents or crises. Given an incident, the system automatically collects and filters relevant information from *Twitter.* However, this work focuses on how to enrich the semantics of *Twitter* messages to improve the incident profiling and filtering rather than detecting sub-events. A research which is similar to our work is presented by Marcus et al. in [9]. A system for visualizing and summarizing events on *Twitter* in real-time, namely *TwitInfo*, is proposed. The system detects sub-events and provides an aggregate view of user sentiment. Sub-events are extracted by identifying temporal peaks in message frequency and by using weighted moving average and variance to detect an outlier as a sub-event. The *Naïve Bayes* classifier is used to analyse the sentiment of messages into positive and negative via *unigram* features. Training datasets are generated for the positive and negative classes using messages with happy and sad *emoticons. Emoticon* is a representation of a facial expression such as a smile or frown, formed by various combinations of keyboard characters and used in electronic communications to convey the writer's feelings or intended tone.

2.3 Sentiment Analysis on Social Networks

There are several research papers discussing sentiment analysis via lexicon-based approaches [10,11]. Meng et al. in [10] presented an entity-centric topic-based opinion summarization framework in *Twitter*. Topic is detected from *hashtags* – human annotated tags for providing additional context and metadata to messages. Target-dependent sentiment classification is used to identify the sentiment orientation of a message. Recent researches in the field of political sentiment analysis are presented by Wang et al. in [11] and Ringsquandl et al. in [12]. A similar work to our approach is introduced in [12]. This work studies the application of the Pointwise Mutual Information measure to extract relevant topics from *Twitter* messages. Unsupervised sentiment classification is proposed; the semantic orientation of word is the most probable class (positive, negative, neutral) of each opinion word according to *synsets* (i.e., synonym) in *WordNet*[1]. The final aspect-level sentiment is determined by a simple aggregation function which sums the semantic orientation of all words in the message that mentions the specific aspect.

3 Proposed Approach

In order to understand whether the activity on *Twitter* can serve as a predictor of the election results, we propose an approach to incorporate sub-event detection and sentiment analysis for each sub-event for predicting user's political preference. The proposed approach consists of three main components: sub-event detection, sentiment analysis and the prediction model. We collected the *Twitter* messages related to the 2013 Australian federal election event to demonstrate our approach. The following information provides details of each component.

3.1 Sub-Event Detection for a Particular Event

The notion of event detection was proposed in our recent work [13] for location-based hotspot emerging events. However, the problem that we address in this paper is how to group a set of micro-blog messages into a cluster (or sub-event) for a particular longer-running event (i.e., an election). The user defines an event by specifying a keyword query. For example, search keywords such as *"election", "Kevin Rudd", "Tony Abbott", "#ausvote" and "#auspol"* are used to collect the data of the 2013 Australian federal election. In the following, we brief the techniques for sub-event detection.

It has three steps as we are not consider the emergence of event. Firstly, the pre-processing was designed to ignore common words that carry less important meaning than keywords and to remove irrelevant data e.g., *re-tweet* keyword, web address and message-mentioned username. Slang word and extensions like "booooored" are replaced by proper English words. The stop words are removed

[1] `http://wordnet.princeton.edu`

and all words are stemmed by using *Lucene 3.1.0 Java API*[2]. Message location identification is conducted in order to understand users' opinions in particular areas. We firstly extract message location from the *geo-tagged* (latitude/longitude) information. If *geo-tagged* information is not available we extract user location in the user profile to query the Australia *Gazetteer* database for acquiring the location's address. Then, if neither of them is available we set user location equal to "Australia".

Secondly, for clustering step, we consider a set of messages where each message is associated with a sub-event. With the number of sub-events being unknown in advance, we applied event detection using hierarchical clustering from our previous work [13] with some modifications. We use a sliding window to divide the messages. The size of the sliding window is defined in time intervals (i.e., one day for our experiment). According to our experiment, the clustering method performs well when using the augmented normalized term frequency and cosine similarity function. The cosine similarity function is used to calculate the similarity between the existing cluster and the new message. Every message is compared with all previous cluster's centroids. The algorithm creates a new cluster for the message if there is no cluster whose similarity to the message is greater than the threshold (α). In order to find the most suitable value for the threshold, we conducted the clustering experiments with different threshold values. Our tests show that when $\alpha = 0.30$ it renders the best performance. The mean is used to represent the centroid of the cluster, which trades memory use for speed of clustering.

