Predicting Political Preference of Twitter Users

Aibek Makazhanov
Nazarbayev University
Research and Innovation System
Astana, Kazakhstan
Email: aibek.makazhanov@nu.edu.kz

Davood Rafiei
Computing Science Department
University of Alberta
Edmonton, AB, Canada
Email: drafiei@ualberta.ca

***Abstract*—We study the problem of predicting the political preference of users on the Twitter network, showing that the political preference of users can be predicted from their interaction with political parties. We show this by building prediction models based on a variety of contextual and behavioural features, training the models by resorting to a distant supervision approach and considering party candidates to have a predefined preference towards their parties. A language model for each party is learned from the content of the tweets by the party candidates, and the preference of a user is assessed based on the alignment of user tweets with the language models of the parties. We evaluate our work in the context of Alberta 2012 general election, and show that our model outperforms, in terms of the F-measure, sentiment and text classification approaches and is in par with the human annotators. We further use our model to analyze the preference changes over the course of the election campaign and report results that would be difficult to attain by human annotators.**

I. Introduction

Today Twitter stands out as one of the most popular microblogging services, where information propagates in no time, and words and actions trigger immediate responses from users. Such an environment is ideal for advertising political views, especially during the heat of election campaigns. Political discourse on Twitter has been studied over the past few years. The focus of studies varied from analyzing social networks of candidates and politically active users [4], [8], to predicting results of actual elections [9], [11], [15]. However, with few exceptions [3], [6], most of the previous work focused on the analysis of individual tweets [2], [9], [16] or aggregate properties of a corpus of tweets [11], [15], and not on the political preference of individual users.

In the present work we address the problem of predicting the political preference of users given the record of their past activity on Twitter. We believe that approaches to political discourse analysis could benefit from the knowledge of political preference of users. For instance, predicted user preferences can be used as noisy labels when evaluating approaches to the community mining in political communication networks [3], [4], and as additional features in methods that focus on the extraction of political sentiments [9], [11]. Also, when used in the context of elections [9], [11], [15], political preference prediction has implications in better understanding changes in public opinion, and possible shifts in popular vote. We set the preference prediction problem in the context of Alberta 2012 general election, and consider a vote for a party to be the best indicator of the political preference of users. More formally, the problem is modeled as a multi-label classification task, where given a user and a set of parties, the goal is to predict which party, if any, the user is most likely to vote for. For each party considered in this work we construct a profile, which includes a ranked list of weighted party-specific topics. Also, for each party we train a binary classifier, using the partisan identifications of the party candidates as the ground truth. For a given user-party pair, such a classifier provides a confidence with which a user can be considered a party supporter. Thus, a party with the highest predicted confidence is considered as the most preferred one.

We evaluate our method on a set of users whose political preferences are known based on the explicit statements (e.g. *'I voted NDP today!'*) made on the election day or soon after. Measuring the performance on a per-party basis in terms of precision, recall and F-measure, we compare our method to human annotators, sentiment and text classification approaches, and to chance. We found that although less precise than humans, for some parties, our method outperforms human annotators in recall. Another experiment, where we analyzed how preferences of users change over time, revealed that politically active, or so called *vocal users* are less prone to changing their preference than users who do not get actively involved, i.e. *silent users*. We also observed that the dynamics of the popular vote shift among silent users closely resembles that of the actual election.

Contributions. Our contributions can be summarized as follows: first, we introduce a notion of a user-party interaction, based on which we propose an interaction-driven approach to the problem of predicting the political preference of users. Second, we explore the dynamics of the election campaign, showing the difference in preference change across different types of users. Third, of the studies concerned with a Twitter-based political discourse analysis, our work is, to the best of our knowledge, the first to report an extensive data cleaning effort. Finally, we evaluate our proposed methods in a real setting with data covering the tweets posted during a recent election and their connections to real events.

Election background. On April 23, 2012 a general election was held in Alberta, Canada to elect 87 members of the Legislative Assembly. The event was highly anticipated[1] as according to polls for the first time since 1971 the ruling Progressive Conservative (PC) party could have lost the election. Since 2009 Wildrose Aliance party (WRA) started to rapidly gain popularity, and by the beginning of the 2012 election they were leading in polls and becoming the main

[1] 57% voter turnout, the highest since 1983: http://www.cbc.ca/news/canada/albertavotes2012/story/2012/04/24/albertavotes-2012-voter-turnout.html

challenger to PC [17]. Two other major parties who nominated 87 candidates, one per each riding, were Alberta Liberals (LIB) and New Democratic Party (NDP). Parties with low polling numbers and popular vote share (e.g. Alberta party) are not considered in this work. To form a majority government a party was required to win at least 44 seats. The election resulted in Conservatives winning 61 seats and defending their ruling party status. Wildrose followed with 17 seats, forming the official opposition. Liberals and NDP won five and four seats respectively. Although WRA had lost almost 10% of the popular vote to PC, their candidates were second in 56 ridings, losing by a tight margin in dozens of constituencies [17].

