# Predicting the Outcomes of Important Events based on Social Media and Social Network Analysis

Lei Wang

Thesis Submitted for the Degree of MPhil

School of Computer Science and Electronic Engineering

University of Essex

April 2020

# Abstract

Twitter is a famous social network website that lets users post their opinions about current affairs, share their social events, and interact with others. It has now become one of the largest sources of news, with over 200 million active users monthly. It is possible to predict the outcomes of events based on social networks using machine learning and big data analytics. Massive data available from social networks can be utilized to improve prediction efficacy and accuracy. It is a challenging problem to achieve high accuracy in predicting the outcomes of political events using Twitter data. The focus of this thesis is to investigate novel approaches to predicting the outcomes of political events from social media and social networks. The first proposed method is to predict election results based on Twitter data analysis. The method extracts and analyses sentimental information from microblogs to predict the popularity of candidates. Experimental results have shown its advantages over the existing method for predicting outcomes of politic events. The second proposed method is to predict election results based on Twitter data analysis that analyses sentimental information using term weighting and selection to predict the popularity of candidates. Scaling factors are used for different types of terms, which help to select informative terms more effectively and achieve better prediction results than the previous method. The third method proposed in this thesis represents the social network by using network connectivity constructed based on retweet data and social media contents as well, leading to a new approach to predicting the outcome of political events. Two approaches, whole-network and sub-network, have been developed and compared. Experimental results show that the sub-network approach, which constructs sub-networks based on different topics, outperformed the whole-network approach.

# Acknowledgments

I would like to express my special appreciation and thanks to my supervisor Professor John Q Gan. You are a tremendous mentor for me. I would like to thank you for encouraging my research and allowing me to grow as a researcher. Your advice on both research as well as on my career is priceless. I would also like to thank my supervisory panel members, Dr. Nigel J Newton and Dr. Michael H Fairbank, for serving as my panel members even at hardship and for your constructive comments and suggestions.

A special thanks to my family. Words cannot express how grateful I am to my mother and father for all the sacrifices that you've made on my behalf. I would also like to thank all my friends who supported me in writing and incented me to strive towards my goal. At the end, I would like to express appreciation to my beloved girlfriend Yanru who spent sleepless nights with me and was always my support at the moments when there was no one to answer my queries.

# Table of Contents

# List of Figures

# List of Tables

# Chapter 1: Introduction

## 1.1 Background

Social network applications play an important role in people's daily life. The rise in social media utilization is rapid: in 2011, approximately 60% of internet users were also users of social media. Social networking sites such as Facebook and Twitter have become powerful marketing and communication tools. Therefore, social network analysis is a hot topic in computer science. Social networks provide a large amount of useful data such as users' opinions. For example, Twitter records many texts and pictures about people's opinions on social events such as political elections. This type of data is very easy to be collected by companies and organisations. They could get ideas of people's choices and preferences based on social network data. However, how to use this data is a problem in the research and application areas. Some data maybe become noisy in social network analysis. Thus, mining social network data in these large data sequences is necessary. Due to complex situations, researchers need more specialised and detailed analysis for different cases. Moreover, each problem has different distributions and features. Thus, the prediction and classification of social networks are worthy of research [42].

The prediction problem in social network analysis has become very important in recent years [1][44]. In traditional machine learning, the model of a prediction algorithm is trained with training datasets. However, data needs to be updated with high frequency in an online environment. Thus, the prediction based on social network data is hard. Traditional machine learning usually focuses on historic data to train the model.

With the explosion of data, more data are generated, so we need to train the model efficiently and quickly. And the model is required to be trained according to online data and improved for every period. Thus, in recent research, prediction based on social network analysis has become an important area.

People create many kinds of data on social networks. How to select these datasets and how to choose the data sequence affect the result of the prediction. Without data selection, the noisy data makes training slow and the prediction performance becomes worse. Therefore, data mining is a very important aspect of the prediction based on social network data.

Researchers found that most online adults use social networking sites, with Facebook as the most popular, followed by LinkedIn, Twitter, and Instagram. Due to the popularity of multiple platforms across a wide range of users, social media has become one of the most popular topics in public relation research. The research of social media could help enterprises understand requirements and demands in the market better, so companies invest and focus on social network analysis. It attracts many new researchers in relevant fields [66].

Social network theories are an important tool to help people to understand and predict the result based on social networks. In social media, many theoretical studies focus on online learning algorithms, graph theory, and machine learning. These researches provide many references about the social network prediction problem [20][24].

Machine learning algorithms build and optimise machine learning models for prediction or other tasks. Through training the new model with social network data, researchers can classify the data. Many of the early studies of social media found some machine learning algorithm was not fit for social network analysis such as I using EM

algorithm to detect the opinions of twitter which is not reflect true result. but it has dramatically changed in the past few years as new algorithms are using social media to analyse data of different applications and predict results in various events.

Graph theory can be used for representing the structure of social networks. It characterises the networks which are constructed by nodes and ties/edges. Graphs visualise social networks, including social media contact and friendships. Representing a problem as a graph can provide a different point of view. Graph theory makes complex problems simple to be represented and graph can be utilised for representing opinion change in social networks. Thus, there are many studies that focus on graph theory about behaviour and opinion changing or prediction [13][15][44].

Companies and organisations could use the above mentioned technology about social media analysis to help meet their goals of development and marketing. Therefore, social network analysis will have a good future and be used in more fields of study [10][17][29].

## 1.2 Research Objectives

It is possible to predict the outcomes of events based on social networks using machine learning and big data analytics. Massive data can be utilised to improve the prediction efficacy and accuracy. However, there are still some challenges in relevant areas such as how to recognise and handle useless information. This thesis tries to tackle the challenges from different perspectives.

The first objective of this research focuses on social media analysis. I investigated a new method to predict the outcome of political events based on social context analysis.

I also proposed a new method for prediction of outcomes of political events based on social media analysis by term weighting.

The second objective focuses on predicting the outcome of political events by social network analysis. I proposed a new method for prediction of outcomes of politic events based on graph-theoretical modelling of retweet links in social networks.

## 1.3 Contributions

The first contribution is the social media based outcome prediction for political events. I proposed a new method for predicting the outcome of political events. It is based on traditional methods, but makes use of neutral tweets. In the existing methods, only positive and negative words are considered in the computing model, however, neutral tweets would affect a candidate's popularity in social networks, as neutral comments can propagandise the relevant candidate and thus may attract more voters to support the candidate. Therefore, the number of tweets related to a candidate, which may not be positive or negative, is considered in the proposed method. The results of the proposed algorithm was compared with existing semantic analysis based methods, showing the advantage of the proposed method.

The second contribution is the proposed term weighting approach to the prediction of the outcome of political events based on social media analysis. This new method provides better key terms selection and thus more accurate popularity predication compared to the method for the first contribution.

The third contribution is the social network based outcome prediction for political events. I proposed a new method for predicting the outcome of political events based on retweets network connectivity analysis. As the existing methods for predicting the outcome of political events are mainly based on semantic analysis and little work has been done to predict the results of political events by applying graph theory, this work explores the application of graph theory in political event prediction. I analysed communities of users based on retweet interaction. Two approaches, whole-network and sub-network, were proposed. Experimental results show that the sub-network approach, which constructs retweet networks for different important topics separately, has advantages over the whole-network approach. All data are collected in English version. The first reason is twitter communicuation usually in English. Even fench news will be boardcast in English. The second reason is language barrier between me and French.

# Chapter 2: Literature Review

## 2.1 Social Network Analysis based on Graph Feature

The earliest forms of the Internet, such as CompuServe, were developed in the 1960s. Primitive forms of email were also developed during this time. By the 1970s, networking technology had improved, and the 1979's UseNet allowed users to communicate through a virtual newsletter. By the 1980s, home computers were becoming more common and social media was becoming more sophisticated. Internet relay chats, or IRCs, were first used in 1988 and continued to be popular well into the 1990s. The first recognisable social media site, Six Degrees, was created in 1997. It enabled users to upload a profile and make friends with other users. In 1999, the first blogging sites became popular, creating a social media sensation that's still popular today. After the invention of blogging, social media began to explode in popularity. Sites like MySpace and LinkedIn gained prominence in the early 2000s, and sites like Photobucket and Flickr facilitated online photo sharing. YouTube came out in 2005, creating an entirely new way for people to communicate and share with each other across great distances.

Social network analysis is the process of investigating social networks through different methods and theories. It characterises social networks in terms of model representation. Social network analysis has its theoretical roots in the work of early sociologists such as Georg Simmel and Émile Durkheim, who wrote about the importance of studying patterns of relationships that connect social actors. Social scientists have used the concept of "social networks" since early in the 20th century to connote complex sets of relationships between members of social systems at all scales,

from interpersonal to international. In the 1930s Jacob Moreno and Helen Jennings introduced basic analytical methods. In 1954, John Arundel Barnes started using the term systematically to denote patterns of ties, encompassing concepts traditionally used by the public and those used by social scientists: bounded groups and social categories [6][7][8].

### 2.1.1 Analysis of Links in Social Networks

Links are a major part of social network. Different nodes are connected by links in the graph. Understanding links of graph could help us understand the graph. Thus, in social network analysis, prediction and characterisation of links is an important topic which attracts many researchers.

Benchettara, Kanawati, and Rouveirol designed a model to predict links in a graph in Supervised Machine Learning applied to Link Prediction in Bipartite Social Networks [12]. They focus on predicting links in a bipartite. Graph and predicting links in a unimodal graph obtained by the projection of a bipartite graph over one of its node sets. Their model uses Jaccard's coefficient to set the metrics and use random walks in the graph. The model sets the indirect metric based on the user's neighbours and their common behaviours. Their result is a clear improvement in the prediction model. They have studied the problem of link prediction in a special type of network obtained by the projection of a bipartite graph over a set of nodes. The writers have introduced new topological metrics that can reflect the likelihood of a link between two nodes that are computed in the dual graph: the graph obtained by the projection of the original bipartite graph but over the other set of nodes. These metrics are used in a dyadic topological supervised machine learning approach for link prediction. Their result

showed that new metrics do enhance obtained results, especially in terms of prediction precision, whether the link to be predicted occurs in the original bipartite graph or in one of the projected graphs. It provides an effective way to predict of links. However, their limitation is it need to test more supervision learning algorithm.

Liben-Nowell and Jon [43] developed approaches to link prediction based on measures for analysing the "proximity" of nodes in a network. They focus on the link prediction problem in social networks. Only few methods considered the proximity element in graph analysis. They introduced the prediction methods based on node neighbourhood with Jaccard's coefficient, rooted page rank and other algorithms. Jaccard's coefficient will measure the common feature and get a score for it with a parameter of the predictor [43]

$$\text{score}(x, y) := \sum_{\ell=1}^{\infty} \beta^{\ell} \cdot |\text{paths}_{x,y}^{\langle \ell \rangle}| \tag{2.1}$$

This algorithm could also have been used to solve the recommendation problem [19][26]. As a result, direct access to information may well confer additional predictive power. They found that information flow and neighbourhood features could improve the performance of link predicting. It provided a new method to predict the link in the social network. However, it still has a limitation because it not tested in more general cases. When the dataset is different, it may not perform better.

Backstrom and Leskovec designed the algorithm of Predicting and Recommending Links in Social Networks [7]. Their research was based on how to predict the link on social network. They tried to improve the performance of predicting algorithms. First, they evaluated several aspects of their algorithm: the choice of the loss function, the choice of the edge strength function, the choice of the random walk restart parameter α, and the choice of the regularisation parameter λ. After that, they also consider the

extension where they suggest a separate edge weight vector depending on the type of the edge [7] by solve problem equation

$$\min_{\omega} F(\omega) = \left\|\omega\right\|^{(2)} + \lambda \sum_{d \in D, l \in L} h(p_t - p_d) \qquad (2.2)$$

where λ is the regularisation parameter that trades-off between the complexity for the fit of the model. Finally, they generated 100 synthetic graphs. The algorithm used 50 of the graphs for training the weights *w*, and tested the algorithm on the other 50 graphs. The results of the graphs showed that when noise increase, the performance of the algorithm drops slowly but works perfectly in noise free environment. The algorithm predicts the links in social networks very well, but how to implement it in big data analysis is one challenge in the future.

### 2.1.2 Graph Theory based Social Network Modelling

Recently, most social network analysis is based on graph theory. Graph theory is where mathematical structures are used to model pairwise relations between objects. A graph is made up of vertices, nodes, or points which are connected by edges or lines. A graph may be undirected or directed, it depends on the different edges between the two nodes associated with each edge. Social network graph theory combines with network theory to analyse network features. It can provide a set of techniques for analysing graphs. Complex system network theory provides techniques for analysing the structure in a system which could represent as a network. Basic graph theory can best represent nodes and behaviour and make computing easier for complex situations. However, a graph is more about the internals of a social network, such as information spread on a social network or the behaviour changes. Social causation of migraine is

one of the hypothesized mechanisms, assuming that social factors, such as the socioeconomic status and social networks, may exert an effect on the level and severity of migraine. A higher prevalence of migraine among low-income or low-education groups has been reported. Additionally, migraine initiation appears to be the dominant mediator of the observed higher prevalence in these disadvantaged groups.[60]

Pennacchiotti and Popescu proposed a model of analysis of topics and network structure [55]. They addressed the task of user classification in social media. The main motivation is that attempting an automatically method infers the values of user attributes. Firstly, the model needs to classify a user by their personal information and context of their tweets. After that, it uses these attributes in the Latent Dirichlet Allocation model to build a model for classifying their group of users. Finally, the result seems good for different types of users. The model can classify their topic of tweets and use their probability distribution to compute results. They presented a generic model for user classification in social media and provided extensive quantitative and qualitative analysis which shows that in the case of Twitter users. Given n classes, each class ci is represented by a set of seed user Si. Each word w issued by at least one of the seed users is assigned a score for each of the classes. The score estimates the conditional probability [55] of the class given the word as follows:

$$\text{proto}(w, \text{c}_i) = \frac{\left|w, S_j\right|}{\sum_{j=1}^{n} \left|w, S_j\right|} \tag{2.3}$$

They tested their model in three different tasks: political affiliation detection, ethnicity and business affinity detection. They found that rich linguistic features prove consistently valuable across the three tasks and show great promise for additional user classification requirements. They described a general machine learning framework for

social media user classification. However, it still has a limitation on detecting a new element of a tweet. It only can be used in the rich information model. New elements will affect the data mining results. It could not fit some new elements in the learning model.

Agrawal and Rajagopalan designed a model based on links in a graph and uses it to classify users of newspapers [3]. Their motivation is that an automatically generated social network within a newsgroup may help information retrieval and text mining applications. This algorithm is used for predicting the behaviour from newsgroup data. It totally uses a graph-theoretical approach to the model. They try to classify user trends in news media. After that, they used the Kernighan-Lin algorithm to measure users' efficiency. It has two kinds of conditions: constrained or unconstrained. They tested their model in gun control, immigration and abortion groups of news. But SVM and Naive Bayes algorithms can't classify two groups in their dataset testing. Their result showed that EV algorithm performs better than the benchmark algorithm. They applied graph-theoretic algorithms to a new domain and did the sensitivity analysis on simulated newsgroup data. Their first limitation is that constrained and iterative method still needs training data. The other limitation is that they should justify why the constrained methods perform much better than the unconstrained ones.

User's attitude of a product is another problem of social network analysis, Leskovec, Huttenlocher, and Kleinberg proposed a model of graph to predict the link in the graph in predicting positive and negative links in online social networks [40]. They studied online social network in which relationship can be either positive or negative. Such a mix of positive and negative links arises in a variety online setting. They introduced related work of the edge sign prediction problem. Firstly, they used a logistic regression classifier to combine the evidence into edge sign prediction, the classification accuracy

and area under the RoC curve. After that, the theory of triad types was added to the model to learn by logistic regression. Finally, they test the global structure for signed link and use the ROC graph to measure their results. They found that the sign of links in the underlying social network can be predicted with high accuracy. They have investigated some of the underlying mechanisms that determine the sign of links in large social networks where interactions can be both positive and negative. However, this algorithm would not be widely used in big data analysis because the logistic regression is limited.

Graph theory analysis is a big part of social network analysis. After reading these papers, I think that graph theory can simplify many problems. It can transform many complex phonemes to the node change problem. This could be an important weak point of graphs. In the future, it will be a limit for many detailed predictions from the algorithm.

### 2.1.3 Prediction of Outcomes of Important Events based on Graph Analysis

There are many researchers studying the prediction of outcomes of social events based on social network analysis. Lu and Kulshrestha [45] investigated the effects of the social network in the 2014 India election. They proposed an augmented contagion analysis model that accounts for the impact of repeated stimuli from adjacent nodes. Their result showed the most popular party among all candidates and the party who utilised social media to enlarge its impact.

Kagan and Stevens [74] used sentiment diffusion forecasting to predict how support or opposition toward a candidate would spread. Their research question is that could people predict the election result using social network data? However, some researchers

claimed that it is not. They applied the diffusion model in the 2014 Indian election. The diffusion model can be used to identify the individuals on social media who are most influential on any topic. Such information provides valuable intelligence to election campaigns, which can use this information to influence the population in various ways. Polls are unable to provide such information. Their results accurately projected the winner of the election and were even able to predict the most influential individual topics on social media. It provided a good case to use social networks to predict the outcome of an election. It also proved that Twitter-based forecasting using AI techniques can beat traditional polling. It still has limitations. They should investigate more cases for this diffusion estimation model. And it may have limited for the data input, because researcher is hard to get these detail data from other research institutes.

Livne et al. studied the use of Twitter by House, Senate and gubernatorial candidates during the 2010 midterm election in the US [42]. They tried to analyse different candidates and suggest a novel use of language modeling for estimating content cohesiveness. First, they collected over 690k documents that they produced and cited in the 3.5 years leading to the elections. Then they used statistical language models on semantic analysis. They set the term weight in the language models. The term computing [42] using a normalisation factor of $\lambda = 0.001$ following:

$$P(t|u) = (1 - \lambda)w^N(t,u) + \lambda P^N(t|D) \tag{2.4}$$

After that, they analysed the density of the group in basic structure analysis. As a result, they found a signification relationship between graph structure and election result by building a model that predicts that a candidate will win with an accuracy of 88%. Their findings show significant differences in the usage patterns of social media which suggest that conservative candidates utilised social medium more effectively. This work showed the relationship between graph and outcome of political events.

However, their analysis only based on US election, it should be tested in more election events.

## 2.2 Social Network Analysis based on user's content

Marketing organisations want to be aware of what people are saying in influential blogs, how the expressed opinions could impact their business, and how to extract business insight and value from these blogs. This has given rise to the emerging discipline of Social Media Analytics, which draws from Social Network Analysis, Machine Learning, Data Mining, Information Retrieval, and Natural Language Processing. The automated analysis of social media raises several interesting challenges [49].

### 2.2.1 Semantic Analysis Methods

Semantic analysis is a large topic. For example, the NLProcessor linguistic parser parses each review to split context into sentences and adding tags on them [8]. For infrequent feature identification, they use association mining [2] because user reviews will have a different story. It is hard to classify different features and different descriptions. Various methods have been developed to analyse the opinion of products, services, events and personality reviews based on social network analysis [67]. Data mining tools already used for opinion and sentiment analysis include collections of simple counting methods in machine learning. Categorising opinion-based text using a binary distinction of positive against negative [25][53][72], is found to be insufficient

when ranking items in terms of recommendation or comparison of several reviewers' opinions [54]. Determining players from documents on social networks has also become valuable as influential actors are considered as variables in the documents [78] when applying data mining techniques on social networks. The idea of co-occurrence can also be viable information.

Hoffmann's research focuses on the latent semantic analysis by unsupervised learning [34]. He wanted to identify and distinguish different contexts of word usage without referring to a dictionary or thesaurus. Probabilistic Latent Semantic Analysis has many applications, most prominently in information retrieval, natural language processing, machine learning from text, and other related areas. He proposed a generative latent model to perform a probabilistic mixture decomposition. The new model is based on latent semantic analysis with probabilistic clustering model. He focused on two tasks to assess the performance of the new model. First task is perplexity minimisation for a document-specific unigram model and noun-adjective pairs. Second task is automated indexing of documents. These experiments clearly demonstrated that the advantages of the new model over standard semantic analysis are not restricted to applications with performance criteria directly depending on the perplexity. He provided a novel statistical method for factor analysis of binary and count data. However, Hoffmann should test more general tasks to prove that his model is better than standard anytime.

