Inferring political preferences of active content consumers in Twitter

by

Jalehsadat Mahdavimoghaddam

Bachelor of Software Engineering, Islamic Azad University of Tehran, 2012

A THESIS SUBMITTED IN PARTIAL FULFILLMENT OF THE REQUIREMENTS FOR THE DEGREE OF

Master of Computer Science

In the Graduate Academic Unit of Computer Science

Supervisor(s):	Patricia A. Evans, Ph.D., Faculty of Computer Science
Examining Board:	Michael W. Fleming, Ph.D., Faculty of Computer Science, Chair
	Joseph D Horton, Ph.D., Faculty of Computer Science
	Donglei Du, Ph.D., Faculty of Business Administration

This thesis is accepted by the Dean of Graduate Studies

THE UNIVERSITY OF NEW BRUNSWICK

May, 2016

Abstract

The growth of user engagement in online social networks has generated a tremendous amount of content regarding various topics. This rich content helps businesses to infer interesting information about public opinions and preferences of OSN users to serve their customers with customized services. Also, this inferred information can be used for different prediction purposes, such as predicting the possible outcome of an election.

Despite the huge increase in the amount of produced content in OSNs, many users tend to consume content on certain topics rather than provide content themselves. Therefore, it is a challenge to discover preferences of content consumers who are silent on a given topic. In this thesis, a novel approach is proposed that predicts personal preferences of content consumers through what they read rather than what they write. In other words, in this study it is shown that only relying on followees to predict preferences of content consumers leads to promising results.

Dedication

Dedicated to my dear parents and beloved sister.

Acknowledgements

I would like to express my sincere appreciation to my supervisor, Dr. Patricia A. Evans, for her advise, continuing guidance, and assistance throughout the course of this thesis. It was an honour for me to work under her supervision. Many thanks also go to my committee members, Dr. Michael W. Fleming, Dr. Joseph D Horton, and Dr. Donglei Du for their constructive comments.

Also, I would like to offer my thanks to my dear friend, Seyed Pooria Madani Kochak, for his valuable suggestions and unlimited support. He was always exceedingly generous with his help throughout my thesis.

Finally, a special thanks goes to my dear parents and beloved sister for their constant support and unconditional love. The contribution of my family have been enormous. I would not go this far without their encouragement and love. There is no word that can show my appreciation for their personal sacrifices throughout my life.

Table of Contents

List of Tables

List of Figures

List of Abbreviations

AC consumers	Active Content consumers, page 1
AC producers	Active Content producers, page 88
TU	Target User, page 27
Ext-Po	Extremely popular users of both level 1 and 2, page 80 and 81
Po-2	Ppopular users of both level 2, page 80 and 81
Po	Popular users of both level 1 and 2, page 80 and 81
Or-2	Ordinary users of level 2, page 80 and 81
Or	Ordinary users of both level and 2, page 80 and 81
Ext-Or-2	Extremely ordinary users of level 2, page 80 and 81

Chapter 1

Introduction

The growth of online social media such as Orkut, Facebook, and Twitter has made users able to communicate with their friends, share their opinions about various topics, and follow the latest news. In social networks, usually people tend to interact with those who have common interests, personalities, or attributes. In other words, based on the homophily principle, a tie between two nodes in a social graph is more likely to be created when those nodes are similar to each other [15]. So, when two nodes in an OSN (Online Social Network) are connected to each other, the assumption is that they have similar opinions.

In OSNs, not only can users express their opinions about services or products, but they can also discuss any subject ranging from food to politics or daily events [21]. Currently, inferring users' interests plays an important role in

many applications, such as information retrieval, search, and recommender systems [33]. Discovering interests of users helps recommender systems of OSNs to assist users with making appropriate selections by suggesting new friends, products, or content [2], so users would not be overwhelmed by the massive amount of information available in OSNs and the huge number of users they can interact with. In addition to users of the OSNs, various businesses can take advantage of recommending techniques. For example, they can examine the comments that customers wrote about a particular product to discover degree of dissatisfaction about their product to improve it. Also, based on each user's needs and preferences businesses can recommend new products that customers were not already aware of [26]. In this way, not only the points of displeasure can be determined and solved, but also the business would be expanded by introducing and selling new products.

Short messages called tweets in Twitter that can be only 140 characters are incomplete, unstructured, linked, and noisy [27]. So, those types of data could not be reliable sources to infer users' preferences. Therefore, social relations, which widely exist in social networks, should be used to increase accuracy of prediction. For example, *friendship*, *mention*, *tweets*, and *retweets* networks contain valuable information that can help predict users' opinions. Not only does considering social relations increase the accuracy of prediction, but it also helps discover the opinions of passive users and content consumers in social networks. Because passive users and content consumers tend to read

content provided by others rather than providing their own, they do not create enough content to be used for prediction. In this case, the only available information to infer preferences of passive users and content consumers is the friendship network.

In a directed online social network, such as Twitter, two people do not have to cooperate in the creation of a link between them; there is no requirement that one user send a friendship request and the other user accept it. In those type of OSNs, unlike Facebook, relationships are not limited to being friends. For instance, user A can follow user B without being followed back. In this case, user B is a "followee" of user A, and user A is a "follower" of user B. In this way, user B did not show willingness to connect to user A. Knowing that direction of the links in an OSN, such as Twitter, reflects the eagerness of users to form relations, there is a question of whether preferences of users who do not produce content should be inferred using their followers' or followees' preferences.

1.1 Thesis objectives

In each online social network, users can have four different roles in terms of providing content. These four roles are as follows:

1. *Active content producers* who create new messages, retweet others' tweets, and share photos frequently;

2. *Semi-active content producers* who create new messages, retweet others' tweets, and share photos from time to time;

3. *Active Content consumers (AC consumers)* who do not provide any content about a given topic, but read content related to that topic;

4. *Passive users* who do not have any interaction activity, such as tweet and retweet, at all or for a long time. Also, they do not follow or they are not followed by many people.

The main objective of this thesis is to develop a method to infer political preferences of common types of users who are AC consumers in a directed online social network, such as Twitter. As discussed earlier, AC consumers tend to read content of other users on a specific topic rather than write about those topics themselves. Hence, AC consumers' personal preferences on a given topic should be inferred from what they read instead of what they write. To discover what types of content AC consumers read that reflects their political preferences, what content their neighbours produce about the topic should be considered. According to the homophily principle, similarities between connected users are more than between two random users. However, AC consumers do not provide enough content on a given topic to attract attention of other users with similar views to follow them. Furthermore, AC consumers usually do not have lots of followers with a large degree of similarity regarding a given topic. In this case, it is important to determine whether considering only followees can predict personal preferences of AC consumers

as well as analysing followers or combination of followees and followers for active and semi-active content producers.

Efforts have been made in other research [3, 20, 21, 13, 24] to infer preferences of popular users instead of common types of users who are active content consumers. However, popular users are followed by millions of users but follow few users. Famous users are more selective about people whom they follow, because their behaviour and the choices that they make are being monitored by the whole world. Thus, studies on those users cannot be generalized for ordinary users as well. Currently there is no research to investigate the effectiveness of using follower and followee networks to infer personal preferences of active content consumers separately.

In some tasks, such as predicting election results or predicting crises before they happen, a larger data sample would lead to more precise predictions. Additionally, predictions done on larger data sets are more generalized. Consequently, it is important to consider preferences of AC consumers, who do not provide any content on a given topic, in making predictions. Similar to AC consumers, non-supporters of a given topic can have effect on the accuracy of prediction tasks. Therefore, to expand the size of the labelled data set and to make precise predictions, in this thesis both supporters and non-supporters of political parties are taken into account. This experiment was conducted on a set of more than 80 thousand Twitter accounts to infer

AC consumers' political orientations.

1.2 Thesis organization

The remainder of this thesis is organized as follows. Chapter 2 presents background, emphasizing social theories in online social networks and predicting missing information covering prediction based on user-generated content and inferring users' interests using social relations. Chapter 3 presents how data sets were created and classified. Also, this chapter shows how followee, follower and followee/follower graphs were generated to be used to predict political preferences of passive users by applying plurality and weighted plurality voting systems. Chapter 4 presents prediction accuracies obtained by using follower, followee and follower/followee networks. This chapter examines how well networks can be used to predict political preferences of AC consumers. Chapter 5 presents conclusions and recommendations for future work.

Chapter 2

Background

2.1 Social theories in online social networks

Currently, many studies [26, 12, 2] have been conducted to answer the question of whether social theories could be applied to social media or not. Their research has led to a positive answer of that question. Social perspectives of an OSN (Online Social Network) can be understood by applying social theories; however, an OSN is a complicated entity. For example, in an OSN such as Twitter, data is quite "big, incomplete, unstructured, linked, and noisy" [27]. As a result, it is not an easy and straightforward task to infer required and useful information from untidy data. In addition, not all users in social networks are active in terms of providing content consistently; they are users who are just active content consumers. Nevertheless, knowledge of social relations widely exists in social media. The data collected contains

useful information that can be used even when text is not available. So, social theories can help to represent, extract, or analyse behaviour of even content consumers to understand social media better.

Many social theories developed by social scientists can assist researchers with analysing social media. For example, Tang *et al.* [28] reported three important social theories, which have been applied to social media by many researchers. Those theories are social correlation theory, balance theory, and status theory. Because most of the studies discussed in this chapter focus on social correlation theory, the general idea of this theory is explained below.

According to [28], Social correlation theory is an important social theory that contains three social processes, homophily, influence, and confounding. One reason that social correlation is important is some links such as friendships in social media can be explained by it. Due to this theory, "attributes or behaviours of adjacent users are close to each other" in a social network. For instance, it is more likely that users who are friends have the same age or support similar political parties, which are an example of homophily. Homophily is one of the social processes that explains social correlation principle [28]. There are many studies [36, 23, 19, 35] that have analysed the existence of homophily in social networks in this decade. Homophily suggests that people have a tendency to interact with those who have common interests, personalities, or attributes [15]. In other words, based on the homophily principle,

a tie between two nodes in a social graph is more likely to be created when those nodes are similar to each other. The other social processes of social correlation are not relevant to this thesis.

Different types of similarities can connect individuals together. Initial studies on social networks by Miller *et al.* [15] showed that similarities among users can be assigned to two major groups. The first group is demographic characteristics like race/ethnicity, age, and sex. The second group consists of psychological characteristics such as intelligence, attitudes, and aspirations. When friendship has occurred, it is more likely that those users of social network will have common interests or political views.

By knowing that the interests of connected users are like each other, the preferences of those which have not been provided about a given topic can be inferred. For instance, Mislove *et al.* [17] showed that missing attributes such as geographic location and schools that users attended could be inferred from users who shared those in Facebook in the same community. Their results indicated that by having the attributes of only 20% of users, the attributes of the remaining users could be predicted with 80% accuracy. Similarly, Abbasi *et al.* [2] demonstrated that among adults, those who have similar political behaviours are more likely to be friends. So, Abbasi *et al.* used preferences of influential neighbours of users, who did not mention attributes such as political view in their profile, through their friendship network to

predict missing attributes. The high accuracy of their prediction results, supported the idea that even the behaviour of passive users of an OSN can be predicted using the information of the people to whom they are connected.

Not only immediate friends, but also the friends of friends can be taken into consideration to extract useful information. As an illustration, Wen *et al.* [35] investigated homophily among one degree friends, two degree friends, and three degree friends in traditional communication media, such as email and meeting calenders. The authors intended to study whether one's interests could be inferred from one's neighbours and surrounding friends. Their work is unique compared with other works, in terms of considering friends with various ties and different degrees of separation. Surprisingly, their investigation showed that two degree friends provided more information than one degree friends. In other words, they showed that although users with one degree of separation were important for inferring users' interests, friends with two degrees of separation had the most impact on inferring user's interests. However, ignoring three degrees of separation friends did not have any pronounced effect on prediction. So, preferences of users who did not provide them can be inferred from both immediate and surrounding friends.

2.2 Predicting missing information

Generally, social media mining tasks can be done by using the different information that exists in OSNs. For example, in a social medium such as Twitter, users can tweet about a variety of topics from trips to politics. Also, they can retweet original tweets of others having the option to add 140 characters to them. Additionally, they can mention their friends' names using the "@" character, and like each others' tweets and retweets. Other useful information that can be obtained from a social medium, such as Twitter, is the direction of links between two users. Such relationships exist in Twitter, where people can have a one way friendship with others. So, one user can follow another user without being followed back.

On the basis of understanding how links are formed, the existing literature is categorized into two main groups. In the first group, in which links are ignored, user-generated content is the only information that is used for different purposes, such as determining polarity of users to determine whether their opinions about a given topic are positive or not. Basically this group takes advantage of tweets that are created by a user, or retweets that are not created originally by this user, but are used by him/her. Also, information such as age, location, and gender, which is provided by users in their profiles, is used in this group of studies. However, missing information of users who do not have enough content can not predicted by applying this group of

research work.

The existing literature categorizes the second group based on the different links that widely exist in social networks. For example in this group, three aspects of online communications are considered:

1. The links that exist between users,
2. The tweets that those users retweeted,
3. The original authors of the tweets.

However, those relations are used both in an undirected network, where there is no difference between followers and followees, and a directed network, where direction of links contains important information. This group of works can be used for prediction task when availability of text data in an OSN is low.

2.2.1 Prediction based on user-generated content

Because of the growing usage of social media, many questions can be answered by gathering and analysing the data provided by people in OSNs rather than traditional methods such as conducting paper surveys [1] or online polls. For Instance, possible election outcome can be predicted by using the tremendous information that is available in OSNs. Thus, automated

tools that can predict the interests and opinions of users based on social networks data can be a suitable replacement for traditional surveys [26]. This replacement leads to more accurate analyses by going through larger samples, which are not counterfeit. Because, usually when people are not aware of being watched, they express their opinions honestly. Consequently, the gathered data are precise.

In a social network such as Twitter, more than 7000 tweets are being posted by users about a wide range of topics every second. Such a massive amount of content can be used for a variety of purposes. For example, Michelson *et al.* [16] used entities in tweets to cluster users based on their interests, and discover their main interest as well. Also, in [8, 22], authors conducted a study in which political alignment of Twitter users was predicted to see whether they would support a left or right wing party. Moreover, Al-Kouz *et al.* [4] applied a new algorithm named Root-Path-Degree to text messages generated by users to infer their topics of interest. In addition, researchers in [1] collected data (posts, comments, and tweets) from Twitter and Facebook, the blogosphere, etc. to analyse human online behaviour and predict their real-world behaviour. Al-Kouz *et al.* [4] showed that this was a challenging task because users' online behaviour was not always similar to what would happen in real world events.

The benefit of using content generated by users that reflect their opinions is

that it is more reliable than inferring interests from friendship networks or similar entities. However, the disadvantages of relying only on content are tremendous. The most highlighted drawback is that content provided in social networks, and especially microblogs such as Twitter, are full of ambiguity. Also, they are fragmentary, immense and growing steadily. Consequently, it is quite a difficult task for machines to determine the semantics of those texts precisely [12]. Additionally, detecting opinions of OSNs users who do not share anything regarding their opinions can not be done by applying classical sentiment analysis techniques. The reason is that those techniques can only be applied when text data are provided for expressing an opinion. Therefore, new approaches that use relations between users of OSNs to understand what they think about a specific topic instead of looking at what they write should be considered [20]. As a result of taking advantage of new models, not only accuracy of classification increases, but also opinions of new users and users who did not express them explicitly can be determined.

