The trouble with free elections is that you never know how they are going to to turn out.

– Vyacheslav Molotov, 1954.

University of Alberta

AN INTERACTION-DRIVEN APPROACH FOR INFERRING THE POLARITY OF COLLABORATIONS IN WIKIPEDIA AND POLITICAL PREFERENCES ON TWITTER

by

Aibek Makazhanov

A thesis submitted to the Faculty of Graduate Studies and Research
in partial fulfillment of the requirements for the degree of

Master of Science

Department of Computing Science

Spring 2013
Edmonton, Alberta

To Aqbota
For making me embrace the point of no return

Abstract

In this thesis we explore the interactions of users of two major information sources, namely Wikipedia and Twitter. In particular, we show that revision histories of Wikipedia articles contain interaction patterns which can be used to build collaboration profiles of editors, and that such profiles can be classified as positive or negative, depending on whether the collaboration was productive or not. The constructed profiles are shown to be useful in a number of related tasks, such as predicting votes in Wikipedia administrator elections and detecting controversial articles.

We further extend the ideas behind collaboration profiles and adapt the approach to the problem of predicting political preference of Twitter users. By considering tweeting on a party-specific topic as a form of a user-party interaction, we show that a record of such interactions, captured per user during an election campaign, can be used to predict the political preference of a user. We analyze how the predicted political preference of different groups of users change over time. Our results, for example, suggest that politically active users are less likely to change their preference during the course of an election campaign.

Acknowledgements

I would like to thank my supervisor Dr. Davood Rafiei who guided me patiently trough my research and gave me an opportunity to work on a number of interesting projects. It was a great pleasure working with him.

I am also grateful to Dr. Renée Elio for the financial support of my research.

For her valuable input in my early work I would like to thank Hoda Sepehri Rad. Various, and sometimes hard to set up experiments, that she used to come up with, were like a manual on research methodology and an additional practice in coding.

Many things I found useful in my research I learned from the following professors, whom I herein thank. Dr. Csaba Szepesvari for his indirect, but substantial influence on my coding style and on my growing love for Python. Doctors Paul Lu and Guohui Lin for excellent introduction into Unix systems and principles of algorithm design. Doctors Grzegorz Kondrak and Denilson Barbosa for introducing me to NLP and social network analysis techniques.

Big thanks go to my friends Agzam, Aibek, Damir and many others who helped to balance my time between work and leisure with a tiny bit emphasis on the latter.

Above all, I thank my family whose constant support I always felt.

Table of Contents

List of Tables

List of Figures

Chapter 1

Introduction

1.1 Thesis Overview

The widespread adoption of Web 2.0 technologies broke the barrier between consumption and production of content. Many users, who used to surf the "static" Web not so long ago, now interact with each other and with web sites by writing blogs and reviews, giving expert advice, sharing multimedia content, and engaging in many other activities of the kind. In this respect, Wikipedia stands out as a particularly fascinating example of collaborative content production.

Over the last decade, a free-to-all encyclopedia enjoyed a tremendous growth. At the moment of writing this thesis, English Wikipedia alone had over four million articles. Needless to say that to achieve this kind of productivity, collaboration among contributors is essential. However, given the sheer number of contributors, disputes and editorial conflicts do tend to happen. Therefore, the ability to infer the polarity of collaborations (productive or counter-productive) between contributors may prove useful in numerous applications involving the analyses of editor interactions, e.g. inter-contributor trust estimation, page-level controversy assessment, analysis of collaborative content production, etc.

The first problem that we address in this thesis is the problem of the inference of the collaboration polarity between Wikipedia contributors. We formulate the problem as a binary classification task, and given a pair of contributors and a history of their interactions, we infer whether they interacted in a productive or a counter-productive manner. The previous research addressed the problem indi-

rectly, in the context of page-level controversy assessment [30, 60], and, to a lesser extent, directly, in the context of the analysis of social networking aspects of collaboration [37]. In either case, collaboration polarities were inferred heuristically, assuming that, in contrast to conflicting editors, contributors who collaborate do not delete or edit each others' work. One of the main reasons to use heuristics is the absence of ground truth, and to the best of our knowledge, no attempt has ever been made to evaluate the results of the inference of collaboration polarities. Apart from the evaluation challenge, the problem presents a conceptual challenge, as no common definition of editor interactions is adopted, and a recent study [27] reports that three consecutive revisions of a page may encode up to more than 30000 distinct interaction patterns. In this respect, a well-adopted assumptions that reverts and delete actions always lead to conflicts [7, 55] and the addition of content constitutes collaborations [37] seem to oversimplify the problem. For this reason, we develop an interaction driven approach to the problem, accounting for the aforementioned challenges.

Whether a page was written in a collaborative manner, or underwent a period of controversy, all of its versions are stored in a so called revision history (see Section 1.4.1). We use revision histories to extract editor interactions[1]. Upon extraction, such interactions are arranged into collaboration profiles that additionally include various statistics on individual activities of editors. To classify collaboration profiles, and subsequently evaluate our method, we resort to a distant supervision approach, using votes casted in administrator elections in lieu of the ground truth. We find that for a given pair of interacting editors, activities of one editor influence the polarity of collaboration to a greater extent than those of the other and even than the pairwise interactions of both editors. With regards to the task of predicting votes in administrator elections, we observe that not all interactions are equally useful, and restricting profiles to very recent or very old interactions may hurt the accuracy of prediction. Also, in line with a previous research [34], we observe that votes are influenced by the transparency of the election process. Lastly, we give an

[1]Given a pair of editors, an interaction between them can be informally defined as an act of co-editing at least one page together.

example of additional application of our method, by reporting the results achieved by Sepehri et al. [48] who used our trained models in the task of automatic detection of controversial articles.

After a successful application of an interaction-driven approach to the problem of predicting votes in non-political elections, we adapt the approach to a political setting, where we address the problem of predicting the political preference of Twitter users. We set the preference prediction problem in the context of Alberta 2012 general election, and formulate it as a multi-label classification task, where given a user and a set of parties, we need to predict which party a user is most likely to vote for. At a user level the problem is relatively understudied, and with few exceptions [14, 19], most of the previous works focused on the extraction of political sentiment from individual tweets (without attribution to a user) [12, 38, 61] or from a corpus of tweets [44,58] (aggregate sentiment). We believe that approaches to political discourse analysis would benefit from the ability to distinguish between users with different political affiliations. For instance, predicted user preferences can be used as noisy labels when evaluating approaches to the community mining in political communication networks. Similarly, methods that focus on the extraction of political sentiment can use predicted political affiliations of users as additional features. Same goes for real-time analysis systems that could, in principle, use trained models to predict user preferences "on the fly". Finally, when used in the context of elections, political preference prediction has implications in better understanding changes in public opinion, and possible shifts in popular vote.

As for the adaptation of the aforementioned interaction-driven approach, it is achieved by drawing an analogy between political discourse on Twitter and editor interactions in Wikipedia. Specifically, treating political parties as abstract entities, we consider tweeting on a party-specific topic as a form of a user-party interaction, and by analogy with collaboration profiles, extract and classify a feature vector for each interacting user-party pair. As we show later in the thesis, doing so allows us to estimate user preferences on a per-party basis, and subsequently, reduce the multi-label classification problem to a one vs. all classification scheme, where a user is considered to support either the most preferred party or no party at all. To

the best of our knowledge, the latter possibility, i.e. predicting the "no preference", is not considered in any of the related works. Moreover, as we again explain later, our interaction definition naturally incorporates conventional definitions of Twitter-based user interactions, i.e. follower/followee relationships, retweets, replies, and mentions, allowing us to overcome limitations of the related works, where the political preference could be predicted only for followers of media [19] and individuals who retweeted political content [14].

As in the case with inferring collaboration polarities, evaluation is one of the major challenges associated with the problem. We semi-automatically construct the evaluation set by identifying users who revealed their preference in tweets, e.g. *"I voted NDP today!"*, soon after they voted. Measuring performance in terms of precision, recall and F-measure, we compare our method to human annotators, to a sentiment analysis based approach, and to chance. Our results suggest that although less precise than humans, for some parties our method outperforms the annotators in recall. Another experiment, where we analyze how preference of users changes over time, reveals that politically active users, or so called vocal minority, are less prone to changing their preference than so called silent majority, users who rarely engage in political discourse on Twitter. Our observations also suggest that changes in the predicted political preference co-occur with campaign-related events discussed heavily in social media.

1.2 Thesis Statement and Contributions

Our thesis statement reads as follows:

> *The history of interactions in a social environment can be used to infer the polarity of the pairwise relationships between individuals and between individuals and more abstract entities.*

At first glance, the connection of the statement to the problems addressed in this thesis may not be obvious. Hence, we provide a further elaboration. First, both Wikipedia and Twitter are, without a doubt, social environments. While some studies [11, 32] argue that follower relationships in Twitter bear a rather distant

resemblance to social ties, the thematic subnetworks, e.g. political followers and retweets networks, were shown [15, 54] to display properties of social networks, such as assortative mixing [43] and preferential attachment [54]. Similarly, the structure of Wikipedia collaboration networks was shown [37] to comply with the social psychology theories of status [21, 35] and structural balance [10, 23].

Second, in the context of both problems we define interactions. In turn, interaction histories are represented by revision histories of pages and volumes of political tweets, respectively in Wikipedia and Twitter.

Finally, both problems may be considered solved, when the polarity of a pairwise relationship between a given pair of individuals (Wikipedia) or an individual and an abstract entity (Twitter) is inferred. A relationship and its polarity, in our case, is either a productive or a counter-productive collaboration in Wikipedia, or the presence or the absence of a preference for a party on Twitter.

The following are the main contributions of this thesis:

1. We have developed two methods for:
   (a) inferring collaborations and disputes between Wikipedia editors,
   (b) predicting political preference of Twitter users.
2. In both methods we have used a common interaction-driven approach.
3. We have experimentally evaluated our methods on the tasks of predicting votes in:
   (a) Wikipedia administrator elections,
   (b) Alberta 2012 general election.

1.3 Thesis Organization

The thesis is organized into four major parts. Part I consists of two chapters. In Chapter 2 we briefly review the related work that addressed the problem of detecting instances of collaboration and dispute in Wikipedia. Chapter 3 contains a brief summary of the research on the extraction of political sentiment from Twitter.

In Part II we describe an interaction-driven approach to the inference of the collaboration polarity between Wikipedia editors. This part consists of three chapters. In Chapter 4 we define editor interactions and collaboration profiles. Along with the basic definitions we provide a detailed description of profile characteristics, and discuss approaches to the classification of profiles. In Chapter 5 we show how collaboration profiles can be used to predict votes in Wikipedia administrator elections. We also investigate how limiting the length of the revision history available for building profiles affects the accuracy of prediction. Lastly, in Chapter 6, we describe the application of collaboration profiles to the task of automatic detection of controversial articles [48].

In the next three chapters, that make up Part III of this thesis, we describe an interaction-driven approach to the problem of predicting political preference of Twitter users. In Chapter 7 we formulate the problem and provide background on Alberta 2012 general election. In Chapter 8 we define user-party interactions, explain the process of building interaction profiles, describe general methodology, and discuss results. In Chapter 9 we analyze how does the predicted political preference of different user groups change over time.

Finally, Part IV concludes our thesis with Chapter 10, in which we provide a brief summary of the work we did and the work that yet needs to be done.

1.4 Background and Terminology

In this section we provide definitions of the concepts related to Wikipedia and Twitter that will be referenced throughout this thesis.

1.4.1 Wikipedia Terminology

Wikipedia is a free multilingual online encyclopedia open for contributions of everyone. Let us describe some of the basics of Wikipedia.

- **A page** is a basic structural element of Wikipedia used to display its content. Pages in Wikipedia are organized into *namespaces* which reflect their purpose. For instance, pages in the *Article:Talk* namespace are dedicated to the discussion

of *articles*, while pages in *User:Talk* namespace are designed primarily for inter-member communications.

- **A section of a page** is a block of content that describes a particular aspect of the topic, much like a chapter of a book.
- **An article** is a page that contains encyclopedic information. Articles are contained in the *main namespace*.
- **The revision history** of a page is a record of all versions of a page. Any saved revision of a page results in a new version being created and the old version being stored in the revision history. We will refer to this process as *editing a page* or *revising a page*. Unless stated otherwise, we will use the terms *revision history* and *edit history* interchangeably. Same goes for terms *revision of a page, edit of a page, version of a page, page version, page revision*, and *page edit*.
- **The revision comment**, or an *edit summary*, is a short description of a revision that typically justifies an edit made to a page.
- **Revert** is a special revision that restores a page to one of its previous versions.
- **Same section edits** are edits made to the same sections of a page.
- **A contributor** is someone who *edits* Wikipedia content. Note, that contributors are not necessarily registered users. Unless stated otherwise, we will use the terms *contributor, editor, Wikipedia user, registered user, Wikipedia member, member of the Wikipedia community* and *wikipedian* interchangeably. Here, the term *Wikipedia community* is usually applied to a group of active contributors.
- **Edit wars**, or *edit warring*, is a series of revisions of the same page by two or more editors aimed at partial or full cancellation of each other's contributions. Edit wars usually result from disputes, where each party tries to defend its point of view and reflect it in the content of a page.
- **Wikipedia Administrator** is a member of the Wikipedia community with a privilege to perform certain special actions, such as blocking and unblocking users from editing, protecting and unprotecting pages from being edited, and performing several other related actions[2].

[2]http://en.wikipedia.org/wiki/Wikipedia:Administrators

- **Barnstar** or a *barnstar award* is a prize for hard work done by a contributor. Contributors are free to reward each other, and no formal procedure is involved.

1.4.2 Twitter Terminology

Twitter is a micro-blogging service that combines features of a social networking application and an instant messenger. One of the distinctive characteristics of Twitter is the widespread usage on mobile devises, which coupled with the extensive network structure, makes Twitter a medium where information propagates very fast. Below we provide some of the Twitter basics.

- **A tweet** is a message sent by a Twitter user. The length of a tweet is limited to 140 characters. Unless otherwise specified, we will use the terms *tweet, microblog, message* and *posting* interchangeably.
- **A follower** is a user who is subscribed to receive messages sent by a given user. Similarly a *followee* is a user whose messages are received by a given user. This relationship is somewhat similar to a "friendship" in a social networking setting, e.g. "friendship" on Facebook.
- **A retweet** is a tweet that is prefixed with a citation (*RT @username*) and contains a message received from a followee in whole or in part.
- **A reply** is a direct message to a follower or a followee.
- **A mention** is a reference (in a tweet) to another user, not necessarily a follower or followee. A user is typically referenced by its screen name which must start with "@" character and consist of no more than 15 alphanumeric characters.
- **A trend** is specific discussion topic.
- **A hashtag** is a label which is used to indicate that a given tweet belongs to a specific trend, i.e. *#politics*. Hashtags must start with the hash symbol "#", hence the name.

Part I

Literature Review

Chapter 2

Detecting Collaboration and Dispute in Wikipedia

2.1 Overview

More than a decade ago Wikipedia emerged as a free-to-all online encyclopedia, a phenomenon that attracted the attention of many researches from various fields. Early attempts to explain the driving force behind the collaborative content production at Wikipedia scale involved interviewing active contributors [8], applying economic theories [13], and visualizing editor interactions [55, 59]. As the encyclopedia continued its rise to prominence, the related research started to focus on the analysis of the underlying processes, such as contributors' acquisition of trust [63] and reputation [1], and their tendency to collaborate with each other or engage in disputes. However, inferring the polarity (a collaboration or a dispute) of editor interactions remains relatively understudied research area, in part, due to well-adopted approximations that treat delete and revert actions as instances of dispute [7,55,60], while addition of content is generally considered a sign of collaboration [27,37]. The need in simplifying assumptions can be explained with the absence of ground truth instances of collaborations and disputes at such a fine-grained (contributor-to-contributor) level. On the contrary, on a (coarser) page level ground truth is available in the form of explicit community-established labels, such as controversial[1] and featured[2] articles. Even intuitively, one would expect certain pages,

[1] http://en.wikipedia.org/wiki/List_of_controversial_articles
[2] http://en.wikipedia.org/wiki/Wikipedia:Featured_articles

such as articles on politics or religion, to express multiple contradictory points of view, providing the evidence of the ongoing or past editorial conflicts. Thus, numerous works addressed the problem of page-level controversy assessment, making observations about contributor-level collaborations and conflicts along the way.

2.2 Page-level Controversy Assessment

The works that focus on the page-level controversy assessment can be conventionally divided into two groups: studies that employ (i) visual, and (ii) quantitative analysis of pages (revision histories). With regards to the former approach, numerous studies proposed various visual representations of revision histories, such as flowcharts [59] and graphs [7, 55]. The goal was to identify distinct patterns of collaboration and dispute through a visual examination and comparison of a number of controversial and non-controversial articles. While Viégas et al. [59] identified patterns of vandalism and edit wars, which in their flowchart-like visualization corresponded to wide gaps and zig-zag shapes, Suh et al. [55] showed that visualizations of "revert graphs", networks in which nodes corresponded to contributors and edges to revert actions, revealed the controversial nature of articles. On the example of the article about Dokdo/Takeshima islands[3], the authors identified patterns of collective edit wars, where whole groups of contributors were reverting edits done by other groups. Using a similar network representation of revision histories, where nodes corresponded to contributors and edges to delete actions, Brandes et al. [7] showed that, as opposed to non-controversial articles, controversial ones could be split into two communities with two contradicting opinions on a given subject. From the visualization of the revision history of the "Telephone tapping" article, the authors observed that the amount of conflict (the number of edges representing deletes) was much greater between than it was within the communities.

To sum up, as exploratory methods the visualization techniques proved to be useful for spotting editorial conflicts and distinguishing between non-controversial and controversial articles, however, given the sheer number of articles, the practical

[3] A disputed islet in the Sea of Japan (East Sea) currently controlled by South Korea, but also claimed by Japan as Takeshima [55]

application of such techniques was simply infeasible. Therefore, a number of works employed approaches based on the quantitative analysis of revision histories.

Analyzing basic characteristics of articles, such as the length of revision histories and the number of edits of corresponding talk pages, Kittur et al. [30] showed that for controversial articles the growth in the number of revisions and talk page edits correlated positively with the growth in the amount of conflict, i.e. the number of received controversial tags. However, rather surprisingly, the increase in the number of unique contributors has been shown to correlate negatively with the increase in the amount of conflict, suggesting that, in contrast to common belief (the more people are involved in a discussion, the more chances it has to end up in a dispute), "having more points of view can defuse conflict" [30].

