SENTIMENT ANALYSIS ON SOCIAL NETWORK USING TWITTER DATASETS

K Karthick
Department of Computer Science, St. Jerome's College of Arts and Science, India

Abstract

This research presents a unique software based on the programming language Twitter API and R. Twitter keywords are searched to get relevant tweets. Twitter APIs and Rs programming may extract these rich-opinion data sets about the contents of tweets, tweet writers, and tweets. This program has been expanded to geographical location search and post-time search in order to gather more complete Twitter feelings about political and economic problems. A new text pre-processing technique is suggested and being explored for Twitter data. The tweets collected may include a range of information about interference in many languages. This research presented for the first time a hybrid model for the categorization of Twitter sentiment. The performance of the Twitter polarity classification will be improved by combining it with a new feature chosen method based on the NRC lexicon and the classic classification algorithms KNN and Nave Bayes. The findings are assessed and verified.

Keywords:

Optimization, Natural Language Processing, Twitter Datasets, Sentiment Analysis

1. INTRODUCTION

Modern technologies are an important way to enhance the efficiency of industrial management by modelling and predicting non-linear and non-stationary processes in various fields of study. A model for connecting system inputs and outputs should be designed in order for the connection between those variables to be revealed [1]. System models and non-stationary system modelling and analysis may roughly describe many systems by simply linear or non-linear system models. For instance, there are numerous processes in the current world that are typically significantly non-linear and fluctuate in time [2]. A Wavelet is a math function which describes either the period domain or the frequency domain of a signal or time series. This has made it possible to use Wavelet theory in many fields, such as signal processing and data modelling. The Wavelet model used for pre-processing signals in nonlinear domains is specified as a mathematical model [3]. Models based on wavelets may be used to show the underlying dynamics of non-linear and non-stationary methods. The wavelet-based model for the projection of monthly rainfall information in India has been used [4] in Taiwan with a wavelet model for prediction of air temperature; [5] the wavelet transformative and wind vector modelling devices have been used [6], and oil prices have been predicted with a wavelet-based model; [7] applied wavelet models for measuring magnet storm disturbance; [8] used wavelet models for predicting water level; the wavelet model can also be used in

There is agreement that the prices of stocks in the short and long term undergo unforeseen swings. An accurate stock market forecast technique may thus assist investors to make a profit from purchasing and selling. However, simulation of the stock market is a challenge because such a financial cycle is a complex process with many factors, such as policy events, the present and future economic conditions, and the sentiments of investors, influencing its performance [11]. There are limits to existing stock market models and forecasting techniques. In particular, typical models cannot handle rapid fluctuations or jumps in stock market systems. It is therefore necessary to create more efficient techniques. An application of wavelet-based models to cope with such serious non-linear processes. The complex system can also be represented by a Wavelet Multi Input Single Output model (WMISO) by decomposing the system input variables into numbers of the new time series at different levels. The least-square method is generally an effective way to assess model parameters for linear WMISO model identification; the non-linear WMSIO model identification is well supported by the minimum orthogonal squares (OLS) Algorithm.

Traditional techniques of bursary analysis typically include regression to simulate the volatility of bursary prices. However, these models are technically flawed, because bond changes and potentially irrational investor behaviour are influenced by political and economic factors, which will render the model and forecast inaccurate. Behavioural economics believes that psychology and behaviour cannot be overlooked when the modelling of stock market volatility changes. However, with the introduction of big data, huge data from the Internet can be used to assist bond model research. In the past, studying investor behaviour was difficult. It was not feasible.

In order to differentiate between Twitter feelings, this investigation will extract tweets and apply a Lexicon-based approach and machine learning method. The feelings index helps us understand the systems in politics and economics.