Vasileios, Hatzivassiloglou, and Kathleen [32] classified the context opinion into positive or negative. They wanted to develop a large system to automatically identify antonyms and distinguish near-synonyms. They proposed a log-linear regression model for conjoined conjunction with the opinion of social media. Combining the constraints across many adjectives, a clustering algorithm separates the adjectives into groups of

different orientations, and finally, adjectives are labelled positive or negative. They achieved 82% accuracy in these tasks, which shows a good result from the experiment. This algorithm could help us analyse the opinion on social media. Their method achieved high precision on identifying adjectives. However, it also should study modern languages such as emoji. Morden language would affect the result for meaning and analysis.

Natarajan [50] studied the analysis of social media based on news. They proposed a new method to analysis users based on social media. First, they used a hybrid approach, which involved the analysis of click through, user tweets, and user Twitter friends lists to build a user profile, to personalise news recommendation. Second, they added a unique new feature of location preference to the news recommendation system to address the importance of temporal dynamics. Third, they allowed users to choose the ratio of popular news vs. trendy news they like. User profile creation architecture could like the following figure:

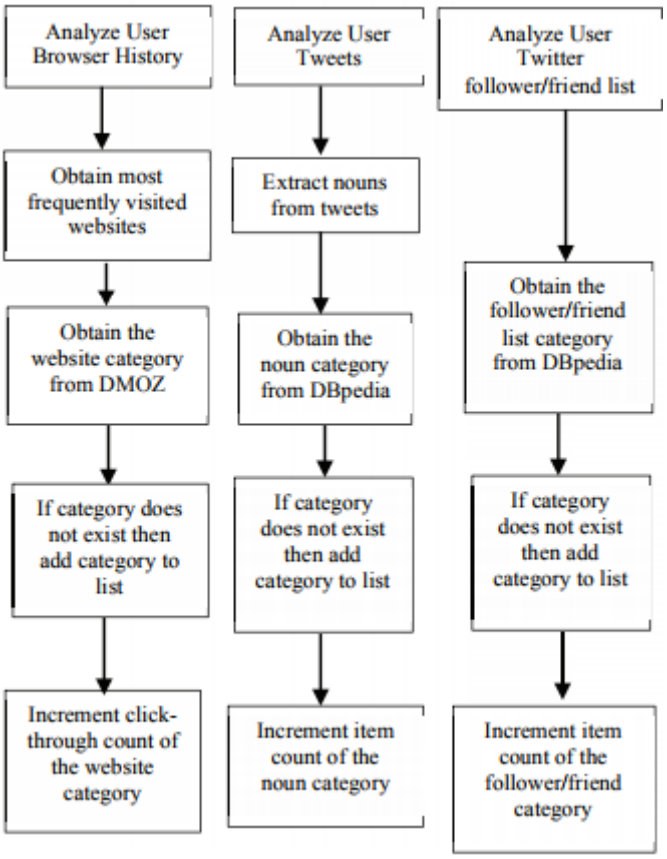


*Figure 1.2.1 Profile creation architecture [50]*

Finally, the resulting system was then evaluated based on user satisfaction and accuracy. The results show that the average user satisfaction increased from 8.6 to 9.4 when location preference is added, while the accuracy of the recommendation system was around 92-95%. They prove that a new method of social media analysis. It produces the characteristics of users in a social network. However, they should do test their method in other applications.

Burnap et al. [57] proposed a prediction model for using Twitter as an election forecasting tool and applied it to analyse the UK 2015 General Election. They assigned a score ranging from –5 to +5 to certain text according to its scale of positive or negative sentiment. Scores were given to words in the dictionary that are known to carry emotive meanings. For example, the score for “love” is 5 and for “hate” is –4. Firstly, they calculated a score for each tweet and produced a list of all tweets with positive and

negative scores. After that, they consolidated the scores for each party and its leader, based on which they predicted the change of seats in the parliament. However, it is not very clear how the scores were consolidated.

Michael focuses on affordable and ubiquitous online communications [68]. Online social media provide the means for flows of ideas and opinions and play an increasing role in the transformation and cohesion of society. Thus, they proposed an opinion formation framework based on content analysis of social media and social physical system modelling. In their opinion formation framework, opinion modelling based on social physics typically focuses on global properties of the modelled system. First, a model for the opinion diffusion network model is assumed. In a Bayesian framework, this corresponds to the definition of an a priori probability density function. The start figure like:

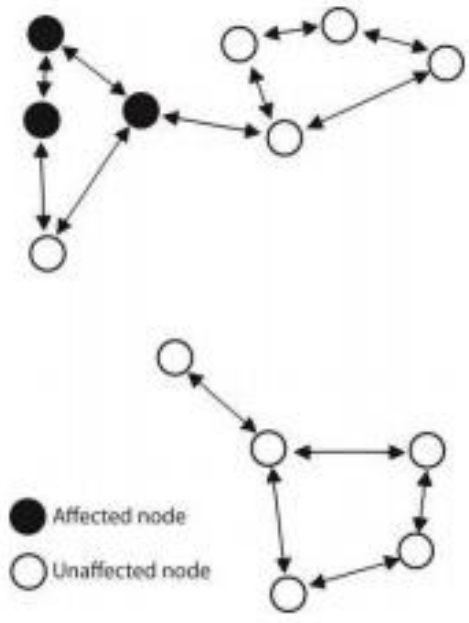


*Figure 2.2 Start network of social media [68]*

They tested their framework in BP oil spill event. Their framework presented a relationship between people and policymakers. It can be used in opinion tracking and

recognises issue-specific and policy-focused arguments. Moreover, they should test their framework in many different cases.

Zhang studied collaborative filtering of context analysis in social media [77]. However, the tradition method cannot deal with the cold-start problem. Thus, they proposed a newly-fashioned scheme - bi-clustering and fusion (BiFu) for the cold-start problem based on the BiFu techniques in a cloud computing setting. First, to identify the rating sources for the recommendation, they introduced the concepts of popular items and frequent raters. After that, they used the bi-clustering technique to reduce the dimensionality of the rating matrix. And to overcome the data sparsity and rating diversity, they employed the smoothing and fusion technique. Finally, BiFu recommends social media contents from both item and user clusters. Experimental results show that BiFu significantly alleviates the cold-start problem in terms of accuracy and scalability. Their new method overcomes the cold-start problem. Moreover, BiFu could be further improved. They should also investigate the item or user similarity calculation and dimension shrink of the rating matrix.

Ghanem[30] compared the language of false news to the real one of real news from an emotional perspective, considering a set of false information types (propaganda, hoax, clickbait, and satire) from social media and online news article sources. They proposed an LSTM neural network model that is emotionally infused to detect false news. Their results emphasized the importance of emotional features in the detection of false information. It is a good way to detect fake news. Future inplement of this method is a  valuable question

### 2.2.2 Prediction of Outcomes of Important Events based on Semantic Analysis

There are many challenges in social media analysis. Influence analysis in social networking big data faces more opportunities and challenges today [56]. Social influence analysis is pervasive throughout society. Social media brings large amounts of attractive opportunities to social influence analysis, but the challenge also comes in fake information and news; too much fake news causes wrong results in prediction [41][64].

Sakaki studied detection of events such as earthquakes in social media [62]. The main motivation is that social media could provide real-time nature. He proposed an algorithm to monitor tweets and to detect a target event. To detect a target event, they devise a classifier of tweets based on features such as the keywords in a tweet, the number of words, and their context. The probability of an event occurrence at time *t* is

$$P_{occur}(t) = 1 - p_f^{n0(1-e^{-\lambda(t+1)})/(1-e^{-\lambda})} \quad (2.5)$$

They can detect an earthquake with high probability merely by monitoring tweets. To classify a tweet into a positive class or a negative class, they use a support vector machine. They prepare three groups of features for each tweet: statistical features, keyword features and word context features. Overall, the classification performance is good considering that we can use multiple tweet readings as evidence for event detection. Location estimation methods such as Kalman filtering and particle filtering are used to estimate the locations of events. It is an application example of social media analysis for event detect. He proved that social media analysis could detect the event in real-time. Moreover, he should test new method in different topics to prove it could work anytime.

Hu focused on the features of semantic analysis based on customer reviews [35]. The number of reviews is larger, which makes it difficult for a potential customer to read

them to make an informed decision on whether to purchase the product or not. They mined the features of the product on which the customers have expressed their opinions and whether the opinions are positive or negative. This information would help manufacturers to improve their product in the future. In their model, part of speech tagging, and frequent feature identification is core to classifying the customer review text. According to the frequent features result, opinion words will be detected by documenting phrases with excellent and poor. They summarised the opinion and feedback of infrequent features of products. They conducted their experiments using customer reviews of five electronics products. The average accuracy for the five products is 84%. They provided a feature-based summary of many customer reviews of a product sold online. However, they only tested in 5 products. It may have limitations in other product reviews.

Another election prediction model based on Twitter was proposed by Cameron et al .[15], in which the authors tried to answer the following questions: What are the links between political information made available through social networks and the voting choices of citizens? Does an online presence and a social media strategy matter? Is online activity an indicator of support and does it influence election results? They tried to analyse the friends and followers of each electorate candidate in social networks. However, most candidate profiles in social networks were not complete. To analyse the relationship between the number of supporters for each candidate in social networks for a specified date and the election result, two regression models were proposed: a linear OLS ordinary least squares model of vote share and a logistic regression model with election outcome as the dependent variable. However, their results showed that their models were not useful in predicting election results based on social networks.

Biao [18] reviewed the recent advances on information diffusion analysis in social networks and its applications. They first shed light on several popular models to describe the information diffusion process in social networks, which enables three practical applications such as influence evaluation, influence maximization, and information source detection. Then, they discuss how to evaluate the authority and influence based on network structures. After that, current solutions to influence maximization and information source detection are discussed in detail.

Mahata [46] developed a classifier to identified posts mentioning intake of medicine by the user. Most of health and drug-related information studies in social media are based on aggregated results from a large population rather than specific sets of individuals. In order to conduct studies at an individual level or specific groups of people, they used a random search for tuning the hyperparameters of the CNN models and present an ensemble of best models for the prediction task.

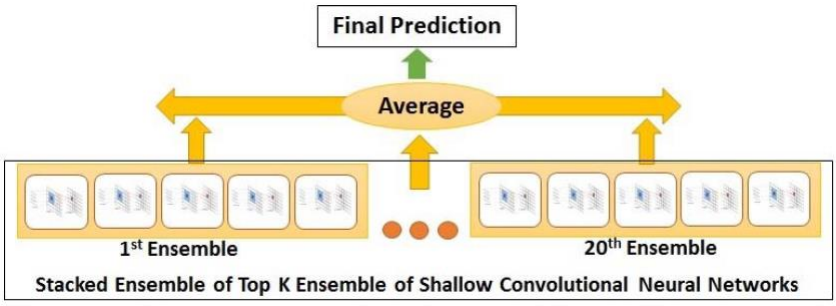


*Figure 2.3 CNN model* [45]

They used 8000 annotated tweets as a training dataset and 2260 additional tweets as development dataset. Their system produced state-of-the-art results, with a micro-averaged F-score of 0.693. However, it is a doubt that covert to other topics.

Tsikerdekis [73] attempted to classify detection methods based on the approaches and identifies factors that, in real-time systems, will impact the effectiveness and efficiency of these methods. Because most identity-deception detection method's

efficiency can vary. They considered identity theft and identify forgery. These involve generating accounts and employing identity management Optimizations are proposed that can limit the computational overhead. Further challenges involving real-time identity-deception detection are discussed.

Pimenta [58] compared different Social Media platforms under different aspects, in order to get the first idea about their suitabilities, advantages and weaknesses in comparison. They monitored the seven major candidates by collecting publicly available data from blogs, Facebook, Twitter and YouTube. They focused on monitoring political candidates in four different Social Media platforms in order to explore the potential of the collected data to investigate these personalities, especially regarding volume, attention and popularity metrics. They found a connection between real-world events and Social Media data and tried out the prediction potential of Social Media regarding the primary election outcomes. They obtained mixed results comparing three primary dates' voting outcomes to the Social Media data collected three days before such dates. Blogs and Twitter approximate very well to the primaries voting percentage. Despite the encouraging results, a deeper analysis of the forecasting power of the Social Media is necessary here.

## 2.3 Social network analysis with Big Data

Machine Learning for big data analysis has been studied in recent years, especially for implementation in a big data environment. There are some machine learning algorithms that show advantages in semantic analysis [13][37].

Tripathy [70] studied review classification in online networks. The reviews and blogs obtained from social networking and online marketing sites, act as an important

source for further analysis and improved decision making. However, these reviews need processing like classification or clustering to provide meaningful information for future uses. Supervised machine learning methods help to classify these reviews. They tested four different machine learning algorithms such as Naive Bayes, Maximum Entropy, Stochastic Gradient Descent, and Support Vector Machine have been considered for the classification of human sentiments. The frameworks as the following figure:

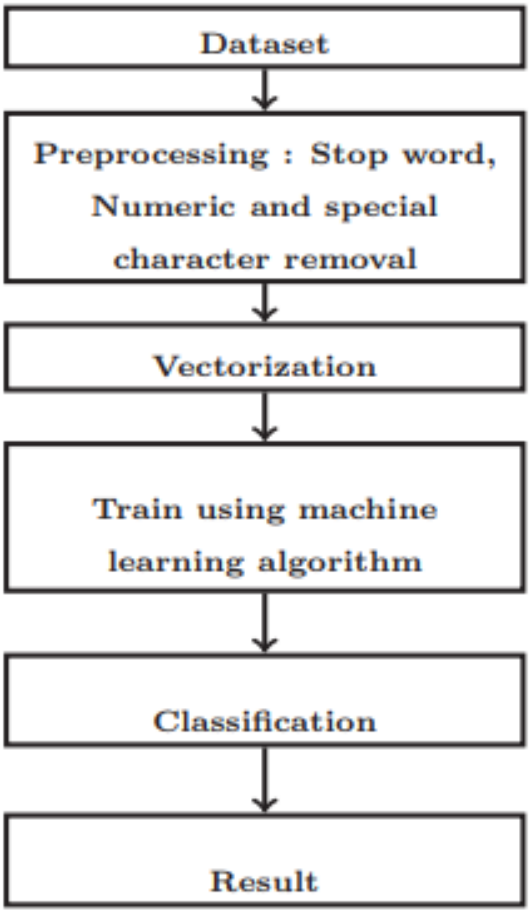


*Figure 2.4 Processing chart of analysis [70]*

The accuracy of different methods was critically examined in order to access their performance on the basis of parameters such as precision, recall, f-measure, and accuracy. They used the IBDM dataset about movie reviewers to exam these algorithms.

These algorithms are further applied using n-gram approach on IMDb dataset. It is observed that as the value of ‘n’ in n-gram increases the classify. Moreover, they showed a framework of machine learning for semantic analysis. However, they should test more datasets.

Casteleiro [16] investigated the feasibility of using word embeddings from Deep Learning algorithms together with terms from the cardiovascular disease ontology as a step to identifying omics information encoded in the biomedical literature. Word embeddings were generated using the neural language models CBOW and skip-gram with an input of more than 14 million PubMed citations corresponding to articles published between 2000 and 2016. Then the abstracts of selected papers from the systematic review were manually annotated with gene/protein names. They sat up two experiments that used the word embeddings to produce term variants for gene/protein names: the first experiment used the terms manually annotated from the papers; the second experiment enriched/expanded the annotated terms using terms from the human-readable labels of key classes. The hypothesis is that by enriching the original annotated terms, it is easier to obtain suitable term variants for gene/protein names from word embeddings. As part of the word embeddings generated from bag-of word and skip-gram, a lexicon with more than 9 million terms was created. Using the cosine similarity metric, a list of the 12 top-ranked terms was generated from word embeddings for query terms present in the generated lexicon. As the terms variants are induced from the biomedical literature, they can facilitate data tagging and semantic indexing tasks. Overall, their study explores the feasibility of obtaining methods that scale when dealing with big data, and which enable automation of deep semantic analysis and mark up of textual information from unannotated biomedical literature.

Cambria [14] studied emotions understanding in AI. Being important for the advancement of AI, emotion processing is also important for the closely related task of polarity detection. The opportunity to automatically capture the general public's sentiments about social events, political movements, marketing campaigns, and product preferences has raised interest in both the scientific community, for the exciting open challenges. This opportunity has made the emerging fields of affective computing and sentiment analysis for distilling people's sentiments from the ever-growing amount of online social data. Semantic computing's hybrid framework for polarity detection as following:

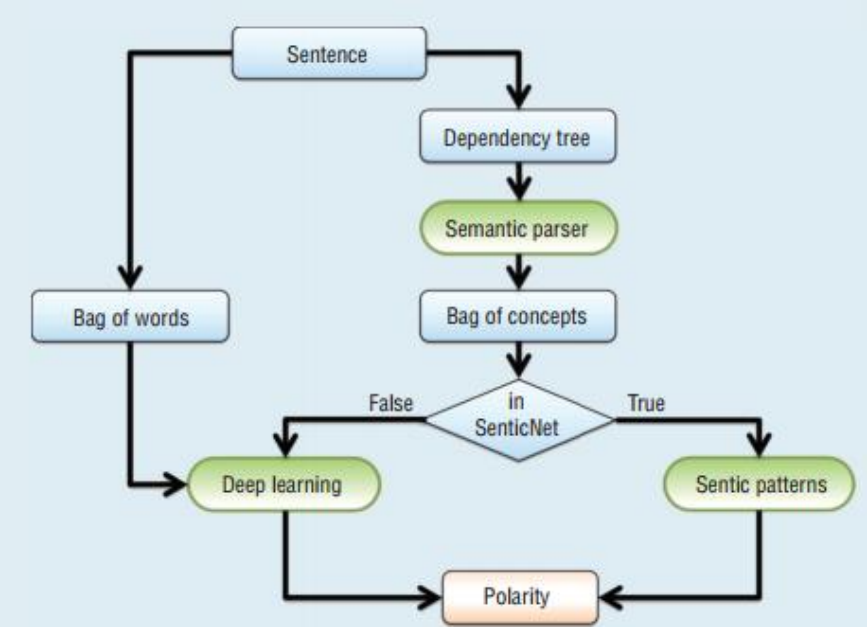


*Figure 2.5 Framework for polarity detection [14]*

Single word expressions, however, are just a subset of concepts, multiword expressions that carry specific semantics. He provided the detailed model of tradition semantic analysis and summarised the current technology of machine learning in semantic analysis.

Arros [4] focused on abstract concepts such as semantic category, writing style, or sentiment. Machine learning models in semantic analysis allow annotating very large

text collections, more than could be processed by a human in a lifetime. Besides predicting the text's category very accurately, it is also highly desirable to understand how and why the categorisation process takes place. To achieve that, they demonstrated that layer-wise relevance propagation technique for explaining predictions of complex non-linear classifiers. They train two word-based ML models, a convolutional neural network and a bag-of-words SVM classifier, the main formula [4] on following:

$$P(w_t|w_{t-n:t+n}) = \frac{\exp\left(\left(\frac{1}{2n}\cdot \sum_{-n\leq j\leq n, j\neq 0} v_{w_{t+j}}\right)^{\mathsf{T}} v'_{w_t}\right)}{\sum_{w\in V} \exp\left(\left(\frac{1}{2n}\cdot \sum_{-n\leq j\leq n, j\neq 0} v_{w_{t+j}}\right)^{\mathsf{T}} v'_{w}\right)} \tag{2.6}$$

This enables one to distill relevant information from text documents without an explicit semantic information extraction step. They further use the word-wise relevance scores for generating novel vector-based document representations which capture semantic information. The measure of model explanatory power and show that, although the SVM and CNN models perform similarly in terms of classification accuracy, the latter exhibits a higher level of explain ability which makes it more comprehensible for humans and potentially more useful for other applications.

In the context of drug discovery, drug target interactions can be predicted based on observed topological features of a semantic network across the chemical and biological space. In order to take into account the heterogeneity of the semantic network, Fu, Ding and Seal proposed semantic link association prediction (SLAP), to predict unknown links between compounds and protein targets in an evolving network [31]. The additional semantic links significantly improved the predictive performance of the supervised learning models. The binary classification model built upon the enriched feature space using the Random Forest algorithm significantly outperformed an existing semantic link prediction algorithm. The whole framework of drug input is as follows:

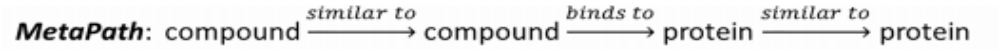


*Figure 2.6 Framework of drug discovery [31]*

In addition to link prediction, Random Forest also has an intrinsic feature ranking algorithm, which can be used to select the important topological features that contribute to link prediction. The proposed framework has been demonstrated as a powerful alternative to SLAP in order to predict DTIs using the semantic network that integrates chemical, pharmacological, genomic, biological, functional, and biomedical information into a unified framework. It offers the flexibility to enrich the feature space by using different normalisation processes on the topological features, and it can perform model construction and feature selection at the same time. However, it could enlarge the edge of semantic analysis.