II. Related Work

Our work is related to the body of research on extracting political sentiment from Twitter. Numerous works employ a two-phase content-driven approach, where at the first phase a set of relevant tweets is identified, and at the second phase the actual sentiments are extracted. Typically, a tweet is considered relevant if it contains at least a term from a list of target keywords, constructed manually [11], [15], [16], or semi-automatically [3], [4], [9]. To identify the polarity of expressed sentiments, various supervised or unsupervised methods are employed. Unsupervised methods rely on the so called opinion lexicons – lists of "positive" and "negative" opinion words, estimating a sentiment polarity based on the positive-to-negative words ratio [11] or just the raw count of opinion words [2]. Supervised methods, on the other hand, train prediction models either on manually labeled tweets [3], [16] or on tweets with an emotional context [9], i.e. emoticons and hashtags, such as *:-)*, *#happy*, *#sad*, etc. Conover et al. [3] took a two-phase approach semi-automatically building a list of 66 target keywords, subsequently extracting more than 250,000 relevant tweets, and training SVM models on unigram features of the tweets. As a ground truth the authors used a random set of 1000 users whose political affiliations were identified based on a visual examination of their tweets. An accuracy of about 79% was reported, which could be boosted up to almost 91% when the features were restricted to hashtags only. A major challenge in applying a similar approach to our work is the data sparsity, and the difficulty in identifying relevant tweets or users and in establishing a ground truth. We use a distant-supervision approach to address those challenges.

Related work also includes interaction-driven approaches to study sentiments expressed implicitly in the form of preferential following [6], [13], and retweeting [4]. We also observed that preferences to *follow* and *retweet* rank among the top discriminative features in detecting a political preference. Our work relates to that of Sepehri et al. [12], who explored the correlation between the interaction patterns of Wikipedia editors and their votes in admin elections. While the authors defined an editor interaction as an act of co-revising of a page by a pair of contributors, we extend the proposed notion to the Twitter environment, and treat tweeting on party specific topics as a form of user-party interactions.

III. User - Party Interactions

Not all postings of users reflect their political preference. One straightforward approach to filter out irrelevant tweets is to identify popular hashtags from the local political trends and consider any tweet containing those tags to be of relevance. However, tweets that use the same hashtag(s) may not be relevant to the same degree. For instance, although both tweets in the example below contain *#Cdnpoli* (Canadian politics) hashtag, the second one is clearly of more relevance to the election.

T1: *Huzzah! Happy birthday to Her Majesty, Queen Elizabeth II, Queen of Canada. Long may she reign! #Cdnpoli*

T2: *A win for @ElectDanielle will be the first step towards the libertarian #CPC regime under Max "The Axe" Bernier. #ABvote #Cdnpoli*

On the other hand, Twitter activities, such as (re)tweets, mentions, and replies, that concern parties or its candidates, may suggest the political preference (or its absence) of users. To detect a preference expressed in this way, for each party we compile a ranked list of weighted terms, or the *interaction profile* of a party, and define a user-party interaction as follows:

Definition 1: Given a user *u*, a party *p*, and a tweet *t*, we say *t* interacts with *p* if it contains at least one term from the interaction profile of *p*. Similarly, *u* interacts with *p* if the user has at least one tweet that interacts with *p*.

Note that we use the term *interaction* for the following two reasons: (i) we assume that collective behavior of users triggers some kind of response from political entities, hence interactions take place; (ii) the term *interaction* makes exposition much more concise, than perhaps a more accurate term *a directed behavior of users towards parties*.

Building interaction profiles. It is reasonable to assume that during a campaign party candidates will use Twitter to advertise central issues for their parties, discuss television debates and criticize their opponents. We aim at utilizing the content resulted from such a behaviour to capture party-specific topics in the form of weighted unigrams and bigrams.

Given a set C of candidates, we associate with each candidate $c \in C$ a document d_c that consists of all postings of c; in our work d_c is modeled as a *bag of words*. Let $D = \{d_c | c \in C\}$, and for each party $p \in P$, $D_p = \{d_c | c \; is \; a \; member \; of \; p\}$. Denote the vocabulary of D with V. To build a language model (LMs) for each party, we use a term-weighting scheme similar to [8] where the Kullback-Leibler (KL) divergence between the LM probabilities of the party and the LM probabilities of all parties gives a measure of importance to terms in the party profiles. To be consistent, we refer to the general corpus as the election corpus and to its LM as the election LM. Similarly we refer to a collection of documents associated with a party as party corpus and its LM as party LM. Term probabilities in the aforementioned language models are calculated using tf x idf scores. The marginal probability of a term $t \in V$ in the election LM is calculated as

$$P(t|D) = \overline{tf}(t,D) udf(t,D)$$

where $\overline{tf}(t,D)$ denotes the average term frequency in a document in D, and $udf(t,D) = df(t,D)/|D|$ denotes the probability of t appearing in a document in D. Here $df(t,D)$ is the document frequency of t in D.

For the party LMs, initial term weights are calculated as

$$w(t|p) = \overline{tf}(t,D_p) udf(t,D_p) idf(t,D)$$

where $idf(t, D)$ is the inverse document frequency of t over the set of all documents D. The obtained weights and scores are then normalized:

$$P^N(t|D) = \frac{P(t|D)}{\sum_{t \in V} P(t|D)}; w^N(t|p) = \frac{w(t|p)}{\sum_{t \in V} w(t|p)}$$

and further smoothed to account for terms that are not observed in the party corpora,

$$w^S(t|p) = (1 - \lambda)w^N(t, p) + \lambda P^N(t|D)$$

with the normalization factor λ set to 0.001. Finally, we calculate the probability of term $t \in V$ in the LM of party p as

$$P(t|p) = \frac{w^S(t|p)}{\sum_{t \in V} w^S(t|p)}.$$

Now we can calculate the KL divergence between probability distributions of party LMs and the election LM as follows:

$$KL_p(P(t, p)||P(t, D)) = \sum_{t \in V} P(t|p) \ln \frac{P(t|p)}{P(t|D)}$$

However, rather than sum, which characterizes the overall content difference, we are much more interested in the individual contribution of each term. Hence the final weight of term $t \in V$ in the interaction profile of party p, or the importance score of the term is calculated as

$$I(t, p) = P(t|p) \ln \frac{P(t|p)}{P(t|D)}$$

The higher the weight of the term the more it deviates from the "common election chatter", and as such becomes more important to the party. As a final step in building the interaction profiles, we chose 100 top-ranked bigrams and unigrams representing party hashtags, Twitter user names of candidates, and the last names and the given names of party leaders. Note that by including candidate and party account names into the profiles we naturally count tweets that mention, reply or retweet candidates as interactions, i.e. our definition is flexible, and conveys standard forms of user activities on Twitter. Moreover, the fact that the profiles are weighed and ranked allows us to weigh and rank interactions. Specifically, given a user *u* and a party *p*, the weight of a *u-p* interaction is calculated as the collective weight of all terms from the interaction profile of *p* found in the interaction. Correspondingly, the average, minimum and maximum ranks are calculated respectively as the average, minimum and maximum ranks of such terms. Table I gives the number of terms in the interaction profiles of the parties and lists some of the top weighted bigrams.

IV. Data Collection and Cleaning

Collecting user accounts. We semi-automatically collected Twitter accounts of the party candidates. As in [8], we retrieved the first three google search results for each candidate name and the *twitter* keyword, and manually verified the accounts. Additionally, if we could not find or verify a candidate account from the search results we looked up the official website of a candidate party. This way we collected 312 Twitter accounts of 429 registered candidates. Of those 252 belonged to candidates of the four major parties considered in this work. To that list we also added the official Twitter accounts of the parties to have a total of 256 accounts. To collect non-candidate accounts we monitored campaign related trends over the course of ten days prior to the election, using a manually selected 27 keywords, such as party hashtags, party names, leader names and general election hashtags. As a result we obtained 181972 tweets by 28087 accounts of which 27822 belonged to non-candidates. Removing accounts with reported communication language other than English, left us with 24507 accounts. For each of these non-candidate and 256 candidate accounts we retrieved up to 3000 tweets. The retrieved content was limited to the tweets posted since March 27, 2012, the official first day of the campaign[2].

Party	Profile size	Top terms
LIB	170	alberta liberals, vote strategically
NDP	157	orange wave, renewable energy
PC	173	premier alison, family care
WRA	182	energy dividend, balanced budget

TABLE I: Basic characteristics of the interaction profiles of the parties

Data cleaning. Benevenuto et al. [1] have shown that spammers often "hijack" popular hashtags and trends to reach a larger target audience. To test this hypothesis, we trained SVM and logistic regression (LR) models based on the dataset and the top 10 features in [1] [3], but to account for the specifics of local trends in our dataset we extracted features for candidate accounts (the only ground truth we had) and added them to the data set as positive examples of non-spam accounts. The features were designed to distinguish the behaviour and the quality of the content generated by the two types of users, i.e. spammers and non-spammers. For instance, spammers tend to include URLs (spam links) in almost every tweet, label their tweets with popular and often unrelated hashtags, and rarely reply to tweets of other users. We performed 10-fold cross validation and found that only one out of 256 candidate accounts was misclassified as spam account. In general, spammers were detected with lower F-measure than non-spammers (in agreement with the original work). The LR classifier performed slightly better than SVM, and F-measure for respective classes were 85% and 94%. However, results of spammers detection across the entire data set were rather unexpected. Out of 24507 accounts 74 or 0.3% were labeled as spam. We manually checked these accounts and did not find any evidence of spam related activity. As a rough estimate of misclassification of spammers as non-spammers we examined a random sample of 244 or 1% of 24433 accounts labeled as non spam. Again we did not find any spammers apart from two accounts that were already suspended by Twitter. From this experiment we drew the following conclusions: First, our model may have been subject to a certain degree of overfitting, given that training data was collected much earlier than our data set and the behavior of spammers might have changed. Second, it could be that local political trends were not attractive enough for spammers and our data set may naturally contain negligibly low number of spam accounts.

[2] http://en.wikipedia.org/wiki/
Alberta_general_election_2012#Timeline

[3] The authors have shown that the top 10 features perform as good as the full set of features.

User group	Accounts	Interactions	Interactions per account
Candidates	256	23195	90.6
LIB	62	7916	127.7
NDP	50	5813	116.3
WRA	73	6029	82.6
PC	71	3437	48.4
P-accounts	24060	311443	12.9
NP-accounts	447	8359	18.7

TABLE II: Data set properties

Third, certain accounts behave like spam accounts and generate content that shares some features with spam tweets. The latter deserves further elaboration. When we were verifying accounts labeled as spam we noticed that some of them represented *media* (e.g. provincial and federal news papers, radio stations), *businesses* (e.g. companies, real estate agencies), and even *official entities* and *organizations*, e.g., City of Calgary. One common characteristic of these accounts is that almost non of them expresses a *personal opinion* of a potential voter. Moreover owners of these accounts often do not or should not have political preference (at least media and official entities). While investigating media bias and influence or support and disapproval of unions is an interesting research direction, in the present work we focus on predicting the political preference of individuals. We will refer to individual accounts as personal or P-accounts, and correspondingly non-personal accounts will be referred to as NP-accounts.

