International Journal of Electrical and Computer Engineering (IJECE)
Vol. 9, No. 5, October 2019, pp. 4452~4459
ISSN: 2088-8708, DOI: 10.11591/ijece.v9i5.pp4452-459

Hybrid approach: naive bayes and sentiment VADER for analyzing sentiment of mobile unboxing video comments

Chaithra V D
Department of Computer Science, CHRIST (Deemed to be University), India

Article Info

Article history:

Received Dec 31, 2018
Revised Apr 24, 2019
Accepted May 3, 2019

Keywords:

Mobile unboxing
Naive bayes
Sentiment analysis
Sentiment VADER
YouTube

ABSTRACT

Revolution in social media has attracted the users towards video sharing sites like YouTube. It is the most popular social media site where people view, share and interact by commenting on the videos. There are various types of videos that are shared by the users like songs, movie trailers, news, entertainment etc. Nowadays the most trending videos is the unboxing videos and in particular unboxing of mobile phones which gets more views, likes/dislikes and comments. Analyzing the comments of the mobile unboxing videos provides the opinion of the viewers towards the mobile phone. Studying the sentiment expressed in these comments show if the mobile phone is getting positive or negative feedback. A Hybrid approach combining the lexicon approach Sentiment VADER and machine learning algorithm Naive Bayes is applied on the comments to predict the sentiment. Sentiment VADER has a good impact on the Naive Bayes classifier in predicting the sentiment of the comment. The classifier achieves an accuracy of 79.78% and F1 score of 83.72%.

Corresponding Author:

Sivakumar R,
Department of Computer Science,
CHRIST (Deemed to be University),
Bengaluru, India.
Email: sivakumar.r@christuniversity.in

1. INTRODUCTION

Social media sites act as a medium for the users to post reviews of the product, services, issues or events. The user generated comments are of great help for people looking for the views of the users who have used the product or services. With the growing popularity of various social media sites YouTube is gaining a lot of attention from the users. YouTube not only provides videos but also a platform for the viewers to express their opinion towards the video in the form of likes, dislikes and comments.

YouTube has variety of user generated and corporate media videos which includes movie trailers, music, TV shows, educational, vlogging and unboxing videos. Unboxing videos are the trending and most viewed among the videos available on YouTube. Unboxing videos are regarding products that are new in the market, like electronic gadgets, mobile phones, clothes and accessories. The concept of unboxing video is a person unboxes the new product, reviews and expresses his/her opinion towards the product as an end user. With the frequent releases of mobile phones by various manufacturer embedded with trending technologies, it becomes difficult to purchase a mobile with technologies worth the money. Review expressed in mobile unboxing video along with the comments from the viewers towards the video guide one to know more about the mobile.

Sentiment analysis of mobile unboxing video comment helps analyzing the user’s reaction towards the mobile phone. The metadata of the video (likes, dislikes, views and comments) express the viewer’s reaction towards the phone. Metadata such as likes and dislikes do not provide detailed sentiment of the viewers, therefore comments are considered to analyze the viewer’s opinion. The comment by the viewers expresses a positive or a negative opinion towards the mobile phone. For studying this unboxing of LG G7 mobile phone is randomly considered. The metadata (likes, dislikes, comments etc.) associated with the LG G7 phone is extracted using an online YouTube scrapper. Hybrid approach which combines both machine learning and lexicon approach is followed to perform the sentiment classification of the comments. The results of the classification will help the viewers to know about the mobile phone and decide is it worth buying or not.

2. THEORETICAL BACKGROUND

2.1. Sentiment analysis

Sentiment analysis [1] is a study for knowing the sentiment expressed in the user's opinion towards a product or an issue. Sentiment expressed by an individual can be positive, negative or neutral. In this paper sentiment is considered to be either positive or negative. Positive sentiment shows that the user has liked the product or agrees with the issue whereas negative sentiment shows the dislike towards the product or disagreement with the issue. Studying the sentiment expressed towards the product answers questions how the product is doing in the market whether it is getting positive or negative response.

2.2. Levels of sentiment analysis

Sentiment analysis can be performed on 3 different levels as shown in Figure 1.