2.2.2 Inferring users' interests using social relations

As already discussed in Section 2.1, users with similarities are more probable to create a friendship link. However, friendship link in Twitter can be a one-way link if only one side of the connection finds himself/herself alike the other side. Therefore, direction of the links have different meaning in OSNs.

The researches presented here are divided to two subgroups. The first sub-

group includes studies that considered different existing links in a social medium without taking their directions into account, and the second one contains works that examined both links and their directions, which allow predictions to be made with improved accuracy.

2.2.2.1 Ignoring direction of links for prediction tasks

In recent studies [37, 27, 9], a wide range of techniques have been applied to different social connections in social media to infer users' opinions about given topics without considering the directions of of links. For example, Chaabane *et al.* [7] proposed a technique through which social media users' private attributes can be inferred. They asserted that the only data they needed from a social medium to guess users' interests is the content they liked in Facebook without any need for their profile or friendship information or the group they belong to. They used the Latent Dirichlet Allocation (LDA) generative model to infer hidden interests by having the names of the other interests that the user previously liked, such as "Justin Bieber" in the music category. So, having the information that users revealed about their music interests, the researchers could guess some of the users' other interests, not necessarily related to music, with 70% accuracy.

Other types of discourses that ignore direction of links conducted by [32, 18, 19], not only considered the content provided by users, but also took users' social connections into account. The rationale was that dealing only

with complex data would not lead to satisfactory results. So, each of these three studies inferred users' interests or determined polarity of either users or posts based on the link structure assumption using both content and relations, such as retweet networks. However, none of those studies considered followees or followers networks (also known as follow network). The results of their works showed that taking social connections into account improved the accuracy of prediction significantly.

Besides the social connections considered in the last group, follow network can affect accuracy of prediction. As an illustration, Tan *et al.* [27] proposed an approach to infer users' sentiments about a specific topic by using two networks: mention and follower graphs. In each network, they considered two types of relations. The first relation is when user A and user B are mutual friends or when user A and user B both mention each other in their tweets. The second relation is when only user A follows user B or user A mentions user B and not vice versa. In order to predict whether a user opinion about a given topic is positive or negative, Tan *et al.* used a directed heterogeneous graph in which nodes are users and tweets. Their results showed that sentiment of users can be predicted more accurately if users are mutual friends or they mutually mentioned each other in tweets. Although the accuracy of predicting a user polarity using one-way mentions or one-way follow links is less accurate than using mutual mentions or friend links, they are still much better predictors of polarity than using two random users. In addition, the

researchers showed that predicting user polarity using the follow graph is more accurate than the mention graph. They argued that the reason that the follow graph is more accurate is because users who follow each other have more similarity than people who mention each other. The other reason for the improved accuracy is that mention links are very sparse, which leads to inaccurate results. So, in the end, they demonstrated that links among users are highly correlated.

Knowing that the follow network is a reliable source of prediction, researchers [23, 9, 12] used the follow network and user-generated content to find users' preferences and polarity. The authors argued that considering both the friendship network and user-generated content led to satisfactory accuracy. Similarly, Speriosu *et al.* [26] considered the followers graph in addition to tweets for polarity classification. They created a graph in which nodes were users were connected based on the follower graph, where tweets were connected to their authors, and word unigrams, word biograms, hashtags, and emoticons were connected to tweets. Thus, not the whole friendship network, but only the follower graph was used in their experiment. The researchers showed that in their implementation, ignoring the follower graph did not change the performance. Their assumption was that the classification accuracy was not enhanced by using the follower graph as a result of ignoring the direction of the links between users, which is important in a social medium such as Twitter.

The benefit of using social relations such as *mention*, *retweet*, and *like* networks, which exist widely in social media, is that not only user generated content is used to find the opinions or polarity of a user/text. Because, as was discussed previously, those messages are full of ambiguity, and especially in a microblog like Twitter, they are too short and incomplete to reflect the exact feelings of their authors. As a result, considering both text, when it is available, and social relations help improve the accuracy of predictions. However, there is no question that passive users, new users and active content consumers in a social medium do not have a large history of likes, tweets, retweets, or mentions on a given topic. Therefore, those techniques can not be useful for those types of users. Nevertheless, Wang *et al.* [33] proposed a method, which is a common approach for recommender systems, that relies on homophily to infer interests of inactive and new users in online social networks, based on their social connections. They tried to infer interests from different social connections such as "retweet," "mention," "follow," and "comment" networks applying a "random-walk algorithm." So, for active users, Wang *et al.* used the content users provided to predict their interests, but for inactive users, they used information of their connection network. The result of their work showed that predictions of users' interests based on retweet and follow networks are more accurate than those using mention and comment networks. Although, in many situations, user-generated content is not available, Welch *et al.* [34] showed that predictions using retweet net-

works are more accurate than the follow network.

However, in a social medium such as Twitter users can connect to each other without any need to send or confirm a friendship request [36]. Consequently, one user may have varying types of people having diverse interests as his/her followers. Therefore, individuals can decide whom to follow due to their own tastes and characteristics. Thus, although they cannot control their followers, they can be selective on their followees. Furthermore, users do not follow others for the same reason that they are being followed. For instance, if user A follows user B either because of the interesting tweets that user B posts, or because he/she finds some similarities between himself and user B, it does not mean that user B feels the same about user A. If so, user B could follow user A as well. Thus, it is not precise to assume that personal preferences of users can be predicted using their neighbours, regardless of direction of the links between them.

Also, the ratios of followers and followees for common types of users who are not active in social media are different with other users. More specifically, passive users and active content consumers do not have a lot of followers, as they do not share content to attract individuals' interests. However, they may have adequate number of followees, as they may follow many accounts to get the most recent updates about a topic of their interest. In this case, the only available social relation to rely on for prediction is the followee

network. Nevertheless, the accuracy of predicting preferences of passive users and active content consumers by using only the followee network may lead to different results than using the combination of followers and followees. Currently, no research has been conducted to answer the question whether political interest of active content consumers who do not have many followers can be inferred from their followees with satisfactory accuracy.

2.2.2.2 Considering semantics of links for prediction tasks

In all of the studies mentioned previously, the researchers treated online social networks like undirected friendship networks, in which both sides of the interaction were equally responsible for that friendship. However, it has been shown by Jeong *et al.* [13, 24] that users of social networks follow without having equal responsibilities. In such a relationship, they follow people without being followed back. Therefore, to understand if those people are following a Target User blindly, favorably, or unfavorably, it is important to classify those followers based on other criteria [13]. For instance, some users follow others because they have the same opinions as the people whom they follow, or sometimes because they just want to read their news, although they disagree with their opinions. So, based on the ratio of tweets, retweets, and profile descriptions, the results of the research showed that followers can be classified into three groups named influential supporters, influential non-supporters, and non-supporters. Similarly, Tan *et al.* [27] showed that the probability that users with the same opinions follow each other is much higher

than connectivity of users with opposite opinions. Nevertheless, friends usually are of the idea that their friends are like themselves, when they have never discussed topics that can be points of disagreement. Consequently, it is quite hard to determine whether two persons have real political tastes in common, or they have other similar characteristics that lead them to think they have similar political views [15].

Currently, various researchers have been conducted to answer different questions by considering the semantics of links in social media. For instance, link prediction is a known area of research which has attracted a huge amount of attention in this decade [39, 25]. Link prediction, is the process of recommending new friends in online social networks [28]. According to [10], link prediction is predicting creation of a link between nodes in a social medium based on link information and attributes of existing objects. In other words, having a set of links in time t, with link prediction techniques, new links at time $t + 1$ will be predicted. For example, in [39] the researchers devised a method to predict new links that will be created in a hybrid social medium like Twitter by analysing link structures without considering content or other attributes. Based on link prediction, social media networks, such as LinkedIn, Facebook, and Twitter, can recommend new connections to users. The recommendations are based on similarities between users which refers to the homophily principle. Because of the link prediction, users can expand their network with new suggested users.

Other studies have tried to answer the question of whether a friendship network is a reliable source of information about users who do not provide it. Before discussing the researches conducted to answer this question, it is worth mentioning Weng *et al.* [36], who asserted that they were the first ones to investigate the existence of homophily in Twitter. They tried to understand if homophily existed between users through followees links in Twitter by answering two questions:

1. Do people with followees relationships have more similarity in topics in comparison with people without that relationship?

2. Do people with reciprocal followees relationships have more similarity in topics in comparison with people without that relationship?

Because the answers to both of the above questions were positive, Weng *et al.* came to the conclusion that homophily between users in a followee graph exists. Similarly, Abbasi *et al.* [3] used Facebook fan pages to infer preferences of popular users, such as Obama, using follower and followee networks separately. The goal of this work was to determine whether followers or followees were more effective in predicting one's preference. The result showed that the degree of homophily between a user and his/her followees was slightly more than a user and his/her followers. Also, Abbasi *et al.* showed that using a combination of followers and followees did not improve the accuracy

of prediction in comparison with just using the followees network.

According to [6], the ratio of followers to followees for popular users, such as politicians, celebrities, newscasters, and journalists, approaches infinity. Because popular users are followed by millions of people, but they do not follow the same number of users. In contrast, the ratio of followers to followees for common types of users approaches 1. The difference between the ratios shows that popular users are noticeably selective of people whom they follow. Nevertheless, individuals with various preferences can follow popular accounts for a wide range of reasons rather than having the same interests. Consequently, it was not surprising that [3] showed the similarity between popular users and their followees was greater than between popular users and their followers. Furthermore, this study cannot be generalized to conclude that *all* users, regardless of their popularity, are more similar to their followees than followers. Also, relying on political preferences, which were clearly demonstrated on fan pages as public attributes, was a solid way of assigning users to political classes. However, a tremendous number of users tend to express their views through their timelines instead of sharing them explicitly as an attribute. Thus, not considering user-generated content in users' timelines would eliminate a large number of users from the experience. Constraining the specific group of people may lead to different accuracy in comparison with the case that more users were involved.

As it is discussed in Section 2.2.2.1, by using link information, accuracy of classification would be increased compared with classical classification techniques that only used user-generated content. Also, in a study conducted by Rabelo *et al.* [20, 21], the authors argued that political tastes of users who did not express them directly through messages could be inferred by using link information. So, they adopted a directed graph in which nodes were users, and links were followers and followees of users in Twitter. Also, collective classification techniques were applied to classify active users' posts to left- or right-wing parties.

In order to obtain a dataset, specific hashtags such as #Obama2012 and #Dems2012 which indicated left and right parties, were defined to classify supporters. Other users who used those hashtags in criticizing context were considered as noise. Hence, 9098 posts provided by 4719 American politicians as authors were collected. Then followers and followees of the authors were gathered, which increased the number of nodes to 97000. The goal was to assign non-authors to "left" and "right" classes. Consequently, because in the expanded graph only 5% of the nodes were labelled, collective classification would not be practical. So, the graph was pruned intensively to only keep the nodes that were strongly connected to authors. In this way, only 1.28% of unlabelled users remained in the graph. Thus, the algorithm started by removing the non-author nodes with fewer connections, since they were less likely to break the graph connectivity. The algorithm stopped when

removal broke the connectivity. By applying the pruning graph, 995 author nodes and 249 non-author nodes remained in the graph.

In [20, 21], due to the lack of number of labelled users, only 5% of the whole data set, the graph had to be pruned heavily. Thus, only users who were highly connected to the labelled users remained in the graph. Consequently, political opinions of most of the inactive users were left unpredicted even after applying the introduced method. Moreover, individuals in a society have varying approaches to discuss one single subject. Some people simply use positive language to reflect their agreement of a topic and others use negative language to criticize it. In both cases, there users are interested in the topics positively or negatively. Therefore, there should be a difference between those people and the ones who do not comment on the topic at all. Furthermore, to get precise understanding about opinions of the society about the given topic, opinions of both supporters and non-supporters should be considered. Omitting users who criticized the topic of investigation from the experiment, will exclude a large portion of valuable information. Thus, the obtained result from the method would not be accurate enough.

In addition, although there are some well known political hashtags which are popular among OSN users, the existence of those hashtags does not prevent users from using other ones to reflect their views. So, a constant list that consists of limited number of hashtags to classify all users will leave many of

the users in the dataset unclassifiable. Similarly, narrowing down the number of classes or defining general classes rather than specific ones that reflect users' preferences may not predict opinion the opinions of users precisely.

The study shows that Rabelo *et al.* [20, 21] are the first ones who considered directed friendship links in Twitter to classify users who did not post their opinions through user-generated content. In addition to novelty, the accuracy of the proposed method was reported at 80% , which was noticeably high. However, a huge amount of users who did not express their preferences clearly whose opinions could be important for some tasks, such as predicting an election result, were excluded from this experience and their opinions left unpredicted. In conclusion, there is a need for a comprehensive labelling of network nodes to avoid the intensive pruning, which leads to a biasing of the test network. As a result of comprehensive labelling, the opinion of more passive users can be inferred.

Chapter 3

Methodology

This chapter presents the method used to infer the political parties that were supported by users. The users considered here are called Target Users (TUs) who are active content consumers, representative of common types of users on Twitter. The process of discovering TUs' political preferences started by creating a main data set (Reference data set) that contained TUs and all of their followers and followees (follow network). Then, all entities in the data set were assigned to the classes that represented users' political views. The process continued by validating the correctness of the labels to make sure the performance of the proposed method was not affected by mislabelled users. Finally, in order to evaluate the effect of data characteristics on accuracy of prediction, four data sets were created based on the "reference data set" that was used to predict the political preferences of TUs. The created data sets helped to form graphs of followees, followers, and combinations of followees

and followers. Then, by applying plurality and weighted plurality voting systems to each generated graph, it can be determined whether only followees can be a reliable source to predict preferences of TUs. Each of these processes is explained in detail in the rest of this chapter.

3.1 Data sets

The population that was used to create data sets was chosen randomly by crawling public Twitter accounts of users, representative of common type of users with less than 2000 followers, who posted frequently about federal Canadian politics and could be labelled as having a party preference. The followers and followees of these users were then crawled.

In Twitter, as discussed in Chapter 2, users can have a two-way friendship or a one-way friendship. In other words, a link between two users can be created without both users agreeing to be friends. Also, the Twitter Application Programming Interface (API) policy allows developers to have access to the most complete information of users' timelines, such as the collection of recent tweets, recent retweets, mentions, as well as the collection of friends' and followers' IDs [29] in a machine-readable JavaScript Object Notation (JSON) format. For the purpose of this thesis, to create the data sets, Twitter was crawled by using Twitter4j [38], which is an unofficial library for the Twitter API.

3.1.1 Collecting TUs and their follow network

The first step to create a data set was selecting TUs whose political tastes would be predicted by the proposed method. "TUs" in this experiment, unlike [3, 20, 21], were the common types of users (ordinary users), whose labels were predicted using their followers and followees. Among all of the existing users in Twitter, individuals selected for this study were those who were interested in "Canadian politics" and specifically one of the federal parties of Canada, "Conservative", "Liberal", or "New Democratic". Also, users were limited to those whose number of followees did not exceed 2000, because at the time the data for this experiment was collected, the Twitter help center [30] determined that every account could follow only 2000 users. However, on October 28, 2015, the follow limit for Twitter increased from 2000 to 5000 [11]. This number could vary for users with different ratios of followers and followees. Accordingly, a follow limit of 2000 was used for TUs.