Removing NP-accounts. We approached the problem by extending the method that we used for spammers detection. For training we used the set of accounts that we annotated during the evaluation of spammers detection. The set contained 161 personal and 25 NP-accounts. To get more NP-accounts for training we extracted the list of Alberta newspapers[4] and searched the dataset for accounts with names matching those in the list. We did the same for the list of Alberta cities[5]. We annotated the extracted set of accounts and obtained a final data set that consisted of 161 P- and 132 NP-accounts. We started with the feature set used in the spammers detection task, but after a 10-fold cross validation, we revised the set, removing some irrelevant features. Also, our observations suggested that personal accounts frequently contain first-person singular pronouns (*I, me, mine, myself*, etc.) and words like *father, mother, wife* or *husband* in the account description. Also account names differ drastically with NP-accounts frequently using location names, abbreviations and business related terms, and P-accounts frequently using person names. To capture these differences in the types of accounts, using the training data, we build unigram language models for the names and the descriptions of P- and NP-accounts. After a feature selection procedure, our final model consisted of 11 features. We classified the entire data set, resulting in 535 out of 24507 accounts being labeled as NP. A visual inspection of these accounts revealed that in 447 cases the model made correct predictions yielding 83% precision. Out of 447 NP-accounts 160 or 36% were associated with media. As a final cleaning step we removed NP-accounts from the data set leaving 24060 accounts. All the experiments reported further were conducted on this cleaned data.

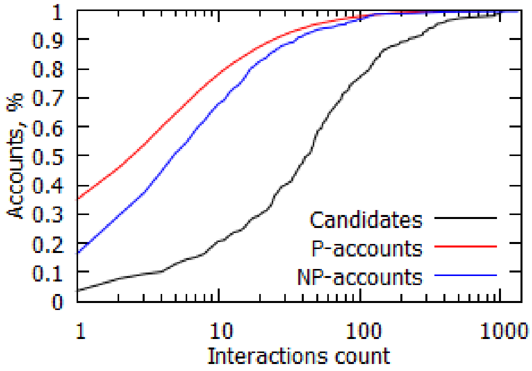


Fig. 1: Cumulative percentage of accounts at each interaction count for candidate, P- and NP-accounts

Discussion. Figure 1 shows the cumulative percentage of accounts at each interaction level for those with at least one election-related tweet, broken down into candidates, P-, and NP-accounts. The plot clearly illustrates the silent majority effect [10], showing that almost half of all P-accounts have produced at most only two interactions. Moreover, it shows that concentration of the silent majority is not even across different target groups, with representatives of media and organizations containing less silent users relative to individuals. Table II contains interaction statistics for different user groups, and statistics for candidates is further split on a per-party basis. We can see a similar trend, with candidates having much more interactions per account than the other two user groups, and owners of NP-accounts having 1.4 times as much interactions per account than individuals, i.e. owners of P-accounts. Considering that the bulk of the interactions of NP-accounts is either news or statements without personal political preference, we conclude that our efforts in data cleaning had payed off and we were able to remove certain amount of noise.

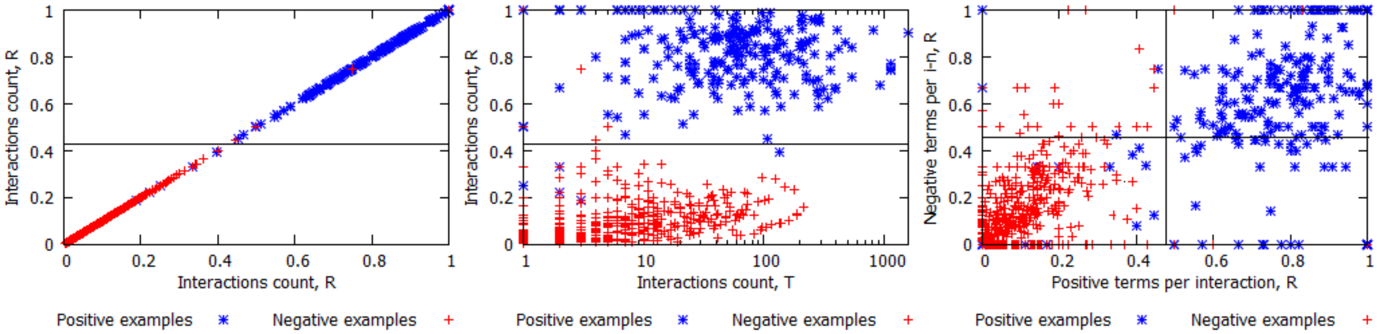
V. Predicting Political Preference

We state the problem of predicting political preference as follows:

> Given a user u and a set of parties P with whom u has interactions, predict which party $p \in P$, if any, u is most likely to vote for in the upcoming election.

To address the problem we employ a one vs. all classification strategy, where for each party we train a binary classifier. Given the interaction statistics of a user-party pair, a classifier trained for the corresponding party provides the confidence with which this user may be considered a supporter of that party. If a user has interacted with multiple parties, confidence scores are compared and the user is said to prefer a party which was given the highest confidence score. Ties are broken randomly. According to our problem statement, a user may prefer none of the parties she interacted with. We use a confidence threshold t, set at 0.5 in our experiments, to detect such cases. Thus, if the confidence provided by a classifier falls below the threshold, we conclude that a user will not support the corresponding party. To train the classifiers, we extract all interactions of the candidates, group them on a per-party basis, and build a feature vector for each group.

