Text Mining Based on Social Media –A Survey

SP.Anitha[1]
MPhil Research Scholar,
Department of Computer Science
Avinashilingam Institute For Home Science And Higher Education For Women
Dr.V.Radha[2]
Head & Professor,
Department of Computer Science
Avinashilingam Institute For Home Science And Higher Education For Women

Abstract—Social Media plays a vital role in the world. These technologies are effectively used to connect friends, family and presently act as effective tool in business communication, vide photos, videos and messages. They basically use tools like Twitter, Facebook, Instagram, Pinterest, Youtube, etc. Social Media is a very useful tool to connect with people and it has some privacy and security. It also provides valuable data in areas of Education, Industries, Web mining, Textmining, Agriculture etc. Among these Text Mining plays a valuable role to Social Media. The Social Media also provides interesting data and also guide in identifying resourceful contexts. Analyzing social media is slightly complicated process. It also provides solution in areas of Machine Learning, Fuzzy and Support Vector Machine (SVM). This survey deals with papers on Text mining

Index Terms-Text mining, Social Media, Support Vector Machine, Machine Learning, Naive Bayes, Sentiment Analysis, Random Forest.

1 INTRODUCTION

Text mining primarily deals on established process to gather valuable information and to examine the data. It has some techniques on sentiment analysis, categorization and entity extraction which are used to extract the information and knowledge from the hidden text content. Text mining technologies are widely applied on research, business strategies, social media data analysis, spam filtering, customer care service, etc. The Text mining also effectively practiced in Military and Government process towards protection of legal, confidential policies, documents, social network post, call center logs, medical records, etc. Text mining also helps in guiding the business applications, trend analysis, sentiment analysis, feature extraction, predictive and opinion mining. Internet, web application would not have been so effective without Text mining. The process of text mining is shown in the figure 1 givenbelow.

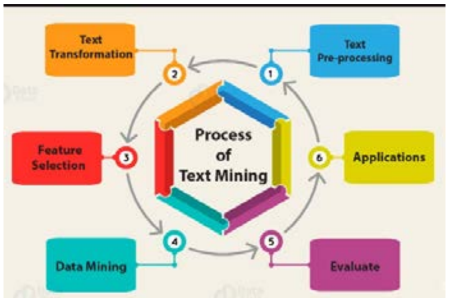

Fig 1: Process of Text mining

2 LITERATURE REVIEW

Ahmad et al. (2017) [1] introduced a new method to analyze the performance of SVM to detect the polarity using different datasets. Detection of polarity was their main objective. They used three datasets for experiments, two from Twitter and one from Internet Movie Database (IMDB) reviews. Their main focus was to increase the performance of SVM. To meet their goal they proposed a frame work consisting of several phases like Preprocessing and Classification.

AkshayAmolik et al. (2016) [2]come out with a method for Twitter sentiment analysis for reviewing movies using machine learning techniques. They focused on detecting and justifying the human sentiments with the help of social media from text. The emotions of human beings were analyzed using text based machine learning algorithms. They used real dataset and the accuracy was observed to be 75% from SVM and 65% from Naive Bayesian.

AnanthiSheshasaayee et al. (2017) [21] came out with a proposal to compare the classification of algorithms in Text mining. Three algorithms are Naive Bayes, Random Forest, and Support Vector Machine was compared with the accuracy. They observed results on the graph of Random forest 88% accuracy, Naive Bayes 90% accuracy and SVM 97% among these the best results were produced by SVM when compared to other algorithms. The study was done by collecting sample dataset containing positive and negative results.

Balakrishnan et al. (2012) [3] discusses about the micro blog data on Twitter. They introduced Synthetic Minority Over-Sampling Technique (SMOTE) to improve the performance of accuracy and analyzed preprocessing of Twitter data. Various classification of algorithm with different dataset was tried using Random Forest. The accuracy level increased to 80% when compared with Naive Bayes which has only 70% accuracy.