The next step after finding the TUs was to collect all their followees' and followers' useful profile information that could help the labelling process. To crawl each followee's and follower's profile, two separate procedures were performed as follows:

1. Obtaining all followers' and followees' IDs from the TUs' profiles;

2. Using Twitter APIs to search through the obtained IDs' profiles to gather information, such as screen names, descriptions, followees and

followers count, and the 200 most recent tweets (at the time that the data set for this experiment was collected, 200 was the maximum number of tweets that Twitter allowed to be crawled).

After the TUs and their follow networks were collected, all of the information was saved in a data set named "reference data set". Then all the entities of this data set had to be assigned to appropriate political classes. More explanation of this data set and classification of TUs and their follow networks can be found in Section 3.1.2.

3.1.2 Reference data set

There are a variety of approaches that users can choose to show their agreement or disagreement with political parties in Twitter. For instance, individuals may use a name of a party in their screen names, such as @Maryndp or @Jesse_liberal. Also, they can promote a particular party in their descriptions with something like "I pray for the NDP to get elected to deliver social democracy and elevate the poor". Moreover, they can tweet or retweet frequently used hashtags in support of one specific party. In all of the mentioned approaches, the labels that are used in descriptions, screen names, or tweets help to distinguish supporters and non-supporters of each political party. Collecting a list of the well known labels or hashtags among supporters and non-supporters of each party assisted classifying the entities in the "reference data set". It is important to note that the process of collecting

hashtags and labelling was different for TUs and follow networks, which will be explained in detail in the following sections.

3.1.2.1 Labelling TUs

As discussed in Section 3.1.1, only ordinary users with fewer than 2000 followees interested in either "Conservative", "Liberal", or "New Democratic" parties had the potential to be selected as TUs. Most of the hashtags that represented each of the mentioned parties were extracted from a website named "poliTwitter" [14]. The rest of the hashtags were extracted from tweets of the followers of popular accounts, such as @CPC_HQ, @liberal_party, and @NDP_HQ. These accounts were official Twitter accounts for the Conservative, Liberal, and the New Democratic parties of Canada respectively. A list of all of the collected hashtags for the three mentioned parties can be found in Table 3.1.

The most reliable way to understand political preferences of users is by examining the words they use clearly in their screen names and descriptions in support of their views. Therefore, if a user reflected the political party that he/she supported in his/her description or screen name by something like "Conservative, Chinese Christian Convert from Atheism, Scientist, Anti-Scientism" or "Jack_CPC", he/she was assigned to the class named Conservative. Thus, because users showed their support for the Conservative party, they were assigned to that class regardless of having any tweets or retweets

Table 3.1: Popular hashtags representing Liberal, Conservative, and New Democratic parties of Canada

Conservative	**Liberal**	**NDP**
# cpc	# lpc	# ndp
# pmHarper	# ptlib	# ptndp
# voteConservative	# lpcldr	# ndpldr
# harper	# canadianLiberty	# ndldr
# cpcYouthConf	# teamTrudeau	# ndleader
# voteHarper	# bldgtheplan	# ndpnow
# newTaxBreaks	# LPCO	# ndprally
	# Trudeau	# votendp
	# ylcjlc	
	# voteLiberal	
	# youngLiberals	
	# liberal	
	# bldgtheteam	

about the Conservative party or any other political topics. However, users could show disagreement with a specific party by saying something like: "Political interest, tired of Conservatives" in their descriptions or "StopHarper" as their screen names. In this case, because they were non-supporters of the Conservative party, their tweets and retweets were investigated to find out if they supported the other two parties or not. In this study, TUs had to be supporters only of the "Conservative", "Liberal", or "New Democratic" parties.

There were many users selected as TUs who were creators of parody accounts. Parody accounts in Twitter are fan or commentary accounts that are gen-

erally made to interest people by humour and sarcasm regarding a specific topic. Furthermore, although those accounts discussed a specific party using the hashtags in Table 3.1, they could not be considered as TUs as they did not reflect actual opinions of a person. Also, many users tended to tweet or retweet frequently about a certain party, which could indicate support for that party. However, if they used a picture attacking that party as their profile or header photo, this showed that they were against, rather than in support of that party. Thus, profiles of the users who could be selected as TUs were checked manually [8, 27] to make sure the TUs were not chosen from non-supporters, or parody accounts. Also, these profiles were checked to validate correctness of the labels. It was important to classify the TUs correctly, because the performance of this prediction method would be affected if TUs were classified incorrectly or TUs did not meet all the determined criteria.

3.1.2.2 Labelling follow networks

Target users selected for this experiment were supporters of the "Conservative", "Liberal", or "New Democratic" parties of Canada. However, TUs' followers and followees could be both supporters and non-supporters of political parties, or could be stating their political philosophy rather than explicit party preference. Also, TUs' follow networks could be interested in American instead of Canadian politics. Therefore, in addition to the hashtags that were used to label TUs, the ones that were popular among supporters and

non-supporters of other parties, or the ones that were used to show political spectrum either in Canada or the USA were considered. In this case, political tastes of a large portion of the TUs' follow network could be understood more clearly. Tables 3.2, 3.3, and 3.4 represent some of the hashtags that were used by Liberal, New democratic, Green, Democratic, Progressive Conservative, Teaparty, and Conservative parties. Also, general hashtags used to express politics like cdnpoli, noIranDeal, billC51, etc. were taken into account to distinguish users who talked about politics from ones who did not have any interests in politics at all. Some of the political parties' names that were used as the labels can be found in Table 3.5. The following algorithm shows all the steps that were taken to label the TUs' follow networks.

- **Step 1:** Create a list named "Labels List", consisting of popular political parties and the labels that show political views of users (Tabel 3.5);
- **Step 2:** Create a list named "Hashtags List", consisting of hashtags represent each class in the "Labels List"(Tables 3.2, 3.3, and 3.4);
- **Step 3:** Search for the hashtags in the "Hashtags List" in a user's description and screen name:
  - **Step 3-1:** If the user used new hashtags not included in the "Hashtag List", update the list;
  - **Step 3-2:** If the classes in the "Labels List" did not reflect his/her political view, add a new class to the "Labels List";
  - **Step 3-3:** Assign the user to an appropriate class;
  - **Step 3-4:** If the user was a supporter of a specific party, change his/her flag to "labelled";
  - Otherwise, go to step 4;

- **Step 4**: Search for the hashtags in tweets and retweets of an unlabelled user;
  - **Step 4-1:** Repeat steps 3-1 to 3-3
  - **Step 4-2:** If the user was a supporter or non-supporter of a specific party, change his/her flag to "labelled";
  - **Step 4-3:** If the user already had a class (from step 3), but his/her flag was "unlabelled", change his/her flag to "labelled";
- **Step 5:** Repeat all these processes for all unlabelled users.

After the initial political classes and hashtags were determined, two procedures were performed to label the TUs' follow network. The first was to classify users who reflected the hashtags showed in Tables 3.2, 3.3, and 3.4 either in their screen names or descriptions. The second procedure was to classify users who had those hashtags reflected in their tweets. In both processes, in order to include as many users as possible, all the non-English descriptions, tweets, and retweets that had the specified hashtags were translated to English from their source languages using Google Translate. To differentiate labelled and unlabelled users in the data set, a flag was assigned to each user with the default value of "unlabelled". This flag changed to "labelled" if the user reflected any interest in politics.

As discussed in Section 3.1.2.1, users could show agreement or disagreement with a specific party in their screen names or descriptions. If users supported a party through their screen names or descriptions, they could be classified without any need to check their tweets or retweets. Thus, their flag changed

Table 3.2: Hashtags used by supporters and non-supporters of Liberal, NDP, and Green parties of Canada and the US Democratic party and the hashtags used to show the political spectrum

Hashtags used by Liberal, NDP, Green, and Democratic party		
# lpc	# liberal	# ptlib
# canadianLiberty	# teamTrudeau	# LPCO
# Trudeau	# ylcjlc	# voteLiberal
# bldgtheteam	# salma zahid	# libcaucus
# lpcdb8	# ruliberals	# bldgtheplan
# uoftliberals	# iamliberal	# LPCOGM
# libCrib	# justinTrudeau	# liberalPrivilege
# harperSucks	# harperBumperSticker	# surpriseHarper
# harperTransitPlans	# stopHarper	# ndp
# ndpldr	# ndldr	# ndleader
# ndprally	# votendp	# ElizabethMay
# greenparty	# greenSurge	# connectTheLeft
# topprog	# democrat	# progressive
# obama	# obamaCare	# thomasMulcair
# orangeCrush	# bcndp	# readyForHillary
# p2	# secularDemocrat	# tpot
# brackObama	# uniteBlue	# mulcair
# abndp	# tm4pm	# ndpnow
# hadEnoughHarper	# GPC	# ptndp
# unseatHarper	# harperBlamesTrudeau	# cpcJesus
# JudyLaMarshFund	# lpca	# youngLiberals
# lpcldr	# YoungLiberalsOfCanada	

to "labelled". However, if users reflected disagreement with a party three steps were taken to classify them. First, those users were assigned to a class that reflected their disagreement, but their flags did not change to "labelled". Then, non-supporters' tweets and retweets were checked to find out if there was any specific party that they supported. Finally, if they supported any

Table 3.3: Hashtags used by supporters and non-supporters of Conservative Party of Canada and supporters and non-supporters of the US Republican Party and the hashtags used to show the political spectrum

Hashtags used by Conservative Party		
# cpc	# voteConservative	# conservative
# cpcYouthConf	# voteHarper	# libertarian
# roft	# JustinOverHisHead	# tcot
# teaParty	# tgdn	# reagan
# conNC	# GOP	# obamaTheEnemy
# nspc	# abpc	# wldrs
# 2A	# harperPartyOfOne	# liberalssucks
# jebBush	# noObama	# pmharper
# obamaScandals	# obamaLies	# ronPaul
# bccp	# nlpc	# nbpc
# cons	# republican	# ccot
# teagan	# pcpo	# wrp
# stopHillary	# tlot	# stopLiberalracism
# wakeUpAmerica		

Table 3.4: Political hashtags in general

Hashtags used by users who talked about politics in general		
# noIranDeal	# cdnpoli	# canpoli
# abvote	# bcPoli	# onpoli
# abpoli	# C51	# IWillVote2015
# vanpoli	# topoli	# stopBill51

other party, their assigned class changed from non-supporter of a party to supporter of another party and their flag changed to "labelled". But, if they did not support any specific party, their initial label was kept for them and their flag changed to "labelled".

Table 3.5: List of the labels that the collected hashtags represent

Political Parties Names		
TeaParty	Republican	Democratic
Liberal	Conservative	NDP
Libertarian	Progressive	Green
Progressive Conservative		

Political preferences of users who did not support any particular party in their descriptions or screen names were discovered from their tweets and retweets. To differentiate the political and non-political tweets, only the tweets that included the political hashtags in the Hashtag List were involved in the labelling process. The approach to having a labelled corpus was a semi-automatic one: the political affiliations of users were indicated by counting the number of hashtags in their posts. The name of the party that was represented by the highest amount of hashtags in a user's tweets was considered as his/her label. However, this approach was not successful in distinguishing supporters and non-supporters of a particular party. More information on how to deal with this problem can be found in Section 3.1.2.3.

One of the important criteria for creating the data sets in this study was to include as many users who discussed politics as possible, because the size of the data set can affect the performance. Consequently, no limits were specified for the number of the political tweets that a user had to have to be

qualified for classification. In other words, in this experiment the only determining factor to classify users was the content of the tweets and not number of the tweets. However, when a user did not have lots of political tweets or retweets, political tweets were a more reliable source from which to infer one's preferences than retweets. For instance, if a user had only one political tweet, which was"I am a lifelong #Liberal and will vote for #justinTrudeau in this election", he/she was classified as Liberal. However, a user with seven retweets sharing general news about Liberals was excluded from this experiment. Also, to increase the size of the data set, unlike [8, 21, 6, 26, 23, 13] that used a limited number of hashtags to assign right, left, positive, or negative classes to users, in this study the Hashtags List was updated dynamically. Thus, if in the process of labelling, any new political hashtags were found, the Hashtag List was updated with the new hashtags. Then, the updated list was used to classify unlabelled users. The advantage of updating the list of hashtags continuously instead of relying on a constant list was that users who did not express themselves using the well known hashtags would not be excluded from this experiment because there was a high possibility that users did not restrict themselves to popular hashtags and used a variety of them instead.

Similar to the Hashtags List that was updated continuously, Labels List had to be updated as well. Although political views of some users did not fit into the pre-defined ones in Table 3.5, they explicitly talked about their

political opinions. Furthermore, a new class was added to the Labels List that represent that user's political taste. As an illustration, if a user reflected disagreement with a party, he/she had to be labelled as non-supporter of that party. So, for example even if a class named "anti-NDP" did not exist in the Table 3.5, it was added to the list. Also if an individual supported more than one party, he/she had to be labelled as supporter of all those parties. For instance, the list was updated with "NDP/Liberal" class to represent users who showed interest to the both parties evenly. Thus, every time that the algorithm ran for unlabelled users, an updated list of hashtags and classes was used to classify the follow network.

3.1.2.3 Validating correctness of labels

It is common in online social networks that both supporters and non-supporters of political parties use the same hashtags in different contexts. For example, non-supporters use the same hashtags as supporters in a sarcastic way to show disagreement and opposition. In contrast, supporters use the same set of hashtags with positive language. Consequently, errors arose when supporters and non-supporters had to be distinguished for a certain party. This problem was mainly for the running parties in Canada and the US and specifically for the Conservative and Democratic parties respectively because usually people tend to criticize the current government rather than supporting a particular party. Furthermore, it was important to use an approach to discriminate supporters and non-supporters of a party who used the same

political hashtags.

Despite the importance of distinguishing supporters and non-supporters, the semi-automatic classifier could not categorize them correctly when both groups used the same set of hashtags. Therefore, non-supporters of a party were mistakenly classified as supporters of that party. For example, if a user used 132, 50, and 40 hashtags representing Conservative, Liberal, and New Democratic parties respectively among 200 tweets, the semi-automatic classifier assigned him/her to the Conservative class. Nevertheless, it was possible that those 132 hashtags were used in disagreement with the Conservative party. So, he/she had to be assigned to the "anti-Conservative" class.

In addition to not differentiating between the supporters and non-supporters successfully, the semi-automatic approach could not recognize ambiguity in tweets, retweets, and descriptions properly. For instance, users who mentioned they were small "l" liberal or showed interest in the liberal arts would be assigned to the liberal class mistakenly. Also, users who called themselves "republican/democrat" were considered to be users interested in politics and the terms were not treated as genuine political information. As a result, ambiguity in tweets led to mislabelling. Consequently, to make sure that contradictions were recognized precisely and users were assigned to proper classes, the profile of each user was checked manually. Thus, not only the 200 most recent tweets and retweets, but also profile icons, background photos,

URLs, older posts, etc. were accessed to validate the assigned labels. It was important to validate the correctness of the assigned labels to avoid the mis-labelled users leading to the wrong predictions and affecting the performance of the method.