[4] http://en.wikipedia.org/wiki/List_of_newspapers_in_Alberta
[5] http://en.wikipedia.org/wiki/List_of_cities_in_Alberta



(a) Relative interactions count

(b) Target vs. relative interactions count

(c) Positive vs. negative terms rel. count

Fig. 2: Distribution of training examples across different features

Building prediction models. We design the features of the models based on the interaction records and the behaviour of users, and build a feature vector for each interacting user-party pair. Clearly if a user does not interact with any party we can make no predictions, and such a user is considered to prefer none of the parties. To capture basic patterns of user-party interactions, we gather the following statistics for each user-party pair: (i) raw count of the interactions, (ii) average weight, (iii-v) average/min/max rank, (vi,vii) average length (in characters, in words), (viii) average frequency (interactions per day), (ix) number of positive terms per interaction, (x) number of negative terms per interaction, (xi) average offset of a party hashtag in the postings of the user, (xii) political diversity, i.e. number of parties the user has interactions with. For positive and negative terms we use an opinion lexicon of 6800 words compiled by Hu et al. [7] and the Wikipedia's list of emoticons[6]. To account for Twitter-specific behaviour of users towards parties, our features also include: (i) number of times a user has retweeted party candidates, (ii) average retweet time, (iii) number of replies a user received from party candidates, (iv) average reply time, and (v) number of party candidates that a user follows. Here retweet time corresponds to the amount of time, in seconds, passed between the moment when the original tweet of a candidate was created and the time the retweet was made. Reply time is calculated similarly. Since a user can interact with multiple parties, and both overall and relative statistics may indicate a party preference, we experimented with features from different *domains*. Let us explain the concept of feature domains. Suppose user u interacted 6 times with party p_1 and 14 times with party p_2, resulting in feature vectors: u-p_1 and u-p_2. In the target, or *T-domain*, the *interactions count* will have its respective per-party value, i.e. 6 in u-p_1, and 14 in u-p_2. In the overall, or *O-domain*, the feature will be calculated as the sum over all parties, and will have the value of 20 in both vectors. In the relative, or *R-domain*, the feature will be calculated as the fraction of its values in *T-* and *O-domains*, i.e. $6/20 = 0.3$ in u-p_1, and $14/20 = 0.7$ in u-p_2. Finally, for some features, such as weight and rank of the interactions, we resort to the absolute difference of feature values in *T-* and *O-domains*; these are referred to as *Δ-domain*. Overall we have defined 51 features. Table III gives the top 10 ranked features based on the information gain (as computed Weka).

Feature	Domain	Type	Avg. rank
Interactions count	R	IB	1.3 ± 0.46
Followees count	R	TS	1.7 ± 0.46
Positive terms per interaction	R	IB	3 ± 0
Retweets count	R	TS	4.1 ± 0.3
Interactions frequency	R	IB	4.9 ± 0.3
Negative terms per interaction	R	IB	6.2 ± 0.4
Interactions weight	R	IB	7 ± 0.77
Followees count	T	TS	8 ± 0
Interactions weight	Δ	IB	9 ± 0.77
Retweets count	T	TS	9.8 ± 0.4

TABLE III: Top 10 features for predicting political preference

Relative statistics turned out to be effective features, as the top 7 ranked features are all based on the relative statistics. Six out of top 10 features are interaction based (IB) and four remaining are Twitter specific (TS) features, as indicated in the *Type* column of the table. In line with previous studies [3], [6], [13], we observe a strong discriminative power of a preferential following and retweeting, as the corresponding features ranked 2nd and 4th in the R-domain, and 9th and 10th in the T-domain respectively. When restricted to the top 10 features only, the prediction models showed better performance on the training data. Therefore, for our experiments we will use models built on these 10 features. Figure 2 shows the distributions of training examples across different feature spaces. Each data point corresponds to a user-party feature vector. As it can be seen from Figure 2a, the *interaction count in the R-domain* is a very strong feature which divides positive and negative examples very accurately. Figure 2b further illustrates this point, showing that while it is hard to separate positive examples from negative ones, judging only by the raw count of interactions, based on the relative count of interactions one can easily split the data. Finally, as it can be seen from Figure 2c, the data points that represent positive and negative training examples are placed (mostly) at the upper-right and bottom-left quadrants respectively, suggesting that, regardless of the term polarity, candidates tend to use more opinion words when they interact with their own parties, and less when they interact with their opponents.

Evaluation of the model. A major evaluation challenge was obtaining the test data. In order to have the ground truth preference of non-candidate users we used the content generated during or after the election, i.e. precisely, everything