Vuong et al. [60] showed that the controversy level of articles was influenced not only by the number of contributors, but also by their tendency to engage in disputes. That being said, the authors increased the role of contributors in detecting conflicts at a page-level by assuming a mutual reinforcing relationship between contributors and articles. A verbatim formulation of the assumption states that "(i) an article is more controversial if it contains more disputes among less controversial contributors; (ii) a contributor is more controversial if s/he is engaged in more disputes in less controversial articles". To implement this, the authors adopted a variation of the famous HITS algorithm [31], treating pairs of conflicting contributors as hubs and articles as authorities. The amount of conflict between contributors was considered to be proportional to the number of words they deleted from each other's revisions. The link structure was represented as a bipartite graph with directed edges running from pairs of conflicting contributors to the corresponding disputed articles.

The authors evaluated their method in terms of top-k precision, recall and F-measure, counting controversially tagged articles in a list of articles ranked by the method. Although the model achieved a precision, on orders of magnitude, higher than a random ranking baseline, simple revision count-based ranking performed equally good and even slightly better for large enough k. This speaks in favor of the previous observation regarding positive correlation between the length of revision histories and controversy in articles [30]. The authors concluded with the analysis

of the ranked lists of articles and contributors, providing evidence of the aforementioned mutual reinforcing relationship. It has been shown that, in agreement with the initial assumption, the "lion's share" of the dispute in the two most controversial articles happened between contributors with low controversy scores.

To summarize, various aspects of inter-contributor disputes have been studied in the context of the page-level controversy assessment. However, due to specificity of the addressed problem, most of the reviewed works overlooked the aspect of collaboration. Although, visualization methods [7, 55] revealed some patterns of collaboration within conflicting groups of contributors, not even superficial analysis of such patterns was attempted. Nonetheless, collaboration patterns were extensively studied in the context of other problems, such as measuring individual contributions in co-authorized works [33, 56], modeling content growth [27], and building and analyzing collaboration networks of editors [28, 29, 37].

2.3 Analysis of Collaboration Patterns

As was mentioned earlier, due to challenges in evaluation, the related work used various simplified assumptions instead of actually inferring collaborations between editors. Assuming that two consecutive revisions (regardless of their type, i.e. insert or delete) constitute an instance of collaboration between corresponding contributors, Keegan et al. [28] demonstrated that collaboration patterns differed across various categories of articles. Specifically, the articles on breaking news possessed greater amount of collaboration and were edited in a more coordinated manner, than articles in other categories. Using a similar definition, but dropping the sequentiality requirement, Laniado et al. [33] considered any two revisions of the same article to result in a collaboration between corresponding editors. The authors showed that active editors who contribute high quality, trusted content [63] tend to collaborate with inexperienced users, thereby helping the newcomers. Other works considered a collaboration to happen between a pair of contributors who restore [37] or complete each other's work by adding new content [27, 37].

To our best knowledge, a study by Maniu et al. [37] is the only work that ad-

dressed the problem of the automatic detection of collaborations and disputes between editors. The authors developed a method for building collaboration networks of editors, where two interacting editors were connected by a directed signed edge, with the sign corresponding to a collaboration (positive) or a dispute (negative). The sign of an edge was determined from a set of positive and negative features, such as the number of words editors inserted, deleted or changed in each other's work, the number of barnstar awards they presented each other with, and the number of supporting or opposing votes they casted for one another in admin elections. As a basis for editor interactions, the authors used the notion of the word level ownership, according to which editors "owned" portions of the text they contributed to a page. An interaction was initiated between two editors, if one of them introduced changes to a portion of the text owned by the other.

Although, due to the absence of ground truth, the method has not been evaluated directly, the work provided some insight into the previous findings [33, 55] from a social networking perspective. The authors built the collaboration network for the editors who had revised articles in the "politics" category, and showed that, like in social networks, distributions of in- and out-degrees of nodes followed a power law. Moreover, the network structure complied with social psychology theory of status [21, 35] and, to a lesser extent, with the theory of structural balance [10, 23]. These findings suggest that, in agreement with observations of Laniado et al. [33], contributors tend to seek collaboration with editors of higher status than themselves. Moreover, as it was shown in [55], it is common to collaborate with others in disputes against the third party, i.e. unite against the common enemy. However, another real world "law of friendship", which states that it is unlikely for two enemies to have a common friend, seems not to apply to Wikipedia, and it is typical for conflicting editors to have a mutual collaborator.

To summarize, various patterns of collaboration and dispute [27] were extensively studied using a range of techniques from visualization [55,59], to application of social network analysis [37].

Chapter 3

Extracting Political Sentiment from Twitter

3.1 Overview

As a large evolving network and an ample source of user generated content, Twitter attracted the attention of researchers from various fields. Early exploratory studies suggested that the nature of follower/followee relationships hardly resembled friendship [25], and that breaking and sharing news was one of the main reasons behind micro-blogging [26]. In this respect, numerous studies on the role of news media on Twitter revealed, that, in terms of high retweet probability [32] and large number of followers [11, 62], media outlets had considerable amount of influence, especially in political trends [3, 19]. That being said, political tweets became highly retweeted and well discussed [14, 15], providing researchers with enough data to analyze political discourse.

Political discourse on Twitter has been extensively studied, with numerous works addressing the problems of measuring the degree of Twitter-involvement of politicians [45, 52], estimating political bias in media [2, 3, 19], detecting and tracking the spread of political abuse [40, 50], and extracting political sentiment (arguably the most well-studied problem in the field). Political sentiment can be expressed explicitly, in the content of tweets, or implicitly, by preferentially following or retweeting users who share same ideological beliefs. Correspondingly, the related work can be divided into two groups: studies that employ (i) content- and (ii) interaction-driven approaches.

3.2 Extracting Implicit Sentiment

Interactions subtly characterize users, and in the political context, they provide valuable clues about users' political preferences and ideological beliefs. In this respect, several studies showed that follower/followee relationships [19, 54] and retweet-based interactions [14] can be used to infer political preferences of users.

Studying a data set of 133 Congresspersons' Twitter accounts, Sparks [54] showed that both inter-politician and public-to-politician follower networks exhibited strong partisan division. The author observed that most of the users chose to follow only one politician, and, as became clear from profile descriptions of the users, such choices often reflected ideological beliefs of the followers. Livne et al. [36] addressed the problem in the context of US 2010 Senate, Congressional and gubernatorial elections, showing that the inter-candidate follower network could be split into two tightly knit communities that consisted predominantly from the same party candidates. These results suggest that "political" follower networks exhibit a certain form of preferential attachment [4], where users prefer to follow those who have similar political views.

Golbeck et al. [19] used this property of follower networks to infer political preferences of the followers of major media outlets. The authors identified users who followed both politicians and news media, and through these common followers propagated known ADA scores of Congresspeople[1] to media outlets. The ranking of news media according to the average ADA scores of their followers agreed with conventional beliefs, and Fox News (FN) and The New York Times (NYT) were placed at the opposite extremes of the ranking[2]. Specifically, the FN followers were deemed the most conservative and the NYT followers ranked second most liberal. An et al. [2] also placed these two media sources at the opposite ends of the ideological scale, showing that the distance, in terms of similarity of the followers sets, between media outlets reflects "a strong tendency of known political dichotomy" [2].

[1] ADA score is an estimator of the ideological leaning calculated by Americans for Democratic Actions (http://www.adaction.org/).

[2] Citing [16, 41, 47] the authors associate Fox News and The New York Times with conservative and liberal leaning media, respectively.

In short, users express implicit political sentiment in their decisions to follow certain politicians or media. However, some users follow several politicians and/or media outlets with different political leanings, which makes it harder to infer their true (the strongest) preference. In this respect, Conover et al. [15] focused on the analysis of political communication networks, i.e. retweet and mention networks. Using a semi-automatically constructed set of 66 political hashtags, the authors built communication networks for users who posted tweets that contained the target hashtags. The retweet network was built by connecting nodes representing authors of the original tweets to nodes representing propagators. The mention network was built in a similar fashion. Each network was clustered using the label propagation algorithm [49] for two labels. Visualization of the clustering revealed that the segregation level was much higher in the retweet network. The authors examined tweets of a random sample of 1000 users, and concluded that the clustering achieved for the retweet network reflected ideological leanings of the member users, whereas the mention network was politically heterogeneous. Moreover, in a follow-up work [14], using the same data set, the authors obtained political affiliations of more than 800 users, and evaluated the clustering algorithm against the ground truth, reporting the clustering accuracy of almost 95%. This suggests that users preferentially retweet only those tweets that agree with their own ideological beliefs, thereby implicitly expressing political sentiment.

To summarize, user interactions have been extensively studied in the context of political discourse analysis. It has been shown, that users express partisan bias in their choices to follow politicians [54] and media [3, 19], and to retweet political content [14, 15]. While interactions are helpful in analyzing user behavior, e.g. preference to follow or retweet certain politicians or media sources, they do not always reflect users' opinions on a specific political issue, e.g. a piece of legislation or an election campaign event. Thus, to extract actual (explicit) political sentiment, a number of content-driven approaches have been developed.

3.3 Extracting Explicit Sentiment

Content-driven approaches typically consist of two phases: topic identification and sentiment analysis. At the topic identification phase relevant tweets are extracted (usually from the popular trends), and at the next phase the general attitude (either positive or negative) towards the topic is inferred. A tweet is considered relevant to a party or a politician, if it contains a hashtag or a keyword associated with that entity. Usually a list of target keywords and tags is constructed manually [44,58,61], or semi-automatically, expanding a set of seeds by means of co-occurrence-based ranking [14, 15], or distant supervision [38].

Once a set of relevant tweets is identified, various sentiment analysis techniques [46] are applied to the content of each individual tweet [12,38,61], aggregate tweets of each user [14], or to the entire corpus [44]. At this stage most of the works employ either supervised or unsupervised methods. Unsupervised methods rely on so called opinion lexicons, lists of opinion words with corresponding sentiments, to estimate the attitude based on the positive-to-negative words ratio [44] or just raw counts of opinion words [12]. More sophisticated approaches employ supervised learning, and train prediction models either on manually labeled tweets [14, 61] or on tweets with emotional context, i.e. smiley faces :) [38]. Such models are usually built on unigram and bigram features extracted from the tokenized content of tweets.

Content-driven approaches found numerous applications. Conover et al. [14] developed a method for predicting political preference of users, and used prediction results to obtain a list of web sites popular among right- and left-leaning users. The authors reported low correlation between Twitter- and traffic-based (general) popularity of sites, and concluded that political campaigns might invest funds more effectively by purchasing advertising on sites that are less popular in general, but more popular among social media users with target political leanings.

Wang et al. [61] employed a classic two-step content-driven approach to develop a real time system for classifying political tweets. The authors evaluated a sentiment classification module of the system on a set of nearly 17 thousand manually

labeled tweets, and reported the accuracy of 59% for a multi-label (positive, negative, neutral and unsure) classification task. While the accuracy was not impressive, emphasis was made on the robustness of the online interface, and in particular, on a real time tweet evaluation feature, which allowed users of the interface to assess the sentiment of selected tweets. In this respect, the authors discussed the possibility for further improvement of the system by iteratively training the sentiment predictor on the freshly labeled data.

A number of works took content-driven approaches to address the problem of predicting political elections. While Tumasjan et al. [58] claimed that mere counts of party mentions in tweets reflected the outcome of German 2009 Federal election, in terms of popular vote, O'Connor et al. [44] and Choy et al. [12] reported low correlation of sentiment classification results with opinion polls and popular vote in US 2008 and Singapore 2011 Presidential elections, respectively. Metaxas et al. [39] put to test the predictive power of Twitter, and in the context of US 2010 Congressional elections showed that sentiment analysis techniques, in particular lexicon-based methods, were only slightly better than chance when predicting election results. Lastly, Gayo-Avello [18] conducted a comprehensive survey of several studies concerned with predicting elections, and suggested a number of improvements, such as considering bias in demographics, using stronger baselines, e.g. incumbency, accounting for sarcasm commonly found in political tweets [20], and addressing the issue of information credibility [40, 50].

Part II

Collaboration and Dispute in Wikipedia

Chapter 4

Building and Classifying Collaboration Profiles

4.1 Editor Interactions

While there is no universal definition of editor interactions, most of the previous works agree that wikipedians interact by means of either co-revising (editing together) same pages or editing different, but closely related pages. For instance, Leskovec et al. [34] define an interaction between two editors as the act of mutual editing of each other's *User:Talk* pages. As these pages are designed specifically for user communications, the proposed definition considers a very strong type of relationships as the basis of an interaction. However, despite having a solid foundation and being straightforward, this definition is limited to talk pages that account for a small portion of Wikipedia. Moreover, constant communication via user talk pages implies acquaintance, which, given a large number of contributors, may be true only for a small portion of editors.

Maniu et al. [37] consider the notion of *word level ownership*, according to which editors "own" portions of the text which they contributed to a page. Based on this notion, an interaction is initiated between two editors, if one of them introduces changes to a portion of the text owned by the other. Unlike the previous one, this definition does not impose limitations on types of the edited pages, nor it assumes acquaintance or any other pre-existent relationships between editors. However, the definition is based on comparison of each consecutive version of a page and attribution of portions of the text to contributors. Given that some articles

may have tens of thousands of versions, thousands of contributors, and long bodies of text, applying the proposed definition on a large scale may be computationally challenging.

In this thesis we adopt the notion of editor interactions proposed in [48].

Definition 4.1 *An interaction happens between two editors e_1 and e_2 if they both edit the same page and their edits are related.*

Definition 4.2 *Two edits are related if they are both applied to the same page and the distance between the edits in terms of the number of intermediate versions is less than a threshold.*

Definition 4.1 is a general definition of editor interactions based on the act of same page co-revision. It is Definition 4.2 which makes the distinction. Indeed, if we define relatedness of edits as revising portions of the text "owned" by others, Definition 4.1 transforms into the previously discussed definition based on the word level ownership [37]. One can also ignore the relatedness clause and focus instead on co-revisions of certain types of pages, which is somewhat similar to the definition proposed in [34]. Hence, due to the flexibility of the relatedness clause, Definition 4.1 is tunable with respect to the application.

We look for a definition of relatedness which is general with respect to the type of co-revised pages, and scalable with respect to the size and number of considered pages. Initially we experimented with the amount of time passed between two edits. Specifically, we considered two edits to be related if the amount of time passed between them was less than a threshold. A problem with this approach is that the average amount of time passed between two edits varies greatly across different articles, making it difficult to specify a single threshold. Given this, we rejected temporal relatedness in favor of Definition 4.2.

While being straightforward, Definition 4.2 is general with respect to the type of co-revised pages. Also, for a small enough relatedness threshold, the definition is applicable to the extraction of interactions on Wikipedia scale. The authors of the original work find the optimal value of the threshold based on the distribution

of distances between related edits. In the following paragraph we briefly describe the procedure.

Intuitively, two edits of the same section are semantically related, and, in all probability, more closely related than two edits of different sections of an article. Hence, for a randomly selected sample of articles the distribution of distances between all pairs of same section edits may be considered as a rough substitute for the distribution of distances between related edits. The authors obtained such distribution for a random sample of 100 articles, and computed the cumulative distribution function. The CDF of the distribution showed that around 42% of same section edits were consecutive (no versions in between), and around half of them were at most one version apart. Based on the analysis of the CDF of the distribution the authors set the relatedness threshold to 34, explaining the choice with the fact that around 90% of all pairs of same section edits were at most 34 versions apart.

Lastly, depending on the application, one may consider *directed* or *undirected* interactions. In this thesis we will work with directed interactions. We assume that, given a pair of related edits, an edit that happened later in the revision history is the *source* of an interaction. We refer to the author of such an edit as a *source editor*. Similarly, an edit that happened earlier is the *target* of an interaction, and it is said to be made by a *target editor*. Intuitively, whoever edits a page later *generates* interactions by editing the previous work of other contributors. Thus, an interaction is initiated by the source editor and directed to the target editor. We will refer to interactions between the source editor S and the target editor T as *S-T interactions*.

4.2 Collaboration Profiles

For a pair of editors who have at least one interaction (according to Definitions 4.1 and 4.2) we define the collaboration profile as follows.

Definition 4.3 *Given two interacting editors S and T, where S is the source and T is the target editor, an $S-T$ collaboration profile is a collection of statistical features which characterize the individual activities of each editor as well as their $S-T$ interactions.*

We deliberately emphasize direction in the definition. In general, an *S-T* profile is different from a *T-S* profile, and the existence of one does not imply the existence of the other.

The *sign* of a collaboration profile can be positive or negative, corresponding to productive or counterproductive collaboration. For brevity, we will sometimes refer to collaboration profiles as *c-profiles* or just *profiles*.

Sometimes it is important to infer collaboration over a certain period. For instance, one may investigate how collaboration prior to a certain event influences its outcome. In this case it would be wrong to include post-event activities and interactions into a profile. We define the *collaboration period* in the following manner.

Definition 4.4 *Collaboration period of a collaboration profile is a period of time to which the information in the profile is restricted.*

It should be noted that a collaboration profile consists of multiple interactions, but all of them are of the same direction and defined over the same relatedness threshold.

4.2.1 Characteristics of Collaboration Profiles

A collaboration profile consists of *individual* features based on *individual activities* of the source and the target editors, as well as statistics calculated based on the *interactions* of two editors and referred to as *pairwise* features.

Table 4.1 contains a list of 18 individual features. Each of these features are calculated for both the source and the target editor. The first three features provide the general merit of editor's contributions. These features were designed to distinguish newcomers from the experienced contributors who try not to engage in disputes and not succumb to trolling, i.e. provocation[1]. While the first two features have self explanatory names, the third one deserves elaboration.

Relative contribution is the collective share of editor's revisions in the set of all articles the editor has revised. Suppose editor e made a total of 10 revisions in 10 articles, i.e. a revision per article on average. Suppose also that each of the

[1] http://en.wikipedia.org/wiki/User:Moreschi/Wikithoughts, _Wikimorality,_Wikiphilosophies

#	Feature
1	Number of pages edited
2	Total number of revisions
3	Relative contribution
4	Average revision size
5	Average response size
6	Average revision time
7	Average response time
8	Number of reverts
9	Number of restored revisions
10	Number of reverted revisions
11	Number of ordinary revisions
12	Agreement in comments
13	Disagreement in comments
14	Agreement in respondent comments
15	Disagreement in respondent comments
16	Average page diversity
17	Average page conflict ratio
18	Co-revised pages share

Table 4.1: Individual features of collaboration profiles

articles had 10 versions. Then the editor's relative contribution will be calculated as $C' = 10/100 = 0.1$

Features 4-11 are designed to provide more detailed statistics on the editorial behavior. In particular, features 4-7 reflect quantitative aspects of editing, while features 8-11 distinguish between several types of revisions. For ease of exposition, let us define the notions of *target* and *response* on the example of a given revision r.