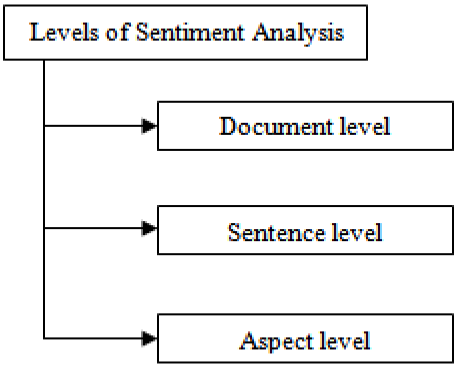

Figure 1. Levels of Sentiment Analysis

2.2.1. Document level

Document level sentiment analysis classifies entire document either as positive or negative. This level of sentiment analysis is used when the entire document is related to a single entity or topic.

2.2.2. Sentence level

Sentence level sentiment analysis classifies each sentence individually based on the sentiment expressed. This level of sentiment classification acts as subjective classification. The classification in sentence level can be expressed as positive, negative or neutral. It differentiates between the subjective and the objective information present in the sentences.

2.2.3. Entity or aspect level

Aspect level concentrates on the opinion, with the idea that an opinion consists of a sentiment and a target. It performs a fine-grained sentiment analysis and differentiates what a user wants and does not want. This level of sentiment analysis is preferred when more than two products reviews are to analyzed or the features of the product has to be analyzed to know what feature is getting more positive response or negative response.

2.3. Techniques of sentiment classification

Sentiment of a document or a sentence can be classified by the following methods as shown in Figure 2.

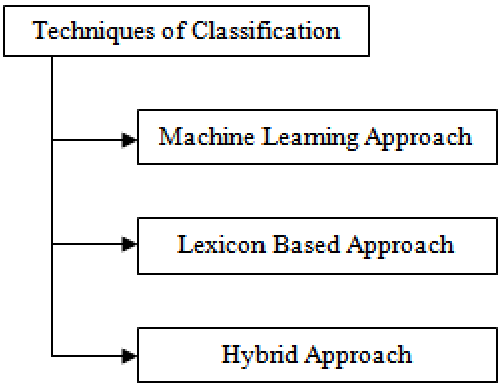

Figure 2. Classification techniques

2.3.1. Machine learning approach

Sentiment classification using machine learning approach is divided into supervised learning and unsupervised learning. Supervised learning way of sentiment classification is used for dataset whereas unsupervised learning methods are used for dataset without the labels. There are many supervised classifiers and the most commonly used classifiers are various probabilistic classifiers and Naive Bayes classifiers.

2.3.2. Lexicon based approach

Sentiment classification using lexicon based has two sub-approaches dictionary based and corpus based. Dictionary based approach depends on the dictionary and searches for seed words in the dictionary. Corpus based approach has a list of words and finds the opinion words in the large corpus in order to extract the semantic orientation.

2.3.3. Hybrid approach

Hybrid approach combines both the machine learning and lexicon based approach in order to perform the sentiment classification.

3. RELATED WORK

The work [2] by Hanif Bhuiyan et al. analyzed the metadata (likes/dislikes/views/comments) in order to retrieve the most relevant and popular video based on the search by using Lexicon sentiment classifier Senti Strength. By applying this approach maximum of 75.4% accuracy is obtained in order to retrieve the relevant and effective videos. Authors Fiktor et al. in [3] proposed Support Vector Machine for classification and a Lexicon method to find the percentage of the sentiment class to understand the character and the performance of Ahok as a governor resulted in an accuracy of 84%. The work by Y.Han et al. in [4] proposed a morphological sentence pattern model which uses aspect based sentiment lexicon part of speech to increase the accuracy of the probability model. The existing model accuracy was increased up to 91.2% and also helps in finding the aspects that are adjacent with respect to part of speech. The paper [5] by S. Rangaswamy et al. proposed a technique to extract and analyze metadata and make decisions by using the video URL. Metadata related to the video is extracted with the stages of categorization, parsing and lookup of the metadata. The work [6] by Vipul et al. use the rule based lexicon model sentiment VADER to analyze the twitter data. Text Blob is used to increase the accuracy and an efficiency of 85-90% is reached.