2. BACKGROUND

People now rely more than ever on the Internet, and the Internet has profoundly affected our everyday lives. In order to contact your friends, for example, people require the Internet, to shop, to browse online sites, or even publish their feelings and pictures on Twitter or other public social media in their everyday life. In a world in which human behaviour and actions leave digital traces, and these traces influence people's everyday lives [10], more particularly, contemporary people live in. Online recordings, online commentary, search engine data, and web browser histories may be part of a digital trail of everyday lives. Online data, search engine information (what individuals look for on Google/Baidu) and web surfing history (internet purchasing records, downloaded records, and bill records) are linked to personal privacy and are not included in the study. Facebook comments, tweets on Twitter, and any online comments may be included in Online Comments. The articles and reviews of other users on the Internet may readily influence people. On Twitter, Facebook and YouTube, text-based data sets may readily be discovered. Twitter is the online and social media platform for "mining opinion mining and sentiment analysis;" it offers a vast variety of text datasets [12]. Some research has been conducted and the textset/comments may be used to analyse and forecast. Numerous studies focus on using commentary and text datasets/comments posted by people on the Web to model and foresee specific information, such as applying sentiment and search queries to predict movies, predicting financial markets online, predicting stock-market volatility online, and using search engine information to detect influenza episodes. Influenza Thus, social media research development offers "a good chance for the public to comprehend their feelings by analysing large-scale data that is rich in opinions." Therefore, comments from one of the most famous social media sites, Twitter, are the primary subject of this study.

Web information includes many types of data, for example, online data, online records, online commentary (twitter, Facebook, YouTube), search engine records, and web browsing history. Web information is also available for web browsers. As previously mentioned, online records, data about search engines and web surfing history are connected to privacy concerns. It is thus difficult to collect and analyse this data lawfully. In addition, online records, data on the search engine and information on web surfing history always lack opinion, are rarely mined or are tiny. These kinds of data cannot completely represent people's feelings and may be hard to utilise for future modelling and prediction. As an international and prominent social media network, Twitter offers information on many matters that may be obtained for various purposes, and its large dataset is an important element since it can be utilised to simulate an actual system. Twitter has sentiment data on stock markets and other financial markets, for instance. The next section discusses in depth the significance of Twitter.

3. TWITTER NETWORK COMMUNICATION

Researchers from many disciplines have attempted over recent years to investigate Twitter from various angles with the rapid growth of the mobile terminal (MT). This section will examine how Twitter information is conveyed and communication and interaction behaviour among Twitter users. Some Twitter users may have more than one account on Twitter. There's one scenario that always occurs. It implies that a person may use multiple profiles to promote their views and play different roles while interacting on various social networks with other Twitter users. This pattern of communication across social networks may produce many Internet connections and datasets which in this study would be useless. The pattern for disseminating Twitter data is helpful to grasp, and Fig.1 illustrates a basic method for disseminating Twitter data.

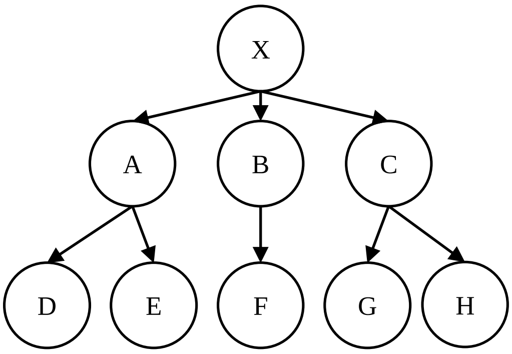

Fig.1. Simple Twitter dissemination process

The most important communication procedure for Twitter is to transmit a tweet. Once a person posts a tweet, the friends of the user may transmit the tweet if they believe it to be interesting or want it to be shown to their friends. Fig.1 illustrates the transmitted connection.

Twitter users have an interactive effect on other Twitter users. Breaking news requires Twitter to make it important through Twitter postings, transmission and debate. Some news items cannot have a variety of societal effects, for two reasons:

- Another news story that is more important than that may take place.
- Few Twitter users participate in the remark or transmission process. There are three quantitative criteria for tweets: the number of sent times, the number of responses given and the number of people visited.

Economic behaviour believes that feelings may affect the behaviour and decision making of people. A major problem in recent years has been the connection between the feelings of the social network and the economy. Tweets about economic issues are increasingly being published. These opinion-rich tweets are distributed through social networks and affect the public's attitude. According to behavioural economics, tweets will affect the global market economy indirectly, either positive or pessimistic.