3.1.3 Mapping the labels

As discussed in Section 3.1.2.2, the goal was to increase the size of the labelled "reference data set" by classifying all followers and followees who expressed their political views. For example, even if users talked generally about politics without having any specific political allegiance were assigned to a class named "politics". Moreover, not only supporters, but also non-supporters of Canadian or American parties were classified. However, labels of TUs were limited to "Liberal", "Conservative", and "New Democratic". Consequently, the "reference data set" could not be used as a training data set to predict TUs' preferences. The existence of labels in the training set that did not match actual labels of TUs would affect the accuracy of prediction. As a result, the following data sets were created based on the "reference data set" as training data sets. Each data set was created with a different strategy to determine how data characteristics would affect accuracy of prediction.

- **Data set 1 ("reduced labels")**: In this data set, most of the labels in the "reference data set" were mapped to the closest classes to "Liberal", "Conservative", and "New Democratic". For instance, by knowing that

the "TeaParty" is in the right of the political spectrum in the USA, users in that class were mapped to the "Conservative" class. Similarly, those in the "anti-Obama" class were mapped to "anti-Liberal/anti-NDP";

- **Data set 2 ("politics as noise"):** In this data set, unlike the "reduced labels", the label named "politics" was considered as noise. So, users who talked generally about politics without having any particular political preference were removed from the data set;

- **Data set 3 ("non-supporters to supporters"):** This data set is created based on the "politics as noise" data set. However, in this data set the rationale was that if someone showed disagreement with a specific party, it meant in an election he/she would vote for other parties. So, non-supporters of a party were considered as supporters of the other two parties.

- **Data set 4 ("Left or right wing"):** This data set is created based on the previous data set. However, in this data set both TUs' and follow network's labels were mapped to the classes named "right" and "left". Therefore, labels close to "Liberal" and "New Democratic" were mapped to the "left" class and the labels close to "Conservative" were mapped to the "right" class.

3.2 Follower, followee, and follower/followee graphs

Twitter is a micro-blogging tool that has attracted the attention of more than 300 million users, who are motivated by the question of "What is happening?" Also, Twitter is known as an increasingly popular tool among politicians who can use it as a user-friendly platform to post about political development and issues. Moreover, many news organizations, political strategists, and bloggers use Twitter to cover politics, the latest election news, political analysis, etc. Therefore, an individual can simply follow an account to get access to the most recent news regarding his/her favourite topics. However, when a user only follows others to read news, without any contribution in producing content, there is not a high possibility that he/she can attract users with similar opinions. Consequently, he/she does not have as many followers as followees regarding a given topic. Furthermore, there is a question of whether preferences of active content consumers can be predicted regardless of the preferences of their followers. To answer this question, the following graphs were constructed for each of the data sets discussed in Section 3.1.2.2.

1. A directed graph G representing part of a social network, created of a set $TU = \{u_{T1}, u_{T2}, ..., u_{Tn}\}$ of TUs and a set $U = \{u_{fr1}, u_{fr2}, ..., u_{frn}\}$ of TUs' followers who are active content producers, and set of edges $E_{U \times TU} = \{(u_{fr}, u_T) \mid \text{user } u_{fr} \text{ follows user } u_T \}$;

2. A directed graph G representing part of a social network, created of a set $TU = \{u_{T1}, u_{T2}, ..., u_{Tn}\}$ of TUs and a set $U = \{u_{fe1}, u_{fe2}, ..., u_{fen}\}$ of TUs' followees who are active content producers, and set of edges $E_{U \times TU} = \{(u_T, u_{fe}) \mid \text{user } u_T \text{ follows user } u_{fe} \}$;

3. A directed graph G representing part of a social network, created of a set $TU = \{u_{T1}, u_{T2}, ..., u_{Tn}\}$ of TUs and a set $U = \{u_{F1}, u_{F2}, ..., u_{Fn}\}$ of TUs' followees and followers who are active content producers, and set of edges $E_{U \times TU} = \{(u_T, u_F) \mid \text{user } u_T \text{ follows user } u_F \text{ or } u_F \text{ follows user } u_T \}$.

After creating the followers, followees, and combination of followers and followees graphs, "plurality" and "weighted plurality" voting systems, explained in Section 3.3, were applied to each graph. As a result, unlike [33, 34] that only combined followers and followees for prediction, in this study preferences of active content consumers predicted using follower, followee, and follower/followee networks separately. Thus, it can be shown whether a followee graph alone is a reliable source to predict personal preferences of active content consumers.

3.3 Predicting opinions of active content consumers

In this research, preferences of TUs were predicted based on a method called "plurality voting system". By applying this method, a TU was assigned to a class that had the highest number of entities. As an illustration, if the plurality voting system was applied to the follower graph of the TU "A" who had 100 followers interested in politics, and 25 of them were "Liberal", 40 were "Conservative", and 35 were "New Democratic", the predicted label of the TU was "Conservative". This method is more formally defined as follows:

$$UniqueClasses(FollowersLabels) = \{C_1, C_2, C_3, ..., C_n\} \tag{3.1}$$

Where C_i was a list representing followers that belonged to class i ($i \in [1, n]$). Also, UniqueClasses grouped followers that belonged to the same class. The label of a TU was predicted as follows:

$$Class(TU_X) = \max\{C_1, C_2, C_3, ..., C_n\}. \tag{3.2}$$

In the following example, the label of the TU_X was predicted using the method described.

$$FollowersClasses = \begin{bmatrix} F_1 & C_2 \\ F_2 & C_3 \\ F_3 & C_1 \\ F_4 & C_1 \\ F_5 & C_2 \\ F_6 & C_3 \\ F_7 & C_3 \\ F_8 & C_1 \\ F_9 & C_2 \\ F_{10} & C_2 \end{bmatrix}$$

Where the FollowersClasses was an array in which the first column represented all followers of the TU_X, who were interested in politics, and the second column represented an assigned class to each follower. From the FollowersClasses array, unique classes that were assigned to followers were extracted as the name of the lists, as shown below, and users who were assigned to these classes were elements of these lists.

$$C_1 = \{F_3, F_4, F_8\}$$
$$C_2 = \{F_1, F_5, F_9, F_{10}\}$$
$$C_3 = \{F_2, F_6, F_7\}$$

After determining the classes and their entities, the final class of the TU_X was the name of the list that had the maximum number of entities. In the

given example, label C_2 was assigned to TU_X, because the list named C_2 had the maximum number of elements:

$$\text{Class(TargetUserX)} = C_2$$

Also, it was possible that, in some cases, more than one class had the maximum number of elements. For instance, in the following array, both of the classes C_1 and C_3 have the maximum number of users:

$$FollowersClasses = \begin{bmatrix} F_1 & C_2 \\ F_2 & C_3 \\ F_3 & C_1 \\ F_4 & C_1 \\ F_5 & C_2 \\ F_6 & C_3 \\ F_7 & C_3 \\ F_8 & C_1 \end{bmatrix}$$

$$C_1 = \{F_3, F_4, F_8\}$$
$$C_2 = \{F_1, F_5\}$$
$$C_3 = \{F_2, F_6, F_7\}$$

In these situations, predicted labels of the TU_X would be combinations of both classes. So, in the example given, the label of the TU_X would be "C_1/C_3".

3.3.1 Weighted Plurality Voting System

As explained in Section 3.3, the predicted label of the TU_X was the class assigned to the maximum number of his/her followers. By applying the plurality voting system on the follower graph, all of the followers were considered to be equally important. Consequently, each user in each class was considered as one vote for that class. However, in the weighted plurality voting system, the assumption was that users with varying numbers of followers had different levels of importance. In other words, the number of followers defined the degree of popularity which led users to have different weights on their votes.

The weighted plurality voting system was similar to the plurality voting system, with slightly different definition of Equation 3.1. In the weighted plurality voting system, C_i in Equation 3.1 was a list of followers' weights that belonged to class i ($i \in [1, n]$). So, elements of list C_i was something like $C_i = \{W_1, W_2, ..., W_m\}$ instead of $C_i = \{F_1, F_2, ..., F_q\}$, where W_j indicated weight of user "j" that belonged to class j ($j \in [1, m]$) and F_x indicated user "x" that belonged to the class x ($x \in [1, q]$).

In this method, appropriate weights were assigned to users by considering their popularity level determined by the number of their followers. The reason that number of followers was a more important factor than the followees in determining popularity level was that in social media a user can be selec-

tive about whom he/she follows. However, users have no control over their followers in terms of their interests, numbers, etc. Consequently, level of popularity of each user was correlated with the number of followers he/she had. Furthermore, the task was to associate users with similar levels of popularity in the same groups. An empirical study was needed to understand which range of followers represented the same level of popularity. After determining the number of groups required to distinguish between popularity levels and assigning users with similar numbers of followers to the same groups, some studies were done to understand if popular or ordinary users had a more significant role in the prediction task. However, in all the experiments the range of the weights were between 0 and 1. Some of the studies that were done to determine how users with different levels of popularity affect the accuracy of prediction are as follows:

1. Assigning the highest weights to the highest degree of popularity and the lowest weights to the lowest degree of popularity;

2. Assigning the highest weight to the lowest degree of popularity and the lowest weight to the highest degree of popularity;

3. Ignoring users with high consistently popularity to find the effect of eliminating popular users on the prediction task.

In each study, preferences of active content consumers were predicted by applying one of the above strategies to their follow networks. Finally, the

obtained accuracies were compared to each other to determine which strategy had a better effect on the performance of the prediction method.

3.3.2 Analysing the final results

One of the most pronounced challenges of classifying users who talked about Canadian politics was due to existence of the Liberal party of Canada. Existence of the Liberal party made classification difficult mainly because people who supported the Liberal party supported some fiscal plans of the Conservative party and some social plans of the New Democratic party as well. In other words, the Liberal party values were closely aligned with the Conservative and New Democratic values. Consequently, assigning users to the Liberal class was not a straightforward task for users who supported both values and leaders of the Liberal party and those of another party from either the left or right wing at the same time. In this case, instead of assigning one user to one party, the user was assigned to all applicable parties. So, for instance a user who was clearly against the Conservative party and encouraged his/her friends through his/her timeline to vote for either the Liberal or New Democratic parties, was assigned to a class named "NDP/Liberal". To determine accuracy of the predicted labels of TUs, the predicted labels were compared to the actual labels of TUs in the data set by the following methods.

1. **Predicted label contains actual label:** If the predicted label was

the combination of some parties rather than a specific party, but it consisted of the actual label of the TU, the prediction was considered to be a correct one;

2. **Predicted label is exactly the same as real label:** If the predicted label was the combination of some parties rather than an exact party, the prediction was considered to be an incorrect one. For Instance, if the predicted label of a TU was "NDP/Liberal" (see Section 3.3) but his/her real label was either "NDP" or "Liberal", that prediction was incorrect.

Finally, by using the above methods, the accuracy of the TUs' predicted labels obtained by using their followees, followers, and combinations of followers and followees could be determined. Furthermore, they indicated whether followees were reliable predictors of personal preferences of TUs.

Chapter 4

Experiment Evaluation

4.1 Data sets

This chapter presents the experiments conducted by applying the methodology discussed in Chapter 3. The experience began by crawling public Twitter accounts during the period between June 18, 2015 and July 6, 2015, after the Alberta provincial election on May 5 and before the Canadian federal election on October 19. The procedure started by defining a subject to monitor (Canadian politics), and it was followed by selecting Target Users (TUs), explained in Chapter 3, Section 3.1.1. Furthermore, 141 users were selected as TUs and their screen names, descriptions, tweets, and followees' and followers' IDs were saved in the database. Then, each TU's followers' and followees' profiles were crawled to obtain the same information gathered about TUs, such as screen names and descriptions, and saved in the database.

There are many limits applied to Twitter to protect it from abuse. For example, the number of the calls and changes a developer using Twitter APIs can make in a day has a specific maximum. Consequently, the most pronounced challenge to crawl followers' and followees' profiles was dealing with the Twitter rate limit policy. According to [31], each authenticated user can send 180 search requests every 15 minutes. Therefore, only 180 users' profiles can be crawled every 15 minutes. In this study, the total number of the followers and followees were 194596 and 83110, including and excluding the repetitive IDs respectively. So, the total time that was needed to crawl all those IDs is as follows:

((83110 users/ 180 request limit) * 15 minutes) / 60 = 115 hours

Not considering the time that was needed to crawl every user's description, tweets, retweets, etc. at least 115 hours were needed to crawl 83110 users. Also, unpredicted errors, such as changing accounts privacy, and a power outage, could increase the total time.

4.1.1 Reference data set

By applying the method described in Section 3.1.2.1, the 141 TUs were assigned to one of the "Liberal", "NDP", or "Conservative" classes. The goal was to choose an equal number of users in each class to prevent the data sets being biased. Thus, the results obtained by applying the proposed method would be more generalised. Table 4.1 shows an even distribution of TUs in

Conservative, Liberal, and NDP classes. Also, the TUs' object IDs in the data set, number of the followers and followees, and labels can be found in Appendix A, Table A.1.

Table 4.1: Distribution of TUs' labels.

Class	**Percentage**
Conservative	34.04%
Liberal	31.21%
NDP	34.75%

Then, by applying the method introduced in Chapter 3, Section 3.1.2.2, each TU's followers and followees were assigned to proper classes. As a result, 29050 users were classified and 54060 were recognized as noise. Among the 29050 classified users in the dataset, 8730 of them reflected their political views either in their screen names or descriptions. Also, 20320 users reflected their political opinions in their tweets and retweets. Distribution of the labels in this data set can be found in Table 4.2.

Table 4.2: Distribution of labels in "reference data set".

Class	**Percentage**
Conservative	31.78%
NDP/PC	0.03%
NDP	14.25%
Liberal	12.27%
ClassicalLiberal	0.07%
AntiOntarioLiberal	0.06%
AntiNSLiberal (anti Nova scotia liberal)	0.03%
LiberalProgressive	0.13%
Green	0.77%
PCPO (PC Ontario)	0.66%
AntiPCPO	0.01%
Pcaa (PC Alberta)	0.14%
antiAlbertaNDP	0.01%
PC	0.61%
PCNS (PC Nova Scotia)	0.02%
PCNB (PC New Brunswick)	0.01%
PCNF (PC NewFoundland)	0.01%
NDP/Liberal	0.64%
Liberal/Conservative	0.29%
Liberal/antiConservative	0.05%
NDP/antiConservative	0.29%
AntiLiberal/antiNDP	0.1%
AntiConservative/antiLiberal	0.05%
AntiConservative/antiNDP	0.02%
AntiC51	0.26%
AntiConservative	12.38%
AntiNDP	0.63%
AntiLiberal	0.91%
AntiObama	0.58%
DemocraticParty	6.11%
AntiDemocraticParty	0.02%
AntiLeftWing	0.02%

Continued on Next Page ...

Class	Percentage
Bccp (BC conservative party)	0.03%
Progressive	0.83%
Libertarian	1.43%
LeftLibertarian	0.01%
AntiLibertarian	0.01%
LeftWing	0.12%
Politics	13.65%
Others	slightly more than 0

4.1.2 The other data sets

According to Chapter 3, Section 3.1.3, since the "reference data set" contained varying labels, the predicted labels of TUs using that data set would not be limited to "Liberal," "NDP," or "Conservative" classes. Therefore, the accuracy of prediction would not be satisfactory. Thus, based on the "reference data set" four other datasets were created named "reduced labels", "politics as noise", "non-supporters to supporters", and "Left or right wing". The description of each data set is as follows:

- **Data set 1 ("reduced labels")**: In this data set, users whose labels could not be mapped to one of the federal "Liberal," "NDP," or "Conservative" parties, were removed. For instance, many of the users, who were labelled in the "reference data set", reflected their political views only about provincial parties of Canada, such as the "Wild Rose" party of Alberta. Although those users showed interest in politics in

general, the name of the party that they supported did not line up with federal parties. Nevertheless, it was seen empirically that supporters of the Wild Rose party supported the federal Conservative party as well. In contrast, individuals who supported the "Progressive Conservative" Party in Newfoundland were in favour of conservative values. However, they did not support Conservative party of Canada federally. Thus, supporters of the Wild Rose party were mapped to Conservative. But, supporters of the Progressive Conservative Party of Newfoundland were mapped to "politics". Additionally, all the other labels beyond Liberal, NDP, or Conservative were mapped to the labels close to those labels. For example, "Progressive" and "Democratic" classes were mapped to "Liberal/NDP" class. The new labels and their distributions in this data set are shown in Table 4.3.