For a revision r, the revision which immediately *precedes* r in the revision history of a page is referred to as the *target* of r and is denoted by $r\prime$. Similarly, the *response* r'' to r is the revision which immediately *follows* r. *Respondent* is a contributor who commits the response.

Given an editor, the average revision size is calculated as

$$L = \frac{1}{|R|} \sum_{r \in R} (l(r) - l(r'))$$

where R denotes the set of all revisions of the editor, and $l(r)$ and $l(r')$ correspond to the lengths, in characters, of a revision and its target respectively. The average re-

sponse size is calculated similarly, with the term $l(r) - l(r')$ in the formula replaced with $l(r'') - l(r)$, where $l(r'')$ denotes the length of a response. Apart from quantifying the extent to which an editor changes the work of others and others change her work, these features may also subtly characterize an editor. For instance, if the average revision size is quite small, e.g. few dozens of characters, then an editor may be "specializing" in so called minor contributions. Such editors mostly fix orthography and punctuation, and as such do not cause or participate in disputes. On the other hand, if the average revision or response sizes are negative, then an editor may be engaged in edit wars.

In the same manner we calculate the average time passed between a revision and its target and response. The average revision time is calculated as

$$T = \frac{1}{|R|} \sum_{r \in R} (t(r) - t(r'))$$

where $t(r)$ and $t(r')$ correspond to the times, in seconds since epoch, when a revision and its target are made. The average response time is calculated similarly as in the case with the revision size. The intuition behind these features is that frequent editing may be the result of the necessity to constantly defend one's point of view in a dispute. Similarly, getting quick responses may suggest that an editor contributes either low quality or controversial content that is usually promptly spotted and edited by others.

In addition to quantitative characteristics, we also explore qualitative distinctions in editorial behavior by classifying revisions into the following four categories.

- *Revert* is a special revision that *restores* a page to one of its previous versions.
- *Restored revision* is the version of a page to which a page was restored by a corresponding revert operation.
- *Reverted revisions* are all versions of a page found between a revert and a corresponding restored revision.
- *Ordinary revisions* are all versions of a page not affected by revert operations.

Intuitively, a revert affects contributors of reverted revisions negatively, because all their work is nullified by the action. In fact, in order to prevent edit wars, the Wikipedia community introduced a special rule, called "Three reverts rule" or "3RR"[2], that restricts performing more than three reverts on a single page within a 24-hour period. Hence, high numbers of reverts and reverted revisions may suggest editor's engagement in edit wars. On the other hand, having many restored revisions has an ambiguous explanation. Sometimes most of such revisions are self-restores, i.e. an editor keeps restoring pages to self-authored revisions. This may suggest involvement in disputes, where an editor constantly attempts to restore pages to versions which reflect her point of view. However, if most of the restores are committed by other editors, an editor may have a good reputation, and her work might be valued by other wikipedians.

The next group of features, see features 12-15 in Table 4.1, was designed to roughly estimate the extent to which an editor agrees and disagrees with others. Specifically, we count occurrences of the agreement and disagreement terms in the revision comments of an editor and those of editor's respondents. We use the list of 77 Wikipedia-specific agreement and disagreement terms compiled by Sepehri et al. [48] from a set of manually labeled revision comments. Few examples of agreement terms are *"add", "fix", "spellcheck", "copyedit"* and *"clarify"*. Similarly, disagreement terms include such words and phrases as *"bias", "uncited", "revert"* and *"see talk page"*.

The editorial behavior may also depend on pages that one edits. To include information about revised pages into c-profiles we designed features 16-17, see Table 4.1. For a given editor, the average page diversity is calculated over the set of all pages P that she revised as

$$D = \frac{1}{|P|} \sum_{p \in P} \frac{e(p)}{v(p)}$$

where $e(p)$ and $v(p)$ are respectively the number of unique contributors who revised page p and the total number of versions of p. Intuitively pages with high fraction of contributors per version seem to be revised in a rather spontaneous manner where

[2]http://en.wikipedia.org/wiki/Wikipedia:Three-revert_ rule#The_three-revert_rule

everybody "do what they want" and no collaboration is intended. On the contrary, pages with lower fraction of editors per version may be maintained by small groups of dedicated contributors who tend to revise such pages in a collaborative manner.

Similarly to diversity, we calculate the average page conflict ratio as

$$C = \frac{1}{|P|} \sum_{p \in P} \frac{c(p)}{v(p)}$$

where $c(p)$ denotes the number of conflicting revisions in page p. We say that revision r is conflicting, if it satisfies any of the following conditions.

- The revision size of r is negative, i.e. $l(r) - l(r') < 0$.
- The comment of r contains more disagreement than agreement terms.
- r reverts at least one related edit, i.e. any other revision within the relatedness threshold.

The intuition here is that some pages may be controversial by nature, such as articles on abortion or euthanasia, and the prevalence of such pages in editor's profile may suggest her *a priori* tendency to get involved in disputes.

Finally, feature 18 (Table 4.1) estimates for each of the interacting editors, the share of pages edited in collaboration with the "profile-mate". This feature is calculated as the ratio of the number of pages which two editors co-revised to the total number of pages revised by each editor. Intuitively, if two editors collaborate intentionally, this quantity should be fairly high for both of them. Note, that, in a strict sense, the feature cannot be considered an individual characteristic, because it is calculated based on the activities of both interacting editors. However, the feature is calculated for *each* of the editors, and as such, it cannot be positioned as a pairwise characteristic either. If anything, this feature should be characterized as a directional characteristic [48], which is calculated for one editor with respect to another. However, given that it is the only feature of the kind, we choose to include it in the broader set of individual features.

Table 4.2 contains a list of pairwise features which are designed to reflect basic properties of a collaboration between two interacting editors. While the first two

#	Feature
1	Number of co-revised pages
2	Number of interactions
3	Average interaction distance
4	Share of consecutive interactions
5	Share of negative interactions
6	Average co-revised page diversity
7	Average co-revised page conflict ratio

Table 4.2: Pairwise features of collaboration profiles

features provide quantitative description of interactions on a general level, features 3-5 are designed to collect more fine grained statistics on interactions. Let us recall that an interaction happens between two editors whose edits are related, and that the relatedness of two edits is measured in the number of intermediate revisions between such edits. Given interaction I with the source edit s and the target edit t, we calculate the interaction distance as

$$d(I) = n(s) - n(t) - 1$$

where $n(s)$ and $n(t)$ correspond to the ordinal numbers of revisions s and t in the revision history of a page. Thus, we use the average interaction distance between two editors as a rough measure of the strength of their interactions. Intuitively, the smaller the distance, the stronger interactions are.

As an alternative measure of interaction strength we calculate the share of consecutive interactions. *Consecutive interactions* are those which have a zero interaction distance. To quantify the amount of dispute between the editors we calculate the share of negative interactions. *Negative interactions* are those that have a *conflicting revision* (see the definition of *the average page conflict ratio* on the previous page) as their source edit.

Finally, the last two features describe the properties of co-revised pages in the same manner as as the corresponding individual features, i.e. average page diversity and conflict ratio.

4.2.2 Classifying Collaboration Profiles

Having built a collaboration profile, the ultimate goal is to infer whether a collaboration between interacting editors was productive or counterproductive. That being said, a profile needs to be classified as positive or negative. The two most common ways to achieve this is to take a supervised learning or a heuristic approach. Some of the previous works take the latter, and consider heuristics based on changes in the topic distribution of the page content [6] or characteristics of the editorial behavior [37]. Although such techniques do not require training data, the evaluation may be challenging, especially if the ground truth is not available.

In our case, due to the lack of the ground truth labels, we resort to a distant supervision approach, and use votes casted in Wikipedia administrator elections as a close approximation of true labels. In other words, our assumption is that votes may be good indicators of the attitude of voters towards candidates and that voter-candidate collaboration may influence a vote. In the next chapter we will assess our assumption on the task of vote prediction.

Chapter 5

Predicting Votes in Administrator Elections

5.1 Motivation

Wikipedia administrators, or admins for short, are editors whom fellow wikipedians granted the power to perform certain special actions. To become an admin an editor must be nominated by another wikipedian or nominate herself. In either case a special Wikipedia page called *Request for Adminship* (RfA) is created. It is on this page, where an admin election takes place.

A promotion to an admin is a peer reviewed process which includes a discussion period, during which Wikipedia editors can comment on the nomination, ask questions and cast supporting, opposing or neutral votes for a candidate. At the end of the discussion period a *bureaucrat*, a member of the Wikipedia community with higher privileges than admin[1], determines whether there is a consensus for promotion. Thus, in a strict sense, the process cannot be considered an election. In fact, in numerous cases candidates, who were supported by more than 50% of voters, did not get promoted. However, we are interested in investigating how editor interactions affect *individual votes*, rather than predicting the final outcome of the process. In this respect, we focus more on the voting part of the process which clearly represents a regular election. Thus, in this thesis, unless stated otherwise, we use terms *a process of promotion to an admin* and *Wikipedia admin election* interchangeably.

In many ways the quality of Wikipedia articles depends on the ability of ad-

[1] http://en.wikipedia.org/wiki/Wikipedia:Bureaucrats

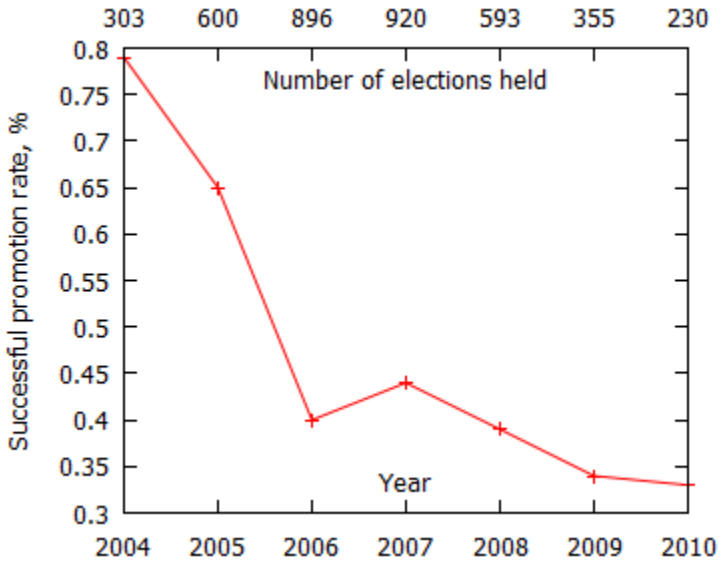


Figure 5.1: Annual rate of success in Wikipedia administrator elections

ministrators to coordinate the process of editing by monitoring articles and their respective discussion pages. Moreover, as a previous study [55] points out, administrators often try to resolve disputes and restore a neutral point of view in controversial articles. In this respect, it is important to understand which characteristics in the behavior and the editing style of adminship candidates account for a successful promotion.

According to Wikipedia statistics[2,3], becoming an admin can be challenging. As it can be seen from Figure 5.1, in seven years the percentage of successful promotions decreased to less than half, from 79% in 2004 to 33% in 2010. Overall out of 3897 elections held in this period 1785 (46%) were successful. While the previous research [9] shows that following the Wikipedia community guidelines and criteria does not always increases likelihood of a promotion and sometimes even hurts candidate's chances, one of the Wikipedia editors names personal revenge among the reasons of receiving negative votes, describing his/her experience of the election process as *"... the firing squad, where editors who broke policy and were rightfully blocked in the past took their opportunity to take revenge without fear of reprisal."*[4].

[2]http://en.wikipedia.org/wiki/Wikipedia:Unsuccessful_ adminship_candidacies_(Chronological)
[3]http://en.wikipedia.org/wiki/Wikipedia: Successful_requests_for_adminship
[4]http://en.wikipedia.org/wiki/User:Dayewalker

Revision time	Editor	Action	Δ	Comment
04:05:25	RJII	Insert	–	*anarcho capitalism removed redundancy*
04:07:29	Infinity0	Revert	-158	*revert to version without weasel words*
04:17:21	RJII	Revert	158	*infinity stop distorting tucker's words and making things up...*
04:24:27	Infinity0	Revert	-158	*shorter more concise version that does not include weasel words*
04:28:33	RJII	Delete	-1121	*anarcho capitalism if you can't represent tucker correct then don't say anything at all*
04:34:32	Infinity0	Revert	1121	*there is no difference in meaning between my version and your version grow up*
04:42:13	RJII	Insert	158	*if you think there is no difference then why do you oppose it...*
04:48:09	Infinity0	Revert	-158	*like i said it's too long hence why the section in the talk page is titled "bloating" also why "unlike anarcho communists"*

Table 5.1: An extract from the revision history of the "Anarchism" article as of February 27, 2006

We claim that a certain level of subjective evaluation may be involved in the election process. In other words, a pair of editors may have a predefined attitude towards each other. This attitude gradually builds up in the process of editing a set of articles together, and may later on influence the vote. The following example clearly illustrates this point.

On February 28, 2006, Wikipedia editor *Infinity0* posted a request for adminship[5]. On March 7, 2006 the request was declined with the candidate receiving five supporting, 32 opposing and six neutral votes. When casting votes, editors usually provide justification such as *"sorry, not enough experience"*, etc. The following is the shortened version of one of such comments made by editor *RJII*.

> *"NO WAY! The kid is OUT OF CONTROL. See edit histories and some of the articles he works on, such as anarchism... It appears to me that infinity's phi-*

[5]http://en.wikipedia.org/wiki/Wikipedia:Requests_for_adminship/Infinity0

losophy is to ban an editor whose edits would otherwise prevail by Wikipedia sourcing policy –so as to preserve the POV that infinity wants presented in an article... EXTREMELY unethical."

The comment clearly points to the past experience of editorial conflicts and counterproductive collaboration between editors *RJII* and *Infinity0*. Indeed, we find the evidence of an edit war between the two editors from a sequence of revisions of the "Anarchism" article presented in Table 5.1. Each row of the table corresponds to a revision of the article by the editor listed in the *Editor* column. The *Action* column contains basic editorial actions performed on the article. The Δ column represents the net change in the length of the article, measured in characters. We can see an alternating pattern with *Infinity0* rather frequently (2-7 minutes between revisions) reverting the revisions of *RJII*, thereby breaking "the three reverts" rule, and even suggesting the latter to *"grow up"*. Needless to say, that when later *Infinity0* sought promotion for an admin, *RJII* casted a strong opposing vote.

In this chapter we explore how collaboration or dispute between editors influences votes casted in admin elections. To this end, we formulate the vote prediction problem, and use collaboration profiles to solve it.

5.2 Problem Formulation

We formulate the problem of predicting votes in admin elections as follows:

1. Given a vote (v, c, t, s) casted by voter v for candidate c at time t, predict sign s of the vote.
2. Information used to predict the sign must be restricted to the period of time from t' till t.

where t' may be varied as a parameter. By default we set t' to the timestamp of the most recent vote of the opposite sign that v casted for c (see Section 5.4 for more detail). If there is no such vote, t' is set to negative infinity. We will refer to the (t', t) period as *pre-vote period*.

5.3 Methodology

We use collaboration profiles to address the vote prediction problem. Specifically, for each voter-candidate pair associated with one or more votes in the data set we build a profile, provided that the pair interacted within the corresponding pre-vote period. We then classify the resulting set of profiles using vote signs as true labels. Thus, we perform two tasks at once. First, we address the problem of vote prediction, and second, we obtain trained models that can be used to classify c-profiles built for any pair of interacting editors.

Finally, few words need to be said about interactions which make up voter-candidate profiles. Although editors may interact while editing any page, we limit interactions only on pages in the main namespace, i.e. encyclopedic articles. We do this mainly because pages in other namespaces, with exception of talk, user-talk and, perhaps, portal pages, are edited in a different manner. For instance, RfA pages in Wikipedia namespace, are edited in a message board fashion, where editors take turns posting questions and comments. Clearly, no collaboration is intended when editing such pages. Other pages are strictly technical, e.g. pages in file, template and mediawiki namespaces. Such pages usually do not involve lengthy discussions and have short revision histories. Hence, articles better reflect patterns of collaboration and dispute than other types of Wikipedia pages, and as such, they are considered a primary source of editor interactions.

5.4 Data Set Description

From the complete Wikipedia data dump dated April 5, 2011, we extracted voting records of 3713 elections. This resulted in a collection of 166432 votes and 9541 user IDs of unique participants (both voters and candidates). Each vote in our data set is represented as a tuple of the form:

(voter ID, candidate ID, timestamp, vote sign)

where the first three attributes uniquely identify a vote. We keep the timestamp of a vote for the following two reasons. First, the same candidate may seek promotion more than once, which may result in multiple occurrences of the same voter-

	Successful elections	Unsuccessful elections	All elections
Number of elections	1731	1982	3713
Number of unique participants	7472	6790	9541
Positive votes	101190	29003	130193
Negative votes	6396	29843	36239
Total votes	107586	58846	166432
Repeated votes	–	–	5601
Changed votes	–	–	1420
MAE of extraction	0.51	0.48	0.49

Table 5.2: Statistics of extracted election data

candidate pair. By keeping the time stamp of each vote we ensure uniqueness of an instance. Second, as we *predict* votes, we need to make sure, that we only consider voter-candidate interactions that happened *before* the time when a particular vote was casted.

Table 5.2 contains detailed statistics of the election data. We present separate statistics for successful and unsuccessful elections, as well as combined statistics for both. In line with Wikipedia statistics, shown in Figure 5.1, we notice that unsuccessful elections are slightly overrepresented in our data set. However, although being rare, successful elections are heavily discussed, with an average of 62.2 votes per election. Also, as one would expect, successful candidates are generally well supported, with 94% of votes in successful elections being positive. On the contrary, an average unsuccessful candidate receives only 29.7 votes per election, with 51% of votes being negative. The difference in vote counts for the two types of elections can be explained with the *snowball clause policy*[6] applied to some nominations, according to which, if a candidate does not stand a *snowball's chance in hell* to get a promotion, a discussion period is interrupted and a request for adminship is declined as early as possible. Thus, such nominations usually receive lower than average number of votes.