Most wrapper approaches are built upon deep learning technologies in debt to their great capacity on the learning of high-order context representation without the requirement of careful feature engineering. As such, deep learning approaches such as convolutional neural networks, long short-term 25 memory (LSTM) networks, and others, have been used extensively in sentiment analysis. In these neural network-based models, only word-level features such as word embeddings are used yet deeper sentence features can be automatically achieved. Mikolov proposed a simple yet effective approach to learn distributed representation of words in 2013. Since then, neural network approaches have been extensively studied for sentiment analysis tasks [22][23].

## 2.4 Open Problems and Challenges

Based on the state of the art review, we could find that there are many limitations and challenges in this area. Collection of complex network data is one problem in social network analysis. Large-scale and high-quality data collection is hard in social network analysis. How to measure the influence and authority in graphs is another challenge. The universality of social media analysis is also a problem. In general, how to apply machine learning and big data analytics to social media analysis is an interesting challenging problem.

# Chapter 3: Twitter Data Collection

The prediction model to be proposed tries to make a prediction of the social event outcome according to social network data analysis. Plenty of data is generated by social networks, but fake information and useless data also come into the dataset, which strongly affects the model accuracy. Therefore, Twitter data collection and processing is important in my research.

## 3.1 Social Media Data Collection

First, I need to verify the data in social network distribution and get the data by using the API of a social network. Then I need to consider which features of data to define the problem and figure out which components to represent the social network data best. Then I need to analyse the data to get the behaviour of users. According to the behaviour of users, I can judge the general opinion of the social network. To collect data for this research, I choosed Twitter. This big social network company provides APIs to access their social networks for getting data, which is easier compared with scanning the webpages. I also learn how to use these APIs to get target data. Their APIs provide many data, with tags included in the context.

The Twitter APIs offer a number of options for public members to gather data from the platform. Researchers can purchase access to the Firehose, Twitter's real-time flow of all new tweets and their related metadata, and those seeking historical data can purchase all relevant public, undeleted tweets from Twitter's archive. However, both of these options can be extremely cost prohibitive. Thus, academic researchers have

largely turned to Twitter's free services, a set of public APIs. The first of these, Twitter's Streaming API, provides tweets in real-time and can be queried using keyword, user ID, and geolocation parameters. When undertaking keyword queries of tweet content, the Streaming API matches keywords in the body of a tweet, the body of quoted tweets, URLs, hashtags, and @mentions. In my case, I collected tweets with "French election" tweets and hashtags. Twitter's documentation suggests that the Stream can return "up to" 100% of all tweets meeting one's query criteria, as long as the relevant tweets constitute less than 1% of the global volume of tweets at any given moment. When that 1% threshold is reached, the API begins to impose rate limits. Twitter's current global volume averages 6,000 tweets per second. However, this figure fluctuates significantly from day to day, hour to hour, and even minute to minute, driven in large part by unpredictable external events such as natural disasters, terrorist attacks, and other shocking or controversial news items. The API does provide rate limit messages, allowing a user to know when rate limits have been imposed, but these do not indicate what types of messages are missing, that is, if there is a systematic character to the tweets that are not captured. Moreover, Twitter's documentation does not suggest that 100% of tweets necessarily to be provided, even when no rate limits are imposed. The second common source of Twitter data, the Search API, is a component of the larger REST API and may be used to query historical data. This option is clearly advantageous for collecting data on issues or events that cannot be easily predicted in advance. However, Search API queries carry significant limitations. First, the Search API only reaches back 6~7 days. Second, each call to the API can return a maximum of just 18,000 tweets, and Twitter limits users to 180 calls every 15 minutes. In addition, queries to the Search API provide matches only to the main text of a tweet. Finally, and perhaps most importantly, Twitter makes it clear that Search

results are non-random. The company states that queries return “top” content, not all relevant tweets, but Twitter does not clarify what constitutes “top” content. All data are in English language because the language barrier between French and me.

## 3.2 Data Processing and Modelling

Another problem in Twitter data collection is how to measure a single user and classify the relationship. I analysed the online data which creates an online context and comments on it. The context of Twitter can provide the behaviour of users to predict the result. Twitter data can reflect the relationship between different users and tweets. Based on retweets tag and user id, I can convert tweets as a social network and classify their topics to predict the outcome of the event. Firstly, I labelled all tweets with different candidates or parties. Based on different candidates, the whole twitter information are classified in different groups. Then I sort and summary the ID of the user based on groups. I summarized the edge of the user and labelled tweets for different candidates inner their group of data. Excel and Python could summarized all user ID and tweets. Based on hashtag of retweet, the retweet ID are summarized. Most of tweets are in the same context which is easy to label. To construct the social network, I count the times of retweets as strengths of edge, which provides more information of the community. In word count, I selected words with high frequency in each candidate and classified them as negative or positive by knowledge.

In whole collection period, more than 40000 tweets collected by twitter API. However, the weekend usually get less tweets than week day. Thus the weekend data are wiped out. After wiped out, there are 38000 tweets in total analysis.

# Chapter 4: Prediction of the 2017 French Election Based on Twitter Data Analysis

## 4.1 Motivation

Twitter is one of the largest social networks, providing a friendly platform for people to express opinions and share views on a variety of topics and issues. Prediction based on Twitter data analysis has drawn much attention in recent years, especially in predicting results of political events. In 2015, the mainstream media polls were wrong in the prediction of the UK general election. Traditional polling methods for election prediction analyse data from questionnaires by a phone call or pedestrian survey and are usually biased in sampling and prediction process as well. Nowadays, social networks provide valuable information for predicting outcomes of social or political events, which may not be obtained from the mainstream media or traditional polls. For example, many people supported Trump in the US presidential election 2016, but they might not be willing to say so in public for some reasons. Thus, prediction based on social media analysis can result in new outcomes, complementary with or even more accurate than traditional prediction poll results.

In Twitter based election prediction it is critical to extract informative keywords or features reflecting true sentiment of voters. In addition, traditional prediction models may not be suitable for the data from social networks. In this paper, a new method for election prediction based on Twitter data analysis is proposed and applied to predict the 2017 French Election.

## 4.2 Method

Term frequency was adopted to get more keywords in our study. According to the frequency of words in the collected tweets, most frequently occurred words with sentimental meanings were considered as features for classifying tweets. However, word frequency usually changes with time. Some words may have high frequency on one day but may not appear in any tweet on another day, such as Macron leaks in the French election.

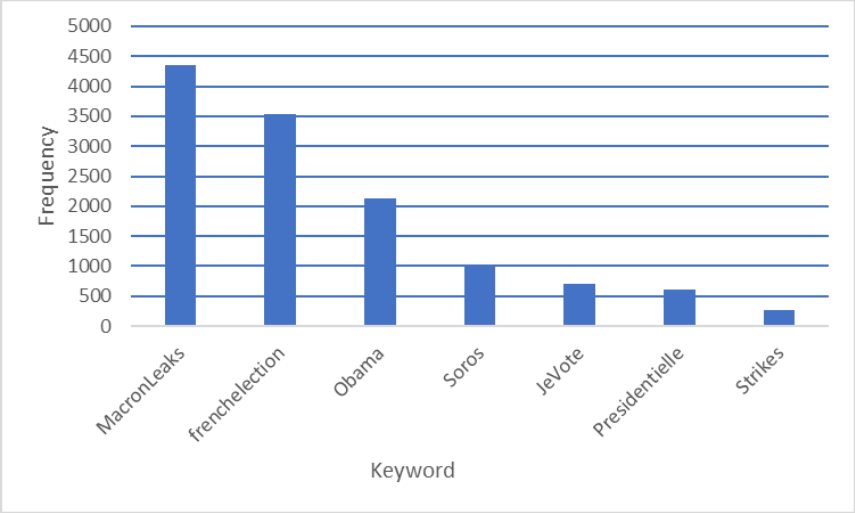


*Figure 4.1 Frequency of keywords in the collected Twitter data*

As an example, Fig. 4.1 shows the words or terms of highest frequencies in the tweets on the last day before the voting day, of which French Election may not be sentimental at all. In our study, data analysis and domain knowledge were combined to select keywords as features. Based on the analysis of the tweets posted before the voting day, the following extra keywords were selected:

Obama: positive for Macron as Obama posted a video to support Macron.

Presidentielle: positive for Macron and negative for Le Pen as presidentielle was used only to describe Macron.

Macron leaks: negative for Macron as it was Fake news to slander Macron.

Soros: negative for Macron as Macron was linked to Soros.

JeVote: positive for Le Pen as it suggested voting for Le Pen in Twitter.

Strike: negative for Le Pen as workers went on strike to protest Le Pen before the election.

Besides, there are some well-recognised keywords related to election being used in our experimental study such as vote, win, fail, and attack.

To calculate the popularity of a candidate in the election, this paper proposes the following formula:

$$popularity(a) = \left[ \frac{pos(a)}{pos(a) + neg(a)} \right] \left[ \frac{N(a)}{N(a) + N(b)} \right] \tag{4.1}$$

where $N(a)$ and $N(b)$ are the number of tweets that are related to candidate $a$ and candidate(s) $b$ respectively, $pos(a)$ and $neg(a)$ are the number of positive and negative tweets for candidate $a$ respectively. As neutral tweets are also considered in the proposed method, the sum of $pos(a)$ and $neg(a)$ is not necessarily equal to $N(a)$. If there are more than two candidates, $b$ represents all the candidates but candidate $a$. To make the sum of the popularities of all the candidates equal to 100%, the popularities are scaled if needed.

The existing methods for election prediction based on Twitter data usually do not consider neutral tweets. However, neutral tweets would affect a candidate's popularity in social networks, as neutral comments can propagandize the relevant candidate and thus may attract more voters to support the candidate. Therefore, the number of tweets related to a candidate, which may not be positive or negative, is considered in our method, as indicated in equation (4.1).

To evaluate the proposed method, the well-recognised Tumasjan's method [71] is compared in our experiments, which calculates the popularity of a candidate as follows:

$$popularity(a) = \frac{pos(a) + neg(b)}{pos(a) + neg(a) + pos(b) + neg(b)} \tag{4.2}$$

where $pos(a)$, $pos(b)$, $neg(a)$, and $neg(b)$ are defined in the same way as in equation (4.1). This method seems to be more reasonable from a mathematical perspective, because the popularities of candidates *a* and *b* add up to 100%. However, it ignores neutral tweets and may be biased to a candidate who is strongly supported by a relatively small group of voters but is not minded by many other voters.

For convenience, the proposed method and the Tumasjan's method are named as Method 1 and Method 2 respectively in the remaining part of the paper.

## 4.3 Experimental Results and Discussion

Method 1 and Method 2 described in the previous section were applied to predict the popularities of candidates in the 2017 French election. Following data collection from Twitter, keywords or terms as features were identified and they were used to classify the collected tweets into positive, negative, or neutral groups for relevant candidates. After that, equations (4.1) and (4.2) were adopted to calculate the popularities of the candidates. After the first round of election on April 23, Francois Fillon and Jean-Luc Melechron dropped out, and the remaining two candidates were Emmanuel Macron and Marine Le Pen. The Twitter data collected during April 24 to May 6 were analysed using the methods described in the previous section to predict the popularities of the two final candidates and thus predict who would be the winner of the 2017 French election on May 7.

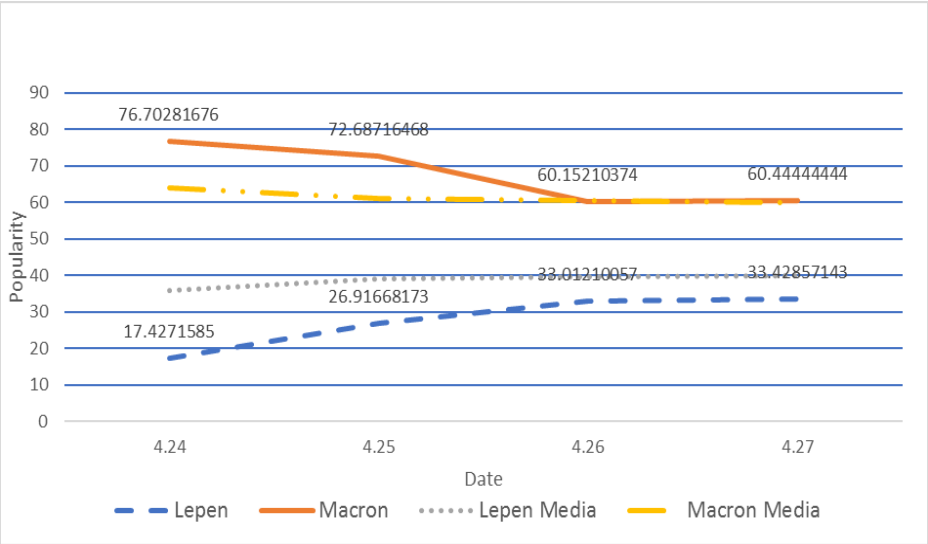


*Figure 4.2 Popularity predicted by Method 1 based on Twitter data during April 24-27*

Fig. 4.2 shows the popularity prediction results of Method 1 based on the data extracted from Twitter during the period from April 24 to 27, in which the solid line in orange represents the predicted popularity of Macron and dashed line in blue the predicted polularity of Le Pen. It can be observed that Macron was more popular than Le Pen in this week. This might be due to that Fillon and Melechorn suggested that their supporters would vote Macron after they lose in the first round. However, Macron's popularity was persistently declining in this week while Le Pen's popularity increased gradually. For comparison purposes, the average mainstream media poll results during the same period (averaged over the poll results of Ipsos, Harris, Ifop-Fiducial, OpinionWay, Elabe and Odoxa) are presented in Fig. 4.2 as well, with dash-dot line in yellow for Macron's popularity and dotted line for Le Pen's. It can be seen that both the values and trends of the popularities predicted by Method 1 and the maintream media polls are quite similar, but the prediction by Twitter data analysis is more dynamic. The popularity of candidate also incloudes another candidate tweets, and some tweets are natural with two candidates. Thus the sum of two candidate is not 100%.

In a similar manner, Fig. 4.3 shows the popularity prediction results of Method 2 based on the data extracted from Twitter during the period from April 24 to 27, which indicates that Macron led on the first two days (24 and 25), but Le Pen was more popular on April 26 and 27.

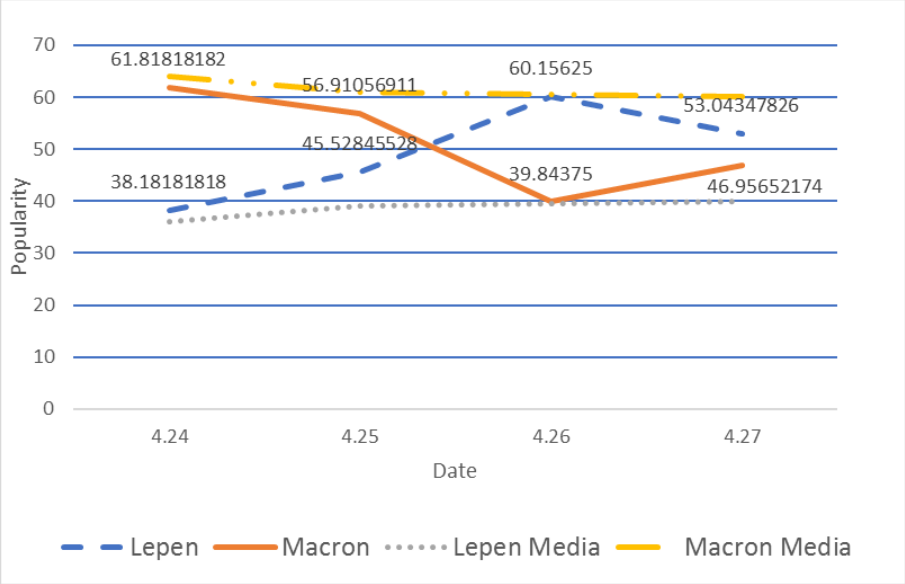


*Figure 4.3 Popularity predicted by Method 2 based on Twitter data during April 24-27*

Fig. 4.4 shows the popularity prediction results of Method 1 based on the data extracted from Twitter during the period from May 1 to 4. It shows that Macron lost the leading position at the beginning of the final week before the election. However, Macron came back to lead since May 2 as Le Pen lost her popularity on May 2 to 4. Comparing with the mainstream media poll results, the prediction by Method 1 has similar trend but indicates a larger gap between the poularities of Macron and Le Pen, which as a matter of fact is closer to the real voting result.

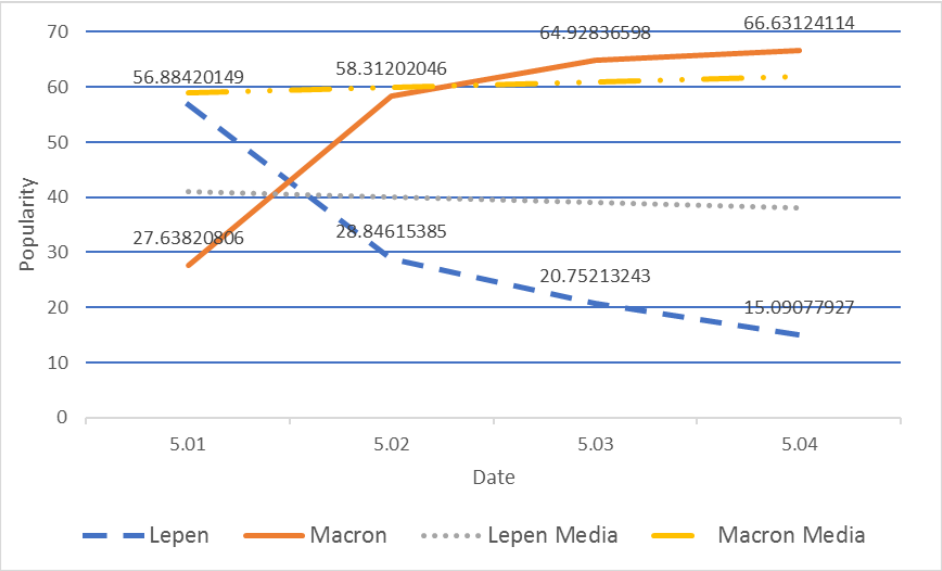


*Figure 4.4 Popularity predicted by Method 1 based on Twitter data during May 1-4*

Fig. 4.5 shows the popularity prediction results of Method 2 based on the data extracted from Twitter during the period from May 1 to 4. Compared with Fig. 4.4, the popularity trends of the two candidates showed in Fig. 4.5 are different. Le Pen led all the time in this week until May 4. On May 2, Macron lost some support by workers. Because there is an activity of workers who supported Le Pen, news and media boardcasted it many times, which enhances the popularity of Le Pen. This was reflected in the results of Method 2 but not of Method 1, because it boardcasts more in natural opinion.

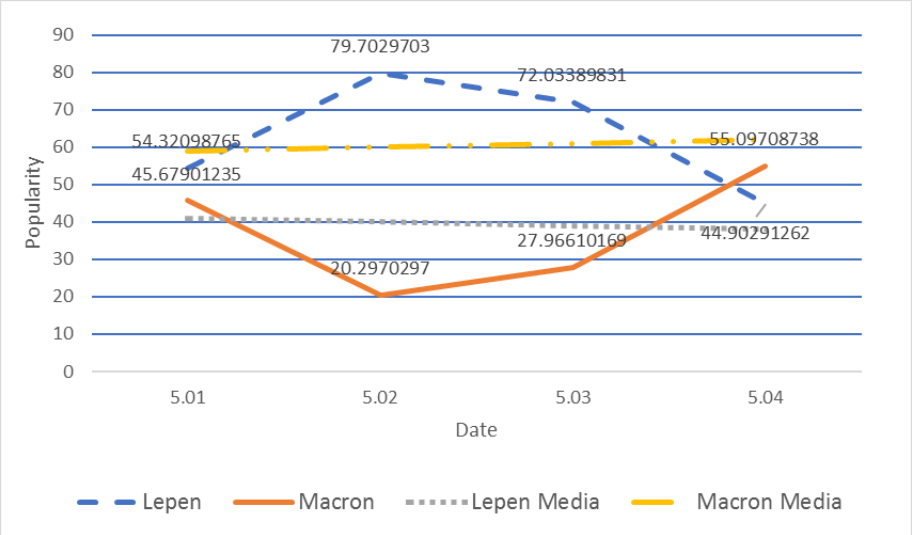


*Figure 4.5 Popularity predicted by Method 2 based on Twitter data during May 1-4*

Fig. 4.6 shows the popularity prediction results of Method 1 and Method 2 based on the data extracted from Twitter on May 6, the final day before election, with "Macron leaks" taken into account. Method 2 predicted that Le Pen would win the final round of election, whilst Method 1 predicted that Macron would lead by a big majority. The main difference between the two methods is whether neutral tweets reflect the popularity of candidates. This figure shows that neutral tweets relevant to specific candidates played an important role in popularity prediction. Although "Macron leaks" is negative for Macron in general, many tweets related to Macron due to "Macron leaks" were actually classified as neutral tweets for Macron due to other positive keywords appeared in these tweets.