The paper [7] by Ashok et al. proposed a framework using the machine learning approaches and sentiment analysis to optimize the search queries for suggesting the restaurants. SVM, Naive Bayes, Random Forest and Maximum Entropy are used to eliminate the unwanted data and consider those data which yields better results to the user search queries. Smitashree Choudhury et al. in [8] proposed an unsupervised lexicon approach to find the polarity of the user comment. Senti Wordnet is used to find the sentiment polarity and list is prepared by negating the words in the comment. Combination of Senti Wordnet and the added list performs results in better categorization of the comments and the outcome is that negative sentiment is poorer than the positive. Authors Amar Krishna et al. in [9] proposed machine learning techniques to analyze the sentiment of the comment in order to identify the trends, seasonality and forecasts. Naive Bayes classification is used to calculate the polarity of comments, decompose () function of R is used to find time series and to give the overall trend. Weka is used for forecasting of 26 weeks. Asad Ullah Rafiq Khan et al. in [10] used Naive Bayes algorithm a machine learning approach to perform multi-label classification. MEKA tool is used for multi-label classification. Naive assumption regarding neighborhood keywords performed well as compare to other experimental setting. The paper work [11] by Rishanki Jain used SAS EM Text Miner software to perform sentiment analysis on movie trailer comments to predict the collection of the box-office. Sentiment scores are calculated and gross trend is plotted which shows that initially people were looking forward to the release, after the movie launched viewers did not enjoy.

Authors Siersdorfer et al. in [12] used Naive Bayes classifier for predicting the comment acceptance. Studies of large dataset using SentiWordNet and YouTube metadata revealed strong dependencies between different sentiments expressed in comments, comment ratings and topic orientation discussed in the video content. The paper [13] by Kaushik proposed a better text-based sentiment model for large dataset with the aim to reduce the complexity of sentiment model. ME modeling technique used for determining the polarity of the comments and is observed that ME system outperforms the Naive Bayes and the performance of the ME technique is slightly inferior to the SVM technique. Work by Aliaksei Severyn et al. in [14] defined a systematic approach for opinion mining and a robust shallow syntactic structure for improving model adaptability. A supervised multi-class classifier is used to carry out the effective opinion mining and kernels are defined to improve the feature vectors. The paper [15] by Turney provides the unsupervised learning algorithms to classify the reviews as recommended or not recommended. Semantic orientation of the phrases is estimated using the PMI-IR algorithm and it attains an average accuracy of 74%.The work [16] by Pang classifies a document not by topic but instead by overall sentiment. To perform this machine learning algorithms like Naïve Bayes, Maximum Entropy and Support Machine vectors are used. In terms of relative performance, Naïve Bayes performs worst and SVMs performs the best.

4. METHODOLOGY

Figure 3 is the proposed framework to perform sentiment analysis of the mobile unboxing video comments. Rule based lexicon approach Sentiment VADER is used to label the extracted data. Labelled data is pre-processed to tokenize, remove stopwords, perform stemming and count vectorizer. Naive Bayes algorithm is applied on pre-processed data to classify the comments as either positive or negative.

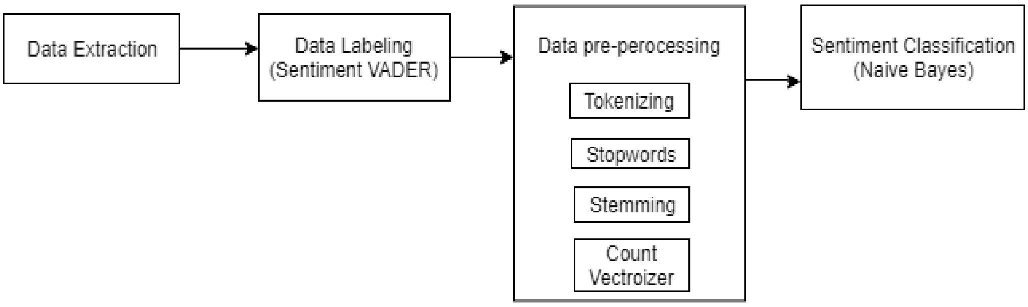

Figure 3. Proposed framework

4.1. Data extraction

The data required for the performing sentiment analysis of the mobile unboxing video is extracted from YouTube. The metadata (likes, dislikes and comments) associated with the video is extracted by using online YouTube scrapper (http://ytcomments.klostermann.ca/). YouTube scrapper extracts all the comments and metadata associated with it. The extracted data are available in the form of CSV and JSON format, CSV format is considered for analyzing the data and classifying the comments.

