A Survey on Predictive Analytics Integrated with and Social Media

Madhusudhan V[1], Shilpa N R[2]

[1]*M.tech, Computer Science and Engineering, REVA ITM, Bangalore, India*

[2] *Asst.professor, REVA UNIVERSITY, Bangalore, India*

---***---

Abstract - *Social media is often used as a very huge source of data that is freely available for mining and analysis. This is often used for sentiment analysis and predictive analysis. Sentiment analysis using social media provides the sentiments of the people based on their posts on social media. It is also used to predict the future situations by predictive analysis. In this paper, we provide the introduction to Social media analytics, predictive analysis and sentiment analysis and provide a survey on various applications of predictive analytics integrated with social media such as the usage of predictive analytics with social media, prediction of the next president to be elected, stock predictions and even predicting the success rate of movies by analyzing tweets and YouTube comments. Predictive analytics with social media is also applied in the field of crime analysis, finding the strength of ties between two people over social media and recommending music and events based on a person's preferences. We finally show that predictive analytics has come a long way with social media integration to predict the future and it will continue to do so with a lot of research still going on in the field.*

KeyWords: Predictive analysis, sentiment analysis, Natural Language Processing.

1. INTRODUCTION

1.1 Social Media Analytics

Social media is now seen as a very rich source of knowledge for data mining as people often update about state of affairs, products and pretty much everything that goes on around the world. Hence, research in social media analytics has improved a lot in the recent times. There are many algorithms developed in the field of social media analytics for text mining from social media such as Twitter and Facebook. These algorithms fall under the category of sentiment analysis where they analyze the sentiments of sentences to be positive, negative or neutral.

Social media is analyzed for several purposes considering that the data from social media is ever increasing and always updated. Social media analytics is used to analyze the sentiments of people towards a particular product, state of affairs or a particular movie. affairs or a particular movie. This analysis is not only performed to analyze what is the current opinion of people on matters but also to analyze what the future may hold as well. Predictive analytics comes into picture to perform such analysis.

Data obtained from social media such as Twitter and Facebook do not adhere to the regular sentiment analysis of text. This is mainly because people do not often express their sentiments only with words and sentences. They also use emoticons and emojis in their tweets to depict their feelings rather than just words. Tweets are often associated with emojis nowadays and hence there is a necessity to use emoji detection and emoticon detection in the analysis of sentiment of data obtained from social media.

1.2 Predictive Analytics

Predictive analytics performs analysis on current data sets obtained from social media or other sources to predict the future. This may be for predicting the future sales of a product or predicting the next president who would be elected in the country or even to predict the stock prices. Hence, it is also termed as the crystal ball to see the future [2] [3]. Predictive analytics is often integrated with social media to obtain the data to be analyzed, which would be obtained from social media. In other words, it is used to observe past and current patterns in order to predict how the pattern could turn out to be in the future such as achieving better business outcomes.

A common example of predictive analytics is the recommendation engines that see the current shopping pattern or music listening pattern of a person and hence predict what they might like next in terms of shopping or listening to the kind of music and recommend such shopping items or songs according to the individual's preferences.

1.3 Sentiment Analysis

Sentiment analysis is the process of analyzing the opinion of the masses. Sentiment analysis is an important part of predictive analysis. It is the basis on which the textual data is analyzed. Sentiment analysis algorithms are broadly classified into two main types.

The first one is Naïve Bayes classification, which is also termed as bag of words approach. This is an approach which emphasizes on the independence of words with respect to each other i.e. there is no dependence between two words and each word is taken up to be analyzed as a separate entity. The second type of algorithm is support vector machines, which uses machine-learning techniques. The model is trained with data so that it can perform analysis on more data that will be fed to it for analysis. If humans label the trained data, it is termed as supervised learning of the model. If the trained data is difficult to find, then unsupervised learning methods are used. They also use parts of speech tagging to identify the sentence and split it based on subject and object rather than using just the bag of words approach [1]

2. RELATED WORK

This section of the paper consists of a brief summary of the previous work done in the grounds of predictive analytics with social media.