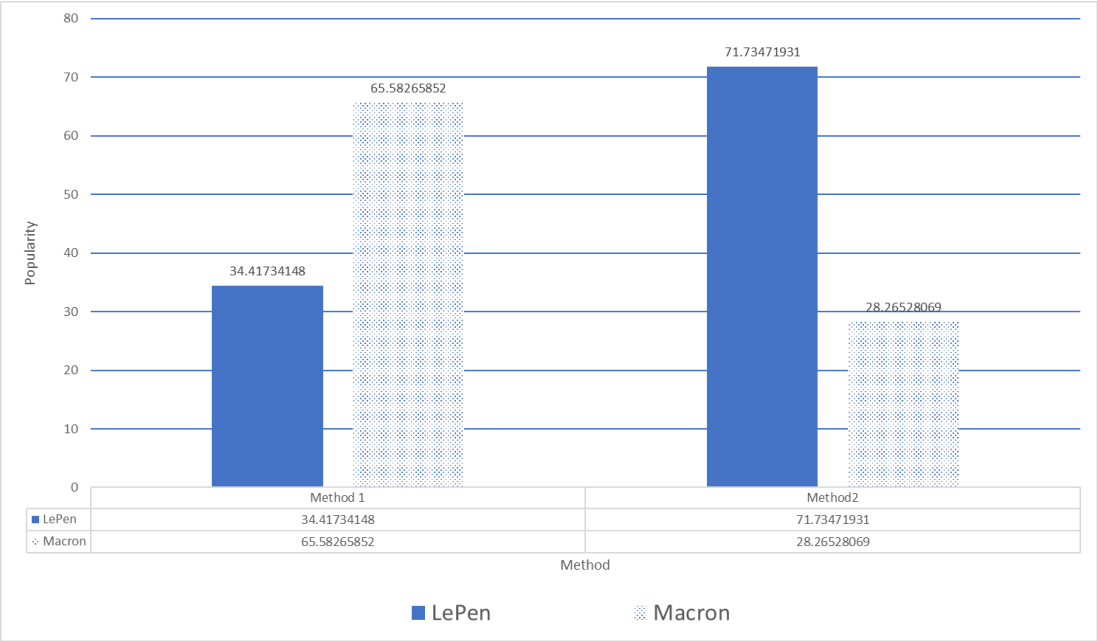


*Figure 4.6 Popularity predicted by Methods 1 and 2 based on Twitter data on the final day before the election*

Fig. 4.7 shows the popularity prediction results of Method 1 and Method 2 based on the data extracted from Twitter on May 6 without the consideration of "Macron leaks". It is interesting that both methods predicted that Le Pen would win the election if "Macron leaks" was ignored, as the Twitter data on the final day contains various highly positive tweets for Le Pen, but many negative tweets for Macron. This shows that "Macron leaks" greatly influenced the election result, demonstrating the importance of selection of keywords or terms.

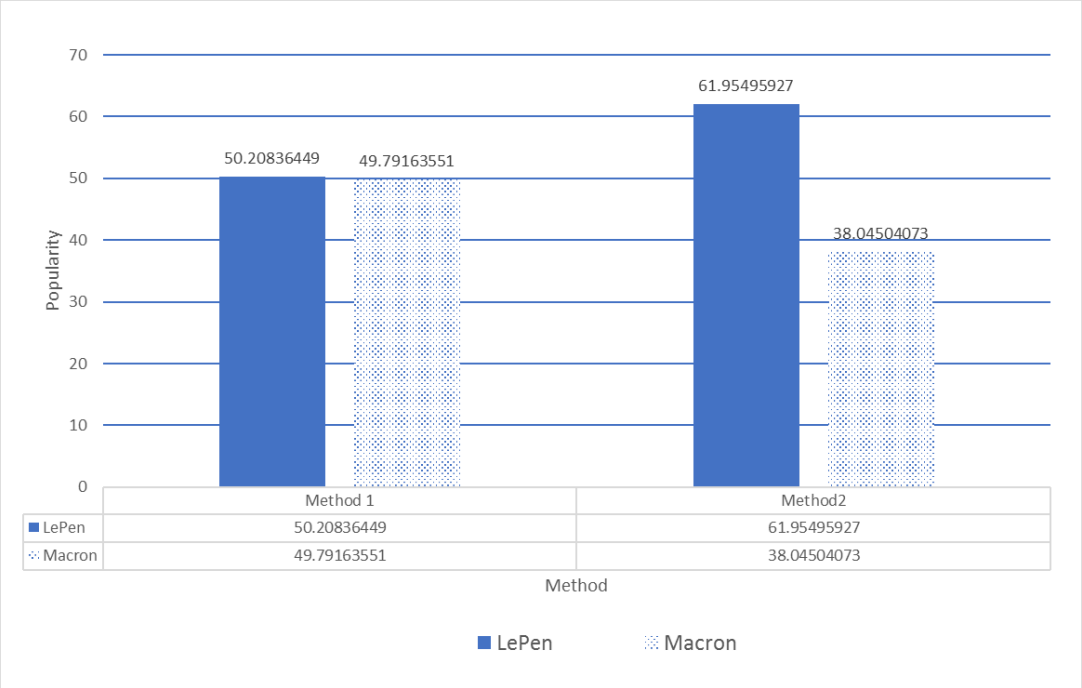


*Figure 4.7 Popularity predicted by Methods 1 and 2 based on Twitter data on the final day before election, without using "Macron leaks"*

The real voting result of the 2017 French presidential election is 66.1% for Macron and 33.9% for Le Pen. Comparing the predicted results of Method 1 and Method 2 on the final day before election with "Macron leaks" considered, the predicted popularity of Method 1 is about 2% different from the real voting result, whilst the prediction of Method 2 is about 38% away from the real voting result and it predicted the winning candidate wrongly. If ignoring "Macron leaks", both Method 1 and Method 2 predicted the wining candidate wrongly, but the popularity prediction of Method 2 is much more away from the real voting result than Method 1.

On the final day before election, Macron got Internet attack named as "Macron leaks". On Twitter, it was the hottest topic just before election, which resulted in many tweets related to Macron, although the mainstream media clarified that "Macron leaks" was fake news. The problem of Method 2 with "Macron leaks" is that it is easily affected by fake events. For example, there were thousands of tweets tagged "Macron

leaks", which were negative to Macron, but there were also many neutral tweets related to Macron due to "Macron leaks" at the same time that were not considered by Method 2. Neutral tweets have propaganda effect in election. Swing voters might most likely post neural tweets, which should not be neglected as they bring in uncertainty for the election. The number of neutral tweets seemed to be a key factor for the higher accuracy of Method 1.

It is noteworthy that the keywords chosen for the two candidates and their sentimental meanings are different. For instance, Barack Obama, who strongly supported Macron, is a positive key term for Macron, but neutral or not a key term for Le Pen. Of course, some keywords are more influential than the others, but it is difficult to weight the importance of these keywords properly based on domain knowledge. Data mining and machine learning techniques may be useful in this aspect. As an example, Fig. 4.8 shows the popularity predicted by Method 1 based on Twitter data on the final day before election using a single keyword respectively, from which we can see that the keyword 'win' or 'vote' alone gave quite accurate popularity prediction compared to the real voting result whilst the other key words gave popularity prediction in more favour of Macron.

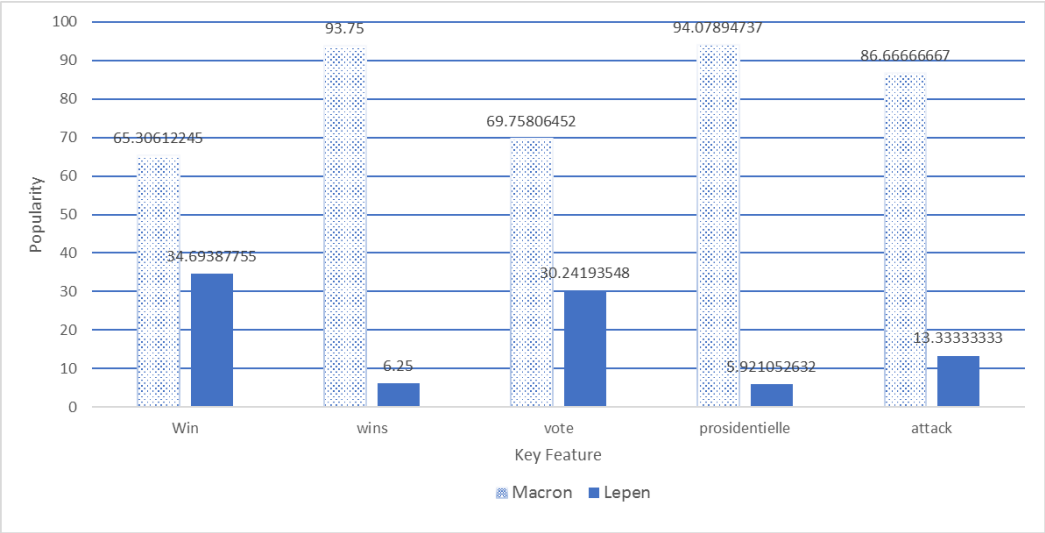


*Figure 4.8 Popularity predicted by Method 1 based on Twitter data on the final day before election using a single keyword respectively*

There are other issues that affect the accuracy of election prediction based on Twitter data analysis. For example, Most Macron's supporters did not attack Le Pen on Twitter, but Le Pen's supporters usually attack Macron. This is difficult to be considered in the Twitter data analysis in general. To summarize the number of tweets, there are more than 30k data in whole prediction period. The number of positive and negative tweets is more than 15k. Others are neutral tweets.

## 4.4 Conclusion

This chapter proposes a new method for candidate's popularity prediction based on Twitter data analysis and thus for election result prediction indirectly. The proposed method considers neutral tweets related to specific candidates, which has been proved to increase prediction accuracy in our case study of predicting the 2017 French election result.

This is a work-in-progress study from the perspective of Twitter data analysis for predicting outcomes of important social or political events. More work will be carried out in our future research to improve the reliability and accuracy of the method for election prediction based on Twitter data analysis.

# Chapter 5: Prediction of the 2017 French Election Based on Twitter Data Analysis Using Term Weighting

## 5.1 Motivation

Traditional methods for prediction poll analyse data from questionnaires by phone calls or pedestrian surveys and are usually biased in sampling and prediction process as well [52]. Nowadays, social networks provide valuable information for predicting outcomes of social or political events, which may not be obtained from the mainstream media or traditional polls. In Twitter based election prediction it is critical to extract informative keywords or features reflecting true sentiment of voters. In this chapter, a new method for election prediction based on Twitter data analysis using term weighting and selection is proposed and applied to the 2017 French Election. In our previous work [76], keywords selection was based on domain knowledge, which may overlook some words or terms that are very informative but not in line with the existing domain knowledge. To combat this problem, data-driven term weighting approaches [47] can be used to provide complementary and quantitative measures for weighting and selecting informative keywords that subsequently can be used for social media analysis based prediction of political events.

## 5.2 Method

After categorising the collected tweets into groups, one for each candidate. Term weighting methods were adopted to weight and select keywords in our study.

According to the weighting scores of words in the collected tweets, the words with the highest weighting scores were considered as features with sentimental meanings for election prediction. However, word weighting scores usually change with time. Some words may result in high scores on one day but may not appear in any tweet on another day, for example, “Macron leaks” in the French election. In our study, data analysis and domain knowledge were combined to select keywords as positive or negative based on the analysis of the tweets posted before the voting day.

In this study, term frequency – inverse document frequency (TF-IDF) is adopted for term weighting, which is a numerical statistic to reflect how important a word is to a document in a collection or corpus of documents [63]. It is often used as a weighting factor in information retrieval, text mining, and user modeling. The equation of term weighting is as follows [47]:

$$TF - IDF_{i,j} = tf \times idf = \frac{t_{i,j}}{\sum_k t_{k,j}} \times lg \frac{N}{|\{d \in D | i \in d\}|} \quad (5.1)$$

where *tf* is term frequency within a document which is the number of times a term occurs in a document, *idf* is inverse document frequency within a corpus, $t_{i,j}$ is the number of times that term *i* appears in document *j*, $\sum_k t_{k,j}$ is the total number of times that all the terms under consideration appear in document *j*, *N* is the total number of documents in corpus *D*, and $|\{d \in D | i \in d\}|$ is the number of documents in which term *i* appears. In this study, a term is a keyword, a document corresponds to a tweet, and a corpus corresponds to a set of tweets relevant to an individual candidate. Term weighting is calculated based on a corpus of tweets relevant to an individual candidate respectively. Based on domain knowledge, if a keyword is positive, its TF-IDF value will be positive, otherwise it is negative. After that, the total weighting score of term *i* in corpus *D* is calculated as follows:

$$TF - IDF_i = \sum_{j=1}^{N} TF - IDF_{i,j} \quad (5.2)$$

where *TF-IDF*$_i$ is the sum of the weighting scores of term *i* in all documents in corpus *D*, representing the weight of term *i* in a group of tweets in this study, whilst *TF-IDF*$_{i,j}$ is the weighting score of term *i* in document *j*.

It has drawn our attention that there are three types of keywords: keywords of type 1 appear in tweets relevant to both candidates and are positive or negative for both candidates; keywords of type 2 appear in tweets relevant to both candidates and are positive for one candidate but negative for the other candidate or vice versa; keywords of type 3 appear only in tweets relevant to one candidate but not in tweets relevant to the other candidate. To investigate which types of keywords are more informative and important for election prediction, this paper proposes to scale the TF-IDF values of these three types of keywords. The scaling factors, *f1*, *f2*, and *f3*, for the three types are determined by a data-driven approach that find their 'optimal' values by making the prediction match the opinion poll result before the voting date. More details about this is given in the Section 5.3.

After term weighting, each selected keyword has a weight, *i.e.*, *TF-IDF*$_i$, represented as a positive or negative score. Based on the weighting scores of keywords selected for a candidate, a score for this candidate is obtained by summing up the scores of all the keywords for this candidate. Because the keywords or terms are weighted and selected from individual candidate's group of tweets respectively, the number of selected terms and their weighting scores are in general different for different candidates.

Based on the scores of two candidates, the proposed method calculates the popularity of a candidate as follows:

$$popularity(a) = \frac{score(a)}{score(a)+score(b)} \quad (5.3)$$

where *score(a)* and *score(b)* are the scores of candidate *a* and candidate(s) *b* respectively. If there are more than two candidates, *b* represents all the candidates except candidate *a*. To make the sum of the popularities of all the candidates equal to 100%, the popularities are scaled if needed.

To evaluate the proposed method, in our experiments it is compared with the method proposed by the authors in 2017 [76], in which the popularity of a candidate is calculated as follows:

$$popularity(a) = \left[ \frac{pos(a)}{pos(a) + neg(a)} \right] \left[ \frac{N(a)}{N(a) + N(b)} \right] \tag{5.4}$$

and the Tumasjan's method [10], in which the popularity of a candidate is calculated as follows:

$$popularity(a) = \frac{pos(a) + neg(b)}{pos(a) + neg(a) + pos(b) + neg(b)} \tag{5.5}$$

where *pos(a)* and *pos(b)* are the number of positive tweets of candidates *a* and *b* respectively, *neg(a),* and *neg(b)* are the number of negative tweets of candidates *a* and *b* respectively, and *N(a)* and *N(b)* are the total number of tweets that relate to candidates *a* and *b* respectively.

## 5.3 Experimental Results and Discussion

As can be seen from Eq (5.1) and Eq (5.2), in term weighting the number of terms under consideration is an important parameter that will affect prediction results. Using the Twitter data on 2nd May 2017 and the mainstream media opinion poll results, the effect of the number of selected terms on the popularity prediction accuracy was firstly investigated. Two cases were tested: First, the top 100 terms with the highest weighting

scores in each group of tweets were used to calculate the popularity of each candidate; Second, all the terms were used to calculate the popularity of each candidate. In this case, all the terms will be scored, not a special list of neutral keywords. There were more than 400 terms which are summarize form all tweets. All word will be counted in it except useless words such as the and a.These data were also used to determine the values of the scaling factors *f1*, *f2*, and *f3* in such a way that the prediction results match the opinion poll results. In our study, various combinations of values of *f1*, *f2*, and *f3* ranging from 0.5 to 1.5 were tested with a changing step of 0.05, and those values that resulted in the best match between the predicted popularity and the opinion poll result were selected.The range is to consider the number of tweets in positive and negative. This small range dose not reduce not lossing the weight of tweets too much.

For convenience, in the remaining part of this chapter the method previously proposed by the authors is named as Method 1, the Tumasjan's method as Method 2, and the methods newly proposed here are named as Term weighting 1 (considering a selected number of terms only) and Term weighting 2 (considering all the available terms) respectively.

Figure 5.1 shows the candidate's popularity prediction results of Term weighting 1, Term weighting 2, Method 1, Method 2, in comparison with the result from main-stream media opinion poll, based on the Twitter data extracted on 2nd May 2017. By Term weighting 1, the popularity of Macron is 50.2% and the popularity of Le Pen is 49.8%. By Term weighting 2, the popularity of Macron is 62.7% and the popularity of Le Pen is 37.3%. The opinion poll showed that the popularity of Macron is 63% and the popularity of Le Pen is 37%. It can be seen that the result from Term weighting 2 is very close to the opinion poll result, indicating that using all the terms achieved good

results whilst using 100 selected terms was not enough to obtain accurate popularity prediction.