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

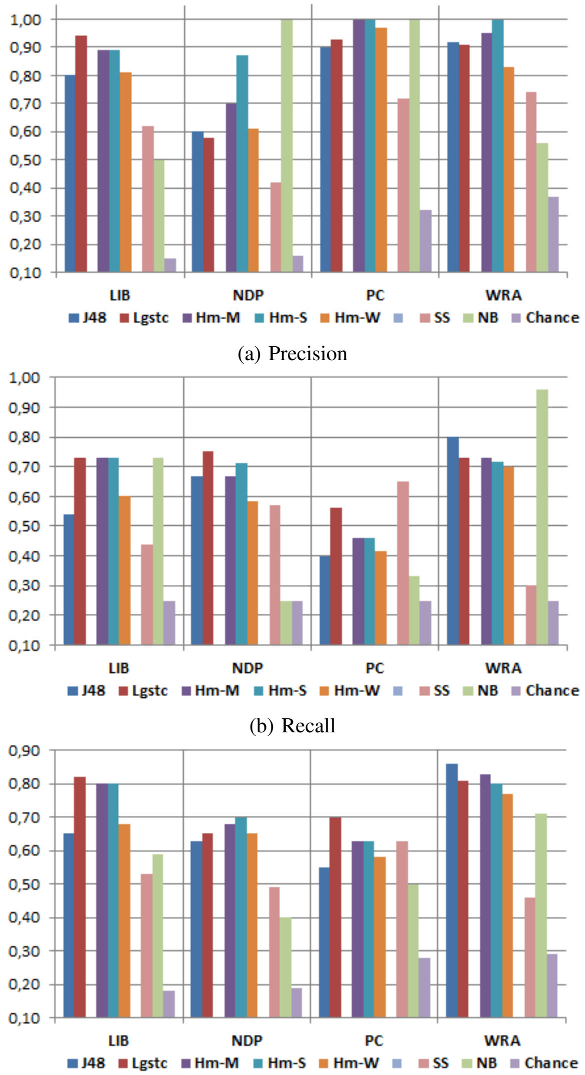
Supported party	Accounts	Interactions	Interactions per account
WRA	28	6883	245.8
LIB	11	1404	127.6
NDP	12	1329	110.7
PC	24	2510	104.6
Total	75	12126	161.7

TABLE IV: Characteristics of the test set

between April 23, 9:00 am, MDT (ballot boxes are opened) and April 26, 11:59 pm, MDT. We searched for the occurrences of words *vote* or *voted* followed or preceded by a *party marker* in a window of three words. Here, a party marker can be a party hashtag, a name of a party leader, a mention of a party account or any known account of a party candidate. This search resulted in a collection of 799 tweets by 681 users. We asked three annotators to classify each tweet in the collection as a statement that either supports certain party or not. Our criterion of support was the clear statement of the fact that vote has been casted or was about to be casted. Retweets of such statements were also counted as signs of support. Cheering for parties, e.g. *vote NDP!*, were asked to be ignored, as the evidence [3] suggests that retweets generally express the agreement with the original tweets. Annotators agreed that 99 out of 681 users had expressed prevailingly supporting statements. In the case of 64 users the agreement was unanimous and for the remaining 35 users two vote majority was achieved. The rate of inter-annotator agreement calculated as Fleiss' kappa [5] was 0.68. Recall that the vote statements were extracted from the content generated after the election. It is possible that users who expressed support for certain parties after the election did not interact with those parties during the election campaign. This was exactly the case for 14 out of 99 users in our test set. We exclude them from the test set, and focus on the remaining 75 users to whom we will refer to as the test users. Table IV shows basic interaction statistics of the test users.

Human evaluation. In order to assess the difficulty of the task on human scale we set an experiment in which we provided the same three annotators with randomly selected interactions of test users. For each test user and each party the user has interacted with we chose up to 50 random interactions out of those that happened before the election. To create equal prediction conditions for humans and classifiers each annotator was given four sets of interactions - one per each party. These sets were provided sequentially, one after another to avoid possibility of comparison. In other words if a test user interacted with four parties, annotators would encounter postings of this user in four different sets of interactions. In such cases, the annotators, just like our classifiers, may predict that the same user will support multiple parties. We use the rate of annotator's agreement with the majority as the analogue of classification confidence.

Baselines. Apart from human annotators, we compare our method with three more baselines. First, we use SentiStrength [14], a lexicon based sentiment analysis tool optimized for informal writing style common to social media services. Under default settings for a given text, SentiStrength provides two scores as respective measures of the strength of positive and negative sentiment expressed in the text. These scores are varied in the range $[+1, +5]$ for positive



(a) Precision

(b) Recall

(c) F-measure

Fig. 3: Preference prediction. Hm-M - majority vote of annotators, Hm-S - strongest annotator, Hm-W - weakest annotator, SS - SentiStrength, NB - Naive Bayes

sentiment and $[-1, -5]$ for negative sentiment. By analogy with our human evaluation experiment, we provide the tool with interactions of each user-party pair. For each interaction the tool returns a sentiment strength score. We sum up these scores and treat the sum as the classification confidence. A resulting sum may be negative, in which case we conclude that SentiStrength predicted no preference. Second, we employ a widely used Naive Bayes text classification method, treating all tweets of a user as a document belonging to a certain class (party), and computing prior and conditional probabilities based on the postings of candidates. Finally, we compare our method to chance. This baseline predicts that user will prefer one of the parties with an equal probability. For each user we randomly generate 1000 predictions and choose the party that was predicted most of the times. Ties are broken randomly.