[6]http://en.wikipedia.org/wiki/Wikipedia:SNOW

As it was mentioned earlier, sometimes candidates make several attempts for a promotion, and it is quite possible that same voters vote for them in different elections. Suppose candidate c seeks promotion three times, and voter v opposes c the first time and supports the candidate in the last two elections. According to our vote representation, this translates into the following three tuples:

(v, c, t_1, –), (v, c, t_2, +), (v, c, t_3, +)

We refer to the vote *(v, c, t_2, +)* as the *changed vote*, because, compared to her previous vote, the opinion of v towards c has *changed* from opposition to support. Similarly, the vote *(v, c, t_3, +)* is referred to as the *repeated vote*, because, by casting this vote, v *repeated* her statement of support for c. We pay attention to these subtleties, because they affect the way we set collaboration periods for voter-candidate c-profiles, which we will explain later in this chapter. In Table 5.2, we report the counts of repeated and changed votes across all elections.

Finally few words need to be said about some parsing difficulties we encountered while extracting the data. The history of admin elections spans multiple years, with one of the first elections being held as early as 2003. In the early days, election pages had no fixed format, making it hard to distinguish vote statements from the rest of the content. The division on *Support*, *Oppose* and *Neutral* sections was optional, and even if a page was divided into sections, it was common to post opposing votes into support section and vise versa. Over the years the election process gradually evolved, and the format of election pages became more structured. However, the recent election pages have *very* extensive voting sections, with multiple editors commenting on a single vote statement and making it hard to identify the actual voter. Nevertheless, for the majority of election pages the final vote tally is usually contained within the first 2-3 paragraphs. As a rough measure of the extraction error, we consider an average deviation of the number of extracted votes from the number of reported votes. More specifically, we report the mean average error (MAE) of the vote extraction process calculated as

$$MAE = \frac{1}{|E|}\sum_{i=1}^{|E|} |e^+ - r^+ + e^- - r^-|$$

Profiles	Full data	Balanced data
Positive	75168	14484
Negative	14484	14484
Total	89652	28968

Table 5.3: Voter-candidate collaboration profiles: Full and balanced data sets

where E is a target set of elections, e^+ and e^- are respectively the numbers of positive and negative votes *extracted* from a particular election page, and r^+ and r^- correspond to the respective numbers of positive and negative votes *reported* on that page.

As it can be seen from Table 5.2, across all elections we have incorrectly extracted 0.49 votes per election, or 1819 votes overall, which corresponds to only 1% of the total number of votes.

5.5 Experiments and Evaluation

Using the optimal value of the relatedness threshold of 34 (see Section 4.1), we extracted almost 5 million voter-candidate interactions. This allowed us to build c-profiles for 89652 votes, which is approximately 54% of all votes. Around 84% of all profiles were associated with positive votes. Because of the imbalanced nature of the data, we also considered a balanced data set, constructed from the full set by randomly removing positive examples until a perfect 50/50 split was achieved. Statistics on both data sets are given in Table 5.3.

For the prediction task, we experimented with several classifiers provided by Weka [22], namely decision tree based Random Forest (RF) and J48, SVM based SMO, and Logistic Regression (LR). Unless stated otherwise, we set parameters of classifiers to their default values.

For comparison purposes, we considered several baselines. The all positive (AP) baseline simply assumes all votes to be positive. Given the imbalanced nature of the full data set, this simple strategy may be highly effective.

The relative edits count (REC) baseline is based on the observation that voters tend to compare candidate's achievements with those of their own [34]. In partic-

ular, the more edits a candidate has compared to those of a voter, the higher the probability of a vote to be positive. Thus, the baseline labels a vote as positive, if a candidate has more edits than a voter.

The last baseline is based on the fact that in a given election the current vote tally is always publicly available, and each voter, as she votes, is aware of the number of positive and negative votes already received by a candidate. As Leskovec et al. [34] show, this knowledge affects a voter, and the higher the fraction of positive votes received by a candidate the more it is likely for the next vote to be positive. That being said, the positive fraction (PF) baseline labels a vote as positive with probability equal to fraction of positive votes in the current tally. The first vote of an election is labeled as positive or negative with equal probability.

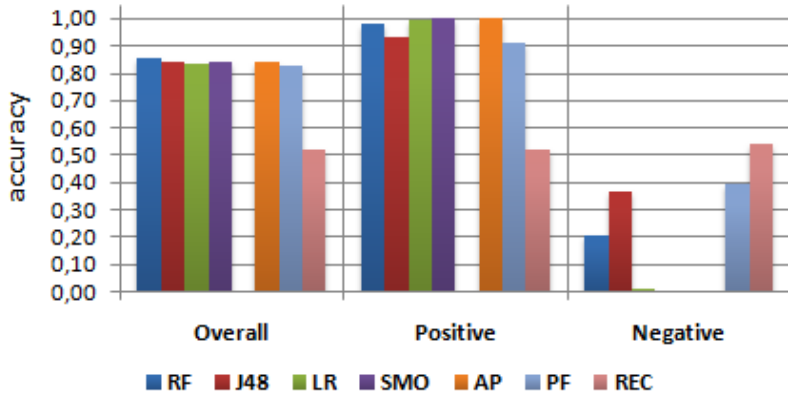
We evaluate the quality of prediction in terms of accuracy, i.e. percentage of correctly predicted votes. For classifiers we report 10-fold cross validation results.

5.6 Results and Discussion

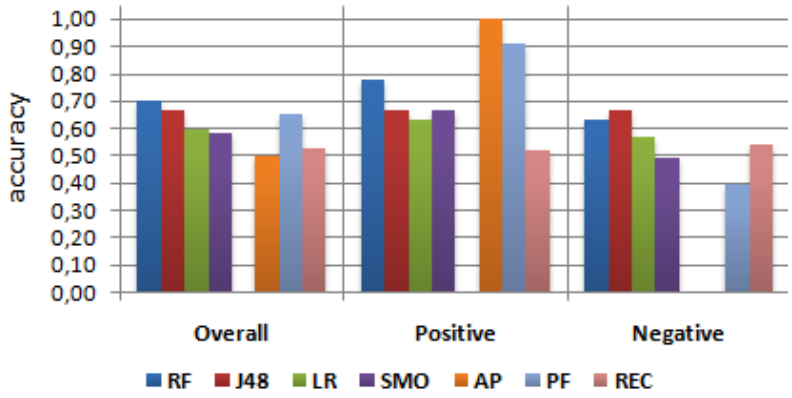
Figure 5.2 shows the results of the prediction for both full and balanced data sets. The overall accuracy and the accuracy of predicting positive and negative votes are shown in three correspondingly labeled clusters. Each cluster is also divided into two parts, separately showing the performance of classifiers and the baselines. The order in which bars are plotted corresponds to the order given in the legend.

As it can be seen from Figure 5.2a, on the full data set classifiers perform only slightly better than baselines. Random Forest classifier shows the highest overall accuracy of 85.5% which is almost 2% higher than the accuracy of all positive baseline, which is 83.8%. The rest of the classifiers show less than 0.5% improvement over this baseline.

Obviously, the AP baseline predicts positive votes with 100% accuracy, because it labels everything as positive. What is surprising that LR and SMO classifiers display the same behavior. While LR labels some votes as negatives (1% accuracy when predicting negatives), SMO behaves exactly like the AP baseline and labels everything as positive. In general all methods, except relative edit count baseline,



(a) Accuracy on full data set



(b) Accuracy on balanced data set

Figure 5.2: Results of vote prediction on full and balanced data sets

achieve above 90% accuracy when predicting positive votes.

Rather unexpectedly, the classifiers showed weaker performance than baselines when predicting negative votes. We explain such a behavior with the imbalanced nature of the data, and to support our claim, we run the experiment again on the balanced data set.

Figure 5.2b shows the accuracy of vote prediction on the balanced data set. As was mentioned earlier, to make the data set balanced we randomly remove positive training examples until we get a perfectly balanced data set. Depending on the random seed, repetition of the procedure results in different sets of positive examples. To account for such fluctuations we obtain 10 instances of balanced data sets. We then test classifiers and baselines on each instance and report the average performance over 10 runs.

Let us first compare the performance of classifiers. In terms of the overall accu-

Rank	Characteristic	Feature type
1	Average page conflict ratio	Individual, target editor
2	Agreement in comments	Individual, target editor
3	Number of pages edited	Individual, target editor
4	Total number of revisions	Individual, target editor
5	Average page diversity	Individual, target editor
6	Number of ordinary revisions	Individual, target editor
7	Average response time	Individual, target editor
8	Number of restored revisions	Individual, target editor
9	Number of reverts	Individual, target editor
10	Number of reverted revisions	Individual, target editor

Table 5.4: Top 10 characteristics of collaboration profiles

racy RF classifier once again shows the best result, gaining 3% improvement over J48, the next best model. When predicting positive votes RF is 11% more accurate than J48 and SMO, both of which correctly predicted 67% of positive votes. All of the classifiers, however, are less effective than PF baseline, which predicts positive votes with 91% accuracy. We will reason about this fact later in the chapter. As for predicting negative votes, all classifiers, except SMO, are more accurate relative to chance, and achieve better accuracy than baselines. This fact supports our intuition that imbalanced nature of the data affects the accuracy of negative votes prediction on the full data set.

To sum up, two decision tree based classifiers, RF and J48, can be distinguished from the rest of the models. While RF predicts positive votes more accurately and is slightly better in terms of the overall accuracy, J48 is more effective, especially on the full data set, when predicting negative votes. Also, as it can be seen from Figure 5.2b, on the balanced data J48 predicts votes with equal per-sign accuracy. Thus, J48 suits best for the unbiased classification of c-profiles, while RF is better at detecting collaboration. Lastly, with a careful parameter tuning, the overall accuracy of RF can be boosted up to 86.9% and 78.1% on the full and balanced data sets respectively [48].

5.6.1 Feature Analysis

Table 5.4 contains 10 of the most important characteristics of the voter-candidate collaboration profiles ranked according to the information gain-based ranking. As it can be seen, the top-10 consists entirely of the individual features calculated with respect to a target editor, i.e. a candidate. This is not surprising, given the fact that we used votes in favor or against candidates to infer their respective collaborations or disputes with voters. While some features, such as *the number of revisions* and *of edited pages*, are considered standard measures of editor contributions and were shown to be used to evaluate candidates [34], other features are often used outside of the election context to characterize an editor's tendency to collaborate with others or engage in disputes. Thus, in line with a previous research [60], the properties, such as *average conflict ratio* and *diversity*, of pages edited by a contributor are shown to be important when inferring the polarity of her collaborations with others. Also, *the number of reverts* is considered an important characteristic (although ranked relatively low as 9th), which is expected, given that reverts were shown to be one of the main "weapons" of edit wars [55]. Even the features like *agreement in comments* and *average response time* subtly characterize interactive aspects of editorial behavior. We therefore, conclude that although our ranking seems to favor individual characteristics of candidates, many features reflect interactive aspects of collaboration, suggesting that votes casted in admin elections are, to certain extent, influenced by the history of pre-vote collaborations and disputes.

It should be noted that in a related study Sepehri et al. [48] also ranked the same set of characteristics of collaboration profiles, using classifier-specific (Random Forest) ranking method [51]. Although the reported ranking has only four features in common with our top 10 characteristics[7], in agreement with our ranking, all of the top 10 characteristics were individual features. Moreover, nine of them were calculated with respect to a target editor, and only one – with respect to a source editor, i.e. a voter. Lastly, although individual characteristics of candidates, such as *total number of revisions* and *relative contribution*, ranked higher,

[7]The features common to both rankings are: total number of revisions, agreement in comments, average response time, and number of reverted revisions.

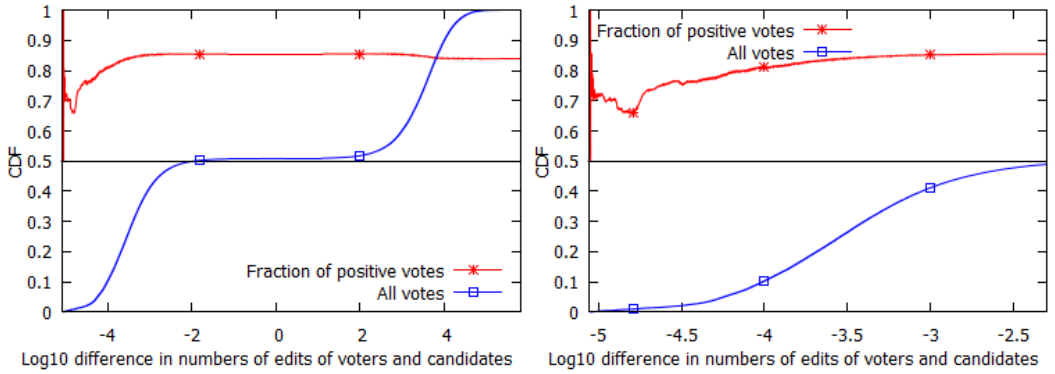
features implying collaborations or disputes, such as *agreement* and *disagreement in comments*, *average times of revision* and *response*, etc., prevailed, supporting the conclusion we made earlier.

5.6.2 Analysis of Baseline Performance

The analysis of the performance of the PF (positive fraction) and REC (relative edit count) baselines on both data sets leads to an interesting observation. First, it seems that both baselines are not affected in any way by the imbalanced nature of the data. Indeed, PF predicts positive and negative votes with 91% and 39% accuracy regardless of the data set. Similarly, on both data sets REC shows 52% and 54% accuracy when predicting positive and negative votes. Such a behavior of PF suggests that voters are indeed affected by the current state of the vote tally, meaning that a voter is less likely to oppose a candidate who, at the moment, is supported by the majority. On the other hand, candidate's having more edits than a voter does not increase her chances of being supported. In fact, REC baseline performs only slightly better relative to chance. Figure 5.3 illustrates our point.

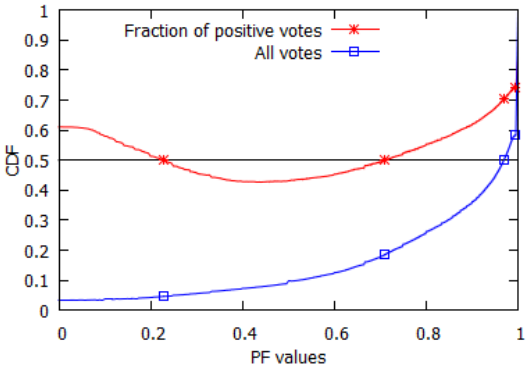
Figure 5.3a depicts cumulative distribution of votes across the absolute difference in edit counts for a respective voter-candidate pair. We refer to this quantity simply as the REC value of a vote. For ease of presentation, we plot signed log10 values instead of actual values of REC. A fraction of positive votes plotted on the figure represents the probability of casting a positive vote given the difference in edit counts. We can see, that for the most part of the distribution the probability of positive vote is about the same as the share of positive votes in the data set, which is $\approx 84\%$. The fluctuations for the very low REC values account for less than 10% of all votes, as can be seen on the rescaled version of the plot, Figure 5.3b. This means that except for a small fraction of voters, who oppose candidates who have much more edits than themselves, candidate's having either more or less edits than a voter has no considerable influence on a vote.

Figure 5.3c shows the cumulative distribution of votes across *PF-values*. We define the PF-value of a vote to be equal to the fraction of positive votes in the election tally at the moment the vote is being casted. A fraction of positive votes

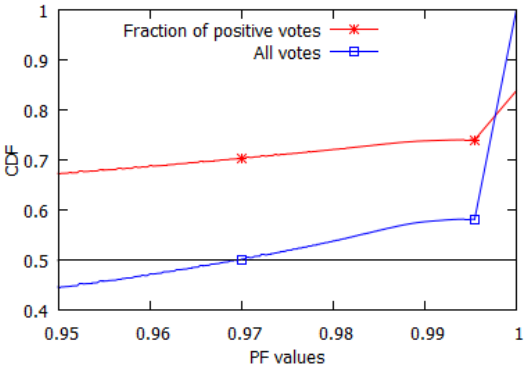


(a) Relative edit count (REC)

(b) REC rescaled



(c) Positive fraction (PF)



(d) PF rescaled

Figure 5.3: CDF of votes across positive fraction and relative edit count values

represents the probability of supporting a candidate given the current PF-value. We can see that less than 5% of voters tend to support (positive vote probability >0.5) a candidate despite the opposition of the majority (PF-value <0.25). The other 17-18% of voters are less likely to support a candidate whose current vote tally is less than 70% positive. Lastly, about 80% of all voters are likely to support a candidate who is already supported by more than 70% of previous voters.

The rescaled plot shown on Figure 5.3d shows the distribution of votes with PF values above 0.95. Note, that half of all votes are casted when the vote tally is more than 97% positive! These votes have more than 70% chance to be positive. Moreover, around 40% of all votes are casted when the vote tally is all positive and these votes have around 97% chance to be positive (83% shown on the figure is the *cumulative* fraction, and the positive ratio of votes which have PF-value of 1 is higher, or 97% to be precise).

From our observations we draw the following conclusions. First, merits of a voter relative to a candidate do not influence the majority of votes. Specifically, the difference in edit counts seems to have a significant effect only on votes casted by about 1% of voters, and the effect on the vast majority of votes is rather insignificant. Second, as an election unfolds, the probability of subsequent votes being positive grows with the increase in support received by a candidate. This fact can be explained by the transparency of the voting process. Because everyone is aware of the votes casted by everyone else, one needs very strong argumentation to oppose a heavily supported candidate. Thus, having editorial conflicts in the past may not be sufficient to oppose strong candidates. This ultimately means, that the format of elections in itself may impose limitations on the predictive power of methods based on the analysis of pre-election collaboration and dispute.

5.7 Manipulating Collaboration Period

Let us recall that according to the formulation of the vote prediction problem, the collaboration period of a profile is set to the pre-vote period of the corresponding vote. In other words, information used in profiles is restricted to a period prior to voting. In most of the cases this period is not limited from below, i.e. the entire history of pre-vote collaboration is used to predict a vote. The question then arises: how does the length of the collaboration period affects the vote prediction accuracy? It could be that recent collaborations, productive or otherwise, influences votes to a greater extent than "long forgotten" interactions. Hence, limiting profiles to recent interactions may improve the accuracy of prediction.

5.7.1 Experiments

To test our hypothesis we set up the following experiment. First, we identify voter-candidate pairs who have interacted a week, a month, six months and a year prior to a vote. We then group such pairs according to these periods. For each pair in each group we build a series of profiles, where each subsequent profile has a longer collaboration period than the previous one. For instance, suppose voter v and can-

didate c have interacted a week before v voted for c. For this pair of editors we build four profiles with respective collaboration periods of one week, one month, six months, and one year. Additionally, we build the fifth profile whose collaboration period is not bounded from below. Now, suppose that we have 1000 such pairs each corresponding to a vote. In this case we will have 5000 profiles, or five sets of 1000 profiles. Profiles in each set correspond to same votes, but have collaboration periods of different lengths. Thus, by classifying profiles in each set and comparing the accuracy of classification, we can assess the impact of the length of the collaboration period on the vote prediction accuracy.