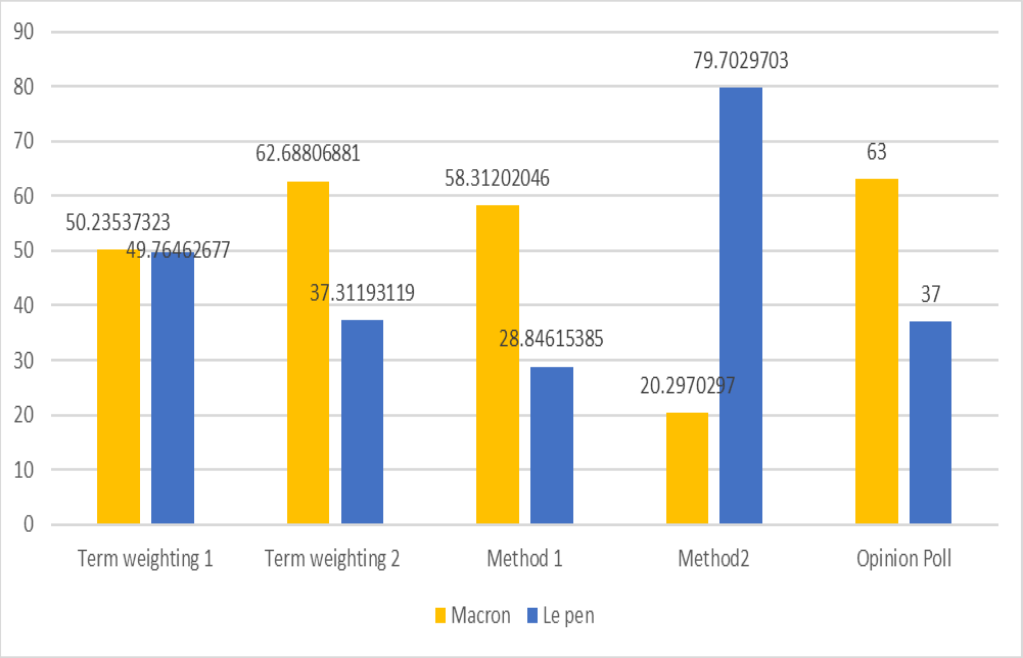


*Figure 5.1 Candidate's popularity predicted by term weighting based on Twitter data on May 2nd*

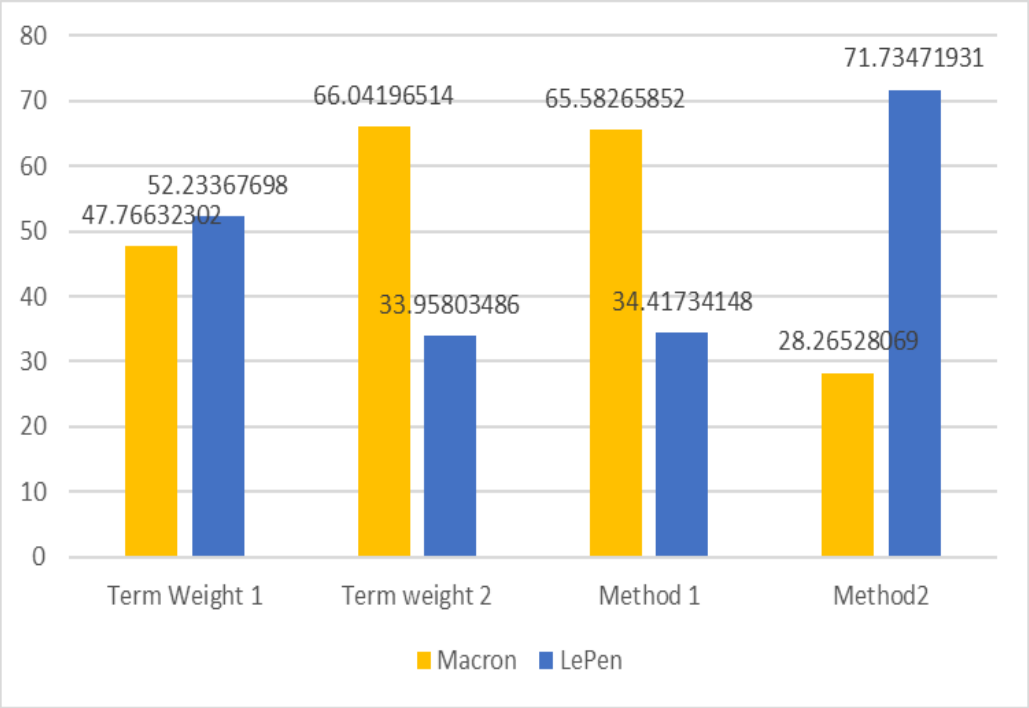


*Figure 5.2 Candidate's popularity predicted by term weighting based on Twitter data on May 6th*

Figure 5.2 shows the candidate's popularity prediction results of Term weighting 1, Term weighting 2, Method 1, Method 2, based on the data extracted from Twitter on 6th May 2017, the final day before the election. Term weighting 2 predicted that the popularity of Macron is 66.0% and the popularity of Le Pen is 34.0%, which are the closest to the real election result, that is, Macron got 66.9% and Le Pen 33.1%. Term weighting 1 and Method 2 got wrong prediction results with Le Pen being more popular.

In order to show the effect of the scaling factors, Figure 5.3 shows the candidate's popularity prediction results of Term weighting 1 and Term weighting 2 with scaling factors, based on the data extracted from Twitter on 2nd May 2017, aiming to match the opinion poll result. Term weighting 2 predicted that the popularity of Macron is 63.0% and the popularity of Le Pen is 37.0%, which matched the opinion poll result as expected. Term weighting 1 also gave predictions very close to the opinion poll result. Obviously, the scaling factors for Term weighting 1 and Term weighting 2 were determined respectively and their values were different.

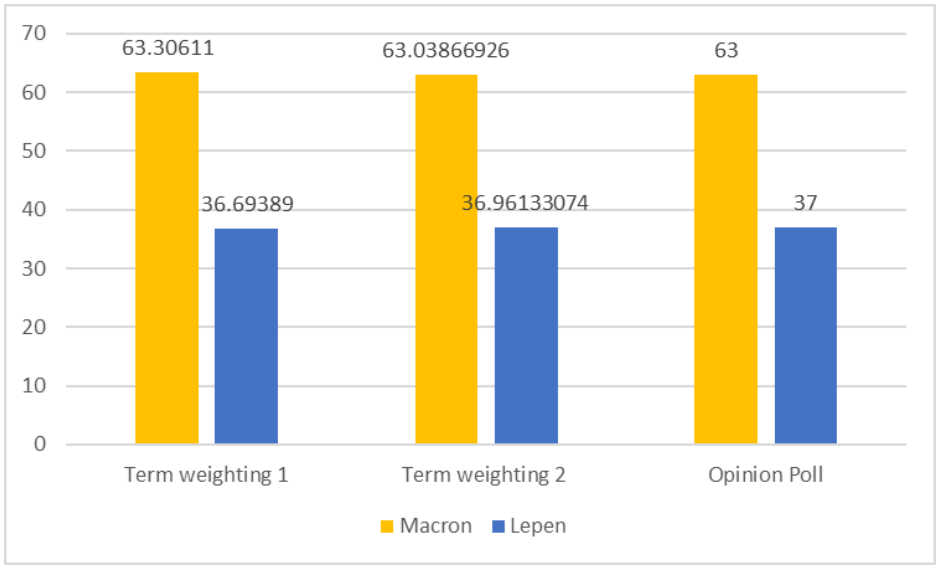


*Figure 5.3 Candidate's popularity predicted by term weighting with scaling factors based on Twitter data on May 2nd*

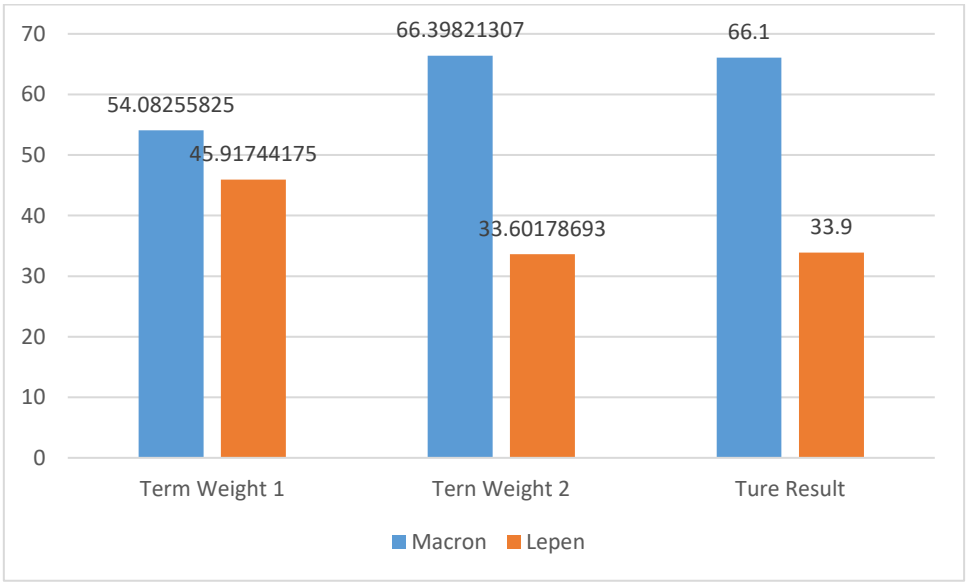


*Figure 5.4 Candidate's popularity predicted by term weighting with scaling factors based on Twitter data on May 6th*

Figure 5.4 shows the candidate's popularity prediction results of Term weighting 1 and Term weighting 2 based on the data extracted from Twitter on 6th May 2017, with the scaling factors obtained using the Twitter data on 2nd May 2017. Term weighting 2

predicted that the popularity of Macron is 66.4% and the popularity of Le Pen is 33.6%, which are the closest to the real election result, that is, Macron got 66.9% and Le Pen 33.1%. It can be seen that the result from Term weighting 2 with scaling factors is closer to the real election result. It improves the result of Term weighting 1 as well, in the sense that Macron's popularity is higher than Le Pen. In this case, the count of all terms may cost one or two days to analysis it. If there are more terms, it need more time.

It is noted that the Twitter data on the final day before voting is very unbalanced. Macron got more tweets on the last day before election, about three times more than Le Pen. Usually, researchers will collect more data or resampling the data to cope with unbalanced data problem. However, more data means more supporters/attackers. Thus, resampling to balance the data will affect the result because it will balance the result at the same time. Increasing the number of tweets will also increase the number of available terms for calculating popularity. Although many of these terms may have small scores, they still contribute to the prediction accuracy. The problem of Term weighting 1 is that the number of terms is balanced artificially. It only used the top 100 terms with the highest scores for each candidate. These selected terms usually have a clear positive or negative sentiment about candidates. Thus, neutral tweets and nearly neutral tweets were not accounted in Term weighting 1 because their scores are low. It is evident that a considerable number of neutral tweets play an important role in achieving higher accuracy for Term weighting 2.

Figure 5.5 shows the average scaling factors for terms of the three types respectively. For Term weighting 1, the average scaling factor for terms of the unique type is 1.4, the highest of all the three types, it is 0.58 for terms that have same sentiment for both candidates and 0.56 for terms that have different sentiment for different candidates. For

Term weighting 2, the average scaling factor for terms of the unique type is 1.07, the highest of all the three types, it is 1.039 for terms that have same sentiment for both candidates and 1.04 for terms that have different sentiment for different candidates. It is interesting that the terms of the unique type got the highest scaling factor. As shown in Figure 5.4, using scaling factors improved Term weighting 2. However, the effect of scaling factors is quite small. Although scaling factors can make prediction results better, the number of terms is still essential for making accurate prediction in Term weighting 2. Scaling factors obviously improved the result of Term weighting 1, in which the terms of the unique type got a much higher scaling factor than the terms of the other two types.

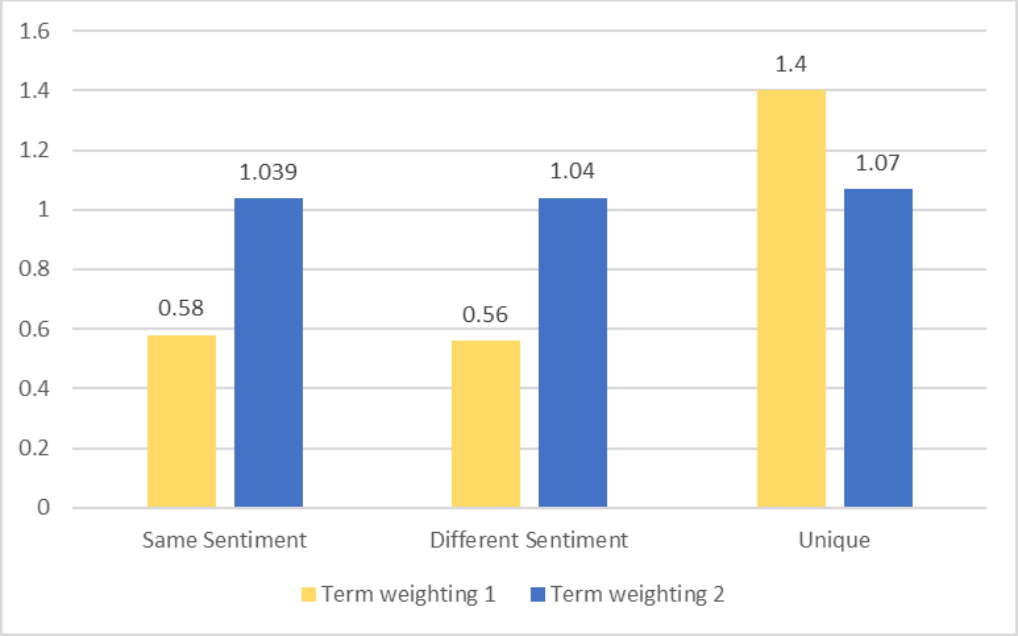


*Figure 5.5 Average scaling factors for terms that have same or different sentiment for both candidates, or candidates or are unique for one candidate only.*

It is also interesting to notice that election prediction based on Twitter data analysis is affected by the sentiment of supporters. For example, most Macron's supporters did

not attack Le Pen on Twitter, but Le Pen's supporters usually attacked Macron. This is difficult to be considered in the Twitter data analysis in general.

## 5.5 Conclusion

This chapter proposes a new method for candidate's popularity prediction based on Twitter data analysis and thus for election result prediction indirectly. The proposed method weights keywords related to specific candidates based on both statistics and domain knowledge including sentimental meanings of keywords, which has been proved to increase prediction accuracy in our case study of predicting the 2017 French election result.

# Chapter 6: Prediction of the 2017 French Election Based on Twitter Network Analysis

## 6.1 Motivation

Twitter is one of the largest social networks, providing a compact platform for people to express opinions and share views on a variety of topics and issues. Twitter data analysis based prediction has drawn much attention in recent years, especially in predicting results of political events.

Existing studies have mainly focused on sentiment analysis of a party or candidate. They neglect the fact that the voters' attitudes and opinions of people may be different depending on specific political topics and in different geographic areas. Moreover, the same voters participating in different discussions may have different political preferences. Secondly, social media may be manipulated by spammers and propagandists. Fake accounts are easy to create and they can be used to amplify the spammers message polluting the data for any observer [11]. In this chapter, we are interested in predicting the result of elections from micro-blog data by incorporating social network analysis and sentiment analysis to detect their political preferences and predict the election results.

## 6.2 Method

Social network is used as a source of data set to predict outcomes of politic events. In this chapter, a social network graph is constructed by interactions among users. In

the graph, one user is represented as a node, and “Retweet” is represented as the edge between nodes. To build a social network graph, edges and nodes can be represented as an N x N matrix, where N is the number of nodes and the values in the matrix represent whether there are connections between nodes. We input retweet matrix data to Gephi, a network visualisation software used in various disciplines, to generate a graph. Gephi is an open-source software for graph and network analysis [4]. Gephi reads nodes and edges data to build a visualisation graph of the network. Gephi can compute average degree, density and cluster coefficient of the social network. After running the Force Altes 2 algorithm, which is a continuous graph layout algorithm for handy network visualisation designed for the Gephi Software [36], the structure of the social network was built to visualise the interactions among Twitter users.

### 6.2.1 Graph Feature

To compute the candidate’s popularity, I computed the strength of their supporter’s community. Thus, social network analysis is necessary. There are several key terms associated with social network analysis: density, clustering coefficient, and degree distribution.

At first, there are some important ideas to indoruce. In the graph, there are some nodes and links. The degree is the number of links with one node.

The density of the social network is an indicator of the general level of connectedness of the social network graph. If every node is directly connected to every other node, it is a complete graph. The density of a graph is defined as the number of links between nodes divided by the number of vertices in a complete graph with the same number of nodes. For directed graphs, the graph density is defined as:

$$D = \frac{E}{N(N-1)} \tag{6.1}$$

where E is the number of links and N is the number of nodes in the graph. The density provides degrees of interaction in a community's social network graph.

Clustering coefficient is a measurement of the degree to which nodes in a graph tend to cluster together. Some important papers [5][21] suggest that in most real-world networks, nodes tend to create tightly knit groups characterised by a relatively high density of ties. If they are in one community, their tend is greater than the average probability of a tie randomly established between two nodes. The clustering coefficient is defined as:

$$Cluster\ coefficient = \frac{3 \times number\ of\ triangles}{number\ of\ all\ triplets} \tag{6.2}$$

where a triplet is three nodes that are connected by either two or three undirected ties. A triangle graph, therefore, includes three closed triplets, one centred on each of the nodes. The global clustering coefficient is the number of closed triplets over the total number of triplets. The global clustering coefficient is designed to give an overall indication of the clustering in the network. The global clustering coefficient is based on triplets of nodes. A triplet consists of three connected nodes. Based on it, I compute interactions in the community.

Centrality is another important feature of social networks. Degree centrality of a node is defined as the number of edges this node has. Closeness centrality of a node is equal to the total distance in the graph of this node from all other nodes.

Term frequency – inverse document frequency (TF-IDF) is adopted for term weighting of edge, which is a numerical statistic to reflect how important a word is to a document in a collection or corpus of documents. It is often used as a weighting factor

in information retrieval, text mining, and user modelling. The equation of term weighting is as follows [67]:

$$TF - IDF_{i,j} = tf \times idf = \frac{t_{i,j}}{\sum_k t_{k,j}} \times lg \frac{N}{|\{d \in D | i \in d\}|} \quad (6.3)$$

where *tf* is the term frequency within a document which is the number of times a term occurs in a document; *idf* is inverse document frequency within a corpus; $t_{i,j}$ is the number of times that term *i* appears in document *j*; $\sum_k t_{k,j}$ is the total number of times that all the terms under consideration appear in document *j*; *N* is the total number of documents in corpus *D*; and $|\{d \in D | i \in d\}|$ is the number of documents in which term *i* appears [39]. In this study, a term is a keyword; a document corresponds to a tweet; and a corpus corresponds to a set of tweets relevant to an individual candidate. Term weighting is calculated based on a corpus of tweets relevant to an individual candidate respectively. Based on domain knowledge, if a keyword is positive, its TF-IDF value will be positive, otherwise it is negative. After that, the total weighting score of term *i* in corpus *D* is calculated as follows:

$$TF - IDF_i = \sum_{j=1}^{N} TF - IDF_{i,j} \quad (6.4)$$

where $TF\text{-}IDF_i$ is the sum of the weighting scores of term *i* in all documents in corpus *D*, representing the weight of term *i* in a group of tweets in this study, whilst $TF\text{-}IDF_{i,j}$ is the weighting score of term *i* in document *j*.

### 6.2.2 Whole-network Method

In the whole network model, edges of twitter retweet social networks are based on retweet tweets. It is different from traditional social networks, in which edges usually are built by the connection of links such as following. However, retweet action is an

instant interaction (maybe without other connections, the user just retweet when saw the tweet). People may retweet a tweet because that they like this tweet or share interesting news/opinions. So, retweet is not a continuous interaction(retweet many times form one user, like follower) in social networks. Thus, I designed a new method to compute edge centrality based on the weighting of edges. First, term weighting is applied to each tweet in the social network:

For tweet *a*:

For each word *t*:

$$Score\ t_{a,t} = \frac{t_{a,t}}{\sum_k t_{k,t}} \times lg \frac{A}{|\{d \in D | t \in d\}|}$$

End

$$Score\ S_a = \sum_0^t t_{a,t}$$

End

where $S_a$ is the score of tweet $a$. $Score\ t_{a,t}$ is the weighting of term $t$ in tweet $a$, $A$ is the number of tweets, $|\{d \in D | t \in d\}|$ is the number of documents in which term $t$ appears. The weighting of an edge is the sum of scores between two nodes. The score of edge *j* is defined as follows:

For edge *j*:

$$\mathrm{Cj} = \sum_0^a Sa$$

End

Based on the score of edge, the centrality of edge weighting is:

$$\mathrm{Cedge, j} = \frac{Cj}{Cmax} \tag{6.5}$$

where $Cmax$ is the largest value of weighting of the single edge obtained in the network under the graph.

Edge centrality can be used to derive a centrality of the whole network. The computing formula is as follows:

$$Cedgenetwork = \frac{\sum_j (Cedge,j)}{E} \quad (6.6)$$

where *Cj* is the centrality value of edge *j* in the graph and *E* is the number of edges in the network. Based on the above defined features, the strength of a supporter's community could be calculated.

In the node level, similarities and social relations are analysed. Thus, nodes are classified based on the name of candidates. Then, every community of each candidate is classified based on tweets attitude. I computed the value of densities of different graphs and different centralities of graphs for each community. Based on structural features of social networks, I designed a model to compute the community strength of the candidate.