Finally, after the clustering is performed, all clusters cannot be assigned as event clusters because they can be private conversations, advertisements or others. A cluster can be considered as sub-event if there is strong correlation between the event location (i.e., location mentioned in the messages) and the user location. For event location identification, we find all terms or phrases which reference geographic location (e.g., country, state and city) from message contents. We simply extract the message-mentioned locations via *Named Entity Recognition (NER)*. We use the *Stanford Named Entity Recognizer* [14] to identify locations within the messages. We also use the *Part-of-Speech Tagging* for *Twitter* which is introduced in [15] to extract proper nouns. We use an extracted terms query into the *Gazetteer* database to obtain candidate locations of the event. We find the most probable location of the event using the frequency of each location in the cluster. The location which has the highest frequency is assigned as the event location. In order to understand what the sub-event cluster is about, we find the set of keywords to represent the sub-event topic. To extract the set of co-occurring keywords, firstly we create a directed, edge-weighted graph. We adopt the smoothed correlation weight function, to calculate the semantic correlation weight between terms. We identify the sub-event topic by extracting the *Strongly Connected Components (SCCs)* from the graph. The details of our algorithm are presented in [13].

[2] http://lucene.apache.org

3.2 Political Sentiment Analysis

In general, opinions can be expressed about anything, such as a product, service, person, topic or event and by any person or organization. Entity is used to denote the opinion target. For example, the targets/entities of messages likes *"As much as you dislike XXX please Australia...Hate YYY more! I beg you"* are *"XXX"* and *"YYY"*. Sentiment analysis can be a supervised approach or an unsupervised approach or a combination of the two. In the supervised approach, the process of labelling training datasets requires considerable time and effort. Collecting training datasets for all application domains is very time consuming and difficult. In this paper, we focus on a lexicon-based approach to perform sentiment classification. However, spotting the target/entity in a micro-blog message is not the focus of this paper. Our method has two steps. First, an opinion lexicon is constructed and then, the opinion is classified, based on a statistical calculation.

For sentiment analysis, the pre-processing is conducted. We performed the part of speech (*POS*) processing from the original messages. We use *Twitter NLP and Part-of-Speech Tagging* proposed by Gimpel et al. in [15] for tagging the messages. Moreover, the *emoticons* are extracted from the messages. Finally, all messages after being tagged are stored in the database.

1) Opinion Lexicon: We used the lexicon dictionary which was introduced in [16]. It consists of 4,783 negative and 2,006 positive, distinct words. However, micro-blog messages are informally written and often contain slang words and abbreviations. The traditional lexicon dictionary does not cover opinion words in micro-blogs. In order to expand the lexicon dictionary, we manually annotated the Internet slang dictionary, downloaded from http://www.noslang.com, into 262 positive and 903 negative slang words. *Emoticons*[3] are also grouped into happy and unhappy facial expressions.

2) Lexicon-based Algorithm: Our algorithm assigns the messages into positive, negative and neutral classes. Given a message, the tasks are divided into three steps: word-level sentiment, aspect-level sentiment and sarcasm identification.

Word-level sentiment: This step aims to mark all opinion words or phrases in the message. Each positive word is assigned an opinion score of +1 while each negative word is assigned the score of −1. We extracted adjectives, adverbs, verbs, nouns, interjections and *hashtags* to assign the opinion score. Also, the happy emoticon is assigned the opinion score of +1 and vice versa. In order to detect a phrase, we applied natural language rules which are shown in Table 1.

In this step, it is important to deal with complex linguistic constructions, such as negation, intensification, diminishes and modality because of their effect on the emotional meaning of the text. Negation and modality are computed in the same way. We defined the rules for negation and intensification as follows. For negation (e.g., "no", "not" and "never"), there are three cases to compute an opinion score (*OS*) of a given phrase.

[3] http://en.wikipedia.org/wiki/List_of_emoticons

Table 1. Natural language rules for phrase detection

Rule	Example
Adverb + Adjective	not good, very sad
Comparative Adverb + Adjective	more offensive, more sincere
Adverb + Verb	not vote, never truth
Intensifier/Diminishes + Adverb	really good, slightly nervous
Modals Verb + Verb	can't promise, can't believe

(1) Negation + Neg. e.g., "not bad"; $OS = +1$
(2) Negation + Pos. e.g., "not good"; $OS = -1$
(3) Negation + Neu. e.g., "not work"; $OS = -1$

Intensifiers (e.g., "very", "really" and "extremely") increase the semantic intensity of a neighbouring lexical item, whereas diminishes (e.g., "quite", "less", "slightly") decrease it. The opinion score of a phrase is computed as follows.

(1) Intensifier + Neg. e.g., "very bad"; $OS = -1.5$
(2) Intensifier + Pos. e.g., "very good"; $OS = 1.5$
(3) Diminishes + Neg. e.g., "slightly mad"; $OS = -0.5$
(4) Diminishes + Pos. e.g., "quite good"; $OS = 0.5$

Aspect-Level Sentiment: In this step we aim to compute the opinion orientation for each aspect/target. For the message likes *"As much as you dislike XXX please Australia...Hate YYY more! I beg you"*, we want to extract a pair of opinion word and the aspect such as {*"dislike"* and *"XXX"*} and {*"hate"* and *"YYY"*} then we can calculate the aspect-level score. We applied an opinion aggregation function to assign the final opinion orientation for each aspect in the message. Each aspect has many names that refer to it, even within the same message and clearly, across messages. For example, {*"Tony Abbott", "Abbott" and "TonyAbbottMHR"*} refer to the same person who is one of the candidates of the 2013 Australian federal election. As extracting the aspect/target in microblog messages is not the focus of this paper, we simply set the aspects of our experiments to two sets of keywords as follows:

A_1 = {*"Tony Abbott","Abbott","TonyAbbottMHR"*},
A_2 = {*"Kevin Rudd", "Rudd", "KRudd", "KRuddPM"*}

Every word opinion score is computed related to its distance to the aspect. The number of words between the current word and the aspect (i.e., the matched keywords in the aspect keyword set) is assigned as the distance of the current word to the aspect. The aspect-level score is computed as:

$$asp\_score(m, A) = \sum_{w_i \in m} \frac{opinion\_score_{w_i}}{min(distance(w_i, a)), a \in A} \tag{1}$$

where m is the message, A is the set of aspect keywords, w_i is the word in the messages m and a is the aspect keyword in A. The aspect sentiment is positive

Table 2. The statistical information of sarcasm messages

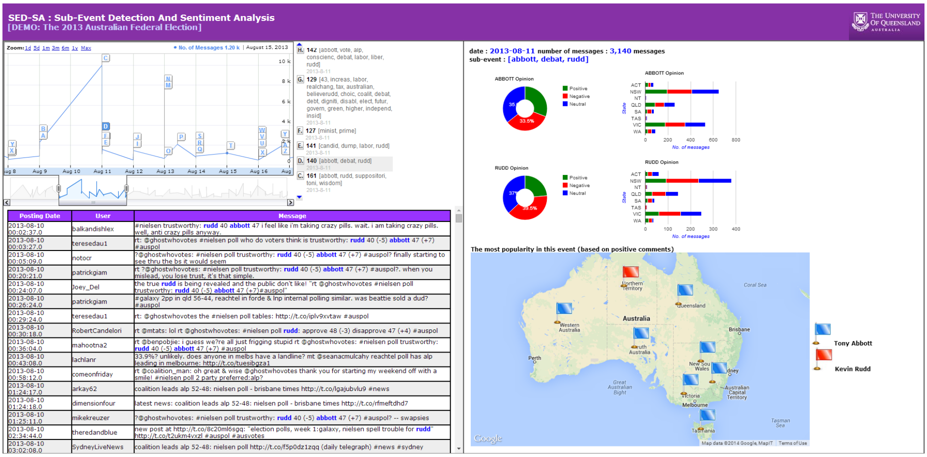
List	Kevin Rudd	Tony Abbott
No. of messages	1,481	3,254
No. of users	959	1,737
No. of users who posted sarcastic messages	48	114
% of users who posted negative sarcasm	100.00%	100.00%
% of users who have the same opinions in every message for a given topic/event	89.58%	92.98%
% of users who have both positive and negative messages for a given topic/event	10.42%	7.02%

if $asp\_score(m, A) > 0$, and is negative when $asp\_score(m, A) < 0$. Otherwise, the aspect sentiment is neutral.

Sarcasm Identification: In addition, micro-blog messages also contain extensive use of irony and sarcasm, which are particularly difficult for a machine to detect [17]. Sarcasm transforms the polarity of the message into its opposite. Negative sarcasm is a message that sounds positive but is intended to convey a negative attitude. Positive sarcasm is a message that sounds negative but is apparently intended to be understood as positive. Watching people's faces while they talk is a good way to pick up on sarcasm. However, it is very difficult to detect sarcasm in writing due to lack of intonation and facial expressions.

In order to understand the sarcastic messages in micro-blogs, we conducted statistical studies. We manually labelled 5,735 messages sent by users around Australia related to one sub-event (i.e., the first debate of the 2013 Australian federal election between *Kevin Rudd* and *Tony Abbott* on 11 August 2013 from 6pm to 9pm). There are 1,481 and 3,254 messages which discussed *Kevin Rudd* and *Tony Abbott*, respectively. The messages are annotated with the polarity being positive, negative or neutral and are also marked as sarcastic messages where applicable. The statistical information for sarcasm is shown in Table 2.

As we can see from Table 2, most users hold negative views on sarcastic messages. Our interest in this task is to mark off whether a message is intended to be sarcastic and assign the polarity of the message. Considering a single message, it is very difficult to classify sarcasm, even for humans. In general, a message like *"XXX: Road is the future of transport! Brilliant."* will be considered as a positive opinion; however, some people in developed countries might think this is a sarcastic message as they have too many roads now. Therefore, the message itself cannot be effective to predict sarcastic message. The previously messaged opinions of the author may help to classify whether the current message tends to be sarcastic or not. However, some people may have different opinions on different topics/sub-events. Based on our observation on sarcasm in micro-blogs we found that most of the micro-blog users have only one opinion on a specific topic or event (89.58% and 92.98% of messages related to *Kevin Rudd* and *Tony Abbott* respectively).