- **Data set 2 ("politics as noise")**: Unlike the "reduced labels" data set, 13.75% of users who talked generally about politics were ignored in this dataset. The distribution of labels in this dataset are shown in Table 4.4;

- **Data set 3 ("non-supporters to supporters")**: In this data set, "anti-Conservative", "anti-NDP", "anti-Conservative/anti-Liberal", "anti-Liberal/anti-NDP", and "anti-Conservative/anti-NDP" labels were mapped to "Liberal/NDP", "Liberal/Conservative", "NDP", "Conservative",

Table 4.3: Distribution of the labels in "reduced label" data set.

Class	**Percentage**
Conservative	35.56%
NDP	14.63%
Liberal	12.41%
Anti-Conservative	12.75%
Anti-Liberal	1.01%
Anti-NDP	0.65%
NDP/Liberal	8.73%
Politics	13.75%
Anti-Liberal/Anti-NDP	0.15%
Liberal/Conservative	0.29%
Anti-Conservative/anti-Liberal	0.05%
Anti-Conservative/anti-NDP	0.02%

Table 4.4: Distribution of the labels in "politics as noise" data set.

Class	**Percentage**
Conservative	41.23%
NDP	16.96%
Liberal	14.39%
Anti-Conservative	14.79%
Anti-Liberal	1.17%
Anti-NDP	0.75%
NDP/Liberal	10.12%
Anti-Liberal/Anti-NDP	0.17%
Liberal/Conservative	0.34%
Anti-Conservative/anti-Liberal	0.06%
Anti-Conservative/anti-NDP	0.02%

and "Liberal" respectively. Nevertheless, the "anti-Liberal" label was not mapped to "NDP/Conservative". The rationale was that the New

Democratic party was on the left of the political spectrum, while the Conservative party was on the right of the political spectrum and the Liberal party was in the centre. Also, a user in the centre could agree with some opinions belong in both the left and right wing parties. So it was reasonable to reclassify someone who was "anti-NDP" to either the Liberal or Conservative classes. But, reclassifying an "anti-Liberal" user to either NDP or Conservative would lead to wrong predictions in most of the cases and decrease the precision of predictions. The distribution of labels in this dataset can be found in Table 4.5;

Table 4.5: Distribution of the labels in "non-supporters to supporters" data set.

Class	**Percentage**
Conservative	41.41%
NDP	17.03%
Liberal	14.41%
Anti-Liberal	1.17%
NDP/Liberal	24.9%
Liberal/Conservative	1.09%

- **Data set 4 ("left or right wing")**: In this data set, 65.96% of TUs labelled NDP and Liberal were mapped to the "left" and 34.04% labelled Conservative were mapped to the "right' class. The distribution of labels in this dataset can be found in Table 4.6.

Table 4.6: Distribution of the labels in "left or right" data set.

Class	Percentage
Right	43.66%
Left	56.34%

4.2 Description of the graphs

To get a better understanding of the follow networks of TUs, three graphs were created by using Gephi [5] representing followers, followees, and the combination of followers and followees. Although these graphs were created for each data set, only the graphs created based on the "non-supporters to supporters" data set were visualized. These generated graphs can be found in Figures 4.1, 4.2, and 4.3.

In the presented graphs, each group of labels was identified by a unique color; orange, red, blue, green, yellow, and purple nodes represent NDP, Liberal, Conservative, NDP/Liberal, Liberal/Conservative, and anti-Liberal respectively. Also, the gray nodes demonstrate TUs whose labels are ignored. Additionally, the sizes of the nodes were determined by the degree of each node. However, as TUs had more quantity and degree in the graphs, in the small figures only Liberal, NDP, and Conservative labels are recognizable. To make the graphs more visual, nodes with degree 1 were pruned from the graph. In other words, users who followed or were followed by only one user were removed from the graphs. The number of the edges and nodes in each

graph can be found in Table 4.7.

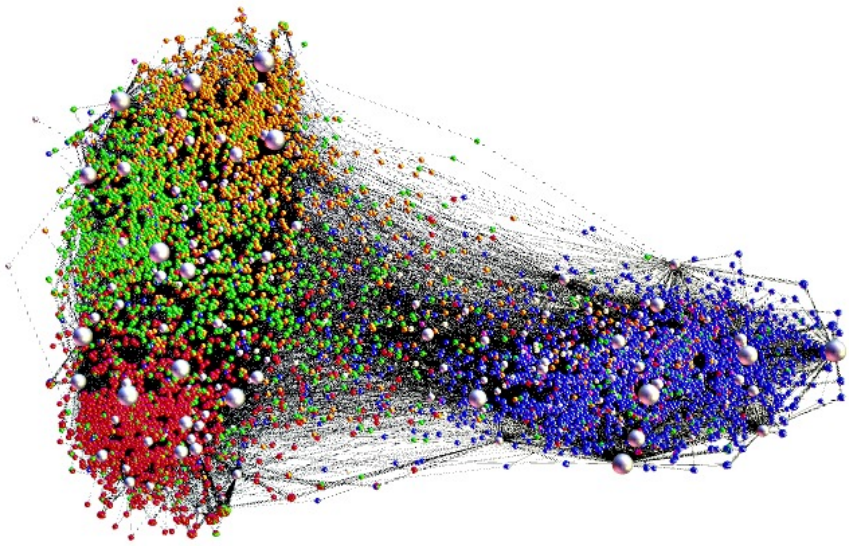

Figure 4.1: Followee graph: orange, red, blue, green, yellow, and purple nodes represent NDP, Liberal, Conservative, NDP/Liberal, Liberal/Conservative, and anti-Liberal respectively. Also, the gray nodes demonstrate TUs whose labels are ignored.

Table 4.7: Number of nodes and edges in each graph.

Graphs	**degree >= 1**		**degree >= 2**	
	Nodes	**Edges**	**Nodes**	**Edges**
Followee	17853	49335	7750	39232
Follower	16686	37994	6414	27722
Followee/Follower	24894	86654	14582	79205

It can be seen in Figures 4.1, 4.2, and 4.3, the density of the links is noticeably high between the nodes labelled NDP and Liberal. It shows that users

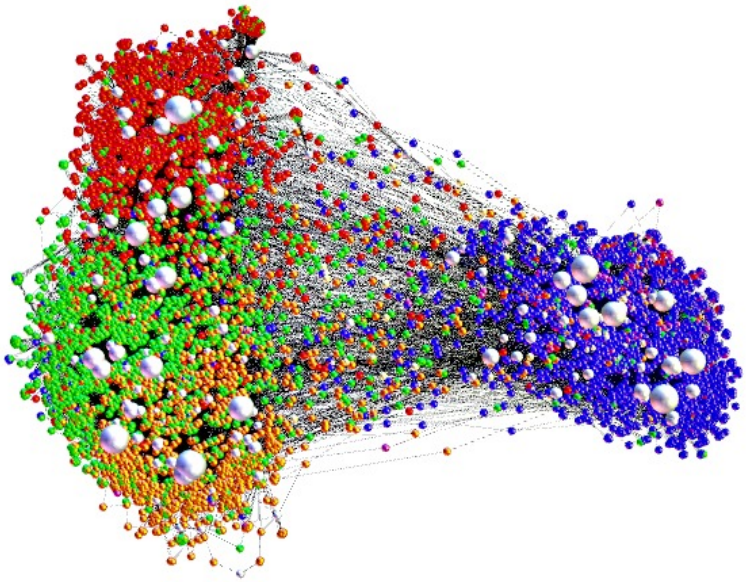

Figure 4.2: Follower graph: orange, red, blue, green, yellow, and purple nodes represent NDP, Liberal, Conservative, NDP/Liberal, Liberal/Conservative, and anti-Liberal respectively. Also, the gray nodes demonstrate TUs whose labels are ignored.

from the left and centre of the political spectrum have a greater tendency to connect to each other. However, the density of the links among the nodes labelled Conservative is higher than between blue nodes and others. This indicates that Conservatives prefer to connect to Conservatives rather than making friends with supporters of Liberal or NDP parties. Also, Figure 4.4 makes it clear that the nodes labelled NDP and Liberals are so tightly knitted that the boundary between them is not recognizable. Furthermore, these figures make the difficulty of assigning users to either the Liberal or NDP classes more obvious.

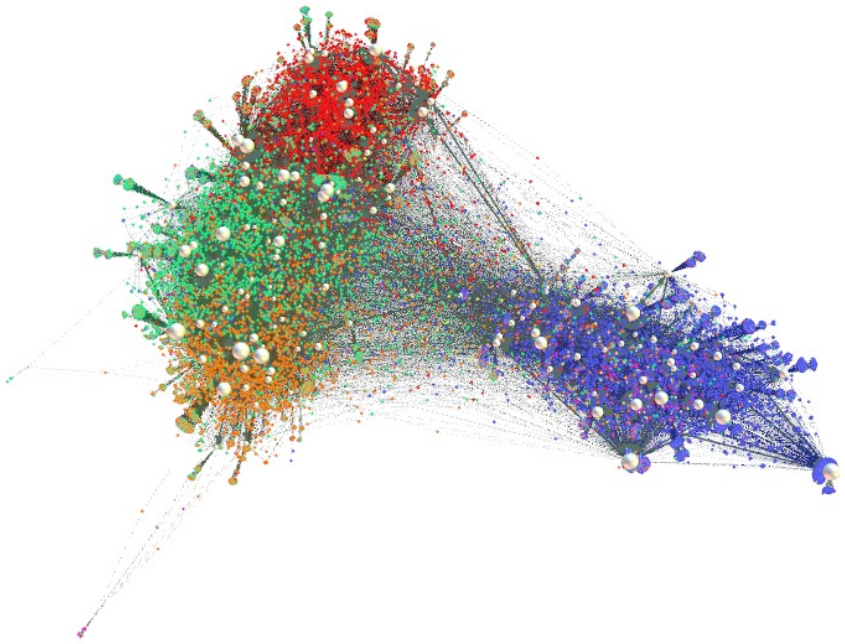

Figure 4.3: Combination of followers and followees graph: orange, red, blue, green, yellow, and purple nodes represent NDP, Liberal, Conservative, NDP/Liberal, Liberal/Conservative, and anti-Liberal respectively. Also, the gray nodes demonstrate TUs whose labels are ignored.

4.3 Results from the plurality voting system method

The first method to predict political views of TUs was the plurality voting system. This method was applied to each data set for followees, followers, and followees/followers graphs separately. To determine the accuracy of the predicted labels, according to Chapter 3, Section 3.3.2, two approaches were taken. In the first approach, a predicted label was considered to be a correct

Figure 4.4: Tightness of NDP and Liberal classes.

one only if it was exactly the same as the label in the data set. But in the second approach a prediction was correct if the predicted label contained the actual label. The accuracy of the results obtained from both approaches is shown in Tables 4.8 and 4.9.

Table 4.8: Accuracy of prediction by applying plurality voting system method, when predicted label is exactly the same as the actual label.

Predicted label **equals** actual label			
	Followee graph	Follower graph	Followee/Follower graph
Data set 1	74.46%	79.43%	76.59%
Data set 2	80.85%	80.14%	80.14%
Data set 3	**75.88**%	**73.04**%	**76.59**%
Data set 4	97.16%	98.58%	99.29%

As shown in Table 4.8, the lowest accuracies were produced from data set

Table 4.9: Accuracy of prediction by applying plurality voting system method, when predicted label contains the actual label.

Predicted label **contains** actual label			
	Followee graph	Follower graph	Followee/Follower graph
Data set 1	76.59%	81.56%	78.72%
Data set 2	82.97%	82.97%	82.26%
Data set 3	**98.58**%	**98.58**%	**98.58**%

1. The labels of the TUs were limited to "NDP", "Liberal", and "Conservative"; however data set 1 contained the "politics" label as well. Therefore, if the majority of a TU's followers and followees were assigned to the "politics" class, the predicted label for that TU would be "politics", which was a wrong prediction compared to the TUs' actual labels. Consequently, considering the "politics" label as noise in data set 2 led to improvements of 6.38% and 1.41% in predictions using followee and follower networks respectively. It is important to note that mainly journalists, news analysts, and news channels, such as "CNN", "BBC", and "CBC", were labelled "politics". So, the difference between improvement of accuracies in followees and followers networks shows that users are more likely to follow news accounts rather than being followed by them. Also, this improvement shows that many users tend to follow a large number of the accounts that spread general political news rather than follow accounts biased in one specific political spectrum.

Similarly, it was expected that mapping the labels assigned to non-supporters

to the labels close to "NDP," "Liberal," and "Conservative" in data set 3 would enhance the accuracy. Nevertheless, accuracy of the predicted labels using followees was improved by 1.41% , but worsened about 6.38% using followers. By comparing Tables 4.4 and 4.5, it can be seen that entities of "NDP/Liberal", "Liberal/Conservative", "Conservative", "NDP", and "Liberal" classes were extended by 14.79% , 0.75% , 0.17% , 0.06% , and 0.02% respectively. Thus, the chance that more TUs were assigned to the "NDP/Liberal" class was increased. So, many TUs were mislabelled thus reducing the accuracy of prediction; a list of all the incorrect predictions in all the data sets can be found in Table 4.10. Thus, as data set 1 led to the lowest accuracy of classification, it was not used in weighted plurality voting system method.

One of the false predicted labels reported for data set 3 in Table 4.10 is "NDP/Liberal/Liberal". However, this false prediction could occur in other data sets as well. The "NDP/Liberal/Liberal" predicted class using followees means a TU followed equal numbers of users labelled Liberal and NDP/Liberal. Furthermore, although the maximum number of followees were explicitly with the Liberal party, the same number of them were interested in both the NDP and Liberal parties. Logically, it can be said that a TU in the "NDP/Liberal/Liberal" class is a supporter of the Liberal party because the maximum number of users who are followed by the TU are Liberal. However, determining accuracy of prediction for all users should be consistent.

So, in approach 1, any predicted labels other than "NDP", "Liberal", and "Conservative" were considered wrong predictions. In contrast, in approach 2 these kinds of predicted labels were considered correct predictions.

Among all the reported accuracies in Table 4.8, applying the plurality vote on data set 4 led to the highest accuracy. As can be seen in Figure 4.4, the boundary between Liberal, NDP, and Liberal/NDP classes is not clear. Thus, limiting the classes to the "left" and "right" prevented the occurrence of many false predictions reported in Table 4.10. Consequently, accuracy of the prediction was increased.