For each candidate:

For positive communities:

For node *i* in community:

Degree Centrality= $C1_{pos,i} = degree\ of node\ N_{pos,i}$

For edge *j* in community:

Weighting Centrality = $C2_{pos,j} = \frac{Cmax - Cj}{Cmax}$

Density= $D_{pos} = \frac{E_{pos}}{N_{pos}(N_{pos}-1)}$

Cluster coefficient= $C3_{pos} = \frac{3 \times number\ of\ triangles}{number\ of\ all\ triplets}$

Community Average Degree = $C1_{pos} = \frac{\sum_i (C1posi)}{N_{pos}}$

Community Weighting Centrality = $C2_{pos} = \frac{\sum_j (1 - C2posj)}{E_{pos}}$

Community strength= $S_{pos} = C1_{pos}f1 + C2_{pos}f2 + C3_{pos}f3 + D_{pos}f4$

For negative communities:

For node *i* in community:

Degree Centrality= $C1_{neg,i} = degree\ of node\ N_{neg,i}$

For edge *j* in community:

Weighting Centrality = $C2_{neg,i} = \frac{Cmax - Cj}{Cmax}$

Density= $D_{neg} = \frac{E_{neg}}{N_{neg}(N_{neg}-1)}$

Cluster coefficient= $C3_{neg} = \frac{3 \times number\ of\ triangles}{number\ of\ all\ triplets}$

Community Average Degree = $C1_{neg} = \frac{\sum_i (C1negi)}{N_{neg}}$

Community Weighting Centrality = $C2_{neg} = \frac{\sum_j (1 - C2negj)}{E_{neg}}$

Community strength= $S_{neg} = C1_{neg}f1 + C2_{neg}f2 + C3_{neg}f3 + D_{neg}f4$

Score of a candidate a score a = $N_{pos} \times S_{pos} - N_{neg} \times S_{neg}$

Popularity of candidate a = $\frac{score\ a}{score\ a + score\ b}\%$

*where C1* to *C3* are different features of the community social network graph. *C1* is degree centrality of whole community's network. *C2* is the new centrality of community networks. *C3* is the clustering coefficient of the graph. *D* is density of the community's social network graph. *f1* to *f4* are scaling factors of features, whose values can be determined by data fitting and they should be in the same range because the significance of each feature is similar. Besides, the number of nodes is the most important feature for a community.

### 6.2.3 Sub-network Method

Sub-network models can be designed in terms of different communities of supporters. There are two kinds of main ideas to compute the strength of the community, based on network structure or based on semantic analysis. However, if only whole network structure of a social network is considered to compute the strength of the community, it may lack consideration of the strength of important edge effect. Some inactive opinions may reflect the hidden opinion of a candidate's policy. If only semantic analysis is considered to compute community strength, it will consider less active users in the social network [39]. Thus, I used Nature Language Processing (NLP) to classify the topics of tweets and analyse the attitude of the community corresponding to each topic using a sub-network approach, which overcomes the limitations of these two methods applied individually.

The process of text categorization in nature language processing includes collecting training data, text semantic analysis, features analysis, training model, validating model. Models of documents include Probabilistic models, Boolean model, and Vector Space model. Semantic analysis uses Named entity recognition and part of speech. To extract features of texts, TD-IDF and cross entropy can help us to reduce dimensionality of features.

I computed the similarity of vectors to measure the similarity of texts. N-gram model is based on a bag of word to judge a sentence. TD-IDF model uses term frequency to classify text. How to extract features of texts is the second step. Usually, the keywords in the text are adopted as features in most text mining process. However, if I select too many words in the text, the dimensionality of features would be very large. It affects the speed of training and precision of the algorithm. Thus, I use TF-IDF, information

gain, cross-entropy and principal components analysis to reduce the dimensionality of features. After that, I need to consider the weight of features. There are many ways to compute the weighting of features, including TD-IDF, position, information entropy, lengths of words and relationship of words.

Firstly, I labelled Twitter data manually (will label data automatically in my future work). TD-IDF and knowledge about the French election help us to find important words in the tweets. The number of words can be reduced by removing stop words. I selected all relevant tweets based on my knowledge and the TD-IDF values. After that, I classified tweets into different topics. Finally, I used the available knowledge to classify the tweets in different communities into different attitudes such as positive or negative scores.

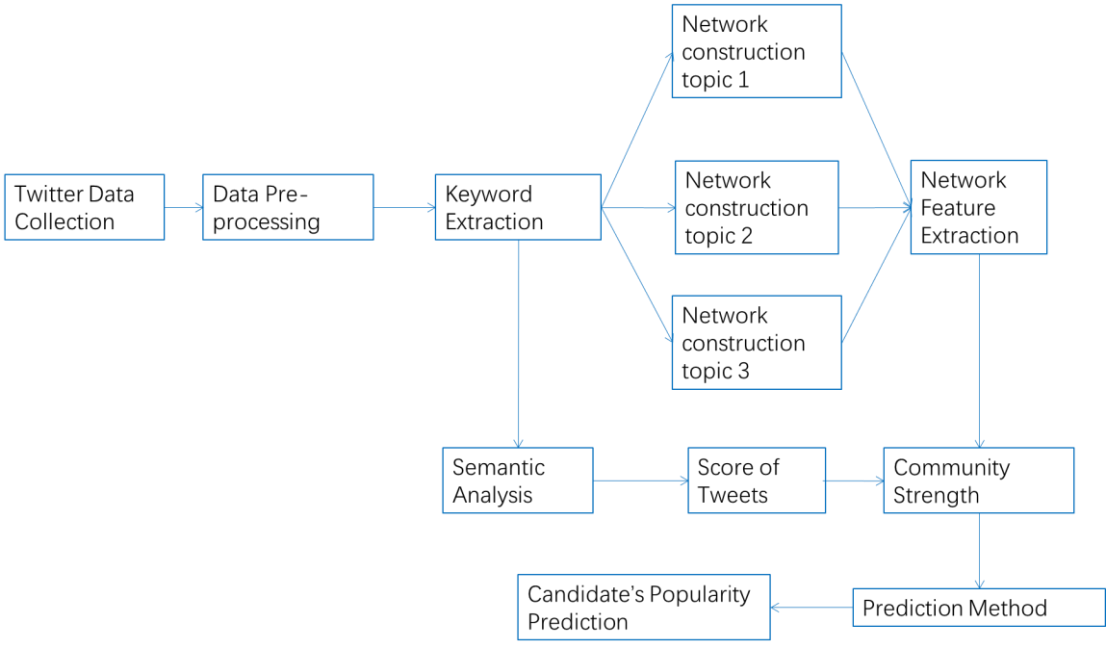


*Figure 6.1 Sub-network model*

As shown in Figure 6.1, I classified the tweets into different topics. The number of topics is determined by the knowledge about the French election. Typical topics include

finance, immigration, defence, etc.. For each topic a sub-network was generated and represented as a sub-graph. Based on these topics, I computed the strength of communities to predict the popularity of the candidate with scores of tweets and network features. Firstly, I built a tweets network for each topic by user ID and retweet action. The users are nodes and retweets are edges. I computed each sub-graph's community strength. I selected density and average degree features to describe the sub-graph. To compute the popularity of a candidate, the steps are as follows:

For each candidate:

For F topic:

For node *i* in community:

Degree Centrality= $C1 = degree\ of node\ N_i$

For edge *j* in community:

Score of edge = Scorej = semantic analysis result

Density= $D = \frac{E}{N(N-1)}$

Attitude= $A = \sum_1^E Scorej\ /E$

Community Average Degree = $C2 = \frac{\sum_i (C1i)}{N}$

Community strength= $S_F = (\mathrm{C1} + \mathrm{C2} + \mathrm{D}) \times N \times A$

Score of candidate $a$= $Score\ a = \sum_1^F S_F$

Popularity of candidate a= $\frac{score\ a}{score\ a + score\ b}$ 100%

where *Score a* is the score of candidate *a*, *F* is the number of topics/communities, *N* is the number of nodes in sub-graph, and *A* is the attitude score of the community. This method considered neutral tweets as well.

In this study, there are five topics of the Twitter data: Defence, Immigration, Attack, Finance, and Neutral. Defence topic is about the policy of national security. For

example, terror attack and budget for police are main concerns of national security. Immigration is a very important topic in recent years because Europe and US are suffering from massive refugees from middle east. One candidate, Macron, supported solving the refugee problem in peaceful way. On the other hand, another candidate, Le Pen, chose a strong way to handle the refugee problem. Attack topic is negative for each candidate. Finance topic is relevant to money. Le Pen supported exiting EU, which many people believe is a nightmare if the French exits the EU. On the other hand, Macron believed that globalisation is good for economic and development. In neutral topics, tweets are not relevant to any policy, but usually positive. For example, “vote Macron to save the French” reflects supporters for Macron. They just show a positive opinion without any reason. Therefore, the neutral topics have less weighting in the semantic analysis.

For each topic, I built a sub-network of users with retweet action. Users are represented by the nodes in the social network. Retweet action links each node in the network. One user was represented by a node in a different sub-network. Thus, the community of a sub-network will be closer than whole-network, although the construction of a sub-network is similar to that of whole-network.

After that, based on the topic of a sub-network, I made a semantic analysis of each tweet. The neutral topic is less important than policy and attack topics. Attack topic has strong negative attitude for each candidate. Thus, attack topic is negative for both candidates. In this study, the semantic factor for Defence, Attack, Finance and Immigration is 1. The semantic factor of neutral topic is uncertain.Therefore, one user could show up in different sub-network because user could discuss different topic in all days. They could make their effect in different ways.

Finally, the community strength of each sub-network is computed by the combination of social network features and semantic analysis results, in a similar way as for the whole-nework.

## 6.3 Experimental Results and Discussion

### 6.3.1 Whole-network Result

To test the whole-network model, I selected Twitter data between April 25th to 29th. For each candidate, positive and negative communities were identified by their tweets. I analysed four communities in one-day Twitter data. I randomly selected half of the data to train the scaling factors. Other data was used for validating the model.

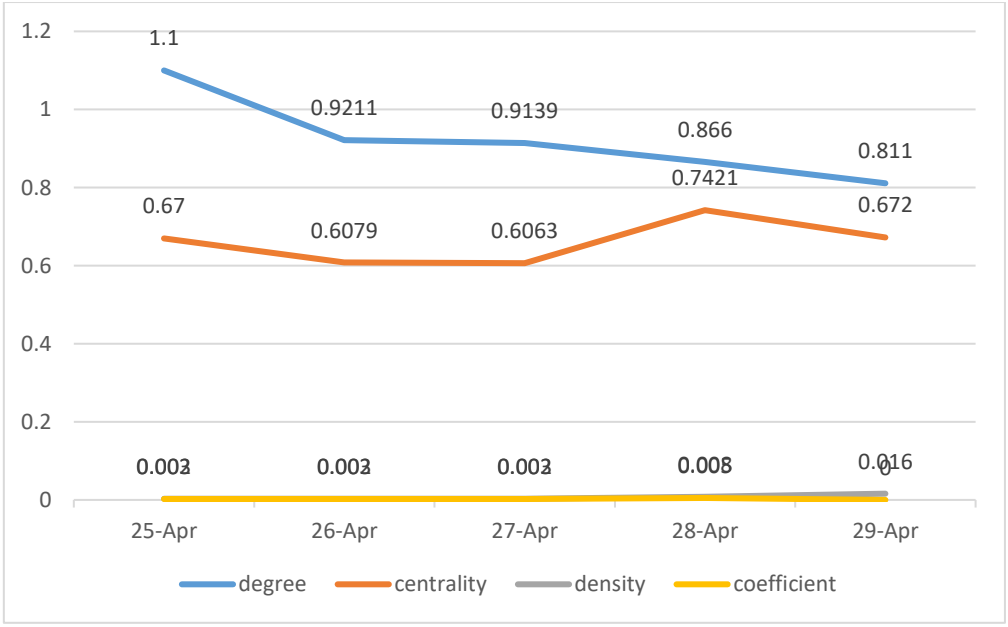


*Figure 6.2 Positive community of Le Pen during April 25-29*

Figure 6.2 shows the trend of positive communities of Le Pen during April 25th to 29th in terms of four social network features respectively. The degree of communities decreased from April 25th to April 29th. The centrality increased on April 28th. And it

can be observed that many nodes linked to one node in this graph because the degree nearly 1 but the density is low. Thus, the density and clustering coefficient are smaller than the average degree. Some center nodes in the graph reflect the supporter's opinion on April 27th. On April 29th, one node provides views in the community, while other nodes discuss it. Thus, density and clustering coefficient are very small.

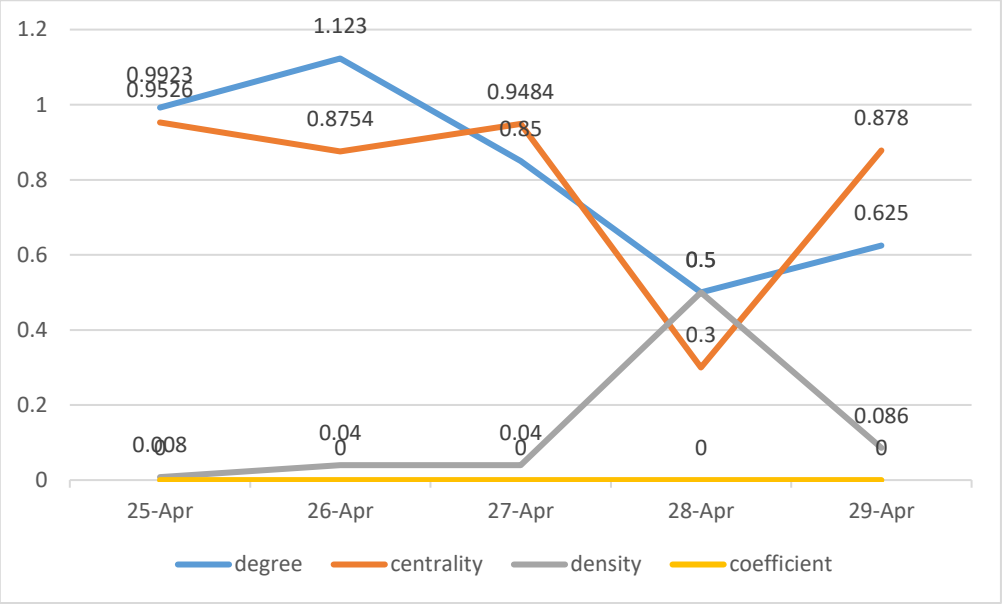


*Figure 6.3 Negative community of Le Pen during April 25-29*

Figure 6.3 shows the trend of negative communities of Le Pen during April 25th to 29th in terms of four social network features respectively. The average degree is around 0.8. The weighting centrality is 0.9526. The density is 0.008. The cluster coefficients are 0 in this period. The negative community of Le Pen got higher weighting centrality. However, one node created one important tweet in the community. The average degree changed a lot on 28th because of only there were two nodes in the community. Usually, there was one main user in the community in this period, which made the cluster coefficient close to 0. A limited number of people expressed their opinion in the negative community of Le Pen. It cannot represent a common opinion in the negative community.

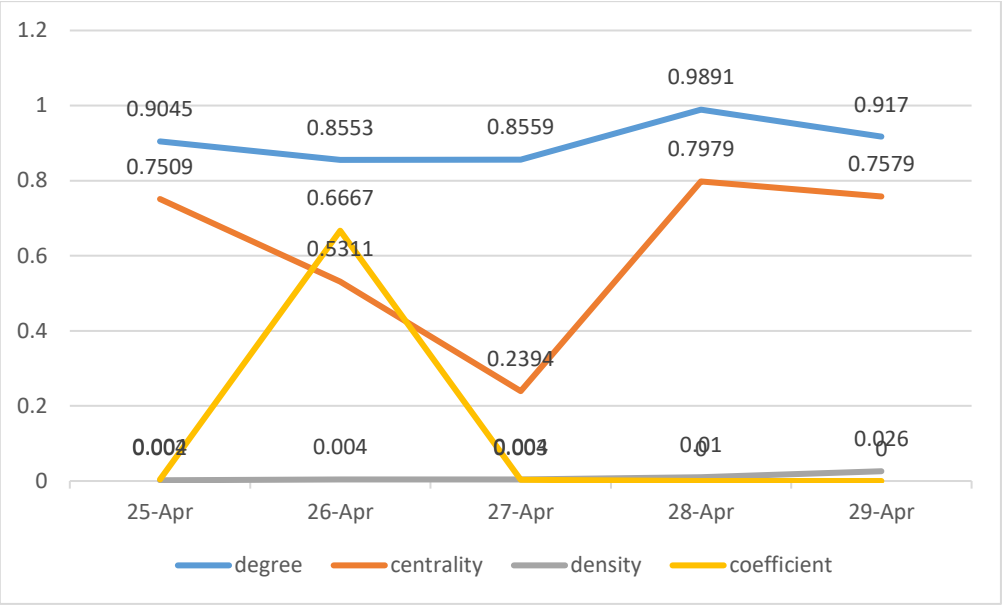


*Figure 6.4 Positive community of Macron in during April 25-29*

Figure 6.4 shows the trend of positive communities of Macron during April 25th to 29th in terms of four social network features respectively. The density and cluster coefficients are very small. The average degree and weighting centrality are very high. There was less interaction in the community. However, the frequency of community interaction was high on April 26th. Thus, cluster coefficient was larger than that of April 25th. On April 27th, a core node provided tweets in the community, but other tweets were not spare in the community. Thus, density deceased on April 27th. Most retweets were based on one node on April 28th to 29th. The retweet action are more popularity in April 28th because Macron annuanced new economic policy for election. The density and cluster coefficients were very small again.

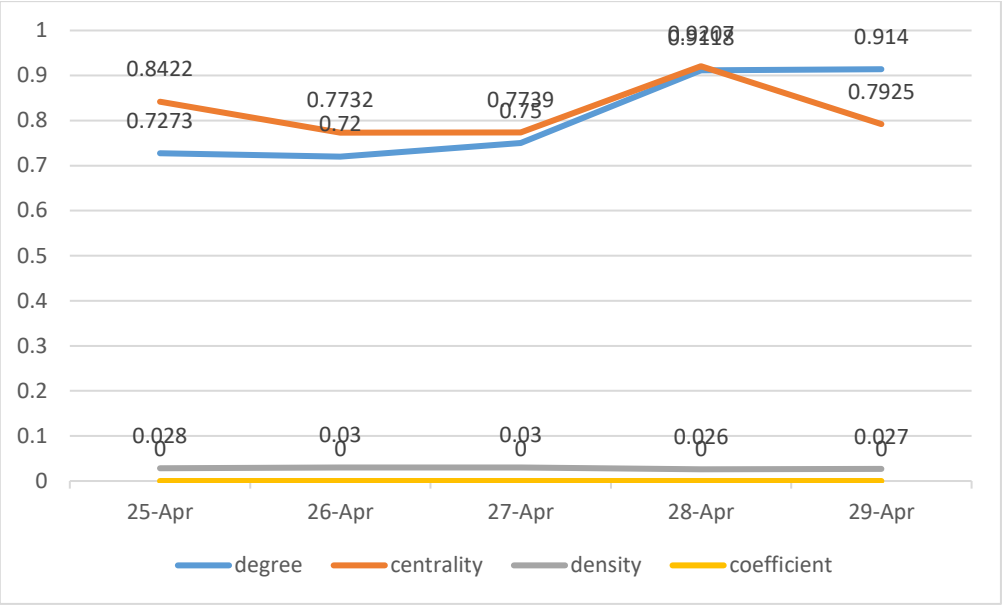


*Figure 6.5 Negative community of Macron during April 25-29*

Figure 6.5 shows the trend of negative communities of Macron during April 25th to 29th in terms of four social network features respectively. The degree and centrality are around 0.8. The density and cluster coefficient are close to 0. There were limited connections between each user. The users did not share their opinions in the community.

I made a linear regression to find values for *f1* to *f4*. The result is *popularity of candidate a* = 298.52D+44.65. However, the error rate is 78%, which is not good. As the values for *f1*, *f2* and *f3* are zero, *f4* is the only non-zero scaling factor in the resulted linear regression model. Because only 5 data points are involved and they are noisy, it is very hard to get an expected result with linear regression. Thus, I selected another way to find appropriate values for *f1* to *f4*, which is based on grid search of values, with those that produce the best accuracy as 'optimal' parameter values.

The values of cluster coefficient and density are very small in the community's features, whose ranges are often smaller than 0.1. If *f3* and *f4* are in the same range as that of *f1* and *f2*, density and cluster coefficient cannot significantly affect the prediction of popularity of candidates. The small density and cluster coefficient present

that most users retweet tweets without feedback, which leads to a small number of users in the retweet networks. The average degree and edge centrality represent the popularity of tweets. If only considering the average degree and edge centrality of the community, interaction in the community is ignored. Density and cluster coefficient are same important as other features, therefore *f3* and *f4* should be in a large range. Thus, scaling factors *f1* to *f4* have been designed in different ranges to balance the weighting of community features in the model. The range of degree centrality scaling factor *f1* is 0 to 10 with step 0.1; the range of weighting centrality scaling factor *f2* is 0 to 10 with step 0.1; the range of density scaling factor *f3* is 10 to 500 with step 10; and the range of cluster coefficient scaling factor *f4* is 10 to 500 with step 10. I searched for possible values of *f1* to *f4* in their range by by fitting the model with opinion poll results. I selected scaling factor values that make the model produce the prediction result closest to opinion poll results of all 5 days. It got minimum variance of opinion poll form April 25$^{th}$ to 29$^{th}$. The obtained values of *f1* to *f4* based on the opinion poll result from April 25$^{th}$ to 29$^{th}$ are presented in the following table:

*Table 6.1 Results of Using Scaling Factors f1 to f4*

| Date | *f1* | *f2* | *f3* | *f4* | Macron's popularity in Prediction | Le Pen's popularity in Prediction | Macron's popularity in opinion poll | Le Pen's popularity in opinion poll |
|---|---|---|---|---|---|---|---|---|
| 4.25-4.29 | 0.9 | 0.6 | 500 | 50 | 66 | 34 | 59 | 41 |

It can be seen that the average value of *f1* to *f4* in the model is *f1*= 0.9, *f2*= 0.6, *f3*= 500, *f4*= 50.