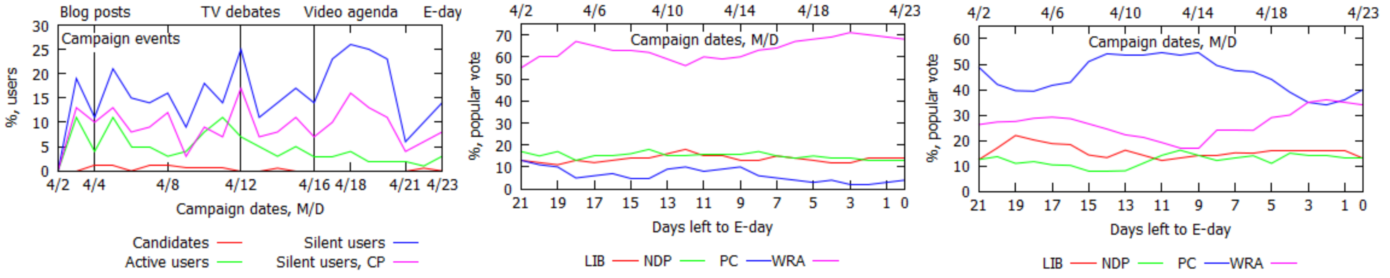
Results. We experimented with both a decision tree based J48 and Logistic regression classifiers. We train one classifier per party, and present the results on a per party basis. As for human annotators, with respect to each evaluation metric, we report results for the strongest and the weakest performing annotators, as well as for the "majority vote". Figure 3 shows the results of the experiment. Each of three plots corresponds to an evaluation metric, and consists of four clusters of bars associated with supporters of four major parties. Each such cluster consists of seven bars, corresponding to the performances of two classifiers, three human annotators, and two baselines. The order of the bars, from left to right, corresponds to that of the legend items. As it can be seen from Figure 3a both classifiers make less precise predictions than the annotators, although LR shows better precision than J48 especially for LIB and PC parties. Moreover, this classifier outperforms the least accurate annotator for LIB and WRA parties. As for baselines, NB makes 100% precise predictions for PC and NDP parties, surprisingly outperforming humans in the latter task. A possible explanation for this could be the fact that test users who support PC and NDP frequently retweet party candidates, literary copying the content, therefore making NB assign higher probabilities to their posts. In terms of recall, classifiers again perform close to human annotators, cf. Figure 3b. It is interesting that for the PC party, both LR and SentiStrength outperform human annotators, and for the WRA party J48 and NB do the same. One of the possible explanations is that a simple quantification of opinion words and estimation of the content similarity employed by the baselines turned out to be more effective (in this particular instance) than a human logic that puts opinion words in the context (sometimes rendering them as neutral) and has no means to compare the content (again, in this particular instance). Similarly, it could be that for the annotators the interactions of some of the users with the PC party did not seem to have meaning, but our learned models could have exploited different features, such as the following and the retweeting preference, interaction frequency, etc. Thus, as we can see, in combination with the content based features, the behavioral features, such as the aforementioned ones, can be very effective. Finally, as shown on Figure 3c, in terms of F-measure, the classifiers outperform all of the baselines. LR shows great performance outperforming all of the annotators for the LIB and PC parties and the weakest annotator for the WRA party. J48 outperforms all of the annotators for the WRA party.

VI. Temporal Analysis

Having a prediction model, we wanted to explore how the predicted political preference of users changes with the progression of the election campaign, and if some group of users are more likely to change their preference than others. To study such changes, we set up an experiment as follows: we choose a window of a fixed number of days, say N, and by sliding this window (one day at a time) over the timespan of the campaign (28 days), we obtain a sequence of time periods, each N days long. We arrange the interactions of each user into these time periods according to their registered tweet time. Suppose user u has interactions in two consecutive periods p_1 and p_2. Let predicted political preference of u for periods p_2 and p_1 be P_2 and P_1 respectively. If $P_2 \neq P_1$, we say that on the given day d, the predicted preference of user u for the current period p_2 has changed compared to the that for the previous period p_1. In order to capture changes in preference for the whole duration of the campaign, we need to repeat this procedure for all consecutive periods. This, however, requires a user to have interactions in all of these periods. Hence, for this experiment we select only those users who satisfy this condition. Experimentally we chose the window of size seven, i.e. we measure changes in the preference of users on a weekly basis. This choice allowed us to identify 3413 users who satisfied the condition of the experiment. Of those 129 were active and 791 were silent users[7]. We consider a user to be active if she contributed at least 10 interactions per day over the course of the campaign. Correspondingly, a user who has engaged in at most 1.5 interactions per day on average is considered a silent user. For the experiment we randomly selected 100 users from each of these user groups.

Discussion. Figure 4a shows the percentage of users for whom on a given date, the political preference for the current period has changed compared to that of the previous period. As one would expect, the preference of candidates did not change during the campaign, apart from negligible deviations of about 1%. On contrary, the preference of the silent users is changing constantly and for certain points in the campaign rather drastically. A further examination revealed that due to a small number of interactions in some periods our method predicted no preference for significant number of users. To account for the interaction sparsity, we repeated the experiment for the silent users under a *constant preference* (CP) setting, where we assume that if for the current period a user was predicted to have no preference, then user's preference did not change since the last period. As seen from Figure 4a, although the CP setting reduced the share of the silent users with changing preference, it is still higher than that of the active users' almost for the whole duration of the campaign. Correspondingly, the share of the active users with changing preference never exceeds 11% and since April 13 (10 days before the E-day) never exceeds 5%. Figures 4b and 4c show the distributions of the popular vote of the active and the silent users for each period on a given date. It can be seen that for both active and silent users, the preference generally changes between WRA and PC parties. In agreement with our findings, popular vote of the active users changes gradually, and sharp transitions mostly occur before April 12. For the silent users, however, shifts in popular vote occur rather sharply for the whole duration of the campaign. From this we conclude that the active users are less likely to change their political preference compared to the silent users. There is an interesting point in Figure 4c about the silent users. The popular vote for PC remains higher than WRA throughout the campaign except for the last few days. In those last few days, the popular vote shows a decreasing trend for PC and an increasing trend for WRA until the two are neck-to-neck 3 days prior to the election day. A CBC poll conducted 3 days before the election day also put the two parties neck-to-neck. However, we see a not-so-strong changing trend starting to show one day before the election with popular vote for PC rising and the popular vote for liberals declining, as a possible sign of strategic voting changing the course of the election. Another point about Figures 4b and 4c is that the popular vote prediction for silent users shows more resemblance to the actual election outcome than active users, with the ordering of the parties: PC, Wildrose, Liberals and NDP from the largest

[7] The rest is moderate users which are ignored here



(a) Political preference

(b) Popular vote of active users

(c) Popular vote of silent users, CP

Fig. 4: Changes of political preference and popular vote over time

to the least number of received votes.