SED-SA : Sub-Event Detection And Sentiment Analysis
[DEMO: The 2013 Australian Federal Election]

Zoom: 1d 5d 1m 3m 6m 1y Max • No. of Messages 1.20 k | August 15, 2013

H. 142 [abbott, vote, alp, conscienc, debat, labor, liber, rudd] 2013-8-11
G. 129 [43, increas, labor, realchang, tax, australian, believerudd, choic, coalit, debat, debt, digniti, disabl, elect, futur, govern, green, higher, independ, insid] 2013-8-11
F. 127 [minist, prime] 2013-8-11
E. 141 [candid, dump, labor, rudd] 2013-8-11
D. 140 [abbott, debat, rudd] 2013-8-11
C. 161 [abbott, rudd, suppositori, toni, wisdom] 2013-8-11

Posting Date	User	Message
2013-08-10 00:02:37.0	balkandishlex	#nielsen trustworthy: rudd 40 abbott 47 i feel like i'm taking crazy pills. wait. i am taking crazy pills. well, anti crazy pills anyway.
2013-08-10 00:03:27.0	teresedau1	rt: @ghostwhovotes #nielsen poll who do voters think is trustworthy: rudd 40 (-5) abbott 47 (+7) #auspol
2013-08-10 00:05:09.0	notocr	?@ghostwhovotes: #nielsen poll trustworthy: rudd 40 (-5) abbott 47 (+7) #auspol? finally starting to see thru the bs it would seem
2013-08-10 00:20:21.0	patrickgiam	rt ?@ghostwhovotes: #nielsen poll trustworthy: rudd 40 (-5) abbott 47 (+7) #auspol?. when you mislead, you lose trust, it's that simple.
2013-08-10 00:24:07.0	Joey_Del	the true rudd is being revealed and the public don't like! "rt @ghostwhovotes #nielsen poll trustworthy: rudd 40 (-5) abbott 47 (+7)#auspol"
2013-08-10 00:26:24.0	patrickgiam	#galaxy 2pp in qld 56-44, reachtel in forde & lnp internal polling similar. was beattie sold a dud? #auspol
2013-08-10 00:29:24.0	teresedau1	rt: @ghostwhovotes the #nielsen poll tables: http://t.co/iplv9xvtaw #auspol
2013-08-10 00:30:18.0	RobertCandelori	rt @mtats: lol rt @ghostwhovotes: #nielsen poll rudd: approve 48 (-3) disapprove 47 (+4) #auspol
2013-08-10 00:36:04.0	mahootna2	rt @benpobjie: i guess we?re all just frigging stupid rt @ghostwhovotes: #nielsen poll trustworthy: rudd 40 (-5) abbott 47 (+7) #auspol
2013-08-10 00:43:08.0	lachlanr	33.9%? unlikely. does anyone in melbs have a landline? mt @seanacmulcahy reachtel poll has alp leading in melbourne: http://t.co/tuesibgza1
2013-08-10 00:58:12.0	comeonfriday	rt @coalition_man: oh great & wise @ghostwhovotes thank you for starting my weekend off with a smile! #nielsen poll 2 party preferred:alp?
2013-08-10 01:24:17.0	arkay62	coalition leads alp 52-48: nielsen poll - brisbane times http://t.co/lgajubvlu9 #news
2013-08-10 01:24:18.0	dimensionfour	latest news: coalition leads alp 52-48: nielsen poll - brisbane times http://t.co/rfmeftdhd7
2013-08-10 01:25:11.0	mikekreuzer	?@ghostwhovotes: #nielsen poll trustworthy: rudd 40 (-5) abbott 47 (+7) #auspol? -- swapsies
2013-08-10 02:34:44.0	theredandblue	new post at http://t.co/8c20ml6sgq: "election polls, week 1:galaxy, nielsen spell trouble for rudd" http://t.co/t2ukm4vxzl #auspol #ausvotes
2013-08-10 03:02:08.0	SydneyLiveNews	coalition leads alp 52-48: nielsen poll http://t.co/f5p0dz1zqq (daily telegraph) #news #sydney

date : 2013-08-11 number of messages : 3,140 messages
sub-event : [abbott, debat, rudd]

The most popularity in this event (based on positive comments)

Fig. 1. A dashboard to display sub-event and sentiment of two specific candidates

Therefore, a reasonable ways to classify sarcastic messages are to consider a specific facial expression (i.e., *emoticon* expression) and to compare them with the author's previous messages in the same topic or sub-event. To address this issue, the *emoticon* expression will be compared with the message polarity. All messages accompanied with an *emoticon* are computed as follows.