After that, I used another half data to validate the model. I set *f1* to *f4* to the above obtained values and computed the popularity of candidates using the validation data. The prediction results on the validation data are presented in the following table:

*Table 6.2 Training Result*

| Date | Macron's popularity in prediction | Le Pen's popularity in prediction | Macron's popularity in opinion poll | Le Pen's popularity in opinion poll |
|---|---|---|---|---|
| 4.25 | 49 | 51 | 59 | 41 |
| 4.26 | 74 | 26 | 60 | 40 |
| 4.27 | 79 | 21 | 60 | 40 |
| 4.28 | 67 | 33 | 59 | 41 |
| 4.29 | 68 | 32 | 61 | 39 |

The model got wrong prediction result on April 25$^{th}$ in the validation, but it worked well on April 26$^{th}$ to 29$^{th}$. The average of community in  In Arpil 25, there still many candidates expect Marcon and Le Pen in twitter. Their tweets made Macron's popularity lower.  The average size of a community is around 17.

I also used these scaling factors to predict the final result of French election 2017. I used Twitter data in May 6$^{th}$ , the last day before the election, to predict the popularity in the final round of election.

The social network of positive community of Le Pen on May 6th is shown in Figure 6.6, with dark colour representing high edge degree. The average degree is 0.855. The weighting centrality is 0.829. The density is 0.001. The cluster coefficient is 0.001. The Density and cluster coefficient are very small, indicating no strong connection among nodes.

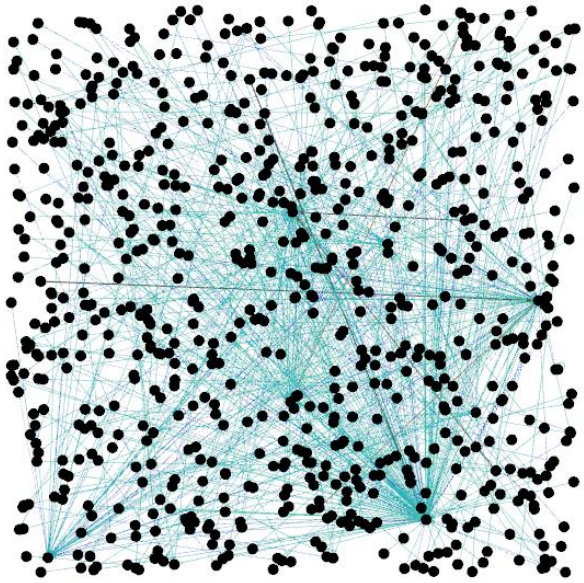

*Figure 6.6 Positive community of Le Pen on May 6th*

The social network of negative community of Le Pen on May 6th is shown in Figure 6.7. The average degree is 0.918. The weighting centrality is 0.851. The density is 0.01. The cluster coefficient is 0. The density and cluster coefficient are very small, and most users retweet from a small number of users.

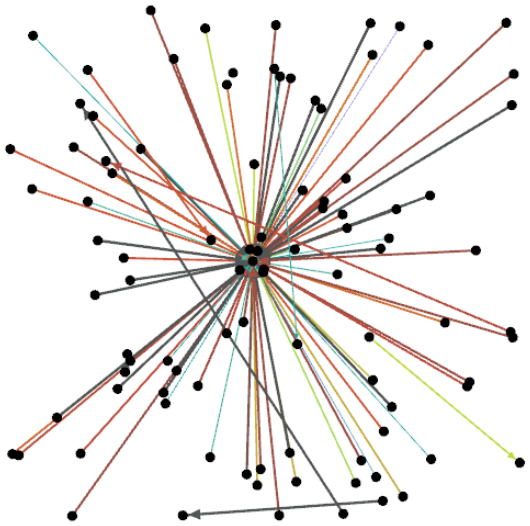

*Figure 6.7 Negative community of Le Pen on May 6th*

The social network of positive community of Macron on May 6th is shown in Figure 6.8. The average degree is 0.948. The weighting centrality is 0.8354. The density is 0. The cluster coefficient is 0.002. Low density and cluster coefficient show that the community is based on few core nodes.

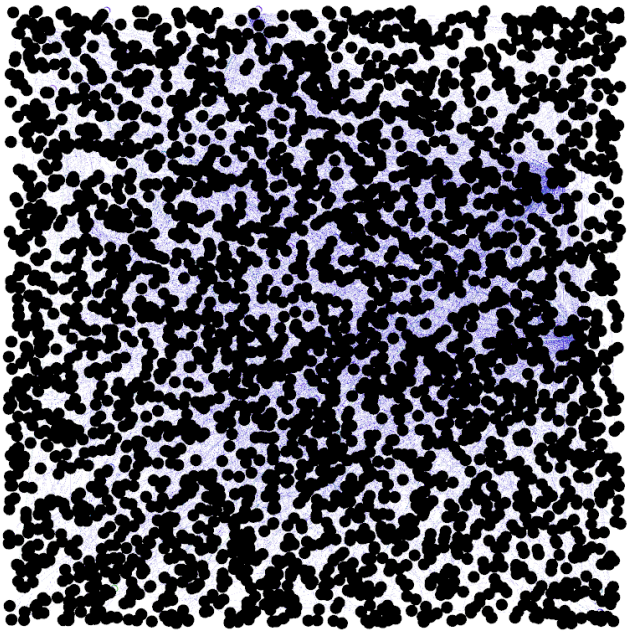

*Figure 6.8 Positive community of Macron on May 6th*

The social network of negative community of Macron on May 6th is shown in Figure 6.9. The average degree is 0.996. The weighting centrality is 0.833. The density is 0.001. The cluster coefficient is 0. Most negative opinion expressed by one main tag “Macron Leak” on May 6th.

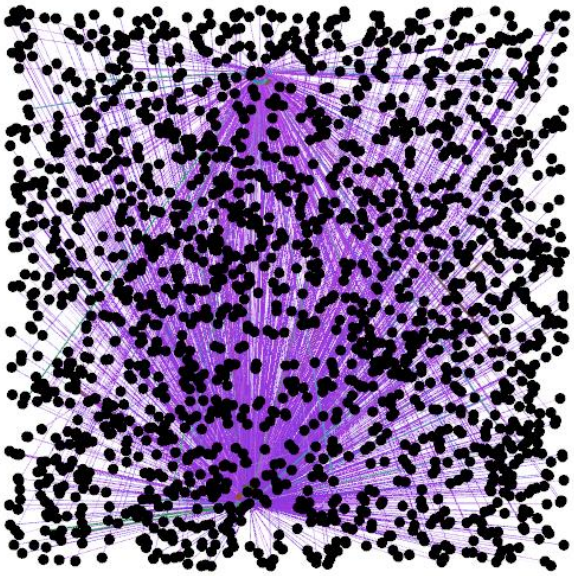

*Figure 6.9 Negative community of Macron on May 6th*

The popularity of Macron is 81% based on the built prediction model, while the popularity of Le Pen is 19%. However, the final opinion poll rate for Macron is 66.1%. The positive community of Macron has a higher number of active users on May 6th. Although community features of the positive community are not higher than negative community, the number of positive nodes still lead to higher Macron’s popularity, because the number of nodes in the community is the most important element in the model. In the next section, better prediction results would be expected using the sub-network approach.

### 6.3.2 Sub-Network Result

In the sub-network model, there are five main topics for each candidate. The neutral topic semantic factor *f1* was determined by the Twitter data from April 27th and 29th, in such a way that the prediction results match the opinion poll results. In our study, various combinations of values of *f1* ranging from 0.1 to 1 were tested with a changing step of 0.1, and those values that resulted in the best match between the predicted popularity and the opinion poll result were selected. It got a minimum variance between opinion poll from April 27th and 29th with a semantic factor of 0.6. The sub-network method predicted the popularity of Macron as 57% and 54% on April 27th and 29th respectively.

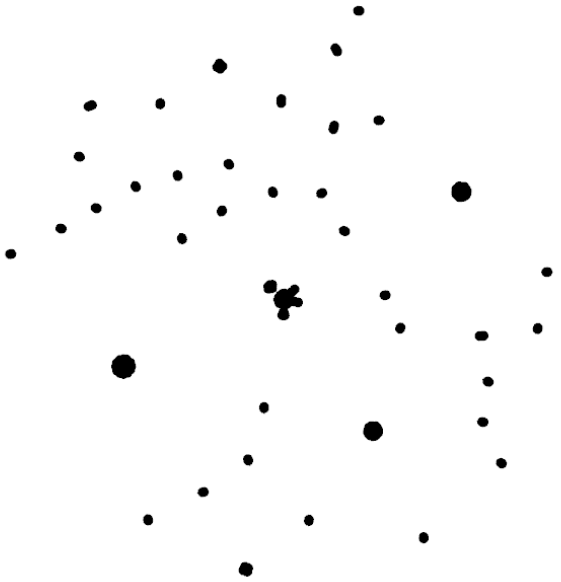

*Figure 6.10 Sub-network of Macron based on Finance topic on May 6th*

The sub-network of Micron based on the finance topic on May 6th is shown in Figure 6.10. There are many small communities in the finance topic. Tax and globalisation are the main concerns. Most people thought that Macron's policy was good for French. It can be seen that the communities under the finance topic lack interaction.

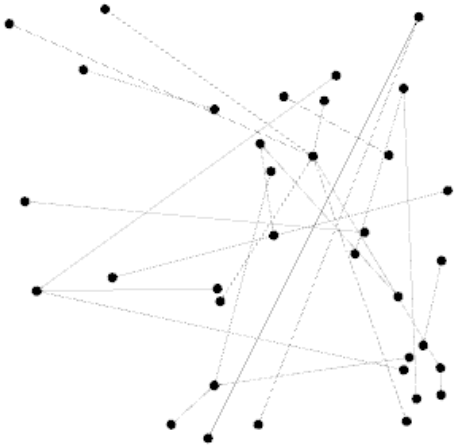

*Figure 6.11 Sub-network of Macron based on Immigration topic on May 6th*

The sub-network of Micron based on the immigration topic on May 6th is shown in Figure 6.11. There were only a few users focusing on the immigration policy of Macron. Macron chose not to reject refugees to enter France. Most of the refugees were from Muslim countries and anti-Muslim community was an active power on Twitter. The interaction of this small topic community is active as shown in the figure.

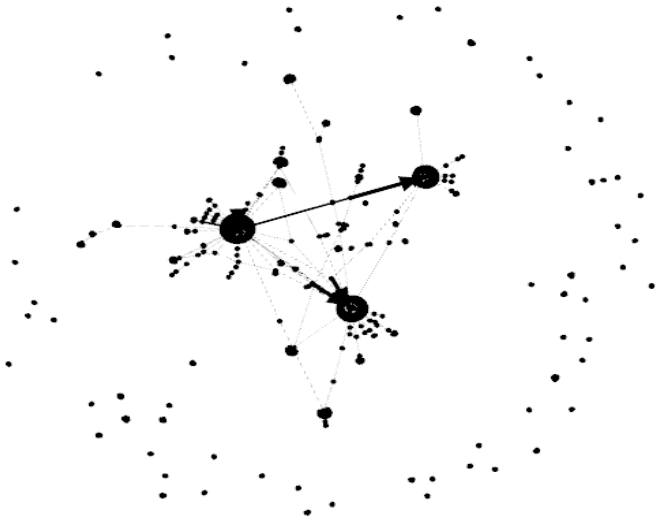

*Figure 6.12 Sub-network of Macron based on neutral topic on May 6th*

The sub-network of Micron based on the neutral topic on May 6th is shown in Figure 6.12. Many people supportted Macron with a positive attitude. There are three main communities on the neutral topic. Each of them suggested and shared news of Macron. Although many small tweets only have 2-3 retweets on it, three main communities gather a large group of users as shown in Figure 6.12.

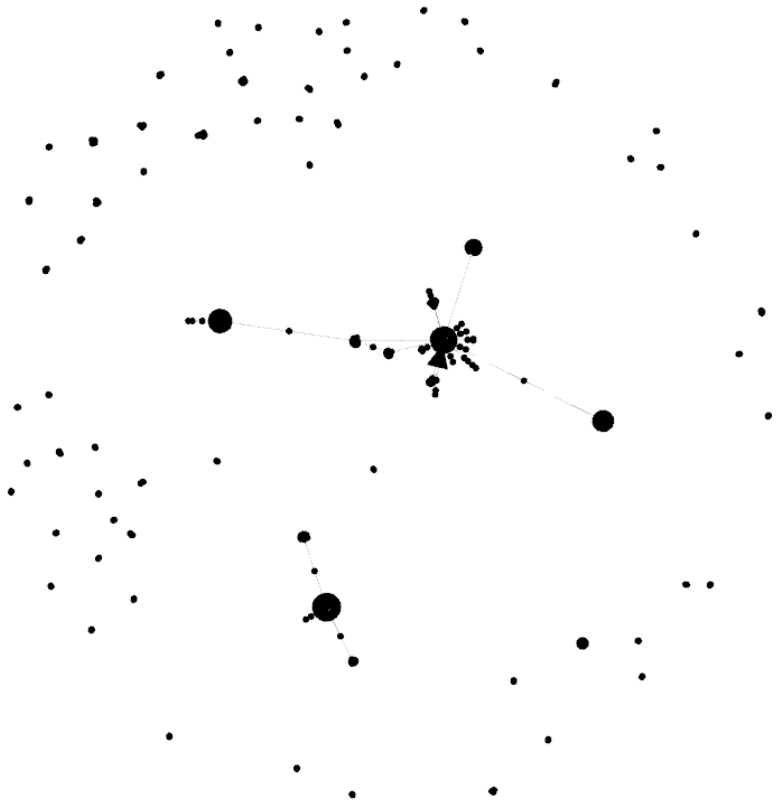

*Figure 6.13 Sub-network of Macron based on Defence topic on May 6th*

The sub-network of Micron based on the defence topic on May 6th is shown in Figure 6.13. There were more interactions in this sub-network. Macron was strongly against terrorist and related organizations. This was positive for Macron. There were 4 main communities linked to each other. Other tweets got 2-6 retweets in this sub-network.

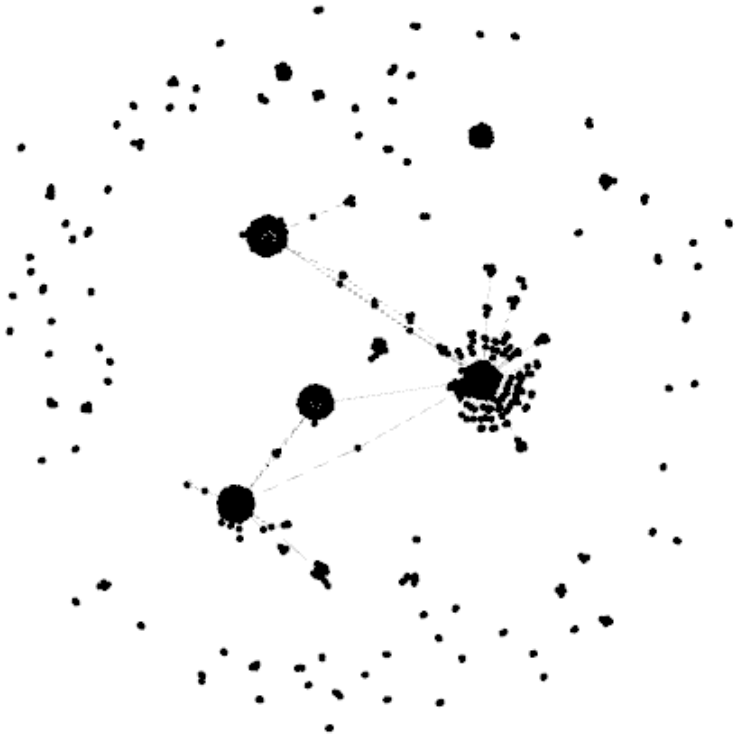

*Figure 6.14 Sub-network of Macron based on Attack topic on May 6th*

The sub-network of Micron based on the attack topic on May 6$^{th}$ is shown in Figure 6.14. Attack is a hot topic for Macron. Many nodes formed 4 big communities under the attack topic. “Macron Leak” was one main tag of attacker’s tweets, which was fake news. Attackers tried to show the relationship between Macron and Wall Street. They believed that poor people cannot benefit from Macron’s policy. Thus, there were more interactions between these communities.

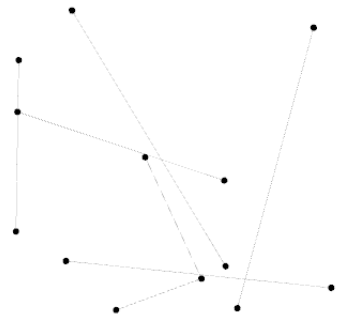

*Figure 6.15 Sub-network of Le Pen based on Finance Topic on May 6th*

The sub-network of Le Pen based on the finance topic on May 6th is shown in Figure 6.15. Only a small number of users were concerned with the finance policy of Le Pen. Compared to other topics, under the finance topic Le Pen got less attention. Users discussed it in very small groups. Most tweets got 1 or 2 retweets on this topic, and most tweets on finance policy of Le Pen were negative because Le Pen supportted exiting from the EU.

*Figure 6.16 Sub-network of Le Pen based on neutral topic on May 6th*

The sub-network of Le Pen based on the neutral topic on May 6th is shown in Figure 6.16. The number of neutral topic tweets for Le Pen is smaller than that for Macron. It reflects that the number of supporters is small. There is no interaction between some small communities. The density and degree are at a low level.

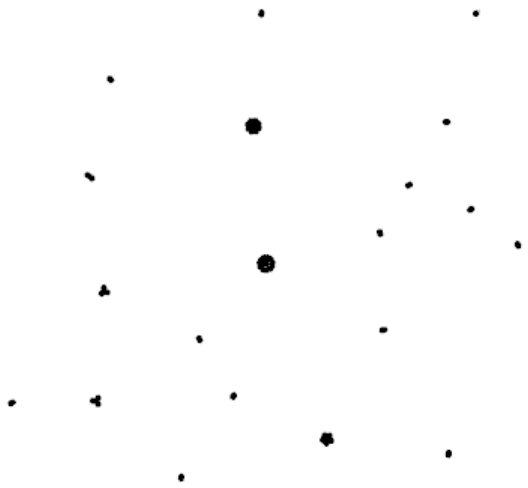

*Figure 6.17 Sub-network of Le Pen based on Defence topic on May 6th*

The sub-network of Le Pen based on the defence topic on May 6th is shown in Figure 6.17. Unlike Macron, under the defense topic the sub-network of Le Pen has less connection. The user number of each community and the strength of the community are also at a low level.

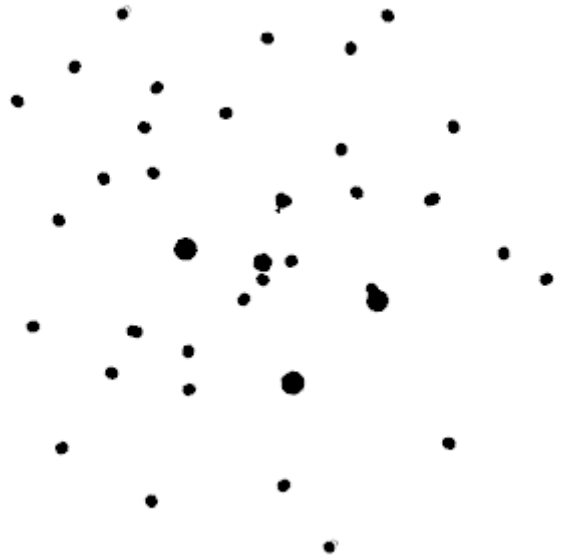

*Figure 6.18 Sub-network of Le Pen based on Attack topic on May 6th*

The sub-network of Le Pen based on the attack topic on May 6$^{th}$ is shown in Figure 6.18. There were a few communities, but they did not make connections to each other. Compared to Macron, the size of this community is the main point of the negative opinion of Le Pen.