Camastra et al. (2015) [8] proposed web page categorization using Machine Learning techniques. Comparisons were made between the performance of supervised and unsupervised techniques with real data set. Detailed discussions were made on web page categorization with semantic graph and Machine Learning. Topographic Error (TE), Combined Error (CE), Quantization Error (QE) was consider to define the perfect accuracy.

ChintanDedhia et al. (2017) [5] has proposed ensemble model for twitter sentiment analysis. They targeted to improve the performance of SVM and Adaboost also used to boost up the performance.

Deep Learning Neural Network (DNN) was imported by Hu et al. (2017) [25] to get high dimension data analysis. They applied DNN for processing sentiment analysis and resulted in getting better performance with enhanced speed. Discussions were initiated as a first step utilizing DNN for sentiment analysis and feature vectors. The performance of DNN seems to be balancing the time with respect to performance. In the second step they have constructed the feature vector. Finally DNN seems to very effective to solve the Data mining problems.

Detailed analysis were conducted on emotion from text and classified as fear, anger, surprise, joy, sad and disgust by Diana Lupan et al. (2012) [7]. They targeted to capture the person emotions while reading news. Natural language processing techniques is used to focus mainly on emotions.

Devika et al. (2016) [15] made a comparative study on various approaches followed in sentiment analysis. It was observed that the following approaches like Machine Learning, Rule Based and Lexicon Based were different from each other with respect to performance. Among these approaches Machine Learning seems to be a better performance on accuracy. Ongoing discussions were done on these approaches to get better results for sentiment analysis.

Dhanalakshmi et al. (2016) [6] worked on opinion mining of student feedback with supervised learning algorithms. Rapid Miner tool was used to find the student feedback. Positive and negative comments were classified and observed that KNN algorithm are best in precision where as Naive Bayes algorithm are best in accuracy.

Discussion were made by ChetashriBhadane et al. (2015) [4] on lexical and machine learning approaches to identify the customer opinion mining and also to detect the mood and they found the accuracy for polarity as well as the aspects classification. They achieved 78% accuracy on the product review using SVM.

Hassan Saif et al. (2015) [10] came out with the proposal on contextual semantics for sentiment analysis of Twitter. SentiCircles concepts were introduced to detect the sentiments in Twitter to show the results for entity-level and tweet-level sentiment detection. Health Care Reform (HCR), Obama-McCain Debate (OMD), Standford Sentiment Gold Standard (STS-Gold) datasets were used for the evaluation.

HeniSulistiani et al. (2017) [23] came out with a method to predict customer loyalty. They made a detailed Comparative Analysis by Feature Selection Method. Their objective was on improving the performance of prediction accuracy and the results. The comparison of the accuracy was listed by predicting the customer loyalty based on Random Forest.

Himank Gupta et al. (2018) [11] proposed on Twitter platform to detect the spam tweets. Different Machine Learning algorithms like Random Forest, Gradient Boosting, Neural Network and Support Vector Machine (SVM) were used to get a better accuracy. They present real data time spam detection from Twitter by collecting 400,000 public tweets.

Hsin-Ying Wu et al. (2014) [13]came out with a proposed technique for Facebook post and comments. Their objective was to satisfy the customer needs through online and customer can directly send feedback for the particular product. Their first step was to collect the Facebook posts and then processed by Chinese Knowledge and Information Processing (CKIP) to segment the known and unknown words. They have separated the words, phrases, Facebook post and terms used by the users. The limitation of their work is to mainly focus on segmentation of the known and unknown words used by the customer.

It was also on Text mining to measure the post and comments by HimanshiAgrawal et al. (2016) [12] from public pages such as Wikipedia on Facebook. The proposal were made on public pages on Non-profit organization, Entertainment websites and concluded that the spam activity cannot be completely remove from the web.

Kim Hammar et al. (2015) [14] reviewed about Instagram from Social Media platform and analyzed the Instagram posts consisting of images and texts with the help of trained data. Natural language processing (NPL) tools are used to formal text and post were classified into separate division.