(1) Pos. message + Neg. *emoticon*; $polarity = -1$

(2) Neg. message + Pos. *emoticon*; $polarity = +1$

If there are no *emoticons* in the message, we compare the aspect opinion score of the current message with the previous messages of the author in the same sub-event and within the same interval of time (i.e., the size of the sliding window). If the current message opinion differed from the overall opinions of previous messages in the same sub-event (i.e., greater than 90%), we change the aspect opinion score of the current message to the same as that for the previous message's opinion. However, if the opinions of the previous messages are divergent, the current message opinion is not changed because it is surmised that this author tends to have different opinions on the same sub-event.

For usability and understanding issues of visualizing the model, we designed a dashboard to display sub-event and sentiment of two specific candidates. Sub-events are presented via *Annotated Time Line Chart* as show in Figure 1 (left) for each day (represented by letters A to Z). The sub-event name is represented by a keywords list described in 3.1. Figure 1 (right) displays how people feel about specific opinion targets for a given sub-event.

3.3 Election Prediction Model

In order to predict the election results, we learn from the professional pollsters. Our prediction model can be divided into two parts; sampling process and user's

Table 3. Minimum sample size for prediction model

State	Enrolment	Twitter users in our dataset	Minimum sample size
New South Wales	4,816,991	13,471	349
Victoria	3,715,925	12,233	270
Queensland	2,840,091	5,360	206
Western Australia	1,452,272	2,630	105
South Australia	1,130,388	2,234	82
Tasmania	362,892	314	26
Australian Capital Territory	265,269	1,683	19
Northern Territory	128,971	268	10
Total	**14,712,799**	**38,193**	**1,067**

vote prediction. The messages since announce Election day (i.e., 4 August 2013) until the day before Election day (i.e., 6 September 2013) were used for predicting the results. Also, we decided to predict the two-party-preferred vote as in Australian politics the candidates will be from the two major parties.

Sampling Process: Since no one can be sure that who will actually vote, the prediction can be approximated by sampling those who will likely to vote. The most important aspect of correct prediction is the selection of a representative. We need to decide who is a particular sample of our prediction and how many people we need to predict. Almost all surveys rely on sampling. This paper analyses a sample of *Twitter* users in Australia. A user account which has *username* contains the words *"news"* and *"TV"* is removed (e.g., "abcnews", "abctv" and etc.) as it is news media account. We compute our sample size by using Cochran's sample size formula [18]. We want to estimate sample size (ss) with 95% confidence and the margin of error no larger than 3%. The formulas used in our sample size calculator are shown as follows:

$$n = \frac{Z^2 p(1-p)}{e^2}, \quad ss = \frac{n}{1+(n-1)/P} \tag{2}$$

where Z is $Z - score$ corresponds to confidence level ($Z = 1.96$ for confidence level 95%), p is the maximum possible proportion (50% is the most conservative assumption), e is the acceptable margin of error (i.e., the amount of error that you can tolerate) and P is the population size. The minimum sample size (ss) for our experiments is 1,067 people. We randomly select the sample users according to the numbers of enrolment by State[4] as shown in Table 3. We only determine the locations of users because *Twitter* users are not required to specify the age and gender in their profile.

User's Vote Prediction: According to the voters' attitudes and opinion may be different depending on the specific political topic and the voters participating

[4] `http://results.aec.gov.au/17496/Website/GeneralEnrolmentByState-17496.htm`

in different discussion events may have different political preference, our predicting model were computed based on the significance of sub-event topics and sentiment scores. The sub-event score is calculated to evaluate the significance of each sub-event topic. The sub-event topic will have a high score if there is a lot of a message of them and many users discussing about it. In this work, sub-event score (SE_e) for a given event topic (e) is defined as:

$$SE_e = \frac{NoOfMessages_e}{NoOfTotalMessages} \times \frac{NoOfUsers_e}{NoOfTotalUsers} \tag{3}$$

All sub-events are ranked based on sub-event scores. In order to determine the voter preference among the candidates, for a given user we compute sentiment score for each candidate (i.e., "Abbott" and "Rudd"). For a given user, Aspect Sentiment (AS) scores are defined as Eq. 4 and 5 for "Tony Abbott" and "Kevin Rudd" respectively.

$$AS_{Abbott} = \frac{\sum_{m=1}^{pos}(asp\_score(m, Abbott) \times SE_m)}{\sum_{m=1}^{neg}(|asp\_score(m, Abbott)| \times SE_m)} \times \frac{C_{Abbott}}{C_{Abbott} + C_{Rudd}} \tag{4}$$

$$AS_{Rudd} = \frac{\sum_{m=1}^{pos}(asp\_score(m, Rudd) \times SE_m)}{\sum_{m=1}^{neg}(|asp\_score(m, Rudd)| \times SE_m)} \times \frac{C_{Rudd}}{C_{Abbott} + C_{Rudd}} \tag{5}$$

where $asp\_score(m, A)$ is the aspect-level score of message m, pos is the number of positive messages, neg is the number of negative messages, C_x is the number of both positive and negative messages of aspect x. If a given user posts only positive messages, we assign the summation of negative messages equal to 1. On the other hand, we assign the summation of positive messages equals to 1 when a user posts only negative messages. The voter preference is defined as the highest score out of the two candidates. If the scores are equal, we randomly selected the user vote. In addition, there is another possibility that people has negative sentiment while he still favour to the candidate however it is very difficult to identify.