A common factor that decreased the accuracy of all the data sets was predicting multi-labels, such as NDP/Liberal, for TUs. Thus, as can be seen in Table 4.9, applying approach 2 caused an improvement in the accuracy of the results. However, this approach was not applied to data set 4, in which the only existing labels were "left" and "right", because in this data set the predicted label was "left/right" only if someone followed the same number of people in opposite sides of the political spectrum. So, it did not make sense to apply approach 2 to this data set; if a TU's label is either "left" or "right", and the predicted label is "left/right", that prediction is always correct.

As shown in Tables 4.8 and 4.10, prediction accuracies obtained from data set 4, left or right wing data set, was higher than other data sets as only

Table 4.10: Common false predictions among followers, followees, and followers/followees graph in each data set.

	Predicted label	Actual label	Occurrence
Data set 1	anti-Conservative	NDP	often
	anti-Conservative	Liberal	
	politics	NDP	
	politics	Liberal	
	politics	Conservative	
	NDP/Liberal	Liberal	sometimes
	NDP/Liberal	NDP	
	Liberal	NDP	rarely
	anti-Liberal	NDP	
Data set 2	anti-Conservative	NDP	often
	anti-Conservative	Liberal	
	NDP/Liberal	Liberal	sometimes
	NDP/Liberal	NDP	
	Liberal	NDP	rarely
	anti-Liberal	NDP	
Data set 3	NDP/Liberal	NDP	often
	NDP/Liberal	Liberal	
	Liberal	NDP	rarely
	anti-Liberal	NDP	
	NDP/Liberal/Liberal	Liberal	
	NDP/Liberal/NDP	NDP	
Data set 4	right	left	sometimes
	left	right	rarely
	left/right	right	

in few cases mis-prediction between left and right labels occurred. Further investigation was done on data set 4, which had strong results, to show how close the voting for both classes was, the "strength of correctness" defined as follows:

$$SC = \frac{RL - OL}{Total\ Number\ of\ followees'/followers'\ Votes} \tag{4.1}$$

Where SC is strength of correctness for an individual prediction, RL is number of followees'/followers' votes to the real label, and OL is number of followees'/followers' votes to the other label. As an illustration, if we had two TUs with the following information:

TU_1: 100 followees consists of 75 left labels and 25 right labels, real label=left, and predicted label=left

TU_2: 100 followees consists of 55 left labels and 45 right labels, real label=right, and predicted label=left

The strength of correctness for the first and second predictions would be: $SC = \frac{75-25}{100}$ and $SC = \frac{45-55}{100}$

To examine the strength of the voting for each predicted class, the average/mean and standard deviation of the individual strengths were calculated (see Table 4.11).

Table 4.11: The mean and standard deviation for the strength of correctness of the predictions for each of the left and right classes of data set 4.

Data set 4	class = Right		class = Left	
	Mean	SD	Mean	SD
Followee	0.57	0.26	0.88	0.15
Follower	0.64	0.25	0.90	0.08
Followee/Follower	0.57	0.26	0.90	0.07

By comparing the standard deviations calculated for the left and right classes in Table 4.11), it can be seen that left class were predicted stronger than right class. The difference between standard deviations of the two classes can show that the left class was more distributed than the right class in this data set. In other words, a large number of users' followees/followers consisted of left labels even if the predicted label was "right". Consequently, strength of correctness for an individual prediction was decreased which caused a higher standard deviation for the strength of correctness of the prediction for the right class.

4.4 Results from weighted plurality voting system method

According to Chapter3, Section 3.3.1, unlike the plurality voting system, in this method the value of votes for each user was not the same. In this experiment, the decisive factor for weight of each vote was the level of popularity of users. The level of popularity for each user was determined by considering the number of users' followers. For example, a user who did not have any followers was categorized as an extremely ordinary user, and a user with 200000 followers was categorized as an extremely popular user. Weights of users based on the level of their popularity was determined by two strategies. The first one was to assign the highest weight to the users with the highest level of popularity and the lowest weight to the users with the lowest degree

of popularity. The second was to reverse the assignment of weights. Therefore, it could be decided whether popular or common types of users had more effect on political opinions of an individual. The method of finding different levels of popularity as well as the accuracies obtained by assigning different weights to the plurality voting system can be found in the following sections.

4.4.1 Finding level of popularity and appropriate weights

As discussed earlier in this thesis, distribution of number of followers determined levels of popularity. Therefore, several bins were created to put users with the same number of followers in the same groups. Thus, number of followers determined intervals of the bins. For instance, if a bin interval was $[0, 500)$ it meant users who did not have any followers, or less than 500 followers could be categorized into this bin. Figure 4.5 shows distribution of the number of followers in different bins, where the number of the bins and their width were determined empirically. As can be seen in the figure, width of the bins for number of followers between 0 and 3500 is 500. However, this width gets larger for more than 3500 followers. It means when users have fewer followers, even 500 more of them can be important in the level of popularity. Nevertheless, when users have more than for example 10000 followers, only every 10000 followers can predict behaviour of different users.

By considering the number of the entities in each bin and the shape of Fig-

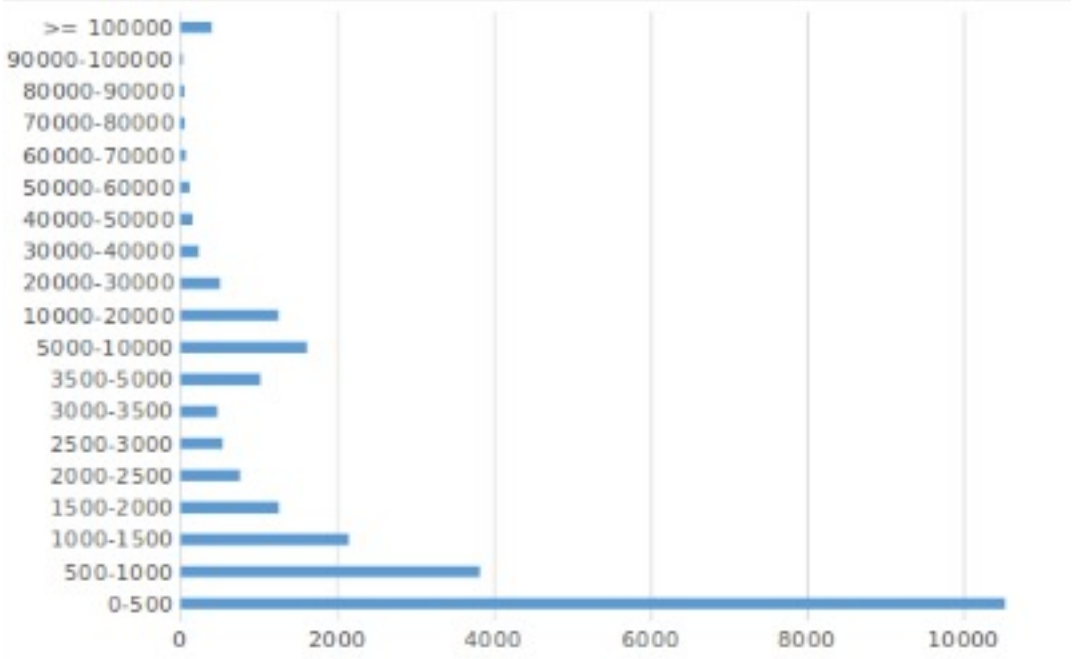

Figure 4.5: This figure represents distribution of the number of followers in different bin. The vertical axis shows the minimum and maximum number of followers in each bin. Also, the horizontal axis shows number of users in each bin.

ure 4.5, it can be seen empirically that 407 of the users who had more than 100000 followers and did not have the "politics" label were mainly politicians, such as "Tom Mulcair", "David Cameron", "Bill Clinton", and "Elizabeth May". Also, most of the users in this category were verified accounts by Twitter. Moreover, the 335 users who had more than 50000 and less than 100000 followers, were political leaders, such as "Jack Layton" or official accounts of specific parties such as "NDP_HQ". Also, many users who reflected their political opinions in their descriptions were categorized in this

group. This group of users had fewer verified accounts in comparison with the previous group. Thus, users who had more than 50000 followers were called extremely popular. However, users with more than 100000 followers were called extremely popular users of level 2, and users with $[50000, 100000)$ follower were called extremely popular users of level 1.

The other accounts that had fewer than 50000 followers did not necessarily belong to celebrities or famous people. For example, 909 of the users who had more than 20000 and less than 50000 followers sent promotional messages, rather than using Twitter for socializing with friends. For instance, they spread some opinions about politics, a technology, etc., or they were strongly against some idea and created an account to reflect their dislike. In other words, people in this category were interested in the news dissemination aspect of Twitter. Similarly, 2875 users who had between 5000 and 20000 followers were not celebrities or famous political figures. However, they were biased in a specific direction. For example, many of them were MPs (Member of Parliament) or leaders of a specific party who made official accounts to spread information regarding their supporting party. Besides, many of the users in this category had lots of followers either because their accounts had existed for a long time, or they shared interesting tweets. Therefore, users with $[5000, 50000)$ followers were called popular users. Users with $[5000, 20000)$ and $[20000, 50000)$ followers had popularity of level 1 and level 2 respectively.

The rest of the unclassified users were common types of users who had a different range of followers. As illustrated in Figure 4.5, numbers of followers in the ranges $[0, 500)$ and $[500, 1000)$ were considered extremely ordinary users of level 1 and 2 respectively. Also, the ranges $[1000, 2000)$ and $[2000, 5000)$ represented ordinary users of level 1 and 2 respectively.

By determining the level of popularity of users, the influence of users with different levels of popularity on prediction was decided by assigning a weight in the range $[0, 1]$ to each group. In the first experiment, higher weights were assigned to higher level of popularity (see Table 4.12), and in the second experiment, lower weights were assigned to higher levels of popularity (see Table 4.15). Accuracy of classifying TUs by applying weights in the first experiment are reported in Tables 4.13 and 4.14. Also, accuracies of predictions in the second experiment are reported in Tables 4.16 and 4.17.

The accuracies reported in Tables 4.16 and 4.17 show that assigning the highest weights to users with the lowest level of popularity led to more accurate results for classification. An explanation of this enhanced accuracy can be found in Section 4.4.2.

To find out whether decreasing the width of each class or determining equal intervals for bins would improve the accuracy of classification, some further

Table 4.12: Different levels of popularity and weights assigned to each group.

Level of popularity		**Boundaries of each class**	**Weight**
Extremely Ordinary	L_1	0 <= n < 500	weight=0.10
	L_2	500 <= n < 1000	weight=0.24
Ordinary	L_1	1000 <= n < 2000	weight=0.36
	L_2	2000 <= n < 5000	weight=0.48
Popular	L_1	5000 <= n < 20000	weight=0.60
	L_2	20000 <= n < 50000	weight=0.72
Extremely popular	L_1	50000 <= n < 100000	weight=0.85
	L_2	100000 <= n	weight=1.00

Table 4.13: Accuracy of classifying TUs when predicted label is equal to the actual label. In this table, accuracies are obtained by assigning the highest weight to extremely popular users and the lowest weight to extremely ordinary users.

Predicted label **equals** actual label			
	Followee graph	Follower graph	Followee/Follower graph
Data set 2	80.85%	77.30%	77.30%
Data set 3	**70.21**%	**72.34**%	**70.21**%
Data set 4	98.58%	99.29%	99.29%

studies were done, modifying the values shown on Figure 4.5. For instance, in one study the intervals of each bin for users who had followers between 0 and 5000 were determined to be 500, and the intervals of the bins for users who had more than 5000 followers were the same as in Table 4.12. In another study, the interval between bins for users having 0 to 5000 followers was 500, and the interval between bins for users who had between 5000 and 10000 was 5000. Also, the interval between bins for users who had between 10000 and

Table 4.14: Accuracy of classifying TUs, when predicted label contains the actual label. In this table accuracies are obtained by assigning the highest weight to extremely popular users and the lowest weight to extremely ordinary users.

Predicted label **contains** actual label			
	Followee graph	Follower graph	Followee/Follower graph
Data set 2	83.68%	81.56%	80.85%
Data set 3	**99.29**%	**98.58**%	**98.58**%

Table 4.15: The highest weight is assigned to users with the lowest degree of popularity and the lowest weight is assigned to the users with the highest degree of popularity.

Level of popularity		**Boundaries of each class**	**Weight**
Extremely Ordinary	L_1	0 <= n < 500	weight=1.00
	L_2	500 <= n < 1000	weight=0.85
Ordinary	L_1	1000 <= n < 2000	weight=0.72
	L_2	2000 <= n < 5000	weight=0.69
Popular	L_1	5000 <= n < 20000	weight=0.48
	L_2	20000 <= n < 50000	weight=0.36
Extremely popular	L_1	50000 <= n < 100000	weight=0.24
	L_2	100000 <= n	weight=0.10

50000 followers was 10000. the intervals of the bins for users who had more than 50000 followers were the same as in Table 4.12. Additionally, in a new study, intervals of the both previous studies were tested. However, none of the experiments led to a better prediction accuracy in comparison with the accuracy reported in Table 4.16.

Table 4.16: Accuracy of classifying TUs by when predicted label is equal to the actual label. In this table accuracies are obtained by assigning the lowest weight to extremely popular users and the highest weight to extremely ordinary users.

Predicted label **equals** actual label			
	Followee graph	Follower graph	Followee/Follower graph
Data set 2	82.26%	78.72%	80.85%
Data set 3	77.30%	73.75%	77.30%
Data set 4	**98.58**%	**97.163**%	**98.58**%

Table 4.17: Accuracy of classifying TUs by when predicted label contains the actual label. In this table accuracies are obtained by assigning the lowest weight to extremely popular users and the highest weight to extremely ordinary users.

Predicted label **contains** actual label			
	Followee graph	Follower graph	Followee/Follower graph
Data set 2	83.68%	80.85%	82.97%
Data set 3	**98.58**%	**97.87**%	**98.58**%

4.4.2 Influence of removing different levels of popularity in classification

As shown in the previous section, decreasing weight of users with higher degree of popularity enhanced the accuracy of classification. This improvement was the motivation for a new question: would omitting users with different levels of popularity affect accuracy of prediction? To answer this question, each time the experiment was conducted, an additional group of users with a certain level of popularity was ignored. Tables 4.18 and 4.19 demonstrate

the effect of excluding each group of users with a different level of popularity on accuracy of classification. Abbreviations used in the "ignored level" column in those tables can be found in the list of abbreviations on page xiv. It is important to note that in every new experiment after the first experiment, not only the specified group of users, but also the previous group were eliminated. As an illustration, the first experiment was done by excluding extremely popular users of level 1 and 2 from the classification. In the second experiment, not only extremely popular users, but also popular users of level 2 were ignored and so on. Finally, only extremely ordinary users of level 1 were kept for classification.

Based on the accuracies reported in Tables 4.16, 4.18, and 4.19 it is clear that assigning the lowest weight to the highest degree of popularity and even ignoring users who had > 20000 followers either did not change the accuracy or improved it slightly in some cases. One reason is because extremely popular and popular users of level 2 only represented a small portion of the data set. Hence, if a user followed even 10 users who were supporters of the same party with more than 100000 followers, their votes were worth: $10 * 0.1 = 1$ vote. Thus, 10 extremely popular users' votes were as valuable as 1 extremely ordinary user vote. However, ignoring users with the number of followers $[0 - 20000)$, and more specifically popular users of level 1, ordinary users of both levels, and extremely ordinary users of level 2, had different influence on accuracy of classification in Tables 4.18 and 4.19.