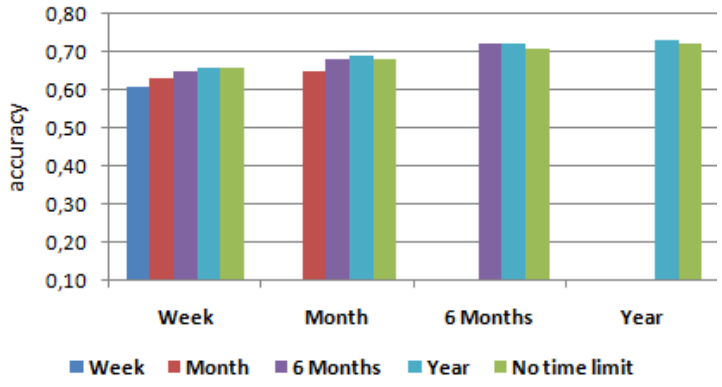
In this experiment we use the Random Forest classifier which achieved the best performance on the general vote prediction task. We perform 10-fold cross validation on each set of profiles and report the overall and per-sign accuracy.

5.7.2 Results

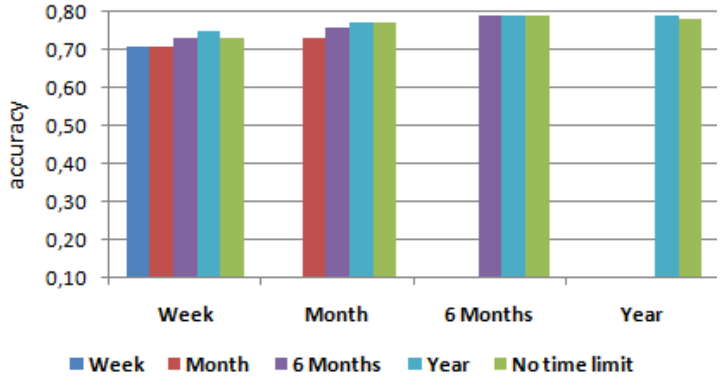
Figure 5.4 shows the results of the experiment. Each plot consists of four clusters of bars associated with four groups of voter-candidate pairs who interacted within the corresponding pre-vote periods. Each bar represents the accuracy of vote prediction on the set of profiles restricted to the corresponding collaboration period. The order of bars, from left to right, corresponds to the order of legend items.

As shown in Figure 5.4a, depending on the starting point, the increase in the length of collaboration period can either boost or drop the overall accuracy of prediction. Let us consider the change in the accuracy over the week cluster. Starting from a week long period, the further we extend the collaboration period, the higher accuracy we achieve. Precisely, the net increase in accuracy over the cluster is 6%, from 60% for a week long to 66% for the unbounded collaboration period. However, looking at the next three clusters, we observe that considering interactions older than a year actually decreases the accuracy of prediction. In all of these clusters the accuracy decreases by 1%, when the collaboration period is extended to more than one year.

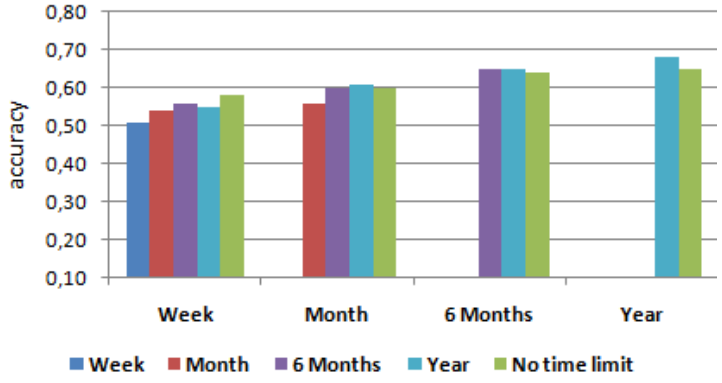
Figures 5.4b and 5.4c show the accuracy of predicting the positive and the negative votes respectively. We can see that the net increase in the accuracy per



(a) The overall accuracy



(b) Accuracy of the positive votes prediction



(c) Accuracy of the negative votes prediction

Figure 5.4: The effect of extending collaboration period

cluster is larger for the negative compared to the positive votes. In the first two clusters the increase is 7% and 4% for the negative votes, compared to 2% and 4% respectively for the positive votes. This suggests that the positive votes are to the lesser extent influenced by the history of collaboration than the negative ones, which is not surprising considering our previous findings about the bias introduced

by the format of elections.

Let us return to our initial speculations that limiting profiles to more recent interactions may improve the accuracy of prediction. The results of the experiment show that it is not entirely true. Restricting profiles to *very* recent interactions, such as a week or a month prior to a vote, decreases the accuracy. On contrary, extending such short periods increases the overall accuracy. However, starting from the period of six month the further extension does not affect the accuracy. Hence, interactions that happened long before voting do not have considerable impact on votes. Moreover, extending the period to more than one year can even hurt the accuracy of prediction, which is especially the case when predicting negative votes.

Chapter 6

Detecting Controversy in Wikipedia Articles

6.1 Motivation

The openness of Wikipedia is considered both its greatest strength and weakness. The content of the encyclopedia is produced by the sheer number of contributors with diverse backgrounds. While collaborative editing makes articles more balanced, reducing regional and cultural bias, it takes time and effort to produce high quality articles that express neutral point of view. Many articles undergo a period of controversy, when the content is repeatedly changed and editors do not agree on certain aspects of the topic. Automatic detection of the ongoing controversy in such articles could potentially benefit both readers and editors of Wikipedia.

As an example of another practical application of collaboration profiles, we briefly describe a method for detecting controversial articles proposed by Sepehri et al. [48]. Collaboration profiles, as defined in this thesis, are used in the cited work to build collaboration networks of articles, and, subsequently, to classify articles as controversial or non-controversial.

6.2 Detecting Controversial Articles

The problem is modeled as a binary classification task, where a target set of articles needs to be classified into controversial and non-controversial classes. The authors evaluate their method on a set of carefully selected 480 articles, 240 per each class.

Each article is associated with a feature vector, which is extracted from the corresponding collaboration network. The authors use a total of 30 structural features of collaboration networks. To name a few, they are numbers of nodes and edges, absolute and relative numbers of triads (three interconnected nodes) and features based on various quantitative characteristics of in- and out-degrees of nodes. The collaboration network of an article is built in the following manner.

Ignoring the contributors who edited an article only once, the authors build collaboration profiles for all pairs of interacting editors. These profiles are then classified as positive or negative using the Random Forest classifier trained on the admin election data. Lastly, the set of profiles is naturally transformed into a directed signed network, where each profile corresponds to a signed edge.

The authors compare their method to several baselines and to existing methods for building collaboration networks [7, 53]. The method based on collaboration profiles achieves $\approx$85% classification accuracy, which is 10% higher than the next best result. When used together with a set of content-based features the method achieves even higher accuracy of 89%.

From the results achieved by Sepehri et al. [48] we conclude that c-profiles can be used to infer collaboration and dispute not only on the editor level, but on the article level as well. Moreover, the fact, that the model trained on voter-candidate collaboration profiles can be used to accurately identify controversial articles suggests, that votes casted in admin elections are, indeed, good indicators of collaborations and disputes between editors.

Part III

Political Discourse on Twitter

Chapter 7

Predicting Political Preference of Twitter Users

7.1 Motivation

Today Twitter stands out as one of the most popular micro-blogging services. On Twitter, information propagates in no time, and words and actions trigger immediate response from users. Such an environment is ideal for advertising political views, especially during the heat of an election campaign. Thus, using the medium as a "social sensor" for gathering and analyzing live feedback from the electorate is arguably the most sought after application of the Twitter-based political discourse analysis. To this end, numerous methods for detecting and mining political tweets have been developed, including real-time systems for extracting political sentiment [61] and tracking political abuse [50]. However, with few exceptions [14,19], most of the previous works did not account for the political preference of users, focusing instead on the analysis of individual tweets (without attribution to a user) [12, 38, 61] or aggregate properties of a corpus of tweets [44, 58].

We believe that approaches to political discourse analysis would 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. Similarly, methods that focus on the extraction of political sentiment can use predicted political affiliations of users as additional features. Same goes for real-time analysis systems that could, in principle, use trained models to predict user preferences "on the fly". Finally, when used

in the context of elections, political preference prediction has implications in better understanding changes in public opinion, and possible shifts in popular vote.

In this chapter we formally state the problem of predicting political preference of Twitter users. We address the problem in the context of political elections, and consider a vote for a party to be the best indicator of the political preference of users. As a case study we use Alberta 2012 general election, on which we provide a brief background in the end of the chapter.

7.2 Problem Formulation

We state the problem of predicting political preference of users as follows:

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

Our problem formulation is set in the context of political elections, and relies on the notion of *a user-party interaction* which informally can be defined as tweeting on a party specific topic. A formal definition of a user-party interaction will be provided in the next chapter together with a detailed description of an interaction-driven approach that we took to solve the preference prediction problem. Lastly, the problem formulation has two key points worth further elaboration.

- *Voting is not always preferring and vice versa*. We treat the two concepts interchangeably, and in the context of elections it is not unreasonable to do so. However, sometimes voters sacrifice their votes, and in order to block a party they least want to see in the legislature, they vote for a party that has better chances to win than does their preferred one.

- *A user may prefer none of the parties she interacted with*. Rather than implying that an individual would always vote for the most preferred party, we decided to account for a "no vote" scenario by introducing a preference threshold. Thus, if the most preferred party does not get a minimum preference score, we assume that a given user will not vote for any party. We will discuss this assumption in more detail in the next chapter.

Party	Abbrev.	Seats won	Popular vote
Progressive Conservatives	PC	61	43.95%
Wildrose Aliance	WRA	17	34.29%
Alberta Liberals	LIB	5	9.89%
New Democratic Party	NDP	4	9.82%

Table 7.1: Results of Alberta 2012 general election

7.3 Background on Alberta 2012 General Election

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, becoming the main challenger to PC[2]. Two other major parties who nominated 87 candidates, one per each riding, were Alberta Liberals (LIB) and New Democratic Party (NDP). Other parties like Alberta party, Evergreen party and independent candidates have had low polling numbers and popular vote shares, and they are not considered in this thesis.

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 popular vote to PC, their candidates were second in 56 ridings, losing by a tight margin in dozens of constituencies. Table 7.1 contains results of the election, along with the abbreviations that are used throughout this thesis to refer to the parties.

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

[2] http://en.wikipedia.org/wiki/Alberta_general_election,_2012

Chapter 8

An Interaction-driven Approach to Preference Prediction

8.1 Motivation

There are a number of ways for approaching the political preference prediction problem. The three main directions are to exploit network structures, content-based clues, or both. The choice often depends on the context in which the problem is set. For instance, Golbeck et al. [19] address the problem of computing political preference among media followers. The authors use follower networks to propagate known ideology scores of politicians to media outlets. Similarly, Conover et al. [14] set the preference prediction problem in the context of general political discourse, and use retweet networks and content-based unigram features to infer political affiliations of users. We set the problem in the context of political elections, and to solve it we adopt an interaction-driven approach developed in Part II of this thesis.

Let us recall that we have successfully used collaboration profiles to predict votes in Wikipedia administrator elections. Of course, admin and political elections differ drastically, as do the media (Wikipedia and Twitter) in which elections are discussed, but the beauty of the approach is that profiles (features) can be changed while the general scheme stays untouched. That is, we still have a single feature vector per each user-party pair, and to build such a vector we need at least one user-party interaction. Another advantage of the approach is that interactions can be used as both indicators of users' relevance to the task, and features for building

prediction models. In other words, if a given user has not posted a single election-related tweet, there is no point in trying to predict her political preference, at least using Twitter[1]. On the other hand, each election related tweet can be considered an interaction, and a record of all interactions produced by a given user may be used to calculate various statistics, and subsequently, to predict user's preference.

The question then arises: how do we define user-party interactions? A straightforward solution would be to consider standard user-to-user interactions (retweets, mentions, and replies) between users and official accounts of parties. This solution, however, limits the size and diversity of a sample of target users, as not everyone who tweets about the election interacts directly with party accounts, and those users who do typically belong to the vocal minority [42]. Therefore, in the next section we propose a more general definition of interactions which covers standard definition and has several other advantages.

8.2 User-Party Interactions

First, let us consider possible Twitter representations of the *"party"* in user-party interactions. Clearly, there is no separate entity for a political party, and the closest analog would be an official party account. However, as we discussed earlier, considering standard interactions with official party accounts may impose limitations and bias. Of course, given a party, it is possible to treat accounts of party candidates and the official party account as one super-account, thereby reducing limitations and bias, but we propose a more elegant solution.

We represent a party as an abstract entity, and with each party we associate a ranked list of weighted terms. We call such a list *the interaction profile of a party* and define it as follows.

Definition 8.1 *Given a party* p*, the interaction profile of* p *is a ranked and weighted list of topics and named entities which are of particular relevance to* p.

[1] Of course, one could always try to exploit structural clues, such as preference to follow certain party candidates, but without ground truth political sentiment expressed in the content, such methods would be difficult to evaluate.

Having found a representation for a party, we can define a user-party interaction as follows.

Definition 8.2 *Given a user* u *and a party* p*, a posting of* u *is called an interaction with* p*, or a* u-p *interaction, if it contains at least one term from the interaction profile of* p.

Definition 8.2 is general with respect to the content of interaction profiles. For instance, if interaction profiles contain names of official party accounts, then retweets, replies, and mentions of party accounts are naturally considered as interactions, because tweets that retweet, reply to or mention party accounts must contain corresponding account names. Same goes for candidate account names. Similarly, including party hashtags provides good coverage of relevant tweets.

The fact that interaction profiles are weighted and ranked allows for establishing properties of interactions, such as weight and various forms of ranks. We define such properties as follows.

Definition 8.3 *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.*

Definition 8.4 *The average, minimum or maximum rank of a* u-p *interaction is calculated respectively as the average, minimum or maximum rank of all terms from the interaction profile of* p *found in the interaction.*

Finally, note that through a single posting, a user may interact with multiple parties. This usually happens when users compare party members, party policies or include multiple party hashtags in a tweet.

8.2.1 Building Interaction Profiles

Let us now describe the method for building interaction profiles of parties. It is reasonable to assume, especially considering the buzz around the importance of social media in politics, that during the election campaign 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 behavior to

capture party-specific topics in the form of weighted unigrams and bigrams. To achieve this we employ the term-weighting scheme proposed in [36].

We associate each candidate $c \in C$ with a document d_c. Such a document consists of all postings of c, and is represented as a Bag of Words model. Correspondingly, we associate each party $p \in P$ with a collection of documents of its members $D_p = \{d_c, \forall c \in p\}$. The entire corpus, therefore, is denoted as $D = \{d_c, \forall c \in C\}$ and its vocabulary is denoted as V.

We proceed to build language models (LMs) for the collection of documents of each party and for the entire corpus. The idea is to calculate Kullback-Leibler (KL) divergence between the LM probabilities of each party and the LM probabilities of the entire corpus. The divergence score for each term in the vocabulary can then be used as a measure of the importance of the term to a given party. This way for each party we are able to produce a ranked and weighted list of party-specific terms. To be consistent, we refer to the general corpus as the election corpus and to its LM as the election LM. Similarly, a collection of documents associated with a party is referred to as a party corpus and its LM as a 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 the collection of documents D, and $udf(t, D) = df(t, D)/|D|$ denotes the probability of t appearing in D. Here $df(t, D)$ is the document frequency of t in D.

For the LMs of parties, initial term weights are calculated as

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

where $idf(t, D) = \log \frac{|D|}{1+df(t,D)}$ is the inverse document frequency of t in D.

Marginal probabilities and weights of terms are then normalized as

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

To account for terms unobserved in party corpora term weights are smoothed as

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

Party	Profile size	Top terms
LIB	170	*alberta liberals* *vote strategically* *municipal heritage*
NDP	157	*orange wave* *renewable energy* *affordable power*
PC	173	*premier alison* *family care* *life leadership*
WRA	182	*energy dividend* *balanced budget* *reality check*

Table 8.1: Basic characteristics of interaction profiles of the parties

where the normalization factor λ is set to be 0.001 as in [36].

Finally, we calculate the probability of term $t \in V$ in the LM of party p as

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

We have experimented with unigram and bigram language models, and in each case, calculated the KL divergence between probability distributions of party LMs and the election LM.

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

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

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

According to this formula, if a term is overrepresented in a party corpus compared to the election corpus it will receive some positive weight. The higher the weight of a term the more it deviates from the "common election chatter" and as such becomes more important to a party. The opposite is also true, and a negative weight

generally means underrepresentation of a term in the corpus of a party compared to that of the election. Such an underrepresentation could be due to a simple lack of interest in a term, or a term could have been avoided intentionally because it hurts the public image of a party. For example, the bigram *climate change* received an importance score of $-3.3*10^{-6}$ in the interaction profile of Wildrose party, while in the profiles of liberals and NDP it scored positively. This score is rather significant given that the average and maximum negative scores in the profile were $-4.9*10^{-7}$ and $-8.3*10^{-10}$ respectively. In fact, this bigram was never mentioned by a single candidate of Wildrose party and received the weight only because of the smoothing. Such a behavior can be explained by the fact that the party leader casted doubts on the science of climate change[2] and was widely criticized for that.

As our final step in building the interaction profiles we need to choose top terms from quite extensive vocabularies of unigram and bigram party LMs. We have experimented with different weight-based ranking techniques, such as choosing terms with above average weights and different top-*k* approaches. We ended up choosing 100 bigrams with top weights and unigrams representing party hashtags, party and candidate account names, and full names of party leaders. Table 8.1 shows the number of terms in the interaction profiles of the parties as well as examples of some of the top weighted bigrams.

8.3 Methodology

Let us recall the formulation of the political preference prediction problem that states: "Given a user u and a set of parties P with whom u has interacted, predict which party $p \in P$, if any, u is most likely to vote for in the upcoming election".

To solve the problem we employ a one vs. all classification strategy where for each party we train a binary classifier. Given interaction statistics of a user, a classifier trained for a certain party provides the confidence with which this user may be considered a supporter of that party. If the provided confidence is lower than a certain minimum, i.e. a preference threshold, we conclude that such user

[2]http://www.cbc.ca/news/canada/calgary/story/ 2012/04/16/albertavotes2012-danielle-smith-climate-change.html

will not support the party. In all our experiments, we set the preference threshold to 0.5. If a user has interacted with multiple parties, confidence scores (only those above the threshold) provided by respective classifiers are compared and such a user is said to prefer the party which was assigned the highest confidence. Ties are broken randomly.