$$UserVote_u = \begin{cases} \text{``}Abbott\text{''} & \text{if } AS_{Abbott} > AS_{Rudd} \\ \text{``}Rudd\text{''} & \text{if } AS_{Abbott} < AS_{Rudd} \\ Random(\text{``}Abbott\text{''}, \text{``}Rudd\text{''}) & \text{otherwise} \end{cases} \tag{6}$$

4 Experiments and Evaluation

In this section, we firstly assess sub-event detection and sentiment analysis methods because both components may affect the final prediction results of our approach. Next, we evaluate our prediction results by computing the Mean Absolute Error (MAE) between the actual and predicted outcomes.

4.1 Dataset and Experimental Setting

A collection of messages posted by Australia-based users (given latitude, longitude and radius) via the *Twitter Search API* service from 4 August 2013 to 8

September 2013 with 808,661 messages with the user's initial event query is used for our experiments. We define an event by specifying the keyword query (i.e., *"#ausvotes13", "#election2013", "#AusVotes", "#auspol", "Kevin Rudd" and "Tony Abbott"*). We decided to choose this period because the election date is announced on 4 August 2013 and people started to discuss about this event. Also, we decided to choose the keywords related to the two candidates because as in Australian politics the candidates will be from the two major parties. Therefore, in this work we will predict the two-party-preferred vote.

For sub-event detection evaluation, we download the ground truth from *The Sydney Morning Herald* website in *Federal Politics* section[5]. It contains 115 real-world events during 4 August 2013 to 8 September 2013.

For sentiment analysis evaluation, we manually labelled 5,735 messages sent by users in Australia related to the first debate event of the 2013 Australian federal election, between *Kevin Rudd* and *Tony Abbott* on 11 August 2013 from 6pm to 9pm. There are 1,481 messages related to *Kevin Rudd* and 3,254 messages referring to *Tony Abbott*. The messages are annotated with a polarity score (positive, negative or neutral) and sarcasm by three local persons who have political knowledge. We assigned the message polarity score which was determined by the majority view of the three annotators.

For prediction evaluation, the messages since announce election day (i.e., 4 August 2013) until the day before election day (i.e., 6 September 2013) were used for predicting the results. We download the election results from *Australian Electoral Commission* website[6]. The two-party-preferred results for all states and territories as a national summary are compared. The four different national opinion polls are also compared with our results.

4.2 Baseline Approaches

In order to evaluate our approach for detecting sub-events in a collection of *tweets*, we compare our approach performance with temporal peaks detection approach in [9]. The authors bin the messages into a histogram by time (i.e., one hour in this paper). Then, the authors calculate a historically weighted running average of message rate and identify rates that are significantly higher than the mean message rate. A window surrounding the local maximum is identified. Finally, top five frequent terms are presented as event name of each peak.

To evaluate our sentiment analysis method, we compare the performance of our method with aspect-based opinion summarization on *Twitter* data in the domain of politics introduced by Ringsquandl et al. in [12] which is the most similar work to ours. Researchers used the opinion lexicon which is presented in [19]. Semantic orientation of a word is the most probable class (positive, negative, neutral) of each opinion word according to *synsets* in *WordNet*. The final aspect-level sentiment is determined by a simple aggregation function which sums the semantic orientation of all words in the message that mentions the specific aspect.

[5] `http://www.smh.com.au/federal-politics/the-pulse-live`

[6] `http://www.aec.gov.au/Elections/Federal_Elections/2013/`

Table 4. The performance of sub-event detection

Method	# of detected events	# of real-life events	# of distinct real-life events	Precision (%)	Recall (%)	F1-Score (%)
Peak detection	19	14	14	73.68	12.17	20.89
Our approach	542	229	79	42.25	68.70	**52.32**

Table 5. The performance of sentiment analysis

Aspect	Polarity	No. of Messages	Baseline (%)			Our approach (%)		
			Prec.	Rec.	F1	Prec.	Rec.	F1
Kevin Rudd (ALP)	Positive	327	32.72	37.15	34.80	70.34	47.92	**57.00**
	Negative	726	18.87	70.98	29.82	54.41	83.51	**65.89**
	Neutral	428	79.44	34.00	47.62	67.76	54.92	**60.67**
Tony Abbott (LNP/Coalition)	Positive	334	38.92	22.03	28.14	62.28	31.09	**41.48**
	Negative	1,624	22.84	72.89	34.79	59.05	74.75	**65.98**
	Neutral	1,296	76.47	45.99	57.43	62.89	62.60	**62.74**

Finally, we evaluate our prediction by comparing the performance of our approach with counting-based approaches [2] for our first baseline. For a second baseline, we adopt the idea from [3] by counting the number of tweets one week before the election day and using only the first message of each user for the prediction. However, we do not incorporate polls data in the second baseline. The third baseline is based on sentiment analysis only. We use the same size of our sample and the same algorithm of our sentiment analysis. We use the sum of sentiment scores for each aspect to predict the user votes. The third baseline is compared in order to see how well the combination between sub-event detection and sentiment analysis improve our results.