3.1 EXTRACTION OF TWEETS

The R language may also assist academics in extracting messages from Twitter. The tweets collected may be saved in Microsoft Excel or Word format. In addition, the R language may mine information from tweets depending on the tweets' location. This kind of information not only reflects the location of Twitter information, but also enables academics to take a complete approach to the diversity of Twitter public opinion. In Twitter information mining, the benefit of the R language is not only because the information is collected comprehensively, but because R can also assist academics with dealing with the retrieved data from Twitter in their own language processing packages. The R language still has Twitter mining shortcomings; R cannot automatically conduct daily extraction work, and Twitter data must be extracted day by day by researchers. R cannot get the author data on each tweet in comparison with Webharvey, and this will affect the study of complicated network systems.

This part is primarily concerned with obtaining tweets and analysing UK stock market tweets using the R word cloud tool. In order to get FTSA tweets, we need to use the Twitter API to extract text from tweets containing "FTSE." Authentication on Twitter implies creating a Twitter app. First, go to https://apps.twitter.com/ and log in to your account on Twitter. Then follow the application name instructions and provide your Twitter API with a short description. The Twitter API also needs a valid website URL. The developer will have a 'consumer' key, a 'consumer' secret, an 'access token' or an 'access secret' once the Twitter API is established. In order to be able to retrieve tweets from Twitter, researchers must register this information.

3.2 PROPOSED MODEL

The size and quantity of digital information has increased considerably with the advent of computer technology. The data sets of people's everyday lives, such as the World Wide Web, have been presented through many platforms. The growth and popularisation of the Internet is thus going to speed up digital information creation and distribution.

There are many types of information on the Web. Information research thus faces a dilemma: an excess of information and a loss of information. Formerly, it was difficult to analyse and interpret huge quantities of information. The loss of information implies that particular information or data in the broad dataset is difficult to locate. A technique is thus required to identify and analyse the particular Web data. The majority of web information is saved in the form of text or corpus. This implies that text data is the web's primary storage type. In view of the size and model of the information, a web-based text-data mining method must be developed.

Web mining consists of web technology, data mining, text mining, processing for natural languages, artificial intelligence, and other technologies implementing data mining algorithms in the data science sector. Web mining is not only a tool for the recovery of information, it also helps to solve Internet data extraction, analysis, modelling and prediction problems. Fig.2 shows the flow diagram of a web mining method.

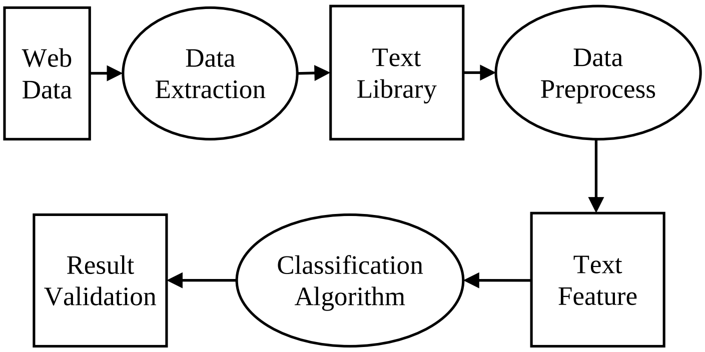

Fig.2. Flow chart of Proposed Twitter data mining

3.3 PRE-PROCESSING

There are a great many challenges to address when researchers want to extract data from social media using the R software. In particular, R provides several applicable functions to get Twitter.com data. However, Twitter users prefer to adopt a diversified approach to communicating their ideas or feelings because of the growth of Twitter. Emoticons and abbreviations, for example, in various languages. Tweets may also include a range of information, such as @someone, links and graphs. R seems to have the ability, however, to occasionally show this information correctly. Some information is unrelated to the study of feelings or even to experimental intervention. Therefore, pre-processing of Twitter data is very important for sentiment analysis.