Xiaotian Jin et al [4] have performed stock prediction based on the combination of statistical method and predictive analytics using social media. They have used SVM regression models in statistical analysis and for the predictive analytics, they have collected millions of tweet data from Twitter and analyzed them for the keywords such as “rally”, “low” which indicate a bearish sentiment and used keywords such as “up”, “good” to indicate bullish sentiment. They have used an N-gram algorithm to perform this operation. The combined result of the sentiment and statistical analysis has proved to predict stock prices effectively. 223 days of data from Twitter is used to train the model whereas the last 30 days of data obtained from twitter is used to analyze the sentiments. They have used open source software LingPipe to perform the sentiment analysis of the tweets.

Stefan Nann et al [5] have predicted stock prices based on Twitter data obtained over six months where they have obtained 2,917,381 tweets by specifying the cash tag i.e. “$”. They used a Naïve Bayes classifier with a conditional independent bag of words methodology along with negation handling and parts of speech tagging and spam removal based on keywords. They showed that a positive return of investment was attained by using only data obtained from social media and predictive analysis. They also stated that a sentiment analysis algorithm with a higher accuracy would provide better results in terms of accuracy in prediction of stock prices.

Andranik Tumasjan et al [6] have used Twitter as the source for the data sets to be mined where they have collected over 1,04,000 tweets. With many people updating their twitter about political issues, they evaluated the sentiments of people to predict the position of the contestants of German elections and hence predict who might be the next chancellor. They used LIWC2007 (Linguistic Inquiry and word count, Pennebaker, Chung and Ireland 2007) to detect the emotions. The LIWC English dictionary is used to measure the sentiments where they have used 12 dimensions of emotions including positive, negative, anger, achievement, sadness and whether they are future oriented or past oriented. The German tweets are translated to English before being processed by LIWC2007. The sentiment analysis performed does not emoticon detection whereas the microblogging services such as Twitter often use emoticons to express the feelings and opinions.They stated that Twitter may finally complement the traditional methods i.e. polls and surveys to be the new method of political forecasting.

Lei Shi et al [7] have used ten million tweets obtained from Twitter as the dataset out of which 90 percent was used to train the model whereas the other 10 percent was used to test. They conducted this from September 2011 until the republican elections in 2012 in the USA to predict the next presidential election results. They used geographical identification method to obtain tweets from different locations i.e. state wise to predict the election results of the state and overall without specific locations to predict the national results. They used features such as retweet count, unique user count and whether the tweets were obtained from promotional or non-promotional accounts. The used a linear regression model to predict the election results and hence proved that it was feasible to predict such results.

Yafeng Lu et al [8] have used predictive analytics from social media data namely from Twitter, Youtube and IMDB reviews to predict the success of a movie as well as the opening weekend earning of the movie. The data was processed and compared with a bag of words approach provided by SentiWordNet [17]. SentiWordNet assigned a value between -1 to +1 to associate a word with its sentimental value. They used a linear regression model to predict the grossing and success of a movie in box office. They have also stated that such framework can also be used in a wide range of applications where the trend values and social media are used as input for predictive analytics.

Dingqi Yang et al [9] have observed that the system of recommendation of movies and music has been a success and they have come up with a way to recommend venues in locations based on tips about the venue and check in information obtained from Twitter on that particular venue. This is performed by obtaining the data set to be mined from LBSN (Location Based Social Networks). There are two kinds of location recommendation systems. The first is generic location recommendation, which recommends a venue that is popular in an area. The second is personalized location recommendation, which recommends location considering the individual's preferences.

Yading Song et al [10] have made a survey on music recommendation systems to analyze the future of music recommendation. In Music Information Retrieval (MIR), there have been many advances in genre classification [18, 19] and instrument recognition [20]. One of the music recommendation methods in collaborative filtering, which recommends songs, based on listening activities of the user and previous ratings to generate a playlist. Another way of music recommendation is content-based model, which uses features such as genre, instrument, rhythm and pitch. The recent kind of music recommendation system comes from emotion-based model which detects the emotion of the song based on the lyrics whether the song is positive, passionate, negative, aggressive or humorous. This is based on sentiment analysis of the song lyrics itself which when matched with similar sentiments of another song, they can be paired into a recommendation category based on emotion.