To train party classifiers we resort to a distant supervision approach, and extract the interactions of all candidates with the parties and build candidate-party feature vectors. We consider candidates to have predefined preferences to their parties, and use their political identifications as the ground truth. Thus, if candidate c interacted with parties $p1, p2$, and her member party, $p3$, three feature vectors will be created, with c-$p1$ and c-$p2$ being negative training examples for respective classifiers, and c-$p3$ being a positive example for the $p3$-classifier.

8.4 Data Set Description

If anything, there was no shortage of attention to the election on Twitter. Over the course of only ten days before the election we collected more than 180 thousand tweets from the campaign related trends. From this initial collection we extracted accounts of users who tweeted about the election. In addition to that we collected candidate accounts and verified their authenticity.

Although abundant and instantly accessible, Twitter data may contain spam. Apart from that, as we will show later, some accounts in the data set may belong to media outlets, organizations, businesses, etc. Predicting political preference of owners of such accounts is out of the scope of this thesis, as the content generated by such accounts does not reflect a personal opinion of a potential voter. Given these specifics of Twitter data, we performed a data cleaning step, during which we, to the extent possible, identified and removed non-personal and spam accounts. In this section we provide a detailed description of data collection and cleaning methodologies.

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

Table 8.2: Properties of the election data set

8.4.1 Data Collection

We semi-automatically collected Twitter accounts of candidates as follows: as in [36], we retrieved the first three Google search results for each candidate's name coupled with the *twitter* keyword, and manually verified the authenticity of the returned accounts. Additionally, we looked up the official party websites of candidates whose accounts we could not find or verify. 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 thesis. To that list we also added the official Twitter accounts of the parties themselves to have a total of 256 accounts. Table 8.2 shows the number of candidate accounts per each party.

To collect non-candidate accounts we monitored campaign related trends over the course of ten days prior to the election. Specifically, we used a list of 27 keywords including party hashtags, e.g. *#ablib, #abndp, #pcaa, #wrp*, party names, leader names, and general election hashtags, e.g. *#abvote, #ableg*, etc. As a result we have obtained 28087 accounts of which 27822 belonged to non-candidates. We removed accounts with reported communication language other than English, which left us with 24507 accounts. For each of these accounts and 256 candidate accounts we retrieved up to 3000 tweets[3]. The retrieved content was limited to the tweets posted since March 27, the official first day of the campaign[4].

[3] 3000 most recent tweets is the limit established by Twitter API

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

Rank	Feature
1	Fraction of tweets with URLs
2	Age of the user account (a number of days since creation till the timestamp of the latest tweet in the data set)
3	Average number of URLs per tweet
4	Number of followers per followees
5	Fraction of tweets replied
6	Number of times the user had replied someone
7	Number of times the user was replied
8	Number of followees
9	Number of followers
10	Average number of hashtags per tweet

Table 8.3: Features used for spam accounts identification

8.4.2 Detecting Spam Accounts

Evidence suggests [40] that election campaigns may be targeted by *political spammers*, i.e. individuals or organizations who, under multiple false identities, spread rumors and fabricated content. Also, it has been shown [5], that in attempt to make their tweets visible to more people, spammers often "hijack" popular hashtags and use popular trends to disseminate spam. Given that Alberta election was an actively discussed topic (at least on a provincial level), we have reasons to believe that our data set may contain spam that contributes nothing but noise to studied interaction patterns. To address this issue we perform a spammer detection procedure.

There are a number of existing approaches for detecting spammers on Twitter. After considering a few of them, we decided to adopt a supervised approach for spammers detection proposed by Benevenuto et al. [5], doing so, in part, due to a well documented and freely available data set[5] of 355 spam and 710 non-spam accounts. According to the original work, when restricted only to features ranked in the top 10, the model performed as good as with the full set of features. Given that, we used only those top 10 features to train our model. These features are listed in Table 8.3 according to the ranking reported in the original work. The features were designed to distinguish behavior and quality of the content generated by the

[5] http://www.decom.ufop.br/fabricio/spammerscollection.html

two types of users, i.e. spammers and non-spammers. For instance, spammers tend to include URLs (spam links) in almost every tweet and label their tweets with popular and often unrelated hashtags. They rarely reply to tweets of other users and, on contrary, most of the non-spammer users never reply to spam. For detailed description of the features readers may refer to the original work.

We trained SVM and logistic regression (LR) models on the data set used in the original work. In order to account for the specifics of the local trends found 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. We performed a 10-fold cross validation and found that only one out of 256 candidate accounts was misclassified as spam account. In general, in agreement with the original work, spammers were detected with lower F-measure than non-spammers. The LR classifier performed slightly better than SVM and F-measure for respective classes was 85% and 94%.

However, results of spammers detection across the entire data set were rather unexpected. Out of 24507 accounts only 74, or 0.3%, were labeled as spam. We looked up these accounts on Twitter and did not find any evidence of spammer 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 was subject to a certain degree of overfitting, especially 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. 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, Canadian Health Coalition, etc.) Some of the accounts had high ratio of URLs per number of tweets, while others had low numbers of generated and received replies. One common characteristic of these accounts is that almost non of them express *personal opinion* of a potential voter. Moreover owners of these accounts (at least media and official entities) often do not or should not have political preference. Of course, investigating media bias and influence or support and disapproval of unions is an interesting research direction, but in this thesis we focus on predicting the political preference of individuals. We will refer to individual accounts as personal, or P-accounts. Correspondingly, non-personal accounts will be referred to as NP-accounts.

8.4.3 Removing Non-Personal Accounts

We approached the problem by extending the supervised learning 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[6] and searched the data set for accounts with names matching those in the list. We did the same for the list of Alberta cities[7], except in this case we looked for occurrence of a name rather than exact match, e.g. *Edmonton* Eskimos. We annotated the extracted set of accounts and obtained a final data set that consisted of 161 P- and 132 NP-accounts.

For the initial experiment we tested features that we used in the spammers detection task. We performed 10-fold cross validation using LR classifier. NP-accounts were detected with 92% precision and 83% recall. For P-accounts precision and recall were 75% and 56% respectively. As we expected, for this task some features lost their relevance. For instance, the age of account became irrelevant as unlike spam accounts neither P- nor NP-accounts get frequently suspended and re-created to have a short age. Also, rather unexpectedly features capturing reply statistics ranked amongst the lowest. As a result of the experiment, out of ten features we

[6]http://en.wikipedia.org/wiki/List_of_newspapers_in_Alberta
[7]http://en.wikipedia.org/wiki/List_of_cities_in_Alberta

Rank	Feature
1	Unigrams used in the names of NP-accounts
2	Unigrams used in the names of P-accounts
3	Unigrams used in the descriptions of P-accounts
4	Unigrams used in the descriptions of NP-accounts
5	Fraction of tweets with URLs
6	Number of followers
7	Average number of URLs per tweet
8	The presence or absence of URL
9	The length (in characters) of the location information
10	Number of followers per followees
11	Average number of hashtags per tweet

Table 8.4: Features used for detection of non-personal accounts

kept five and removed those that contributed little or no information gain.

Based on our analysis of data available from account profiles we introduced ten more features. Some of them we filtered out during the feature selection process. Our final model consisted of the 11 features listed in Table 8.4 and ranked according to the information gain-based ranking. The intuition behind the first four features is 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. The features 8-9 were introduced because, in their profiles, NP-accounts tend to provide URL and exact location, as opposed to short forms such as *AB, Alberta, Canada* or *US* frequently used by P-accounts. The remaining features were adopted from [5].

We classified the entire data set, and as a result 535 out of 24507 accounts were 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 including popular online blogs and editions. As our final cleaning step we removed NP-accounts from the data set leaving a total of 24060 accounts.

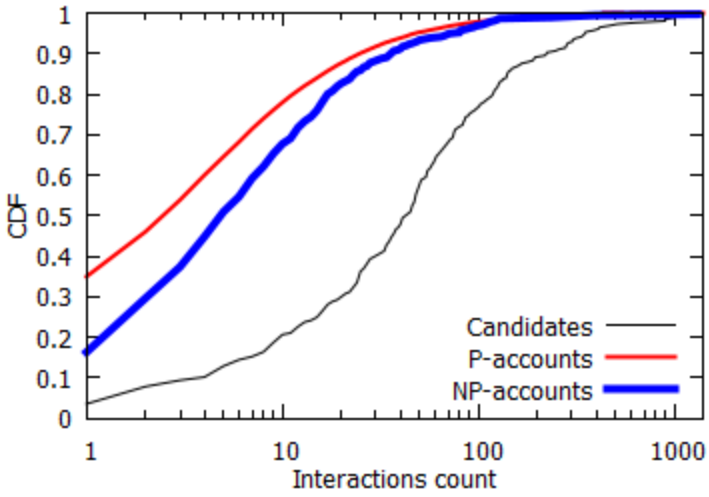


Figure 8.1: CDFs of the net amount of interactions generated by candidates, P- and NP-accounts

8.4.4 Results of Data Cleaning

Figure 8.1 shows the cumulative distribution functions of the net amount of interactions produced over the course of the campaign by candidates, P-, and NP-accounts. The plot clearly illustrates the silent majority effect [42], showing that almost half of P-accounts have produced at most only two election related tweets. Moreover, it shows that concentration of silent majority is not even across different target groups, with NP-account owners having less silent users relative to individuals.

Needles to say that candidates engaged in more interactions, with about 20% of candidate accounts producing more than 100 election-related tweets. For other two groups of accounts this ratio hardly reaches 5%. However, from Table 8.2 we can see that owners of NP-accounts have 1.4 times as much interactions per account than individuals. Considering that the bulk of them is either news or statements without personal political preference, we conclude that our efforts in cleaning the data had payed off, and we were able to remove certain amount of noise.

8.5 Building Prediction Models

We design features of the models based on the behavior of and the content generated by users, distinguishing between interaction-based and Twitter-specific features.

Domain	Value in $u - p_1$	Value in $u - p_2$
T	6	14
O	20	20
R	0.3	0.7
Δ	-14	-6

Figure 8.2: An example calculation of values of the *interactions count* feature over different domains

All features, including Twitter-specific, are extracted in the context of interactions, meaning that we build one feature vector per each interacting user-party pair. If a user did not interact with any party we make no predictions, and such a user is considered to prefer none of the parties.

Due to the fact that a user can interact with multiple parties, multiple feature vectors may contain statistics calculated for the same user but with respect to different parties. In order to provide prediction models with *overall* statistics, calculated for all parties with whom a user has interacted, and *relative* statistics, calculated for a *target* party in relation to all parties, we define features over different *domains*.

Let us consider the example shown on Figure 8.2. Suppose user u interacted 6 times with party p_1 and 14 times with party p_2. First of all we create two feature vectors: u-p_1 and u-p_2. In the target, or *T-domain*, the *interactions count* feature will have its respective per-party value in each feature vector, 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 same value of 20 in both feature vectors. Lastly, 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. For some features, such as interactions weight and rank, we also use a variation of relative domain, the Δ-*domain*, in which a feature is calculated as the absolute difference of its values in *T*- and *O-domains*. Overall, counting all features, defined over all domains, our prediction models use 51 features.

Feature	Domain	Description
Political diversity	O	Number of parties a user has interacted with
Interactions count	T, O, R	Domain total
Interactions weight	T, O, R, Δ	Domain average
Average rank of interactions	T, O, R, Δ	Domain average
Minimum rank of interactions	T, O, R, Δ	Domain average
Maximum rank of interactions	T, O, R, Δ	Domain average
Words per interaction	T, O, R	Domain average
Characters per interaction	T, O, R	Domain average
Interaction frequency	T, O, R	Interactions per day, domain avg.
Positive terms per interaction	T, O, R	Domain average
Negative terms per interaction	T, O, R	Domain average
Party hashtag position	T	Offset in words, domain avg.

Table 8.5: List of interaction-based features

8.5.1 Features Based on Interactions

Table 8.5 contains a list of interaction-based features with corresponding domains. We designed these features to account for the quantity and quality of user-party interactions. The intuition behind quantitative features, such as *interactions count* and *frequency*, is that users would probably tweet more about parties they prefer, and less, or not at all, about other parties. Also, while politically active users might interact with several parties in fairly equal proportions, the frequency of interactions with a preferred party may be higher. To further discriminate interactions with a preferred party from the rest of interactions, we introduce qualitative features, such as *interactions weight* and *rank*. The intuition here is that users would probably tweet about issues that are more important to their preferred parties, making corresponding interactions to receive higher weight and rank than interactions with other parties. To capture possible content difference in the target and overall domains we introduce features like the *number of words and characters per interaction* and the *position of the party hashtag*. The intuition behind these features is that users may post longer tweets about parties they prefer. Also, when labeling tweets with multiple party hashtags, users may do it in order of preference. Finally, to account for the explicit sentiment expressed in tweets, we calculate the average number of positive

Feature	Domain	Description
Retweets count	T, O, R	Number of times a user has retweeted party candidates, domain average
Retweet time	T, O, R	Domain average
Replies count	T, O, R	Number of replies a user has received from party candidates, domain average
Reply time	T, O, R	Domain average
Followees count	T, O, R	Number of party candidates followed by a user

Table 8.6: List of Twitter-specific features

and negative terms per interaction. To this end, we use an extensive (6800 words) opinion lexicon compiled by Hu et al. [24] and the Wikipedia's list of emoticons[8], i.e. smiley faces.

8.5.2 Twitter Specific Features

Table 8.6 contains a list of Twitter specific features. We designed these features to account for the Twitter specific behavior of users and structural clues hidden in the follower-followee and *retweet* networks. As the previous research suggests [14,15], retweet networks of politically active users display distinctive structural patterns, and may be divided into communities according to political alignments of users. In other words, retweetting party candidates may be considered as a strong support indicator, as users very seldom retweet those whose political views do not coincide with their own. *Replies* and *followees count* are considered for the same reason.

As in the case with interaction-based features, we also considered frequency-based distinctions. In particular, we added *retweet* and *reply time* as the estimators of relative frequency of these actions. Thus, if users do retweet candidates from different parties, these features may capture a tendency, where candidates of a preferred party are retweeted more frequently. Here retweet time corresponds to the 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.

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

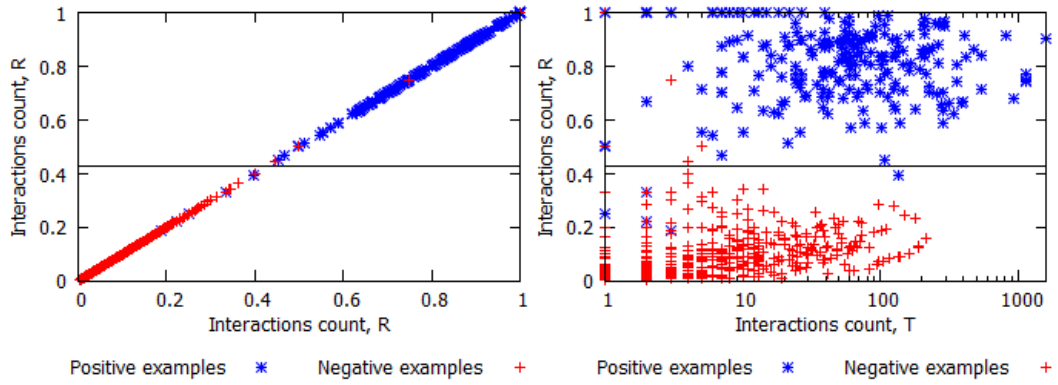
Feature	Domain	Type	Avg. rank
Interactions count	R	IB	1.3 ± 0.46
Followees count	R	TS	1.7 ± 0.46
Positive terms per interaction	R	IB	3 ± 0
Retweets count	R	TS	4.1 ± 0.3
Interactions frequency	R	IB	4.9 ± 0.3
Negative terms per interaction	R	IB	6.2 ± 0.4
Interactions weight	R	IB	7 ± 0.77
Followees count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

Table 8.7: Top 10 features for predicting political preference

8.5.3 Feature Selection

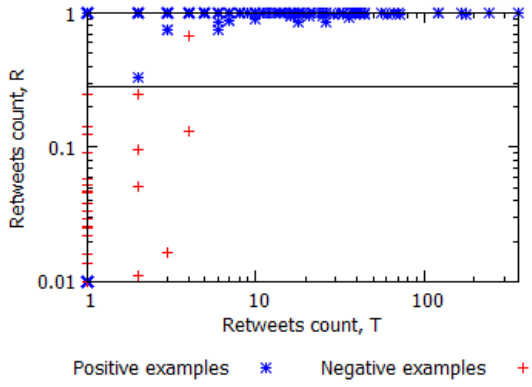
Using Weka [22] we performed an information gain-based feature ranking on the training data, and identified the top 10 features listed in Table 8.7. As we performed a cross validated ranking procedure, the rank received by a feature is averaged over the number of folds, which is ten in our case. Comparing feature values in T- and O-domains turned out to be effective, as all features ranked in top 7 are based on relative statistics, and *interaction weight* in Δ-domain has ranked 9th. Moreover, almost all features based on the overall statistics contributed zero information gain, suggesting that party-specific features are more effective. 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.

Figure 8.3 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 8.3a, *relative interactions count* is a very strong feature, which divides positive and negative examples very accurately. The distribution of training examples over per-party and relative counts of interactions, depicted on Figure 8.3b, provides more support of this claim. As it can be seen, although an average user interacts more with a preferred party than with other parties, there is a considerable overlap of positive and negative examples across per-party interactions count. On the contrary, using relative interactions count, one could easily classify the training

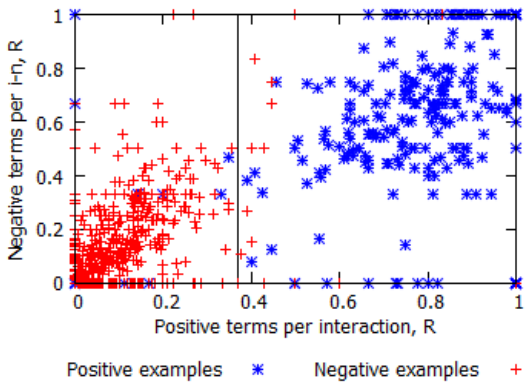


(a) Interactions count, R

(b) Interactions count, T / Interactions count, R



(c) Retweets count, T / Retweets count, R



(d) Positive term p.i., R /Negative terms p.i., R

Figure 8.3: Distribution of training examples across different feature spaces

examples, labeling as positive every example for which the value of the feature is equal to or greater than 0.43. In line with the previous research [14, 15], from Figure 8.3c we observe that candidates rarely retweet candidates from other parties. In fact, 94% of negative training examples have a zero per-party retweets count. Finally, Figure 8.3d depicts a rather interesting distribution of the training data across the relative usage of positive and negative terms. It seems that in the training data the frequent usage of opinion words, regardless of the expressed sentiment, distinguishes positive examples from negative ones. However, the data is more accurately dividable across the horizontal axis, suggesting that in general interactions with a preferred party are more positively worded. When restricted only to the top 10 features, the prediction models showed better performance on the training data. Hence, for our experiments we will use models that use only the top 10 features.