NadeemAkhtar et al. (2013) [16] used GEPHI tool, it is open source software for network and graph analysis to study social network. They have collected the dataset from April 2009 through Facebook. They focus on sub-graph of Facebook social network which contains high-degree nodes. In their study the sub-graph file was converted to Comma-Separated Values (CSV) format. The number of friendship is higher, and then they found there is little direct relation among these high-degree nodes. Friendships were high but their relationships were less on the Facebook graph. The performance was also improved in social network analysis (SNA) with the help of GEPHI tool.

Neethu M et al. (2013) [17] analyzed public opinion about the electronic products using sentiment analysis in Twitter with the help of machine learning techniques. The datasets were collected with the help of tweets from April 2013 to May 2013. The data were analyzed using sentiment analysis which automatically annotated positive or negative comments. Preprocessing was done to avoid slang words and misspelling in tweets.

Po-Wei Liang et al. (2013) [18] made some analysis on sentiment messages by proposing architecture to automatically generate in the micro blogging post. They found positive and negative opinion messages. It was also concluded that Machine Learning techniques performed well.

Rajeswari et al. (2017) [19] came out with a classified text using Naive Bayes classifier and KNN classifier. The focus was on accuracy and performance using Rapid Miner with datasets of students. Naive Bayes showed a better accuracy of 66.67% when compared with KNN showing an accuracy 38.89 which was also analyzed.

Shuhufta Fatima et al. (2017) [9] came out with a method for categorization of text documents. They have tested the documents by partitioning the documents at ratio of 60-40 and 80-20. This resulted in increasing the accuracy level on 80 training documents and 20 testing documents. They mainly focused on the time consumed.

Suge Wang et al. (2009) [22] proposed a feature selection method based on fisher's discriminant ratio for text sentiment classification. Their focus was to determine positive and negative reviews. They automized the text sentiments with positive (thumbs up) or negative (thumbs down). Four types of feature selection method were proposed by them with the help of Support Vector Machine (SVM). They also adopt three kinds of measures like Precision, Recall and F1 value using the data collected from 2006 January to March 2007. The output seems to be 578 positive reviews and 428 negative reviews from Chinese text reviews 1006 and about 11 kinds of car trade marks.

Vijayarani et al. (2016) [20] made attempts on preprocessing techniques on Text mining to solve problems. They concluded that Text mining research would be highly useful to analyze Social Media contents.

Yang et al. (2015) [24] tried to achieve better performance using Twitter with sentiment analysis. Naive Bayes classifier used to get high level accuracy and they also made attempts of chi-square to get more information. Their main target was to get high level accuracy on performances.

3 CONCLUSION

The survey was conducted on various Social Media's like Twitter, Facebook, Instagram, Pinterest, Youtube, etc. Data mining techniques is effectively used for this study. Various methods and techniques is studied which has its own advantages and disadvantages. The survey also reveals that no specific technique can be proposed for Text mining also concluded that better performance on accuracy can be attained by boosting the algorithms.

REFERENCES

[1] Ahmad, Munir, &ShabibAftab.Analyzing the Performance of SVM for Polarity Detection with Different Datasets.International Journal Modern Education and Computer Science.

[2] Amolik, A.,Jivane N.,Bhandari M., Venkatesan M. Twitter Sentiment Analysis of Movie Reviews using Machine Learning Techniques. International Journal of Engineering and Technology (IJET), 2016, Vol.7 No.6.

[3] BalakrishnanGokulakrishnan, PavalanathanPriyanthan, ThiruchittampalamRagavan, NadarajahPrasath, AShehanPerera, Opinion Miningand Sentiment Analysis on Twitter Data Stream, International Conference on Advances in ICT for Emerging Regions-ICTer ,2012.

[4] C.Bhadane, H.Dalal, and H.Doshi, Sentiment analysis: Measuring opinions, ProcediaComput.Sci., vol.45, no. C, pp. 808-814, 2015.