We wanted to see if major changes in predicted political preference occur spontaneously or on certain dates that have some significance to the campaign. From the social media highlights of the campaign[8] we found descriptions of the following events, which were extensively discussed in blogsphere, on Facebook and Twitter: (i) *Blogposts* by Kathleen Smith (April 2) and Dave Cournoyer (April 4) criticizing WRA party. (ii) Television broadcast of party leaders' *debates* (April 12). (iii) *YouTube video* titled *"I never thought I'd vote PC"* (April 16), asking people to vote strategically against WRA party. The vertical lines on Figure 4a represent these events together with their occurring dates. As it can be seen, the highest change in the preference of silent users occurred on April 12, the day of the TV broadcast of the party leaders' debate. It is interesting that for the active users the peak change in preference occurred one day before the debates. This could be the case where discussion of the event was more interesting than the event itself. In the case with blogposts and video the rise in the preference change rate occurs only on the next day after events took place. Twitter discussion of these events might have had the "long term" effect gaining more popularity on the next day and influencing the predictions for the next period.

VII. Conclusions

We studied the problem of predicting political preference of users on the Twitter network. We showed that the generated content and the behaviour of users during the campaign contain useful knowledge that can be used for predicting the political preference of users. In addition, we showed that the predicted preference changes over time and that these changes co-occur with campaign related events. We also compared the preference change of silent users to that of active users. One future research direction is to investigate patterns of strategic voting. For the particular election campaign studied here, strategic voting was a widely discussed issue, and we found numerous evidence of it on Twitter, while developing our method. For instance: *just voted PC lesser of two evils but i still feel like i need a shower. #abvote #yyc #yeg* Also, at some point in the campaign special trends started to emerge. These trends used hashtags like #nowrp, #strategicvotes, #strategicvoting, etc. We believe studying the behaviour of users engaged in these discussions can help improve our models, leading to more accurate preference predictions.

Acknowledgements This research was supported by the Natural Sciences and Engineering Research Council.

References

[1] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida, "Detecting Spammers on Twitter," Jul. 2010. [Online]. Available: http://ceas.cc/2010/papers/Paper%2021.pdf

[2] M. Choy, M. L. F. Cheong, M. N. Laik, and K. P. Shung, "A sentiment analysis of singapore presidential election 2011 using twitter data with census correction," *CoRR*, vol. abs/1108.5520, 2011.

[3] M. Conover, B. Gonçalves, J. Ratkiewicz, A. Flammini, and F. Menczer, "Predicting the political alignment of twitter users," in *SocialCom/PASSAT*, 2011, pp. 192–199.

[4] M. Conover, J. Ratkiewicz, M. Francisco, B. Gonçalves, F. Menczer, and A. Flammini, "Political polarization on twitter," in *ICWSM*, 2011.

[5] J. Fleiss *et al.*, "Measuring nominal scale agreement among many raters," *Psychological Bulletin*, vol. 76, no. 5, pp. 378–382, 1971.

[6] J. Golbeck and D. L. Hansen, "Computing political preference among twitter followers," in *CHI*, 2011, pp. 1105–1108.

[7] M. Hu and B. Liu, "Mining and summarizing customer reviews," in *KDD*, 2004, pp. 168–177.

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

[9] M. Marchetti-Bowick and N. Chambers, "Learning for microblogs with distant supervision: Political forecasting with twitter," in *EACL*, 2012, pp. 603–612.

[10] E. Mustafaraj, S. Finn, C. Whitlock, and P. T. Metaxas, "Vocal minority versus silent majority: Discovering the opionions of the long tail," in *SocialCom/PASSAT*, 2011, pp. 103–110.

[11] B. O'Connor, R. Balasubramanyan, B. R. Routledge, and N. A. Smith, "From tweets to polls: Linking text sentiment to public opinion time series," in *ICWSM*, 2010.

[12] H. S. Rad, A. Makazhanov, D. Rafiei, and D. Barbosa, "Leveraging editor collaboration patterns in wikipedia," in *HT*, 2012, pp. 13–22.

[13] D. Sparks, "Birds of a feather tweet together: Partisan structure in online social networks, presented at the 2010 meeting of the midwest political science association."

[14] M. Thelwall, K. Buckley, G. Paltoglou, D. Cai, and A. Kappas, "Sentiment strength detection in short informal text," *JASIST*, vol. 61, no. 12, pp. 2544–2558, 2010.

[15] A. Tumasjan, T. O. Sprenger, P. G. Sandner, and I. M. Welpe, "Predicting elections with twitter: What 140 characters reveal about political sentiment," in *ICWSM*, 2010.

[16] H. Wang, D. Can, A. Kazemzadeh, F. Bar, and S. Narayanan, "A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle," in *ACL (System Demonstrations)*, 2012, pp. 115–120.

[17] wikipedia.org, "Alberta general election, 2012: http://en.wikipedia.org/wiki/Alberta_general_election,_2012."

[8] http://blog.mastermaq.ca/2012/04/28/alberta-election-social-media-highlights/