4.3 Evaluation

In this section, we evaluate the performance of our sub-event detection, sentiment analysis and the prediction approaches. For sub-event detection, we compare the precision, recall and F1-score against the peak detection baseline.

$$Precision_{event} = \frac{\#detect\_realworld\_events}{\#total\_detect\_events}, \tag{7}$$

$$Recall_{event} = \frac{\#distinct\_detect\_realworld\_events}{\#total\_realworld\_events} \tag{8}$$

There is more than one detected event can relate to the same real-world event, then they are considered correct in terms of precision but only one event is considered in counting recall. In order to evaluate the performance of our sentiment analysis method, we compare the the *Precision*, *Recall* and *F1-Score* of each polarity category against the aspect-based baseline.

$$Precision_{opinion} = \frac{T}{C}, \quad Recall_{opinion} = \frac{T}{L} \tag{9}$$

Table 6. MAE for comparing election results with three baselines (%)

State	**Election result**		**Baseline1**		**Baseline2**		**Baseline3**		**Our method**	
	ALP	**LNP**	**ALP**	**MAE**	**ALP**	**MAE**	**ALP**	**MAE**	**ALP**	**MAE**
NSW	45.65	54.35	37.94	7.71	44.81	0.84	42.60	3.05	43.11	2.54
VIC	50.20	49.80	37.11	13.09	40.41	9.79	39.00	11.20	38.48	11.72
QLD	43.02	56.98	42.44	0.58	51.35	8.33	45.05	2.03	45.56	2.54
WA	41.72	58.28	37.11	4.61	44.80	3.08	41.38	0.34	41.02	0.70
SA	47.64	52.36	33.21	14.43	46.88	0.76	40.94	6.70	42.38	5.26
TAS	51.23	48.77	26.35	24.88	35.00	16.23	35.11	16.12	38.40	12.83
ACT	59.91	40.09	38.23	21.68	46.58	13.33	42.61	17.30	45.54	14.37
NT	49.65	50.35	35.11	14.54	58.06	8.41	38.08	11.57	42.74	6.91
Average				12.69		7.60		8.54		**7.11**
National	46.51	53.49	37.23	9.28	55.64	9.13	41.69	4.82	42.08	**4.43**

Table 7. MAE for comparing election results (National) with opinion polls (%)

Firm	**Date**	**ALP**	**LNP**	**MAE**	**Remark**
Morgan (multi) [20]	4-6 Sep 2013	46.50	53.50	1.01	
ReachTEL [21]	5 Sep 2013	47.00	53.00	0.49	
Newspoll [22]	3-5 Sep 2013	46.00	54.00	0.51	excludes Northern Territory
Essential [23]	1-4 Sep 2013	48.00	52.00	1.49	
Our approach		42.08	57.92	4.43	

where T is the number of correct classified messages in one opinion category, C is the number of messages classified in one opinion category and L represents the number of the true labelled messages in one opinion category. Finally, we evaluate our prediction results by computing the Mean Absolute Error (MAE) between the actual and predicted outcomes.

Table 4 shows the *Precision*, *Recall* and *F1-Score* of the sub-event detection of our approach against the peak detection baseline. In Table 4, we can observe that our approach can effectively detect real-world events which is significantly larger than the baseline. The baseline can detect smaller number of events because it considers only the temporal peaks in *tweet* frequency. Some events might not be frequently posted on social networks. On the other hand, our approach detects many duplicated events such as the first debate event. There are many different topics discussed during the debate which can cause many clusters when we perform the clustering process. However, our approach outperforms the baseline method by 31.43%. Table 5 represents the performance of the sentiment analysis of our approach against the baseline. It can be seen that our approach can effectively classify the micro-blog messages with a *F1-Score* which is significantly higher than the baseline in the same domain of politics.

Table 6 illustrates the performance of our prediction method against the three baselines. It can be seen that by incorporating sub-event detection and sentiment analysis can effectively improve the prediction accuracy in both state and national levels. In addition, it can correctly predict five out of eight states and territories with smallest error and only 4.43% error for national level. Table 7 presents the performance of our approach against the four different national opinion polls. It can observe that our method comes close to traditional polls with the same trend.

In our study, the incorporating between sub-event detection and sentiment analysis achieved better prediction results than the three baselines. It might suggest that the discussions of sub-event topics that user had engaged in influenced their voting. Also, it can be seen that *Twitter* is able to reflect underlying trend in a political campaign. Even if people who use social media are not completely representative of the public, the amount of attention paid to an issue is an indicator of what is happening in society. Our approach allows researchers to surface user opinions of the social sphere at different time points to determine a view of sentiment for a given event. Also, it turns out that what people say on *Twitter* is a very good indicator of how they will vote.