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 8.8: Characteristics of the test data for predicting political preference

8.6 Experiments and Evaluation

A major evaluation challenge we faced was to obtain 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. everything between April 23, 9:00 am, MDT (when 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 party marker can be a party hashtag, a name of a party leader, a mention of a party account or any known account of party candidates. 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 supporting or a void statement. 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. The following are the examples of two types of statements:

Supporting: *I cast my vote for PC!! Get out and make your choice #yegdt #yeg*

Void: *Why would anybody actually vote WRA it baffles me*

Annotators agreed that 99 out of 681 users had expressed prevailingly supporting statements. Out of those 99, for 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 [17] for three annotators was 0.68

Let us 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.

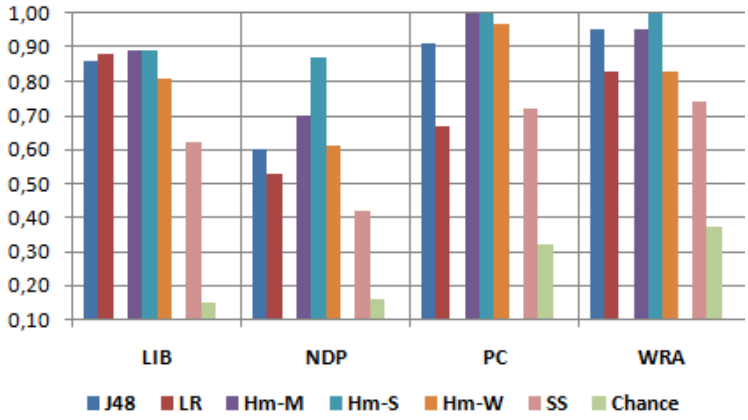
This was exactly the case for 14 out of 99 users in our test set. Clearly, without interactions we were unable to make predictions for these users, and had to exclude them from the test set. We focus on the remaining 75 users to whom we refer to as test users. Table 8.8 shows basic interaction statistics of test users on a per party basis.

8.6.1 Human Evaluation

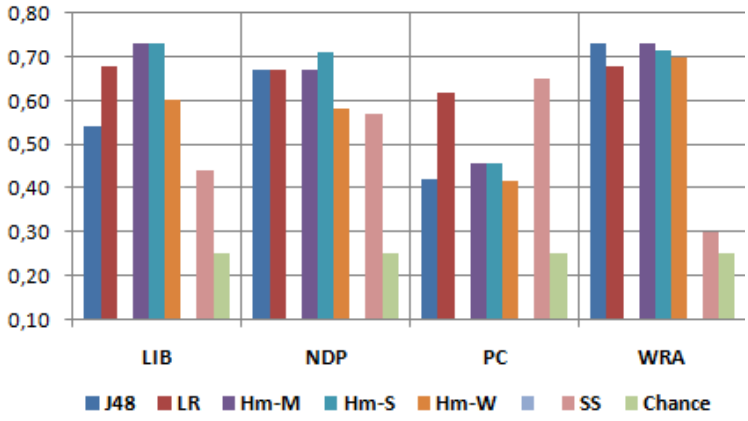
In order to assess the difficulty of the task on human scale we set up an experiment in which we provided the same three annotators with randomly selected interactions of test users. For each test user and each party the user has interacted with we chose up to 50 random interactions out of those that 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. For instance, if annotator A agreed with the two others in four observations out of 10, A's confidence is calculated as $4/10 = 0.4$.

8.6.2 Baselines

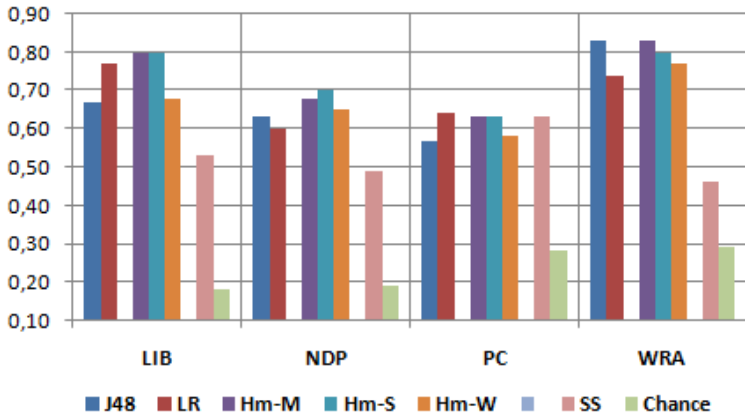
Apart from human annotators, we compare our method with two more baselines. As one such baseline, we use SentiSntrength [57], a lexicon based sentiment analysis tool optimized for informal writing style, common to social media services. Under default settings for a given text, SentiSntrength provides 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 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



(a) Precision



(b) Recall



(c) F-measure

Figure 8.4: Results of predicting political preference

sum as the classification confidence. Resulting sum may be negative, in which case we conclude that SentiSntrength predicted no preference. Note that this is equivalent to setting the confidence threshold to zero. Finally, we compare our method to chance, by expecting a given user to support all parties with equal probability.

Thus, for each user we randomly generate 1000 predictions and choose the party that was predicted most of the times. Ties are broken randomly.

8.7 Results

We have experimented with a number of classifiers provided by Weka [22], and based on the performance chose a decision tree based J48 and Logistic regression classifiers. 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 8.4 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 bars, from left to right, corresponds to that of legend items. Here labels Hm-M, Hm-S, and Hm-W correspond to human majority vote, the strongest, and weakest annotators, respectively, and SS stands for SentiStrength.

As it can be seen from Figure 8.4a both classifiers make less precise predictions than the annotators, although J48 shows better precision than LR especially for PC and WRA. Moreover, this classifier outperforms the least accurate annotator for LIB and WRA. In terms of recall, classifiers again perform close to human annotators. It is interesting that for PC, both LR and SentiStrength outperform human annotators, with SentiStrength outperforming LR on 3%. This behavior can be explained by the fact that for annotators interactions of some users with PC did not seem to have meaning, but the machine learning algorithm could have exploited different features, including followers share, interaction frequency, etc. Of course, the annotators had no means to assess those features. Finally, in terms of F-measure, classifiers outperform both baselines, as shown in Figure 8.4c. LR shows great performance outperforming all of the annotators for PC and the weakest annotator for LIB. J48 outperforms both the weakest and strongest annotators for WRA, achieving F-measure equal to that of human majority vote.

Chapter 9

Temporal Analysis of the Predicted Political Preference

9.1 Motivation

In this chapter we show how the predicted political preference of users changes with the progression of a campaign. By doing so we answer the following two questions. First, how does preference change for different types of users? Second, do changes in predicted preference correlate with the campaign events, especially those discussed extensively in social media? With regards to the first question we want to check if the preference of vocal users, who often tweet about elections, is less likely to change compared to that of silent users, who rarely discuss elections on Twitter. Similarly, it would be interesting to know if "bursts" in Twitter activity, that reportedly [61] happen during televised debates or other campaign related events, have any effect on predictions of political preference.

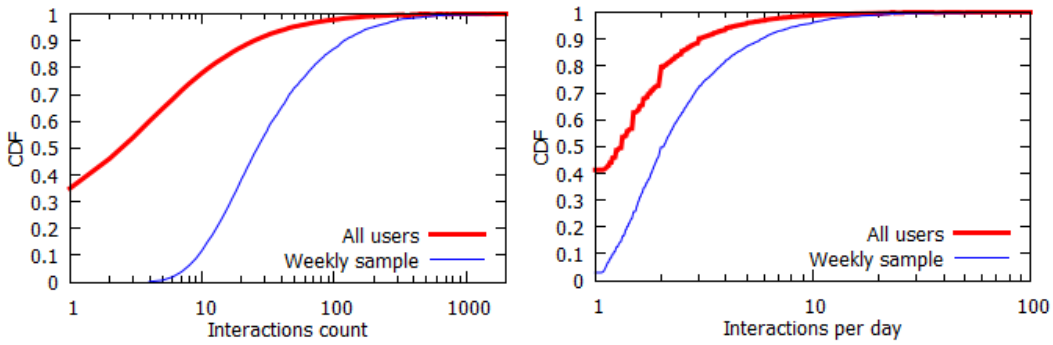
9.2 Experiments

To track changes in the predicted preference we set up the following experiment. We choose a window of fixed number of days, let us say N. By sliding this window (one day at a time) over the timespan of the campaign (28 days), we obtain a sequence of $28 - N + 1$ time periods, each N days long. For a window of size $N = 1$ these periods do not overlap. Otherwise $\forall N \in [2..27]$ there's an overlap of $N - 1$ days between any two consecutive periods. 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 period p_2 end on day d and start on day $d - N + 1$. Period p_1 correspondingly should end on day $d-1$ and start on day $d-N$. Let the predicted political preferences 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 $p_2 \in [d-N+1, d]$ has changed compared to the predicted preference for the previous period $p_1 \in [d - N, d - 1]$. In order to capture user's change of preference for the duration of the campaign we need to repeat this procedure for all $28 - N + 1$ 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.

In order to obtain a decent sized sample of vocal and silent users, we need to carefully chose the size of the sliding window. Clearly, choosing small window sizes results in more fine grained analysis; e.g. for a window of size one we can study daily changes of predicted political preference. However selecting a small window also limits quantity and variety of users available for the experiment. In fact, in the entire data set there were only 60 users who had interactions on each day of the campaign. Needles to say that they were all active users who contributed 426.8 interactions per account and 15.2 interactions per day on average. 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. We will refer to this sample of users as the weekly sample.

Next, we need to chose a Twitter-involvement measure, to define vocal and silent users. The previous research [42] used raw counts of relevant tweets (interactions in our case), and considered users who posted only one tweet to be silent, and those who contributed at least 50 to be vocal users. Figure 9.1 shows CDFs of raw and per-day interaction counts. As it can be seen from Figure 9.1a not a single user from the weekly sample has less than four interactions. In general, due to requirements of the experiment, sampled users have more interactions than all users. Thus, raw count-based definitions do not work with our experimental setting. To



(a) (b)

Figure 9.1: CDFs of raw and per-day interaction counts calculated for all users and the weekly sample

this end, we propose frequency-based Twitter-involvement measure. As it can be seen from Figure 9.1b, about 30% of sampled, and 60% of all users engaged in at least 1.5 interactions per day. Similarly, about 10% of the weekly sample and less than 5% of all users have more than 10 interactions per day. Thus, we define silent and vocal users as those who contributed at most 1.5 and at least 10 interactions per day, respectively. We do not further classify the rest of the users, and refer to them collectively as moderate users. As a result, we divided the weekly sample into 129 vocal, 791 silent, and 2493 moderate users. For the experiment we randomly selected 100 users from each of these user groups.

9.3 Results

9.3.1 Political Preference

Figure 9.2 shows percentage of users for whom on a given date, the predicted preference for the current period has changed compared to that for the previous period. In this respect, our initial experiments with silent users revealed that due to low counts of interactions in some periods our method predicted “no preference” for a significant number of users. Controversially, for vocal and moderate users “no preference” predictions were much more rare. Thus, to ensure equal experimental conditions, and to account for the interaction sparsity introduced by splitting the data into periods, we run the experiment for silent users under a constant prefer-

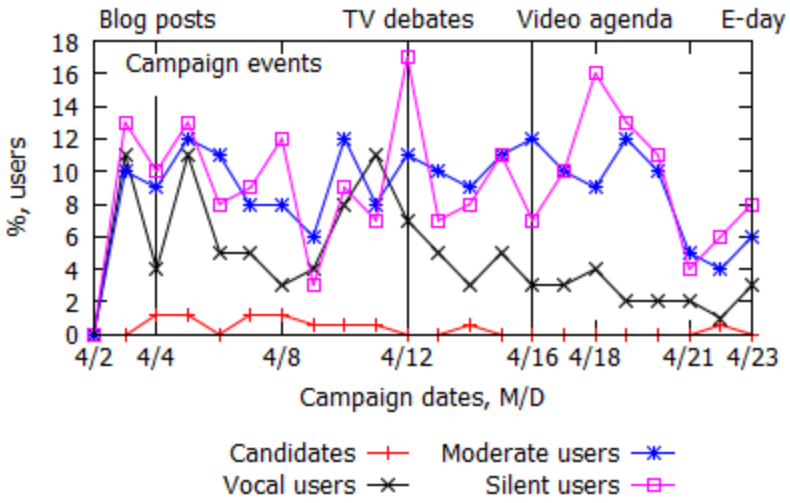


Figure 9.2: Changes in political preference

ence 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. In other words, a no preference prediction for the current period is replaced with the prediction made for the previous period.

To check the correctness of the experimental setup we show the dynamics in change of preferences of party candidates. Indeed, as one would expect, the preference of candidates did not change during the campaign, apart from negligible deviations of about 1%. Closest to this, was the preference change rate of vocal users, which never exceeded 11%, and after April 13 (10 days before the E-day) never exceeds 5%. On contrary, the preference of silent users is changing constantly and for certain points in the campaign rather drastically, hitting the peak (17%) in the middle of the campaign and having another major change (16%) five days before the election. The preference of moderate users displays a flatter, and for a period of about ten days (4/10 – 4/20), almost periodically alternating change rate, with a deep dive and a small rise towards the end of the campaign. This diving-rising pattern, which happens 2-3 days before the election, is observed for all groups of users (except candidates). Similarly, all groups of users exhibit spiky pattern in the preference change rate over the first 3-4 periods. This trend deserves further investigation, as it may suggest that change in preference does not happen spontaneously, but co-occur with certain events.

9.3.2 Campaign Related Events

Now let us check if there is any significance in dates on which major changes in the predicted political preference occur. From the social media highlights of the campaign[1] we found descriptions of the following events, which were extensively discussed in blogsphere, on Facebook and Twitter: (i) *Blogposts* by Kathleen Smith[2] (April 2) and Dave Cournoyer[3] (April 4) criticizing WRA. (ii) Television broadcast of party leaders' *debates* (April 12). (iii) *YouTube video* titled *"I never thought I'd vote PC"*[4] (April 16), asking people to vote strategically against Wildrose party.

The vertical lines on Figure 9.2 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. Given that this was a scheduled and anticipated event, it is reasonable to expect that a lot of live Twitting about the event occurred on that day. Thus a lot of interactions may have been produced influencing the predictions for that period. The trend of the preference change for moderate users also went up on that day, although not to its highest point. It is interesting that for vocal users the peak change in preference occurred one day before the debates. This could be the case that discussion of the anticipated event was more interesting than the event itself.

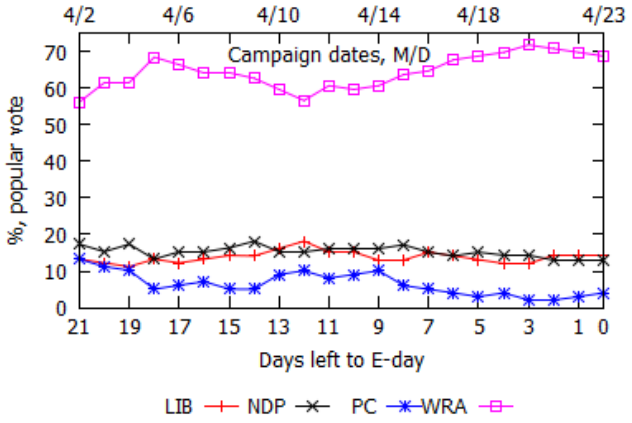
In the case with blogposts and video the rise in the preference change rate occurs only on the next day after events took place. Twitter discussion of these events might have had the "long term" effect gaining more popularity on the next day and influencing the predictions for the next period. This is different in case of moderate users, for whom one of the peak changes occur on the exact same day with video release. To sum up, changes in the predicted preference may precede, follow, or co-occur on the exact same day with an event. There could be one to two days deviations, but overall, trends seem to be influenced by campaign-related events, and changes in preference seem not to happen spontaneously.

[1] 2012 Alberta Election: Social Media Highlights: http://blog.mastermaq.ca/2012/04/28/alberta-election-social-media-highlights/

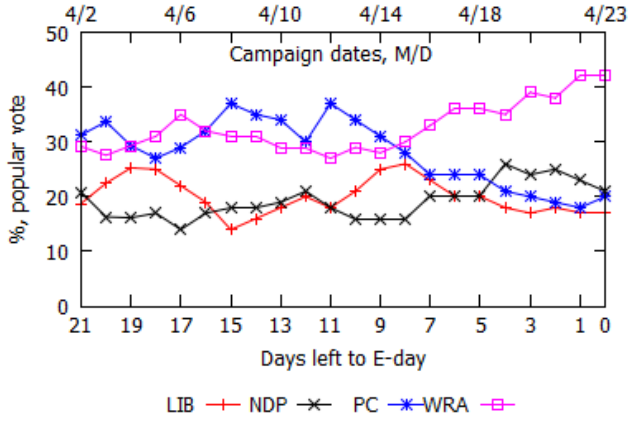
[2] http://www.kikkiplanet.com/ pruned-bush-confessions-of-a-wilted-wild-rose/

[3] http://daveberta.ca/2012/04/ danielle-smith-wildrose-candidates/

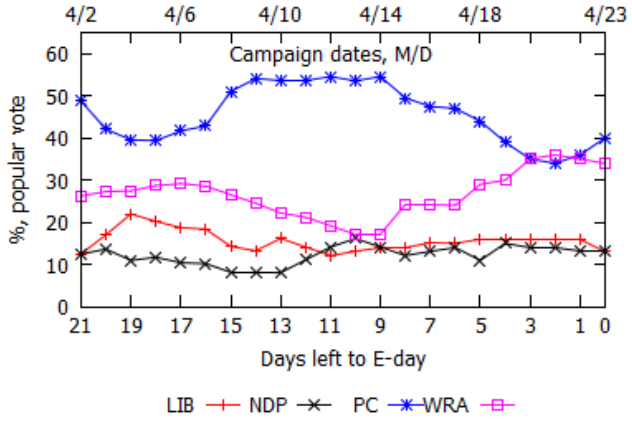
[4] http://www.youtube.com/watch?v=rPR84Gn1d9I



(a) Vocal users



(b) Moderate users



(c) Silent users

Figure 9.3: Changes in popular vote

9.3.3 Popular Vote

Figure 9.3 shows the distributions of the popular vote of each group of users for each period on a given date. It can be seen that in all three case, the preference

generally changes between WRA and PC parties with, increase of the popular vote for one party resulting in a decrease for the other. It is interesting how for different groups of users shifts between these two parties increase towards PC with the decrease of Twitter-involvement of a group. Indeed, through the whole duration of the campaign, vocal users seem to favor WRA, and PC is placed at the opposite extreme. For moderate users the two parties seem to split the votes in alternating fashion (recall alternating pattern of the preference change), with WRA starting and continuing to win approximately a week before the election. Lastly, silent users seem to favor PC all the way up to three days before the election, where there is a tiny shift towards WRA. This is in line with a CBC poll conducted three days before the election, which also placed the two neck-to-neck. However, we see a changing trend starting to show one day before the election, with popular vote for PC rising and the popular vote for liberals declining. This seems like a possible sign of strategic voting changing the course of the election. Lastly, the popular vote prediction for silent users shows more resemblance to the actual election outcome than that of the two remaining groups. Parties are placed in the actual winning order, and popular vote percentages are close to the actual ones (see Table 7.1).