In view of the 4 deficiencies in Twitter's pre-processed material, many interfaces and languages remain unrecognised. Because Twitter is a public online social network globally, it may be used by everyone. Some Twitter users choose to utilise their own language or to use English to convey their thoughts. Instead of just using English. Researchers will collect all the English material from Twitter to address this issue.

3.4 FEATURE SELECTION

TF shows the number of text words in a document. The content, format, and length of text are varied. All of these variables will affect the TF value and normalisation is the typical way of addressing this issue. If the text function contains numerous stop words (the, an, my...) and the high frequency of occurrence of these words increases the weight of stop words in TF, then this will affect the outcome of classification. Finally, the TF findings rely heavily on stopping words being deleted.

3.5 SENTIMENT ANALYSIS

The aim and purpose of the Twitter text data lexical sentiment analysis is:

- Identify Twitter feeling polarity.
- Determine the percentage of various emotions in the content of Twitter.

Note that the neuronal feeling is neither supported nor opposed to the Twitter content and sometimes it appears as news. Some news may not have any surface feeling elements, but readers will be invited to decide for themselves. Since this information is not handled reliably, the neuronal feelings are not taken into account by scientists. It is essential to be aware that Twitter has many distinct emotions, which is the emotion that has the greatest percentage according to its main emotions. Then, R is applied to the feeling that Twitter shows the results in positive and negative percentages. Finally, the lexicon-based approach to categorising Twitter emotions will be applied.

3.6 KNN CLASSIFIER

Eight emotion index findings for each Twitter data will be provided by the NRC lexicon on Twitter emotion analysis. The NRC lexicon can more precisely identify the data on Twitter not only in three types of feelings (positive, negative and neural), but also in eight kinds (Anger, Anticipation, Disgust, Fear, Joy, Sad, Surprise, and Trust).

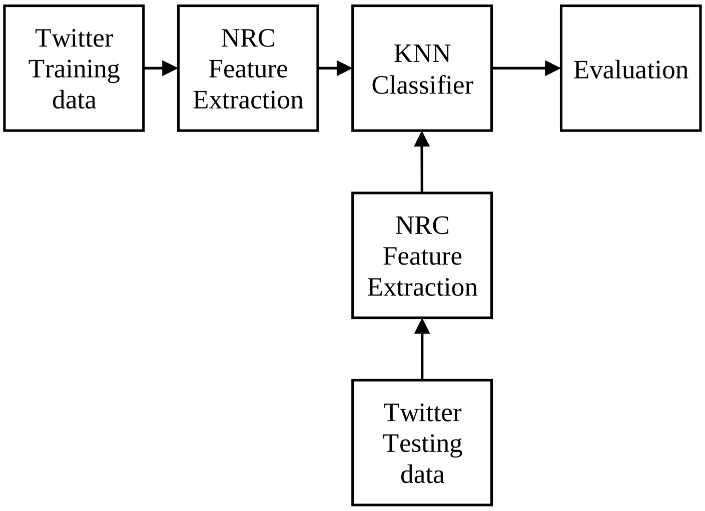

Fig.3. Process of KNN classifier

The purpose of our KNN classifier is to automatically identify the polarity of Twitter based on the training data set KNN (Positive or Negative). The 8 Twitter emotions are utilised for categorization as eight-dimensional numerical characteristics. As such, our key feature vectors will be Twitter's emotional data.

The initial stage in Fig.3 is to choose a training and text dataset and a category label should be added to the texts chosen (positive or negative). The NRC emotion index and other materials determine the function. In step three, it is important to identify the *K* value of the KNN algorithm, which is not described in detail. Next, by calculating the closest k^{th} Euclidean distance from test data to training data, the classifier identifies the group.

4. PERFORMANCE EVALUATION

Researchers will analyse the common performance index accuracy, reminder, and F-1 score of the model to assess classification outcomes for various artificial learning classificators. The confusion matrix may be used to obtain these values. We have developed two experiments to elucidate the algorithms of NRC KNN and NB. The data sets for testing and training are randomly selected and we do 10 classifiers. The performance of the KNN classification NRC is shown in Table.1.