5 Conclusions

In this paper, we studied a problem of predicting elections based on publicly available data on social networks, like *Twitter*. An effective method of predicting election results is proposed. An approach to detecting sub-events and performing sentiment analysis over micro-blogs in order to predict user preferences is also presented. Extensive experiments are conducted to have evaluated the performance of our approach on a real-world *Twitter* dataset. The proposed approach is effective in predicting election results against the given baselines and comes close to the results of traditional polls. In future work, we will further consider the sarcasm identification and analysis. More studies on the credibility will be conducted in order to remove disinformation and spamming.

Acknowledgement. This paper is partially supported by the Australian Research Council Discovery Project ARC DP140100104.

References

1. O'Connor, B., Balasubramanyan, R., Routledge, B.R., Smith, N.A.: From tweets to polls: Linking text sentiment to public opinion time series. In: ICWSM, pp. 122–129 (2010)
2. Tumasjan, A., Sprenger, T.O., Sandner, P.G., Welpe, I.M.: Predicting elections with twitter: What 140 characters reveal about political sentiment. In: ICWSM, pp. 178–185 (2010)
3. Sang, E.T.K., Bos, J.: Predicting the 2011 dutch senate election results with twitter. In: EACL Workshop on Semantic Analysis in Social Media, pp. 53–60 (2012)
4. Makazhanov, A., Rafiei, D.: Predicting political preference of twitter users. In: ASONAM, pp. 298–305 (2013)

5. Jungherr, A., Jurgens, P., Schoen, H.: 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". Soc. Sci. Comput. Rev. 30(2), 229–234 (2012)
6. Metaxas, P.T., Mustafaraj, E., Gayo-Avello, D.: How (not) to predict elections. In: SocialCom/PASSAT, pp. 165–171 (2011)
7. Gayo-Avello, D.: I wanted to predict elections with twitter and all i got was this lousy paper - a balanced survey on election prediction using twitter data. CoRR abs/1204.6441, 1–13 (2012)
8. Abel, F., Hauff, C., Houben, G.J., Stronkman, R., Tao, K.: Semantics + filtering + search = twitcident. exploring information in social web streams. In: HT, pp. 285–294 (2012)
9. Marcus, A., Bernstein, M.S., Badar, O., Karger, D.R., Madden, S., Miller, R.C.: Twitinfo: aggregating and visualizing microblogs for event exploration. In: CHI, pp. 227–236 (2011)
10. Meng, X., Wei, F., Liu, X., Zhou, M., Li, S., Wang, H.: Entity-centric topic-oriented opinion summarization in twitter. In: KDD, pp. 379–387 (2012)
11. Wang, H., Can, D., Kazemzadeh, A., Bar, F., Narayanan, S.: A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In: ACL (System Demonstrations), pp. 115–120 (2012)
12. Ringsquandl, M., Petkovic, D.: Analyzing political sentiment on twitter. In: AAAI Spring Symposium: Analyzing Microtext, pp. 40–47 (2013)
13. Unankard, S., Li, X., Sharaf, M.A.: Emerging event detection in social networks with location sensitivity. World Wide Web, 1–25 (2014)
14. Jurafsky, D., Martin, J.H.: Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition. Prentice Hall (2000)
15. Gimpel, K., Schneider, N., O'Connor, B., Das, D., Mills, D., Eisenstein, J., Heilman, M., Yogatama, D., Flanigan, J., Smith, N.A.: Part-of-speech tagging for twitter: Annotation, features, and experiments. In: ACL, pp. 42–47 (2011)
16. Hu, M., Liu, B.: Mining and summarizing customer reviews. In: KDD, pp. 168–177 (2004)
17. González-Ibáñez, R., Muresan, S., Wacholder, N.: Identifying sarcasm in twitter: A closer look. In: ACL, pp. 581–586 (2011)
18. Cochran, W.G.: Sampling techniques. Wiley, New York (1977)
19. Wilson, T., Wiebe, J., Hoffmann, P.: Recognizing contextual polarity in phrase-level sentiment analysis. In: HLT/EMNLP, pp. 347–354 (2005)
20. RoyMorgan: Two party preferred voting intention (%)., http://www.roymorgan.com/morganpoll/federal-voting/2pp-voting-intention-recent-2013-2016 (accessed: July 7, 2014)
21. ReachTel: Two party preferred result based on (2010), election distribution, https://www.reachtel.com.au/blog/7-news-national-poll-5september13 (accessed: July 7, 2014)
22. NewsPoll: Two party preferred, http://polling.newspoll.com.au.tmp.anchor.net.au/image_uploads/130922 (accessed: July 7, 2014)
23. Essential: Two party preferred, federal politics – voting intention, http://essentialvision.com.au/documents/essential_report_130905.pdf (accessed: July 7, 2014)