Table 4.18: Accuracy of classifying TUs when the predicted label is equal to the actual label. In this table accuracies are obtained by ignoring users with different degrees of popularity from the experiment. Ext-Po-2 represents extremely popular users of level 2 and Ext-Or-1 represents extremely ordinary users of level 1. Accuracies which are better than the ones reported in table 4.16 are shown in bold.

Predicted label **equals** actual label				
Ignored group	**Data sets**	**Followee**	**Follower**	**Followee/Follower**
Ext-Po	Data set 2	82.26%	78.72%	**81.56**%
	Data set 3	**78.01**%	**74.46**%	**78.01**%
	Data set 4	98.58%	97.16%	98.58%
Po-2	Data set 2	82.26%	78.72%	**81.56**%
	Data set 3	**78.01**%	**74.46**%	**78.01**%
	Data set 4	98.58%	97.16%	98.58%
Po	Data set 2	81.42%	**79.43**%	80.14%
	Data set 3	**78.01**%	**76.59**%	**78.01**%
	Data set 4	98.58%	97.16%	98.58%
Or-2	Data set 2	81.56%	**80.14**%	**78.72**%
	Data set 3	**79.43**%	**75.88**%	77.30%
	Data set 4	98.58%	97.16%	98.58%
Or	Data set 2	82.26%	**79.43**%	78.72%
	Data set 3	**80.14**%	**78.72**%	**78.01**%
	Data set 4	98.58%	97.16%	98.58%
Ext-Or-2	Data set 2	81.56%	**79.43**%	79.43%
	Data set 3	**80.14**%	**78.01**%	**78.72**%
	Data set 4	98.58%	97.16%	98.58%

By comparing Tables 4.16 and 4.18, it can be seen that the accuracy of classification was enhanced for all data sets overall, except for data set 4. In data set 4, the accuracy of classification in all the cases was the same.

Table 4.19: Accuracy of classifying TUs when predicted label contains the actual label. In this table accuracies are obtained by ignoring users with different degrees of popularity from the experiment. Ext-Po-2 represents extremely popular users of level 2 and Ext-Or-1 represents extremely ordinary users of level 1. Accuracies which are better than the ones reported in Table 4.17 are shown in bold.

Predicted label **contains** actual label				
Ignored group	**Data sets**	**Followee**	**Follower**	**Followee/Follower**
Ext-Po	Data set 2	83.68%	80.85%	82.97%
	Data set 3	98.58%	97.87%	98.58%
Po-2	Data set 2	83.68%	80.85%	82.97%
	Data set 3	98.58%	97.87%	98.58%
Po	Data set 2	82.26%	80.85%	81.56%
	Data set 3	97.87%	97.87%	97.87%
Or-2	Data set 2	82.26%	**81.56**%	79.43%
	Data set 3	**99.29**%	97.87%	97.8723%
Or	Data set 2	82.97%	80.85%	79.43%
	Data set 3	**99.29**%	97.87%	98.58%
Ext-Or-2	Data set 2	82.26%	80.85%	80.14%
	Data set 3	**99.29**%	**98.58**%	98.58%

However, when users with $[0 - 20000)$ followers were ignored, only in a few cases did accuracy of classification using followee graphs drop slightly. This comparison is shown graphically in Figure 4.6. Unlike Table 4.18, it can be seen that in Table 4.19 the overall accuracy was decreased in comparison with Table 4.17. According to the results of Tables 4.18 and 4.19, there is no doubt that users with a range of $[0 - 20000)$ followers had more effect on accuracy in comparison with extremely popular users and popular users of level 2. This is because users who had followers in this range had more

quantity and weight, which made their votes more valuable.

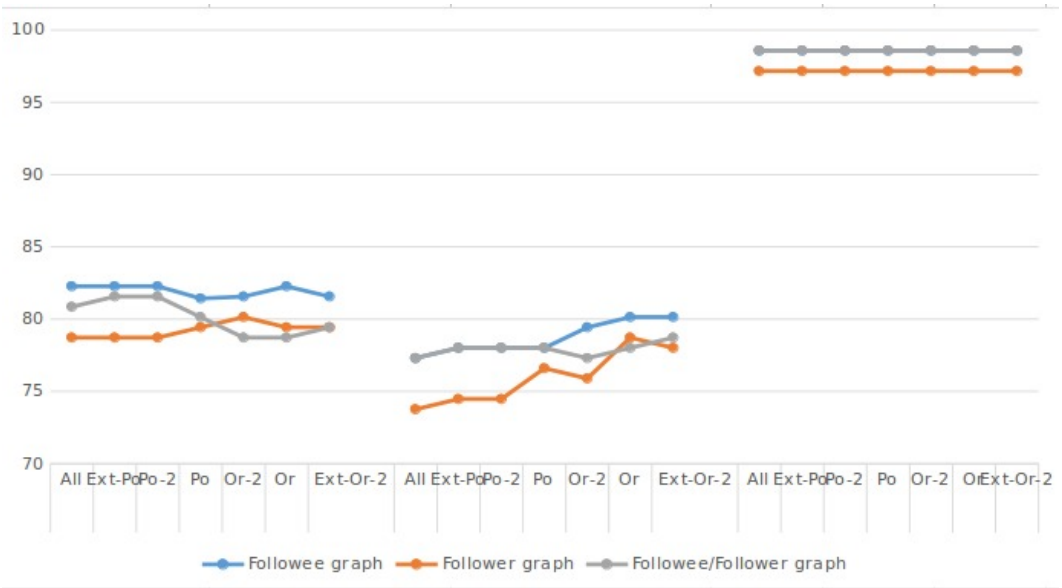

Figure 4.6: In this figure, accuracies obtained by ignoring users with different levels of popularity are compared. Blue, red, and gray lines represent followee, follower, and followee/followers graphs respectively. Also, the first group on the left shows accuracies of data set 2, the second group in centre shows accuracies of data set 3, and the last group shows accuracies of data set 4. The very first point in each group represents accuracy when all the users are included in classification.

Improved accuracy of classification, by ignoring users who had $[0, 20000)$ followers in data set 2 and decreased accuracy for the same data set in Table 4.18 can be argued by the existence of multi-party labels. For example, in data set 2 it seems a lot of users were labelled "NDP/Liberal" and were considered as noise when the predicted label had to be equal to the actual label. Consequently, ignoring those users increased the accuracy (see Ta-

ble 4.18). However, when the predicted label has to contain the actual label, those multi-labels are helpful in correct classification. Thus, ignoring them will decrease the accuracy. Moreover, improvement of accuracy by removing ordinary users and extremely ordinary users of level 2 in both tables for data set 3 can be explained by the existence of non-supporters. In Table 4.18 the explanation of the improved accuracy is the same as for the improvement of accuracy in data set 2. However, this improvement in Table 4.19 can indicate that labels of many users who have $[500 - 5000)$ followers are labelled incorrectly, or they are mapped incorrectly from non-supporters to supporters. Thus, removing them led us to the highly accurate classification of 99.29% .

By comparing all the reported results, it can be concluded that only including the 10530 users who had less than 500 followers led to satisfactory accuracies. As in Figure 4.5, a large portion of users are classified in the $[0 - 500)$ class. Also, the weight of a vote of each user in that category is worth 1, which is the maximum weight that a vote can have in this experiment. Thus, it is not surprising that having only those users can lead to satisfactory accuracies.

Knowing that only users who have less than 500 followers can predict the political views of AC consumers is highly beneficial. Because, as discussed in Chapter 2, Section 2.1, the nature of data in Twitter makes the data preprocessing step and classification relatively time consuming and expensive. Regardless of the difficulty of preparing data sets, many prediction tasks,

such as predicting an election outcome, has to be done in a short amount of time and continuously. Prediction should be done continuously because the trend of changing data in social media is so fast, and as the election time gets closer, more debates and advertisements are developed that may change the possible outcome. Hence, when a huge amount of data is available for the prediction task, and quickness and accuracy of prediction are important, large groups of annotators should get involved in the pre-processing and classification processes, whether they are done manually or automatically. Human annotators are needed even in automatic classifiers, because correctness of classification have to be tested whether entities are labelled manually or automatically. Furthermore, having a small set of users likely to provide good predictions decreases the needed time and cost noticeably.

4.5 Discussion

As seen in previous sections, applying the weighted plurality vote system using different weights had varying effects on accuracy of predictions in data sets 2 and 3. However, accuracy of data set 4 did not change noticeably during all the weight analyses. One possible reason that accuracy of data set 4 did not change noticeably is that it only contained "left" and "right" labels. By not having varying labels, the number of the possible errors in prediction are limited. Also, analyses of weights make more sense when changing the

weights causes users to be classified in more than two classes. Moreover, it was seen in all the accuracies reported in this chapter that if the predicted label contains the actual label instead of being exactly the same, the accuracy of prediction improved. However, higher degree of accuracy in this case is not a guarantee of precision. In other words, a multi-label class, such as "NDP/Liberal", does not predict precisely if in an upcoming election a user assigned to that class will vote Liberal or New Democratic Party. Furthermore, the predicted label containing the actual label should be considered as a true classification only if the political wing of a user is the goal and not the specific party that users support. Similarly, it can be argued that the accuracies obtained from data set 3 are not precise, because if users should vote for one the three parties, and the only information which is available for users is that they are against a specific party, this information does not indicate that they are supporters of the two other parties. It is possible that they do not support any of the other parties in the election.

In this chapter, political parties of AC consumers were predicted by different accuracies using their followers, followees, and followers/followees networks. Comparison of the accuracies obtained from each network using data sets 2 and 3 can be found in Figure 4.7. As already discussed, if a predicted label contains the actual label, the accuracy would be high but precision would be low. Consequently, accuracies in Figure 4.7 are obtained when the predicted labels are equal to the actual labels.

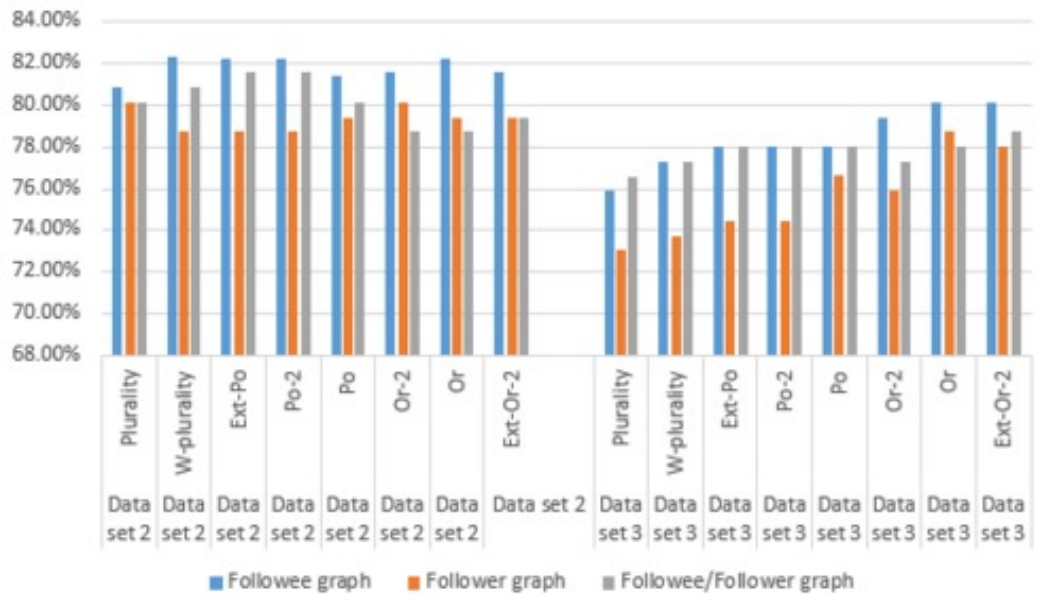

Figure 4.7: Comparing accuracy of prediction obtained by using followers, followees, and followers/followees networks, when predicted label has to be equal to the actual label.

As can be seen in Figure 4.7, accuracies obtained using followee networks to which the plurality and weighted plurality vote systems are applied, and ignoring different levels of popularity are higher than those obtained from the other two networks generally. However, in data set 3 when the plurality vote method was applied, the combination of followers and followees led to a better accuracy by 0.71% . Nevertheless, it should be kept in mind that none of these accuracies are obtained by predicting political tastes of real AC consumers. This is because to be able to evaluate the proposed method, TUs were chosen from active users who shared their opinions clearly and regularly

through their Twitter accounts. Therefore, TUs shared enough content to attract other users' attention to be followed by them. In contrast, content consumers rarely express their opinions regarding a certain topic, but tend to follow news by following others. Thus, they usually have more followees without many followers. However, those AC consumers may be followed by accounts created for advertising purposes or people who want to expand the size of their networks without considering similarities. As a result, if TUs were actual AC consumers, the only case in Figure 4.7 in which combination of followees and followers networks led to a better accuracy would not occur. Similarly, accuracy of prediction using follower networks would be relatively decreased. Also, as discussed in Chapter 2, Section 2.2.2.1 users can be selective about their followees, but they do not have any control over their followers. In conclusion, in the case that user-generated content for AC consumers was insufficient, their political preferences can be predicted by considering only their followee networks.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

The goal of this thesis was inferring political preferences of active content consumers (AC consumers), representative of ordinary users on Twitter. Because those AC consumers do not have enough content to reflect their preferences regarding a given topic, based on the homophily principle, preferences of their neighbours can be used for this purpose. However, AC consumers' neighbours consist of more followees than followers, because AC consumers do not provide an adequate amount of content on a certain topic to attract many users with similar tastes. Consequently, they would not have enough followers to make accurate predictions possible. However, those AC consumers tend to follow many users who have common interests to get access to their favourite topics. As a result, they have more followees than followers

in their profiles that can be used for prediction. Thus, the objective was to find whether personal preferences of AC consumers can be inferred by using their followees with satisfactory accuracy.

The experiments were conducted by creating four data sets, in which the population was selected randomly from the crawled Twitter public accounts whose political preferences could be determined, representative of common types of users who had less than 2000 followers. Therefore, the entities of data sets were ordinary active content consumers (TUs) and their followers and followees (follow network). The entities in each data set were labelled by applying different strategies to determine the influence of data characteristics on the performance of the proposed method. In order to assign as many users as possible to political classes, both supporters and non-supporters of political parties were classified. Also, to increase the size of the labelled data, the defined list of the political hashtags was updated continuously by discovering new hashtags that represented a political party. In this case, not only the users who used the known hashtags, but also users who used new hashtags beyond the list could be classified. By having each data set labelled, graphs of followers, followees, and the combination of followers and followees can be formed for prediction. Then, plurality and weighted plurality voting systems were applied to each graph to determine which network predicted personal preferences of AC consumers more accurately.

The experiments revealed that by assigning appropriate weights to users' follow networks, their followees can be used to predict their preferences with 82.26% accuracy. With similar settings, followers and combination of followers and followees predict preferences of AC consumers with 78.72% and 80.85% accuracies respectively. All these results are obtained when labels are used strictly, with non-supporters not mapped to rivals' supporters and the prediction is only correct if the predicted label is equal to the real label. The results show both followees and followers are effective in predicting one's preferences. However, using followees led to more accurate results.

In this experiment, the accuracy of using followees to predict one's preferences is not immensely better than using followers or the combination of followers and followees. However in this study, to be able to evaluate the proposed method, TUs were chosen from active content producers (AC producers) instead of AC consumers. Thus, those AC producers expressed their points of view clearly through their posts and attracted lots of users with similar opinions. In this case, preferences of AC producers' followers were closer to TUs in comparison with those of AC consumers' followers. In both cases, either if AC producers or AC consumers were used as TUs, the accuracy of using followees was better than that of using followers. This result may be due to TUs not having control over their followers, but being able to select their followees themselves.