Table.1. Performance of NRC KNN classifier

Experiments	Precision	Recall	F1
1	0.8333	0.6452	0.7273
2	0.85	0.5	0.6296
3	0.6957	0.5517	0.6153
4	0.8077	0.6563	0.7241
5	0.64	0.5714	0.6038
6	0.9048	0.5135	0.6552
7	0.7391	0.6538	0.6939
8	0.7222	0.5	0.5909
9	0.8182	0.5625	0.6667
10	0.84	0.5833	0.6885

The restriction of the NRC-based classifier is that huge quantities of Twitter-labelled data are needed to improve classification performance. So labelled tweets on a particular subject are necessary to address a new issue in Twitter sentiment analysis.

5. CONCLUSION

In this research, we have simply studied machine learning in Complex Network analysis on sentiment analysis and data visualisation. Researchers cannot investigate these ideas in detail, considering the time constraints. In the future, the KNN classifier will be used for US presidential tweets, FTSE 100 tweets for a more reliable modelling and foreclosure sentiment index, when we have enough training samples. The investigation shows an improved KNN classifier, which is expected to perform better for text classification tasks.

REFERENCES

[1] B. Heredia, J. Prusa and T. Khoshgoftaar, Taghi, "Exploring the Effectiveness of Twitter at Polling the United States 2016 Presidential Election", *Proceedings of IEEE 3rd International Conference on Collaboration and Internet Computing*, pp. 283-290, 2017.

[2] M. Khatoon, W. Mehjabin and S. Chinthamani, "*Sentiment Analysis on Tweets", *IGI Global Publisher, 2019.

[3] Bala Durga Dharmavarapu and Jayanag Bayana, "Sarcasm Detection in Twitter using Sentiment Analysis", *International Journal of Recent Technology and Engineering*, Vol. 8, No. 4, pp. 1-13, 2019.

[4] S.K. Bharti, R. Naidu and K.S. Babu, "Hyperbolic Featurebased Sarcasm Detection in Tweets: A Machine Learning Approach", *Proceedings of IEEE International Conference on India Council*, pp. 1-6, 2017.

[5] Pang Bo and Lillian Lee, "Opinion Mining and Sentiment Analysis", *Foundations and Trends in Information Retrieval*, Vol. 2, No. 1-2, pp. 1-135, 2008.

[6] A. Rajadesingan, R. Zafarani and H. Liu, "Sarcasm Detection on Twitter: A Behavioral Modeling Approach", *Proceedings of ACM International Conference on Web Search and Data Mining*, pp. 97-106, 2015.

[7] Mondher Bouazizi and Tomoaki Ohtsuki, "Sarcasm Detection in Twitter", *Proceedings of International Conference in Global Communications*, pp. 331-334, 2015.

[8] E. Riloff, A. Qadir, P. Surve, L. De Silva, N. Gilbert and R. Huang, "Sarcasm as Contrast between a Positive Sentiment and Negative Situation", *Proceedings of Conference on Empirical Methods in Natural Language Processing*, pp 704-714, 2013.

[9] P. Reyes and T. Veale, "A Multidimensional Approach for Detecting Irony in Twitter", *Proceedings of International Conference on Language Resources and Evaluation*, pp. 1-30, 2012.

[10] I. Habernal and J. Hong, "Sarcasm Detection on Czech and English Twitter", *Proceedings of 25th International Conference on Computational Linguistics*, pp. 121-127, 2014.

[11] R.J. Kreuz and R.M. Roberts, "Two Cues for Verbal Irony: Hyperbole and the Ironic Tone of Voice", *Metaphor and Symbol*, Vol. 10, No. 1, pp. 21-31, 1995.

[12] Debanjan Ghosh, Weiwei Guo and Smaranda Muresan, "Sarcastic or Not: Word Embedding to Predict the Literal or Sarcastic Meaning of Words", *Proceedings of International Conference on Empirical Methods in Natural Language Processing*, pp 1003-1012, 2015.

[13] B. Pang and L. Lee, "*Opinion Mining and Sentiment Analysis*", Now Publishers, 2008.