Jennifer Golbeck et al [11] have used predictive analytics with social media to predict the personality of a person. The data obtained to analyze are taken from Facebook where they analyze it for five specific personality traits. These specific personality traits are openness, agreeableness, conscientiousness, extroversion and neuroticism. If a person is having higher percentage of openness from the result of the analysis, it means that the person tends to be artistic and sophisticated whereas a person having a higher percentage of neuroticism states that the person tends to be more anxious and sensitive. The details of users such as their liked activities and preferences such as their favorite movies, tv shows were obtained along with the "About me" information of the person and status updates of the person were taken into account as well. They have used the LIWC tool to analyze the text. They performed linear regression analysis for each personality factor and hence make up the five personality features. They also state the example of Facebook advertisements that a person gets based on the kind of interests he or she has.

Trip Kucera et al [12] from the Aberdeen Group have used predictive analytics to improve the business outcomes by analyzing the customer behavior and insights of customers. The result was a twofold sales increase in their marketing campaigns. By the usage of predictive analytics, they have improved their marketing offers towards the respective clients. The data of customers were obtained from expenditure history and their behavior towards marketing data as well as textual data from social media. They also specify that predictive analytics is a largely growing field with a lot of research in terms of marketing and sales prediction going on making predictive analytics more attractive even for smaller companies.

Eric Gilbert et al [13] proposed and implemented an approach to predict the ties of a person with another person using data over Facebook and compared them with a questionnaire to the people on Facebook over a web application and found out that they predicted the ties between two people as either strong or weak with 85 percent accuracy. They state that it can be used to identify the relations and tie strength between coworkers of an organization. They have used seven dimensions to analyze the tie strength including intensity, intimacy and emotional support. They have used features such as the number of wall words exchanged, inbox messages exchanged, days since last communication, whether their wall posts and inbox messages are having positive or negative emotions. Thirty-two such variables are used to predict the tie strength. The final tie strength is calculated by a combination of the seven dimensions, which are further divided into thirty-two variables.

Venkata Rama et al [14] have presented a platform called FAST (Forecast and Analytics of Social Media and Traffic) to predict traffic using social media and online news sources. The data source taken was within one hour of the time to predict. They have created a separate 5-minute and a 3-day forecasting window. They compared the results with the results obtained from Al Jazeera.

Anthony J. Corso et al [15] have performed predictive analytics towards crime using social media and geographical information system. With the events being reported on social media, the geographical location along with that is taken up which is usually obtained via Twitter or Facebook. They collected around 2,250,000 tweets based on a location by using a latitude and longitude bound for the tweets. The tweets obtained based on the location and based on time series provided a good predictive approach to analyze crime in a location.

Marco Balduini et al [16] have proposed an approach to predict the events a person might like to attend based on the location. They have performed this by using the digital footprint i.e. data posted by the user on the social media as to the events that the person would attend. The user's profile is analyzed to detect the kind of event or venue he or she would like to attend and hence the event is recommended to them. The time window is measured and events within the time window of 3 hours are provided to the user. A social listener listens to all the events posted over the social media within the time and based on the person's social media tweets, similar events are posted to the user.

3. CONCLUSION

Social media analytics and predictive analytics with text data as the source to be analyzed use sentiment analysis as the basis of classification. The algorithm used in the sentiment analysis plays a major factor in the accuracy of the results itself. We observe that predictive analytics integrated with social media is used in various ways in simple words is a crystal ball to look into the future with surprisingly good accuracy. It is used in predicting many issues ranging from the prediction of the next president to be elected to predicting the success rate of a movie. Briefly, we conclude that predictive analytics has come a long way in the field of predicting many things and it will continue to do so with social media being a very rich and huge source of data that can be analyzed. We predict that the research in predictive analytics will go on a long way improving in many aspects and it will be a widely used research topic.

REFERENCES

[1] Walaa Medhat, Ahmed Hassan and Hoda Korashy, "Sentiment Analysis and Applications: A survey," Ain Shams Engineering Journal, vol. 5, issue 4, pp. 1093-1113, December 2014.