Rabelo *et al.* [20, 21] tried to infer preferences of users who did not express them explicitly by using their link information. However, because the size of the labelled data in their experiment was small, the authors had to prune their graph heavily to keep only the higher degree nodes. Consequently, the performance of their prediction model was only satisfactory if the unlabelled nodes in the graph were highly connected to the labelled nodes. As a result, the personal preferences of some individuals could not be inferred by this model. Unlike [20, 21], a larger labelled data set was used in the method in this thesis, which helps in predicting every active content consumers' preferences regardless of the strength of their connectivity.

5.2 Future work

This study suggests some interesting directions for future work. For example, to generalize the applicability of the proposed method to the real world, the preferences of real active content consumers on one subject should be predicted using this technique, with the actual preferences of AC consumers being determined through surveys and interviews and compared with the predicted preferences. The intuition is that the achieved results would be similar, but experiments would have to be executed to provide a clearer idea of the applicability of this method to real data.

Another expansion of this work to enhance efficiency would be using an au-

tomatic classifier for users' tweets, since the pre-processing and classification processes are the two most time consuming parts of each data analysis task. The time needed for these processes would grow if the size and complexity of the data increased. In these situations, classifying users manually would make this model inefficient. However, in the research reviewed the classifiers used a constant list of hashtags for classification. But, to expand the size of the labelled data, there is a need for a classifier that can recognize new hashtags that represent the defined subject automatically and update the list continuously. Therefore, even users who did not use the defined hashtags in the list to express their views can be assigned to proper classes. Additionally, the automatic classifier should be able to distinguish users for and against the determined subject, because not only the supporters, but also the non-supporters contain information useful for prediction. Currently, considerable research on sentiment analysis of text in online social networks has been done. Therefore, by integrating the results of these studies and the technique created in this research, a large number of users who discussed the subject positively or negatively can be classified in the shortest amount of time.

Bibliography

[1] Mohammad-Ali Abbasi, Sun-Ki Chai, Huan Liu, and Kiran Sagoo, *Real-world behavior analysis through a social media lens*, Social Computing, Behavioral-Cultural Modeling and Prediction, Springer, 2012, pp. 18–26.

[2] Mohammad Ali Abbasi, Jiliang Tang, and Huan Liu, *Scalable learning of users' preferences using networked data*, Proceedings of the 25th ACM conference on Hypertext and social media, ACM, 2014, pp. 4–12.

[3] Mohammad Ali Abbasi, Reza Zafarani, Jiliang Tang, and Huan Liu, *Am i more similar to my followers or followees?: analyzing homophily effect in directed social networks*, Proceedings of the 25th ACM conference on Hypertext and social media, ACM, 2014, pp. 200–205.

[4] Akram Al-Kouz and Sahin Albayrak, *An interests discovery approach in social networks based on semantically enriched graphs*, Advances in Social Networks Analysis and Mining (ASONAM), 2012 IEEE/ACM International Conference on, IEEE, 2012, pp. 1272–1277.

[5] Mathieu Bastian, Sebastien Heymann, Mathieu Jacomy, et al., *Gephi: an open source software for exploring and manipulating networks.*, ICWSM **8** (2009), 361–362.

[6] Carolina Bigonha, Thiago NC Cardoso, Mirella M Moro, Marcos A Gonçalves, and Virgílio AF Almeida, *Sentiment-based influence detection on twitter*, Journal of the Brazilian Computer Society **18** (2012), no. 3, 169–183.

[7] Abdelberi Chaabane, Gergely Acs, Mohamed Ali Kaafar, et al., *You are what you like! information leakage through users interests*, Proceedings of the 19th Annual Network & Distributed System Security Symposium (NDSS), Citeseer, 2012.

[8] Michael D Conover, Bruno Gonçalves, Jacob Ratkiewicz, Alessandro Flammini, and Filippo Menczer, *Predicting the political alignment of twitter users*, Privacy, Security, Risk and Trust (PASSAT) and 2011 IEEE Third Inernational Conference on Social Computing (SocialCom), 2011 IEEE Third International Conference on, IEEE, 2011, pp. 192–199.

[9] Hongbo Deng, Jiawei Han, Hao Li, Heng Ji, Hongning Wang, and Yue Lu, *Exploring and inferring user–user pseudo-friendship for sentiment analysis with heterogeneous networks*, Statistical Analysis and Data Mining: The ASA Data Science Journal **7** (2014), no. 4, 308–321.

[10] Lise Getoor and Christopher P Diehl, *Link mining: a survey*, ACM SIGKDD Explorations Newsletter **7** (2005), no. 2, 3–12.

[11] Ian Anderson Gray, *Twitter follow limits*, `http://iag.me/socialmedia/guides/do-you-know-the-twitter-limits`, 2015.

[12] Xia Hu, Lei Tang, Jiliang Tang, and Huan Liu, *Exploiting social relations for sentiment analysis in microblogging*, Proceedings of the sixth ACM international conference on Web search and data mining, ACM, 2013, pp. 537–546.

[13] Kwang-Yong Jeong, Jae-Wook Seol, and Kyung Soon Lee, *Follower classification based on user behavior for issue clusters*, Proceedings of the First International Conference on Advanced Data and Information Engineering (DaEng-2013), Springer, 2014, pp. 143–150.

[14] Trevor May, *politwitter*, `http://politwitter.ca/page/canadian-politics-hash-tags`.

[15] Miller McPherson, Lynn Smith-Lovin, and James M Cook, *Birds of a feather: Homophily in social networks*, Annual review of sociology (2001), 415–444.

[16] Matthew Michelson and Sofus A Macskassy, *Discovering users' topics of interest on twitter: a first look*, Proceedings of the fourth workshop on Analytics for noisy unstructured text data, ACM, 2010, pp. 73–80.

[17] Alan Mislove, Bimal Viswanath, Krishna P Gummadi, and Peter Druschel, *You are who you know: inferring user profiles in online social networks*, Proceedings of the third ACM international conference on Web search and data mining, ACM, 2010, pp. 251–260.

[18] Debora Nozza, Daniele Maccagnola, Vincent Guigue, Enza Messina, and Patrick Gallinari, *A latent representation model for sentiment analysis in heterogeneous social networks*, Software Engineering and Formal Methods, Springer, 2014, pp. 201–213.

[19] Federico Alberto Pozzi, Daniele Maccagnola, Elisabetta Fersini, and Enza Messina, *Enhance user-level sentiment analysis on microblogs with approval relations*, AI* IA 2013: Advances in Artificial Intelligence, Springer, 2013, pp. 133–144.

[20] Juliano CB Rabelo, Ricardo BC Prudêncio, Flávia Barros, et al., *Using link structure to infer opinions in social networks*, Systems, Man, and Cybernetics (SMC), 2012 IEEE International Conference on, IEEE, 2012, pp. 681–685.

[21] Juliano CB Rabelo, Ricardo CB Prudêncio, and Flávia A Barros, *Leveraging relationships in social networks for sentiment analysis*, Proceedings of the 18th Brazilian symposium on Multimedia and the web, ACM, 2012, pp. 181–188.

[22] Delip Rao, David Yarowsky, Abhishek Shreevats, and Manaswi Gupta, *Classifying latent user attributes in twitter*, Proceedings of the 2nd international workshop on Search and mining user-generated contents, ACM, 2010, pp. 37–44.

[23] Fengyuan Ren and Ye Wu, *Predicting user-topic opinions in twitter with social and topical context*, Affective Computing, IEEE Transactions on **4** (2013), no. 4, 412–424.

[24] Jae-Wook Seol, Kwang-Yong Jeong, and Kyung-Soon Lee, *Follower classification through social network analysis in twitter*, Grid and Pervasive Computing, Springer, 2013, pp. 926–931.

[25] Ehsan Sherkat, Maseud Rahgozar, and Masoud Asadpour, *Structural link prediction based on ant colony approach in social networks*, Physica A: Statistical Mechanics and its Applications **419** (2015), 80–94.

[26] Michael Speriosu, Nikita Sudan, Sid Upadhyay, and Jason Baldridge, *Twitter polarity classification with label propagation over lexical links and the follower graph*, Proceedings of the First workshop on Unsupervised Learning in NLP, Association for Computational Linguistics, 2011, pp. 53–63.

[27] Chenhao Tan, Lillian Lee, Jie Tang, Long Jiang, Ming Zhou, and Ping Li, *User-level sentiment analysis incorporating social networks*, Proceed-

ings of the 17th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2011, pp. 1397–1405.

[28] Jiliang Tang, Yi Chang, and Huan Liu, *Mining social media with social theories: A survey*, ACM SIGKDD Explorations Newsletter **15** (2014), no. 2, 20–29.

[29] Twitter, *Twitter developers*, `https://dev.twitter.com`.

[30] ———, *Twitter help centre*, `https://support.twitter.com/articles/68916`.

[31] ———, *Twitter rate limits*, `https://dev.twitter.com/rest/public/rate-limits`.

[32] Jinpeng Wang, Wayne Xin Zhao, Yulan He, and Xiaoming Li, *Infer user interests via link structure regularization*, ACM Transactions on Intelligent Systems and Technology (TIST) **5** (2014), no. 2, 23.

[33] Tingting Wang, Hongyan Liu, Jun He, and Xiaoyong Du, *Mining user interests from information sharing behaviors in social media*, Advances in Knowledge Discovery and Data Mining, Springer, 2013, pp. 85–98.

[34] Michael J Welch, Uri Schonfeld, Dan He, and Junghoo Cho, *Topical semantics of twitter links*, Proceedings of the fourth ACM international conference on Web search and data mining, ACM, 2011, pp. 327–336.

[35] Zhen Wen and Ching-Yung Lin, *On the quality of inferring interests from social neighbors*, Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining, ACM, 2010, pp. 373–382.

[36] Jianshu Weng, Ee-Peng Lim, Jing Jiang, and Qi He, *Twitterrank: finding topic-sensitive influential twitterers*, Proceedings of the third ACM international conference on Web search and data mining, ACM, 2010, pp. 261–270.

[37] Huan Xu, Yujiu Yang, Liangwei Wang, and Wenhuang Liu, *Node classification in social network via a factor graph model*, Advances in Knowledge Discovery and Data Mining, Springer, 2013, pp. 213–224.

[38] Yusuke Yamamoto, *Bworld robot control software*, `http://twitter4j.org//`, 2015, [Online; accessed June/July-2015].

[39] Dawei Yin, Liangjie Hong, and Brian D Davison, *Structural link analysis and prediction in microblogs*, Proceedings of the 20th ACM international conference on Information and knowledge management, ACM, 2011, pp. 1163–1168.

Appendix A

Description of Target Users

Table A.1: List of all the Object IDs, followers and followees numbers, and labels of Target Users.

Numbers	**FollowersNum**	**FolloweesNum**	**Lable**
1	281	380	NDP
2	1393	1232	NDP
3	970	1218	NDP
4	2062	2120	NDP
5	82	74	NDP
6	177	348	Cons
7	62	62	NDP
8	1061	46	Cons
9	261	582	Liberal
10	571	448	Liberal
11	287	935	NDP
12	467	572	Liberal
13	98	103	Liberal

Continued on Next Page ...

Numbers	FollowersNum	FolloweesNum	Lable
14	183	331	Liberal
15	28	155	NDP
16	332	557	NDP
17	209	593	NDP
18	406	433	NDP
19	199	208	NDP
20	180	74	NDP
21	60	86	NDP
22	108	294	NDP
23	169	656	Liberal
24	132	98	NDP
25	66	112	NDP
26	247	324	Liberal
27	320	422	NDP
28	393	218	NDP
29	25	169	NDP
30	41	78	Conservative
31	492	1979	Liberal
32	447	1436	NDP
33	732	876	NDP
34	1795	1849	Liberal
35	1140	1396	Liberal
36	497	829	NDP
37	108	201	NDP
38	488	59	NDP
39	346	327	NDP
40	102	300	NDP
41	77	193	Liberal
42	188	193	Liberal
43	90	228	Liberal
44	545	1038	Conservative
45	344	604	Conservative
46	1461	1216	Conservative

Continued on Next Page ...

Numbers	FollowersNum	FolloweesNum	Lable
47	318	323	Conservative
48	412	1084	Conservative
49	302	503	Conservative
50	906	820	Conservative
51	728	690	Liberal
52	719	1540	Liberal
53	868	1993	Conservative
54	713	385	NDP
55	218	87	Liberal
56	278	913	NDP
57	256	349	NDP
58	685	1896	NDP
59	474	869	NDP
60	131	230	NDP
61	781	745	NDP
62	925	1211	Liberal
63	768	1071	Liberal
64	500	676	Liberal
65	981	1518	Liberal
66	916	993	Liberal
67	106	455	Conservative
68	435	519	Liberal
69	557	876	Liberal
70	622	1149	Liberal
71	676	721	NDP
72	400	697	NDP
73	861	1997	Liberal
74	845	728	Liberal
75	49	97	NDP
76	440	507	NDP
77	306	542	Liberal
78	159	378	Liberal
79	609	551	Conservative
80	243	232	Conservative

Continued on Next Page ...

Numbers	FollowersNum	FolloweesNum	Lable
81	1745	1848	Conservative
82	1577	657	Conservative
83	53	83	Conservative
84	355	767	Conservative
85	635	1963	Liberal
86	180	725	NDP
87	202	696	NDP
88	74	191	NDP
89	376	909	NDP
90	108	231	Liberal
91	353	570	Liberal
92	185	280	Liberal
93	901	1451	Conservative
94	194	165	Conservative
95	328	1049	Conservative
96	126	146	Conservative
97	1576	1859	Conservative
98	709	1635	Conservative
99	223	193	Conservative
100	421	1855	Conservative
101	338	258	Conservative
102	568	667	Conservative
103	1030	859	Conservative
104	141	292	Conservative
105	220	208	Conservative
106	59	111	Conservative
107	74	229	Conservative
108	68	133	Liberal
109	1083	1988	Conservative
110	547	672	Conservative
111	137	575	Liberal
112	608	994	NDP
113	1305	1766	Liberal

Continued on Next Page ...

Numbers	FollowersNum	FolloweesNum	Lable
114	1740	642	Liberal
115	1104	1996	Liberal
116	295	325	Conservative
117	1882	1963	NDP
118	453	173	NDP
119	1266	782	Liberal
120	586	792	Conservative
121	490	739	Conservative
122	417	489	Conservative
123	271	635	Conservative
124	1071	1674	NDP
125	443	342	Liberal
126	108	424	Conservative
127	1312	1077	Liberal
128	669	979	Liberal
129	1307	1088	Conservative
130	429	494	NDP
131	1456	1715	NDP
132	508	386	Liberal
133	932	1222	Liberal
134	1174	502	NDP
135	5000	2000	Conservative
136	1068	1729	Liberal
137	2318	1954	Conservative
138	2083	1801	Conservative
139	181	453	Conservative
140	970	1911	Conservative
141	1263	1732	Conservative

Vita

Candidate's full name: Jalehsadat Mahdavimoghaddam
University attended: Islamic Azad University of Tehran, North-Branch
Degree awarded: Bachelor of Computer Software Engineering 2012