In agreement with our findings, popular vote of vocal users changes gradually, and sharp transitions mostly occur before April 12. For silent users, however, shifts in popular vote occur rather sharply for the whole duration of the campaign. From this we conclude that vocal users are less likely to change their political preference compared to silent users. Moreover, as the election day draws closer, less and less vocal users are likely to change their preference.

Part IV

Conclusions

Chapter 10

Conclusions and Future Work

10.1 Summary of the Findings

In this thesis we addressed the problems of (i) inferring the polarity of collaborations between Wikipedia editors, and (ii) predicting the political preference of Twitter users. We took an interaction-driven approach to solve both problems. In both cases we defined interactions and designed statistical features to characterize pairs of interacting entities. Let us begin with the summary of findings regarding the problem of inferring the polarity of collaborations.

First, with regards to the predictive power of collaboration profiles, rather surprisingly, we found that individual features, calculated with respect to a target editor, were the most discriminative. In other words, while one would expect pairwise, interaction-based features to have more weight in the inference of the collaboration polarity between two editors, individual characteristics of one of them were assigned the highest weights. However, a closer examination of the top ranked characteristics revealed, that many of them reflected a general tendency of a target editor to collaborate productively or counter-productively with other contributors, including, but not limited to her "profile-mate". For instance, the feature *agreement in comments* that ranked second most discriminative, reflects an editor's overall tendency to agree with contributions of others. Similarly, the first ranked feature, *average page conflict ratio*, estimates the extent to which a contributor has a priori tendency to engage in disputes, by editing controversial articles. Hence, we conclude that votes casted in admin elections are, to certain extent, influenced by the

history of pre-vote collaborations and disputes. Moreover, as Sepehri et al. [48] showed, collaboration profiles can be used to accurately identify controversial articles. This suggests that votes casted in admin elections are, indeed, good indicators of collaborations and disputes between editors.

Second, as we evaluated our method on the problem of predicting votes in admin elections, we made several observations about voting behavior of editors. In particular, we investigated how the age of interactions affects votes, and concluded that instances of collaboration and dispute that happened long before voting did not have considerable impact on votes. Also, in line with a previous research [34], we observed that transparency of admin elections influenced votes. In other words, due to the fact that everyone is aware of the votes casted by everyone else, one needs very strong argumentation to oppose a heavily supported candidate. Thus, the probability of subsequent votes being positive grows with the increase in support received by a candidate.

With regards to the problem of predicting political preference of Twitter users, we preformed two tasks: the general preference prediction and the analysis of changes in preference. In the general preference prediction task we compared our method with human annotators and a sentiment analysis-based approach. We found that content-driven methods had relatively low recall. In particular, human annotators, sometimes performed worse than our method. This suggest that content alone may not be enough to predict political preference of Twitter users. The ranking of features supports this observation, showing that out of the top 10 features only two (*positive* and *negative terms per interaction*) were content-based. Also, in line with the previous research, we found that the features based on preferential following [54] and retweeting [14, 15] had a great predictive power, as they ranked 2nd (*relative followees count*) and 4th (*relative retweet count*), respectively. Overall, we found our interaction-driven approach to be effective, and the intuition behind some of the interaction-based features to be correct. For instance, the features, such as *relative interaction frequency* and *relative interaction weight* were introduced to distinguish between preferences of users who had fairly equal amounts of interactions with several parties. The two features were ranked as 5th and 7th respectively,

suggesting that they served their purpose, perhaps not as the most important ones, but rather as the features used for additional and occasional clarifications.

Temporal analysis of the predicted preference revealed that politically active users were less likely to change their preference during the election campaign. It was also shown that significant changes in the predicted preference co-occurred with, and sometimes, preceded or followed campaign related events, such as leader debates, video agenda, etc. Lastly, the trend of the popular vote shift displayed an interesting behavior, with the final vote split for the less Twitter-involved group bearing the closest resemblance to the actual results of the election. Thus, while for vocal users the trend of popular vote shift did not look like the actual trend at all, for silent users the trend, over the last few days before the election, very closely resembled polling results, and the final vote split was very close to the actual one.

10.2 Future Work

Certainly, there is room for improvement in our approach. Definition-wise, interactions of Wikipedia editors can be made less threshold-dependent, by incorporating content-based relatedness metrics, such as word level ownership [37] and edit longlevity [1]. Similarly, as many existent topic identification techniques [14, 38, 61], static interaction profiles do not account for dynamic nature of Twitter, where everyday new issues are discussed, and subsequently, new hashtags are created. Hence, our approach needs mechanisms for on the fly profile updates, especially, if used in a real-time political preference prediction system, which is one of potential applications of our method.

Another direction for future research could be identifying strategic voting behavior. For the particular election campaign considered in this thesis strategic voting was a widely discussed issue. On Twitter it was discussed under special trends, such as #nowrp, #strategicvotes, #strategicvoting, etc. We plan to study the content and behavior of users engaged in these discussions in order to introduce necessary corrections to our method or develop a new method for recognition of strategic voting behavior.

Bibliography

[1] B. Thomas Adler and Luca de Alfaro. A content-driven reputation system for the wikipedia. In *WWW '07*, pages 261–270, New York, NY, USA, 2007. ACM.

[2] Jisun An, Meeyoung Cha, Krishna Gummadi, Jon Crowcroft, and Daniele Quercia. Visualizing media bias through twitter, 2012.

[3] Jisun An, Meeyoung Cha, P. Krishna Gummadi, and Jon Crowcroft. Media landscape in twitter: A world of new conventions and political diversity. In *ICWSM*, 2011.

[4] A. L. Barabasi and R. Albert. Emergence of scaling in random networks. *Science*, 286:509–512, 1999.

[5] F. Benevenuto, G. Magno, T. Rodrigues, and V. Almeida. Detecting Spammers on Twitter. July 2010.

[6] Petko Bogdanov, Nicholas D. Larusso, and Ambuj Singh. Towards community discovery in signed collaborative interaction networks. In *ICDMW'10 workshop*, pages 288–295, Washington, DC, USA, 2010. IEEE Computer Society.

[7] Ulrik Brandes, Patrick Kenis, Jürgen Lerner, and Denise van Raaij. Network analysis of collaboration structure in wikipedia. In *WWW '09*, pages 731–740, New York, NY, USA, 2009. ACM.

[8] Susan L. Bryant, Andrea Forte, and Amy Bruckman. Becoming wikipedian: transformation of participation in a collaborative online encyclopedia. In *Proceedings of the 2005 international ACM SIGGROUP conference on Supporting group work*, GROUP '05, pages 1–10, New York, NY, USA, 2005. ACM.

[9] Moira Burke and Robert Kraut. Mopping up: modeling wikipedia promotion decisions. In *Proceedings of the 2008 ACM conference on Computer supported cooperative work*, CSCW '08, pages 27–36, New York, NY, USA, 2008. ACM.

[10] D. Cartwright and F. Harary. Structural balance: a generalization of Heider's theory. *Psychological Review*, 63(5):277–93, 1956.

[11] Meeyoung Cha, Hamed Haddadi, Fabrício Benevenuto, and P. Krishna Gummadi. Measuring user influence in twitter: The million follower fallacy. In *ICWSM*, 2010.

[12] Murphy Choy, Michelle L. F. Cheong, Ma Nang Laik, and Koo Ping Shung. A sentiment analysis of singapore presidential election 2011 using twitter data with census correction. *CoRR*, abs/1108.5520, 2011.

[13] Andrea Ciffolilli. Phantom authority, self-selective recruitment and retention of members in virtual communities: The case of wikipedia. *First Monday*, 8(12), 2003.

[14] Michael Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer. Predicting the political alignment of twitter users. In *SocialCom/PASSAT*, pages 192–199, 2011.

[15] Michael Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flammini. Political polarization on twitter. In *ICWSM*, 2011.

[16] Stefano DellaVigna and Ethan Kaplan. The fox news effect: Media bias and voting. *The Quarterly Journal of Economics*, 122(3):1187–1234, 2007.

[17] J.L. Fleiss et al. Measuring nominal scale agreement among many raters. *Psychological Bulletin*, 76(5):378–382, 1971.

[18] Daniel Gayo-Avello. "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, 2012.

[19] Jennifer Golbeck and Derek L. Hansen. Computing political preference among twitter followers. In *CHI*, pages 1105–1108, 2011.

[20] Roberto González-Ibáñez, Smaranda Muresan, and Nina Wacholder. Identifying sarcasm in twitter: a closer look. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: short papers - Volume 2*, HLT '11, pages 581–586, Stroudsburg, PA, USA, 2011. Association for Computational Linguistics.

[21] R. Guha, Ravi Kumar, Prabhakar Raghavan, and Andrew Tomkins. Propagation of trust and distrust. In *Proceedings of the 13th international conference on World Wide Web*, WWW '04, pages 403–412, New York, NY, USA, 2004. ACM.

[22] Mark Hall, Eibe Frank, Geoffrey Holmes, Bernhard Pfahringer, Peter Reutemann, and Ian H. Witten. The weka data mining software: An update, 2009.

[23] Fritz Heider. Attitudes and cognitive organization. *The Journal of Psychology*, 21:107–112, 1946.

[24] Minqing Hu and Bing Liu. Mining and summarizing customer reviews. In *KDD*, pages 168–177, 2004.

[25] Bernardo A. Huberman, Daniel M. Romero, and Fang Wu. Social networks that matter: Twitter under the microscope. *CoRR*, abs/0812.1045, 2008.

[26] Akshay Java, Xiaodan Song, Tim Finin, and Belle L. Tseng. Why we twitter: An analysis of a microblogging community. In *WebKDD/SNA-KDD*, pages 118–138, 2007.

[27] David Jurgens and Tsai-Ching Lu. Temporal motifs reveal the dynamics of editor interactions in wikipedia. In *ICWSM*, 2012.

[28] B.C. Keegan. Breaking news on wikipedia: dynamics, structures, and roles in high-tempo collaboration. In *Proceedings of the ACM 2012 conference on Computer Supported Cooperative Work Companion*, pages 315–318. ACM, 2012.

[29] Brian Keegan, Darren Gergle, and Noshir S. Contractor. Do editors or articles drive collaboration?: multilevel statistical network analysis of wikipedia coauthorship. In *CSCW*, pages 427–436, 2012.

[30] Aniket Kittur, Bongwon Suh, Bryan A. Pendleton, and Ed H. Chi. He says, she says: conflict and coordination in wikipedia. In *CHI '07*, pages 453–462, New York, NY, USA, 2007. ACM.

[31] Jon M. Kleinberg. Authoritative sources in a hyperlinked environment. *J. ACM*, 46(5):604–632, September 1999.

[32] Haewoon Kwak, Changhyun Lee, Hosung Park, and Sue Moon. What is twitter, a social network or a news media? In *Proceedings of the 19th international conference on World wide web*, WWW '10, pages 591–600, New York, NY, USA, 2010. ACM.

[33] David Laniado and Riccardo Tasso. Co-authorship 2.0: patterns of collaboration in wikipedia. In *Proceedings of the 22nd ACM conference on Hypertext and hypermedia*, HT '11, pages 201–210, New York, NY, USA, 2011. ACM.

[34] J. Leskovec, D. Huttenlocher, and J. Kleinberg. Governance in Social Media: A case study of the Wikipedia promotion process. In *Proceedings of the 4th Annual Conference on Weblogs and Social Media (ICWSM 2010)*, 2010.

[35] Jure Leskovec, Daniel Huttenlocher, and Jon Kleinberg. Signed networks in social media. In *Proceedings of the 28th international conference on Human factors in computing systems*, CHI '10, pages 1361–1370, New York, NY, USA, 2010. ACM.

[36] Avishay Livne, Matthew P. Simmons, Eytan Adar, and Lada A. Adamic. The party is over here: Structure and content in the 2010 election. In *ICWSM*, 2011.

[37] Silviu Maniu, Bogdan Cautis, and Talel Abdessalem. Building a signed network from interactions in wikipedia. In *DBSocial'11 workshop*, pages 19–24, New York, NY, USA, 2011. ACM.

[38] Micol Marchetti-Bowick and Nathanael Chambers. Learning for microblogs with distant supervision: Political forecasting with twitter. In *EACL*, pages 603–612, 2012.

[39] Panagiotis Takis Metaxas, Eni Mustafaraj, and Daniel Gayo-Avello. How (not) to predict elections. In *SocialCom/PASSAT*, pages 165–171, 2011.

[40] P.T. Metaxas and E. Mustafaraj. From obscurity to prominence in minutes: Political speech and real-time search. *WebSci10: Extending the Frontiers of Society On-Line. http://bit. ly/h3Mfld*, 2010.

[41] MondoTimes. The worldwide news media directory. available at http://www.mondotimes.com.

[42] Eni Mustafaraj, Samantha Finn, Carolyn Whitlock, and Panagiotis Takis Metaxas. Vocal minority versus silent majority: Discovering the opionions of the long tail. In *SocialCom/PASSAT*, pages 103–110, 2011.

[43] M. E. J. Newman and Juyong Park. Why social networks are different from other types of networks. *Phys. Rev. E*, 68:036122, September 2003.

[44] Brendan O'Connor, Ramnath Balasubramanyan, Bryan R. Routledge, and Noah A. Smith. From tweets to polls: Linking text sentiment to public opinion time series. In *ICWSM*, 2010.

[45] Panagiotis Panagiotopoulos and Steven Sams. An overview study of twitter in the uk local government. In *Transforming Government Workshop*. Brunel University, London, 2012.

[46] Bo Pang and Lillian Lee. Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2):1–135, January 2008.

[47] R. Puglisi. Being the new york times: The political behaviour of a newspaper. available at ssrn: http://ssrn.com/abstract=573801.

[48] Hoda Sepehri Rad, Aibek Makazhanov, Davood Rafiei, and Denilson Barbosa. Leveraging editor collaboration patterns in wikipedia. In *HT*, pages 13–22, 2012.

[49] Usha N. Raghavan, Réka Albert, and Soundar Kumara. Near linear time algorithm to detect community structures in large-scale networks. *Physical Review E*, 76(3):036106+, September 2007.

[50] Jacob Ratkiewicz, Michael Conover, Mark Meiss, Bruno Gonçalves, Alessandro Flammini, and Filippo Menczer. Detecting and tracking political abuse in social media. In *ICWSM*, 2011.

[51] David M. Reif, Alison A. Motsinger, Brett A. McKinney, James E. Crowe Jr., and Jason H. Moore. Feature selection using a random forests classifier for the integrated analysis of multiple data types. In *CIBCB*, pages 1–8, 2006.

[52] Mark Senak. Twongress: The power of twitter in congress. white paper, 2010, eyeonfda.com.

[53] Hoda Sepehri Rad and Denilson Barbosa. Towards identifying arguments in wikipedia pages. In *WWW'11*, pages 117–118, New York, NY, USA, 2011. ACM.

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

[55] Bongwon Suh, Ed H. Chi, Bryan A. Pendleton, and Aniket Kittur. Us vs. them: Understanding social dynamics in wikipedia with revert graph visualizations. In *Proceedings of the 2007 IEEE Symposium on Visual Analytics Science and Technology*, VAST '07, pages 163–170, Washington, DC, USA, 2007. IEEE Computer Society.

[56] Libby Veng-Sam Tang, Robert P. Biuk-Aghai, and Simon Fong. A method for measuring co-authorship relationships in mediawiki. In *Proceedings of the 4th International Symposium on Wikis*, WikiSym '08, pages 16:1–16:10, New York, NY, USA, 2008. ACM.

[57] Mike Thelwall, Kevan Buckley, Georgios Paltoglou, Di Cai, and Arvid Kappas. Sentiment strength detection in short informal text. *JASIST*, 61(12):2544–2558, 2010.

[58] Andranik Tumasjan, Timm Oliver Sprenger, Philipp G. Sandner, and Isabell M. Welpe. Predicting elections with twitter: What 140 characters reveal about political sentiment. In *ICWSM*, 2010.

[59] Fernanda B. Viégas, Martin Wattenberg, and Kushal Dave. Studying cooperation and conflict between authors with history flow visualizations. In *Proceedings of the SIGCHI conference on Human factors in computing systems*, CHI '04, pages 575–582, New York, NY, USA, 2004. ACM.

[60] Ba-Quy Vuong, Ee-Peng Lim, Aixin Sun, Minh-Tam Le, and Hady Wirawan Lauw. On ranking controversies in wikipedia: models and evaluation. In *WSDM'08*, pages 171–182, New York, NY, USA, 2008. ACM.

[61] Hao Wang, Dogan Can, Abe Kazemzadeh, François Bar, and Shrikanth Narayanan. A system for real-time twitter sentiment analysis of 2012 u.s. presidential election cycle. In *ACL (System Demonstrations)*, pages 115–120, 2012.

[62] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He. Twitterrank: finding topic-sensitive influential twitterers. In *WSDM*, pages 261–270, 2010.

[63] Honglei Zeng, Maher A. Alhossaini, Li Ding, Richard Fikes, and Deborah L. McGuinness. Computing trust from revision history. In *PST '06*, pages 1–1, New York, NY, USA, 2006. ACM.