[5] Dedhia, C and Ramteke, J. Ensemble model for Twitter Sentiment Analysis. International Conference on Inventive Systems and Control 2017.

[6] Dhanalakshmi V.,Dhivya Bino, Saravanan A.M, Opinion mining from student feedback data using supervised learning algorithms, International Conference on Big data and Smart City,2016.

[7] Diana Lupan, MihaiDascalu, Stefan Trausan-Matu, Philippe Dessus, Analyzing Emotional States Induced by News Articles with Latent Semantic Analysis, International Conference, 2012.

[8] F.Camastra, A.Ciaramella, A.Placitelli, and A.Staiano, Machine Learning-based Web Documents Categorization by Semantic Graphs, ResearchGate Publisher, DOI: 10.1007/978-3-319-18164-6-8, June-2015.

[9] Fatima,S.,&Srinivasu,B.Text Document Categorization using support Vector Machine. International Research Journal of Engineering and Technology (IRJET).2017, Vol.4 Issue 2.

[10] H.Saif, Y.He, M.Fernandez, and H.Alani, Contextual semantics for sentiment analysis of Twitter, Inf. Process. Manag., vol.52, no.1, pp. 5-19,2015.

[11] Himank Gupta, Mohd. Saalim Jamal, SreekanthMadisetty and MaunendraSankarDesarkar, A Framework for Real-Time Spam Detection in Twitter, International Conference, 2018.

[12] HimanshiAgrawal, RishabhKaushal, Analysis of Text Mining Techniques over Public Pages of Facebook ,International Conference on Adavance Computing, 2016.

[13] Hsin-Ying Wu, Kuan-Liang Liu, CharlesTrappey for Understanding Customers Using Facebook Pages: Data Mining Users Feedback Using Text Analysis. International Conference on Computer Supported Cooperative Work in Design 2014.

[14] Kim Hammar, Shatha Jaradat, Nima Dokoohaki, Text Mining of Instagram data, International Conference, 2015.

[15] M.D.Devika, C.Sunitha, and A.Ganesh, Sentiment Analysis: A Comparative Study on Different Approaches, in Procedia Computer Science, 2016, vol. 87, pp. 44-49.

[16] NadeemAkhtar, HiraJaved, GeetanjaliSengar for Analysis of facebook social network. International Conference on Computational Intelligence and Communication Networks 2013.

[17] Neethu M, S, Rajasree R, Sentiment analysis in Twitter using Machine Learning Techniques, 4th ICCCNT, 2013.

[18] Po-Wei Liang, Bi-Ru Dai, Opinion Mining on Social Media, International Conference on Mobile Data Management, IEEE , 2013.

[19] Rajeswari R.P., Kavitha Juliet, Dr.Aradhana, Text Classification for Student Data Set using Naïve Bayes Classifier and KNN Classifier, International Journal of Computer Trends and Technology 43(1) (2017), 8-12.

[20] S. Vijayarani, R.Janani, Text Mining :A survey, International Journal of Innovative Research in Computer and Communication Engineering, 2016.

[21] Sheshasaayee A. &Thailambal G. Comparison of Classification Algorithms in Text Mining. International Journal of Pure and Applied Mathematics 2017, Vol.116 No.22 pp 425-433.

[22] Suge Wang, Deyu Li, Yingjie Wei, Hongxia Li for A Feature Selection Method Based on Fisher's Discriminant Ratio for Text Sentiment Classification 2009.

[23] Sulistiani, H., &Tjahyanto, A. Comparative Analysis of Feature Selection Method to Predict Customer Loyalty.Journal of Engineering, Vol.3, No.1, 2017.

[24] Yang, Ang, et al. Enhanced Twitter Sentiment Analysis by using Feature Selection and Combination 2015.

[25] Z.Hu, J.Hu, W.Ding, and X.Zheng, Review Sentiment Analysis Based on Deep Learning, in 2015 IEEE 12th International Conference on e-Business Engineering, 2015, pp. 87-94.