[2] "A Survey of Prediction Using Social Media", Sheng Yu and Subhash Kak, Department of Computer Science, Oklahoma State University Stillwater, Oklahoma, U.S.A. 74078.

[3] "Predictive Analytics: Bringing The Tools To The Data", An Oracle White Paper September 2010.

[4] "Enhanced Stock Prediction using Social Network and Statistical Model", Xiaotian Jin, Defeng Guo, 2014 IEEE Workshop on Advanced Research and Technology in Industry Applications (WARTIA)

[5] "Predictive Analytics On Public Data – The Case Of Stock Markets", Stefan Nann, Jonak Krauss, Detlef Schoder, ECIS 2013 Completd Research. Paper 102.

[6] "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment", Andranik Tumasjan, Timm O. Sprenger, Philipp G. Sandner, Isabell M. Welpe, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

[7] "Predicting US Primary Elections with Twitter", Lei Shi, Neeraj, Agarwal, Ankur Agrawal, Rahul Garg, Jacob Spoelstra.

[8] "Integrating Predictive Analytics and Social Media", Yafeng Lu, Robert Kr¨uger, Student Member, IEEE, Dennis Thom, Feng Wang, Steffen Koch, Member, IEEE, Thomas Ertl, Member, IEEE, and Ross Maciejewski, Member, IEEE.

[9] "A Sentiment-Enhanced Personalized Location Recommendation System", Dingqi Yang, Daqing Zhang, Zhiyong Yu, Zhu Wang, Proceedings of the 24th ACM Conference on Hypertext and Social Media. Pages 119-128.

[10] "A Survey of Music Recommendation Systems and Future Perspectives", Yading Song, Simon Dixon, and Marcus Pearce, The 9th International Symposium on Computer Music Modeling and Retrieval (CMMR), 2012.

[11] "Predicting Personality with Social Media", Jennifer Golbeck, Cristina Robles, Karne Turner, Proceeding CHI '11 Extended Abstracts on Human Factors in Computing Systems, Pages 253-262.

[12] "Predictive Analytics for Sales and Marketing", Trip Kucera, David White, January 2012, Aberdeen Group.

[13] "Predicting Tie Strength With Social Media", Eric Gilbert and Karrie Karahalios, CHI '09 Proceedings of the SIGCHI Conference on Human Factors in Computing Systems. Pages 211-220.

[14] "Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment", Andranik Tumasjan, Timm O. Sprenger, Philipp G.

Sandner, Isabell M. Welpe, Proceedings of the Fourth International AAAI Conference on Weblogs and Social Media.

[15] "Predicting US Primary Elections with Twitter", Lei Shi, Neeraj, Agarwal, Ankur Agrawal, Rahul Garg, Jacob Spoelstra.

[16] "Integrating Predictive Analytics and Social Media", Yafeng Lu, Robert Kr¨uger, Student Member, IEEE, Dennis Thom, Feng Wang, Steffen Koch, Member, IEEE, Thomas Ertl, Member, IEEE, and Ross Maciejewski, Member, IEEE.

[17] "A Sentiment-Enhanced Personalized Location Recommendation

System", Dingqi Yang, Daqing Zhang, Zhiyong Yu, Zhu Wang, Proceedings of the 24th ACM Conference on Hypertext and Social Media. Pages 119-128.

[18] "A Survey of Music Recommendation Systems and Future Perspectives", Yading Song, Simon Dixon, and Marcus Pearce, The 9th International Symposium on Computer Music Modeling and Retrieval (CMMR), 2012.

[19] "Predicting Personality with Social Media", Jennifer Golbeck, Cristina Robles, Karne Turner, Proceeding CHI '11 Extended Abstracts on Human Factors in Computing Systems, Pages 253-262.

[20] "Predictive Analytics for Sales and Marketing", Trip Kucera, David White, January 2012, Aberdeen Group.

[21] "Predicting Tie Strength With Social Media", Eric Gilbert and Karrie Karahalios, CHI '09 Proceedings of the SIGCHI Conference on Human 0Factors in Computing Systems. Pages 211-220.