From the above sub-network figures, it can be seen that the communities in the sub-networks of Macron were larger and more interactive in general. Following the steps of the sub-network approach for calculating the popularity of candidates, the score of Macron is 126, and the score of Le Pen is 84. Thus, the popularity of Macron is 67.1%, and the popularity of Le Pen is 32.9%. It is really close to the final voting result on the election day. The real voting result of the 2017 French presidential election is 66.1% for Macron and 33.9% for Le Pen. Comparing the predicted results of the whole-network method and the sub-network method on the final day before election, the predicted popularity by the sub-network method is about 1% different from the real

voting result, whilst the prediction by the whole-network method is about 15% away from the real voting result.

*Table 6.3 Final Prediction*

| | Training set result of the popularity of Macron | Final day predicted popularity of Macron | Opinion poll of popularity of Macron in Training set | Real result of the popularity of Macron |
|---|---|---|---|---|
| Whole-network method | 66% | 81% | 59% | 66.1% |
| Sub-Network method | 55.5% | 67.1% | 59% | 66.1% |
| Tumasjan's method | | 28% | | 66.1% |

## 6.4 Conclusion

On the final day before the election, Macron got an Internet attack named "Macron leaks". On Twitter, it was the hottest topic just before the election, which resulted in many tweets related to Macron, although the mainstream media clarified that "Macron leaks" was fake news. The problem of the whole-network method with "Macron leaks" is that it was easily affected by the hottest event. There were thousands of tweets tagged "Macron leaks", which were negative to Macron. However, most of it refers to a small number of nodes. It shows that there was not any strength between fake news and twitter users. But there were many neutral tweets related to Macron due to "Macron leaks" at the same time that were not considered by the whole-network method. Neutral tweets have propaganda effect on the election. Swing voters might most likely post neutral tweets, which should not be neglected as they bring in uncertainty for the election.

The sub-network method classified tweets in terms of different topics. It provides more detailed semantic factor rather than a scaling factor for the network feature. The

sub-network method could analyse different topic networks, which provides more network features and semantic factors. Using sub-networks we can deeply analyse the structure of social networks to predict the final result more accurately. In the whole-network method, the interactions in the communities are not classified. The whole-network is easily affected by massive interactions, which may lead to failure in the predict the outcome of events. For more discussion, the graph features may hint some information of election result. A higher centred graph shows a wide boardcast of information which may lead to win. A uni-centred graph means the information is not widely boardcast. Only few people discuss it which is not enough exposure for candidate. The candidate should expect a large centred graph in the election.

# Chapter 7: Conclusions and Future Work

## 7.1 Conclusions

This thesis proposes three methods for candidate's popularity prediction based on Twitter data analysis and thus for election result prediction indirectly.

The first proposed method considers neutral tweets related to specific candidates, which has been proved to increase prediction accuracy in my case study of predicting the 2017 French election result. It is a direct method to consider the neutral opinion in the tweets. However, it is a simple way to improve the accuracy of the current method.

The second proposed method weighs keywords related to specific candidates based on both statistics and domain knowledge including sentimental meanings of keywords, which has been proved to increase prediction accuracy in our case study of predicting the 2017 French election result. In more cases with large data and context, the domain knowledge and manually labelled method may not work. An automatic semantic analysis algorithm is developed for weighting keywords to handle large data.

The third proposed method is based on retweet of twitter users to specific candidate's topics in the election. It classifies the topics by knowledge and frequent words in tweets. It classifies the users to different communities and beads on community strength based on retweet networks constructed using graph theory, which has been proved to increase prediction accuracy in my case study of predicting the 2017 French election result.

## 7.2 Future Work

There are some limitations in the proposed methods in this thesis. In this section, I explore and discuss some potential future directions for the prediction of social events based on social media and network analysis.

This thesis uses a naive way to classify negative and positive opinions in the tweets. This could be improved by classifying tweets using a rating system. In this way, each tweet will have a score, based on which the model for popularity prediction could be more accurate.

There are still many challenges in semantic analysis especially in social events. The positive and negative opinions can be interfered with by the campaign slogan. The same words may represent different opinion to candidates. For example, the Trump slogan is “make US great” which can reduce the “great” score in the prediction. Thus, the semantic analysis model should have the ability to classify which tweets are real support and what is just a slogan to assign different scores. The combination of different topics also is a big challenge in semantic analysis.

# References


[1] C.C. Aggarwal, "An introduction to social network data analytics," *Social Network Data Analytics*, vol. 1, pp. 1-2, 2011.

[2] R. Agrawal, and R. Srikant, "Fast algorithms for mining association rules," *Proceedings of the 20th International Conference on Very Large Data Bases*, 1994, pp. 487-499.

[3] R. Agrawal, S. Rajagopalan, R. Srikant, and Y. Xu, "Mining newsgroups using networks arising from social behavior," *Proceedings of the 12th International Conference on World Wide Web*, 2003, pp. 529-535.

[4] L. Arras, F. Horn, G. Montavon, K. R Müller., and W. Samek. "What is relevant in a text document?: An interpretable machine learning approach," *PloS ONE*, vol 12, pp. 181-142, 2017.

[5] G. Avello, PT. Mustafara. "Limits of electoral predictions using twitter," *Proceedings of the 5th International AAAI Conference on Weblogs and Social Media.* 2011

[6] L. Backstorm, J. Kleinberg, "Romantic partnerships and the dispersion of social ties: a network analysis of relationship status on facebook," *Proceedings of the 17th ACM conference on Computer supported Cooperative Work & Social Computing*, 2014, pp. 831-841.

[7] L. Backstrom, and J. Leskovec, "Supervised random walks: rredicting and recommending links in social networks," *Proceedings of the 4th ACM International Conference on Web Search and Data Mining*, 2011, February, pp. 635-644.

[8] L. Backstrom, J. Kleinberg, R. Kumar, and J. Novak, "Spatial variation in search engine queries," *Proceedings of the 17th international conference on World Wide Web*, 2008, pp. 357-366.

[9] M. Bastian, S. Heymann.and, M. Jacomy, "Gephi: an open source software for exploring and manipulating networks," *Proceedings of the 3rd International AAAI Conference on Weblogs and Social media*, 2009, pp. 361-362.

[10] H. Becker, M. Naaman, L. Gravano. "Learning similarity metrics for event identification in social media," *Proceedings of the 3rd ACM International Conference on Web Search and Data Mining*. 2010, pp. 291-300.

[11] A. Bermingham and A.F.Smeaton, "On using Twitter to monitor political sentiment and predict election results," *Proceedings of Workshop on Sentiment Analysis where AI meets Psychology*, 2011, pp. 2-8.

[12] N. Benchettara, R., Kanawati, and C. Rouveirol, "Supervised machine learning applied to link prediction in bipartite social networks," *Proceedings of the 3rd Social Networks Analysis and Mining,* 2010, pp. 326-330.

[13] P. Burnap, R. Gibson, L. Sloan, R. Southern, and M. Williams, "140 characters to victory?: using twitter to predict the UK 2015 general election," *Journal of Electoral Studies*, vol. 41, pp. 230-233, 2016.

[14] E. Cambria. "Affective computing and sentiment analysis," *IEEE Intelligent Systems*, vol. 31, pp. 102-107, 2016

[15] M.P. Cameron., P. Barrett, and B. Stewardson, "Can social media predict election results? evidence from new zealand," *Journal of Political Marketing,* vol. 15, pp. 416-432, 2016.

[16] M.A. Casteleiro, G. Demetriou, W.J. Read, M. J. F., Prieto, D. Maseda-Fernandez, G.Nenadic, and R.Stevens, "Deep learning meets semantic web: a feasibility study with the cardiovascular disease ontology and PubMed citations," *Online Drugs Licensing System,* pp. 1-6, 2016.

[17] J. Cannady. "Artificial neural networks for misuse detection," *Proceedings of the National information systems security conference.* 1998, pp. 368-81.

[18] B. Chang, T. Xu, Q. Liu, and E. H. Chen. "Study on information diffusion analysis in social networks and its applications," *International Journal of Automation and Computing*, pp. 1-26, 2018.

[19] G. Chowdhury, *Introduction to Modern Information Retrieval.* Facet Publishing. 2010.

[20] A. Clauset, C. Moore, M E J. Newman, "Hierarchical structure and the prediction of missing links in networks," *Nature*, vol. 453, pp. 98-101, 2008.

[21] B. O. Connor, R. Balasubramanyan, B.R. Routledge, and N.A. Smith, "From tweets to polls: linking text sentiment to public opinion time series," *Proceedings of the 4th International Conference on Weblogs and Social Media,* 2010, pp. 122–129.

[22] E.M. Cody, A.J. Reagan, P.S. Dodds, and C.M. Danforth, "Public opinion polling with Twitter," *arXiv preprint,* arXiv:1608.02024, 2016.

[23] E.M. Daly, & M. Haahr, "Social network analysis for information flow in disconnected delay-tolerant manets," *Transactions on Mobile Computing,* vol. 8, pp. 606-621, 2009.

[24] E.M. Daly, M. Haahr, "Social network analysis for routing in disconnected delay-tolerant manets," *Proceedings of the 8th ACM International Symposium on Mobile Ad Hoc Networking and Computing,* 2007, pp. 32-40.

[25] KL. Dave, D. Pennock, "Mining the peanut gallery: opinion extraction and semantic classification of product reviews," *Proceedings of the 12th International Conference on World Wide Web, Budapest, Hungary*, 2003. pp. 519-528.

[26] B. De Longueville, R.S. Smith, and G. Luraschi, "Omg, from here, I can see the flames! a use case of mining location based social networks to acquire spatio-temporal data on forest fires," *Proceedings of the 2009 International Workshop on Location based Social Networks*. 2009, pp. 73-80.

[27] D. Gayo-Avello, P.T. Metaxas, and E. Mustafaraj, "Limits of electoral predictions using Twitter," *Proceedings of the 5th International Conference on Weblogs and Social Media*, 2011, pp. 178–18.

[28] M. Girvan, E.M.J. Newman, *"*Community structure in social and biological networks," *Proceedings of the National Academy of Sciences*, 2002, pp. 7821-7826.

[29] J. Golbeck and D. Hansen, "A method for computing political preference among Twitter followers," *Journal of Social Networks*, vol. 36, pp. 177-184, 2014.

[30] B. Ghanem, P. Rosso, and F. Rangel, 2020. "An emotional analysis of false information in social media and news articles." *ACM Transactions on Internet Technology*, 20(2), 1-18, 2020.

[31] G. Fu, Y. Ding, A. Seal ,B. Chen, Y. Sun, & E. Bolton, "Predicting drug target interactions using meta-path-based semantic network analysis," *BMC Bioinformatics*, vol. 17, pp. 160, 2016.

[32] V. Hatzivassiloglou, K.R. McKeown, "Predicting the semantic orientation of adjectives," *Proceedings of the 8th Conference on European chapter of the Association for Computational Linguistics* 1997, pp. 174-181.

[33] B. Highton, "Updating political evaluations: policy attitudes, partisanship, and presidential assessments," *Journal of Political Behavior*, vol. 34, pp. 57–78, 2012.

[34] T. Hofmann, "Unsupervised learning by probabilistic latent semantic analysis," *Machine Learning*, vol. 42, pp. 177-196, 2001.

[35] M. Hu, and B. Liu, "Mining and Summarizing Customer Reviews," *Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 2004, pp. 168-177.

[36] M. Jacomy, T.Venturini, S. Heymann, and M. Bastian, "ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software," *PloS one*, vol. 9, pp. 986, 2014.

[37] C. Jiang, W. Tao, A. Ralph, P. Joseph and H. Chapel. “A distributed decision tree algorithm and its Implementation on big data platforms,” *Proceedings of 2016 IEEE International Conference Data Science and Advanced Analytics,* 2016, pp. 752-761

[38] D. Jurgens, T. Finethy, J. McCorriston, Y. T. Xu, D. Ruths, “Geolocation prediction in twitter using social networks: a critical analysis and review of current practice,” *Proceedings of the 9th International Conference on Weblogs and Social Media,* 2015, vol. 15, pp. 188-197.

[39] M. Latapy, T. Viard, C. Magnien “Stream graphs and link streams for the modeling of interactions over time,” *Social Network Analysis and Mining,* vol. 8, pp. 61, 2018.

[40] J. Leskovec, D. Huttenlocher, and J. Kleinberg, “Predicting positive and negative links in online social networks,” *Proceedings of the 19th International Conference on World Wide Web,* 2010, pp. 641-650.

[41] H.T. Le, G.R. Boynton, Y. Mejova, Z. Shafiq, and P. Srinivasan, “Revisiting the american voter on twitter,” *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems,* 2017, pp. 4507-4519.

[42] A. Livne, M P.Simmons, E. Adar, et al. “The party is over here: structure and content in the 2010 election,” *Proceedings of the 5th International Conference on Weblogs and Social Media,* 2011, pp. 17-21.

[43] D. Liben-Nowell, and J. Kleinberg, “The link-prediction problem for social networks,” *Proceedings of Conference on Information and Knowledge Management,* 2003, pp. 556-559.

[44] D. Liben-Nowell, J. Kleinberg. “The link-prediction problem for social networks,” *Journal of the Association for Information Science and Technology,* vol. 58, pp. 1019-1031, 2007.

[45] D. Lu, A. Shah, A. Kulshrestha “India's TwitterElection,” *Stanford University Social and Information Network Analysis,* 2014.

[46] D. Mahata, J. Friedrichs, R. R. Shah, J. Jiang, “Detecting personal intake of medicine from Twitter,” *IEEE Intelligent Systems,* vol. *33*, pp. 87-95. 2018

[47] L. Man , Tan C.L., Su, J. and Lu, Y. “Supervised and traditional term weighting methods for automatic text categorization,” *IEEE Transaction of Pattern Analysis and Machine Intelligence,* vol.31, pp721-735, 2009.

[48] P. Melville, W. Gryc, and R. D. Lawrence. "Sentiment analysis of blogs by bombining lexical knowledge with text classification," *Proceedings of the 15th*

*ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Paris, France,* 2009, pp.1275-1284.

[49] P. Melville., V. Sindhwani, R.D Lawrence, E. Meliksetian, Y Liu, P.Y. Hsueh, and C. Perlich, "Machine learning for social media analytics," *Watson Research Center, IBM.* 2010.

[50] S. Natarajan, M. Moh, "Recommending news based on hybrid user profile, popularity, trends, and location," *Proceedings of 2016 International Conference on Collaboration Technologies and Systems,* 2016, pp. 204-211.

[51] J. O'Madadhain, J. Hutchins, P. Smyth. "Prediction and ranking algorithms for event-based network data," *ACM SIGKDD Explorations Newsletter,* vol. 7, pp. 23-30, 2005.

[52] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining," *Proceedings of the 7th International Conference on Language Resources and Evaluation,* 2010, pp. 1320-1326.

[53] B. Pang, and L. Lee, "Seeing stars: exploiting class Relationships for sentiment categorization with respect to rating scales," *Proceedings of the Association for Computational Linguistics,* 2005, pp. 115–124.

[54] B. Pang, L. Lee, et al., "Opinion mining and sentiment analysis," *Foundations and Trends R in Information Retrieval*, pp. 1-135, 2008.

[55] M. Pennacchiotti, and A. M. Popescu, "A machine learning approach to Twitter user classification," *Proceedings of International Conference on Web and Social Media*, 2011, pp. 281-288.

[56] S. Peng. G. Wang. D. Xie. "Social influence analysis in social networking big data: opportinities and challenges," *IEEE Network,* vol. 31, pp. 11-17,2017.

[57] B. Pete, G. Rachel, S Luke, S. Rosalynd, and W. Matthew, "140 characters to victory?: using Twitter to predict the UK 2015 general election," *Journal of Electoral Studies,* vol. 41, pp. 230-233, 2016

[58] F. Pimenta, D. Obradovic, A. Dengel. "A comparative study of social media prediction potential in the 2012 U.S. republican presidential preelections," *Proceedings of 3rd International Conference on Cloud and Green Computing,* 2013. pp, 226-232.

[59] E. Paulis, (2020). "Using social network analysis (sna) to study members and activists of political parties." *Bulletin de méthodologie sociologique: BMS,* 147-148(1-2), 13-48, 2020

[60] Y. J. Pan , Y. C. Chen, S. R. Lu, K. D. Juang and S. J. Wang, "The influence of friendship on migraine in young adolescents: a social network analysis." *Cephalalgia,* 2020.

[61] Q. Qian, M. Huang, J. Lei, X. Zhu, "Linguistically regularized lstms for sentiment classification," *arXiv preprint,* arXiv:1611.03949.

[62] T. Sakaki, M. Okazaki, and Y. Matsuo, "Earthquake shakes Twitter users: real-time event detection by social sensors," *Proceedings of the 19th International Conference on World Wide Web,* 2010, pp. 851-860.

[63] G.Salton, and C. Buckley, 1988. "Term-weighting approaches in automatic text retrieval," *Information Processing & Management*, vol. 5, pp. 513-523.

[64] E.T.K. Sang, J. Bos, "Predicting the 2011 Dutch senate election results with Twitter," *Proceeding in EACL Workshop on Semantic Analysis in Social Media*, 2012 pp. 53–60

[65] M. A. Sato, "Fast learning of on-line EM algorithm," *published by ATR Human Information Processing Research Laboratories,* 1999.

[66] J. Scott. *Social Network Analysis*, Sage, 2017.

[67] A. Segatori, F. Marcelloni, and W. Pedrycz, "On distributed fuzzy decision trees for big data," *IEEE Transactions on Fuzzy Systems*, vol. 99, pp. 1, 2017.

[68] P. Sobkowicz, M. Kaschesky, and G. Bouchard, "Opinion mining in social media: modeling, simulating, and forecasting political opinions in the web," *Proceedings of Government Information Quarterly*, 2012,vol. 29, pp. 470-479.

[69] K. S. Tai, R. Socher, C. D. "Manning, improved semantic representations from tree-structured long short-term memory networks," *arXiv preprint,* arXiv:1503.00075.

[70] A. Tripath, A. Agrawal and S.L. Rath. "Classification of sentiment reviews using n-gram machine learning approach," *Expert Systems with Applications*, vol. 57, pp. 117-126. 2016.

[71] A. Tumasjan, T. Sprenger, P.G. Sandner, and I.M. Welpe, "Predicting elections with Twitter: what 140 characters reveal about political sentiment," *Proceedings of the 4th International AAAI Conference on Weblogs and Social Media,* 2010, pp. 178–185.

[72] P. Turney, "Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews," *Proceedings of the Association for Computational Linguistics (ACL), Philadelphia, US*. 2002, pp. 417–424.

[73] M. Tsikerdekis, “Real-time identity deception detection techniques for social media: Optimizations and Challenges,” *IEEE Internet Computing*, vol. 22, pp. 35-45, 2017.

[74] K. Vadim., A. Stevens, and V.S. Subrahamian, “Using Twitter sentiment to forecast the 2013 Pakistani election and the 2014 Indian election,” *IEEE Intelligent Systems*, vol. 30, pp. 2-5, 2015.

[75] D.J. Watts and S. Strogatz. “Collective dynamics of 'small-world' networks,” *Nature*,vol. 393, pp. 440-442, 1998.

[76] L. Wang, and J.Q. Gan, 2017. “Prediction of the 2017 French election based on Twitter data analysis,” *Proceedings of the 9th Computer science and Electronic Engineering Conference (CEEC),* Colchester, UK, pp. 492–499.

[77] D. Zhang, C. HSU, M. Chen, Q. Chen, N. Xiong, and J. Lloret, “Cold-start recommendation using bi-clustering and fusion for large-scale social recommender systems,” *IEEE Transactions on Emerging Topics in Computing*, vol. 4 pp. 239-250, 2014

[78] D. Zhou, I. Councill, H. Zha, and C. Giles, “Discovering temporal communities from social network documents. data mining,” *Proceedings of the 7th IEEE International Conference*, 2007, pp. 745 – 750.

# Appendix: Papers published

L. Wang, and J.Q. Gan, "Prediction of the 2017 French election based on Twitter data analysis," *Proceedings of the 9th Computer science and Electronic Engineering Conference (CEEC)*, Colchester, UK, 2017, pp. 492–499.

L. Wang, and J.Q. Gan, "Prediction of the 2017 French election based on Twitter data analysis using term weighting," *Proceedings of the 10th Computer science and Electronic Engineering Conference (CEEC)*, Colchester, UK, 2018, pp. 231-235.