In this paper unboxing of LG G7 mobile is randomly considered to study the sentiment associated with the video comments. The online YouTube scrapper is used to extract 6248 comments and the metadata associated with the video. The extracted dataset consists of 14 metadata like user id, user name, date, timestamp, likes, comment etc., which provides the details of the user name, time and date when the comment was made and also provides details of the number of likes and dislikes for the comments. With all these metadata available only the comment is concentrated to study the sentiment expressed by the user towards the mobile phone being unboxed.

4.2. Data labeling using sentiment VADER

LG G7 unboxing video dataset does not contain output attribute that labels the comment as positive or negative in order to train a supervised classifier. A rule based lexicon approach Sentiment VADER (Valence Aware Dictionary for sEntiment Reasoning) [17-18] is applied to get the output attribute that labels the comments as positive, neutral or negative. Sentiment VADER not only assigns polarity to the comment but also assigns the intensity value. Unlike polarity based method which classifies the sentences as positive, negative or neutral with the values 1, -1 and 0 respectively, the valence or intensity based method considers the intensity value associated with the words to range from -1 to 1.

For example, the word “good” and “excellent” will have same polarity of 1 in polarity based approach, whereas in valence based “excellent” is considered to be more positive than “good” so the intensity value for excellent will be more than the intensity value of good.

In this paper Sentiment VADER which is a rule based approach is applied using the python package SentimentIntensityAnalyzer. This assigns the intensity of individual comment to what extent is positive, negative and neutral. Along with the 3 values it also provides the compound value which is the normalized value associated with the comments that range from -1 to 1. Cut-off value are chosen for the compound value where the values more than 0.2 are made 1 and values which are less than -0.2 are made -1 indicating positive and negative comment respectively. The values between 0.2 and -0.2 are made 0 making neutral comments. Increase in the range (-0.2 and 0.2) will increase the neutral comments and decrease the positive and negative comments. It becomes difficult to train the classifier with less negative and positive comments. Therefore, the range is considered to be 0.2 and -0.2.

4.3. Data pre-processing

Comments are pre-processed to apply the machine learning algorithm Naive Bayes. The comments with the neutral value 0 are removed and comments with values 1 and -1 are considered to pre-process and apply the Naive Bayes algorithm. Data is pre-processed to remove stopwords, stem the words to the root word using the Porter Stemmer algorithm. Count vectorizer is applied to convert the words as vectors.

4.3.1. Tokenizing

Tokenizing is considered as an important pre-processing step in classifying a text data. Training an algorithm to classify text data by using entire document or a sentence is very hard. So, it is necessary to tokenize the sentence into words and train the classifier with the positive and negative words.

4.3.2. Stopwords

These are the words that do not add too much value in classifying it as either positive or negative word. They are not necessary in classifying the sentence or document so stopwords of English (a, as, is, the etc.,) are removed from the sentences.

4.3.3. Stemming

Stemming is a process of finding the root word of the words. This is done by using the Porter Stemmer algorithm. This reduces the time taken by the algorithm in training all the tense of the word into either positive or negative.

4.3.4. Count vectorizer

Count vectorizer helps in converting the words into vectors which a machine learning algorithm can easily understand. Applying count vectorizer on the data creates a matrix of vectors for the words present in the dataset.

4.4. Sentiment classification using naive bayes

Supervised machine learning algorithm Naive Bayes [19-21] is applied on the pre-processed data to predict the sentiment of the comments. The classifier is trained with data and tested on the unseen data or the testing data. The dataset is divided in the ratio of 7:3 with 70% data for training and 30% for testing. From the data all the neutral comments are removed making it a binary classification. For predicting the comment as either positive or negative the probability of predicting the class has to be 50% each. If the accuracy of predicting single class is more than 50%, SMOTE algorithm is applied in order to balance the probability of predicting a sentiment to 50%. The balanced dataset is then applied with Naive Bayes algorithm to predict the sentiment of the comments. The performance of the Naive Bayes classifier is calculated using the confusion matrix. The confusion matrix gives the count of the correctly and incorrectly predicted positive and negative comments by the classifier.

5. RESULTS AND DISCUSSION

Sentiment VADER applied to label the comments with the range value of 0.2 and -0.2 results effectively labeling the comments. Table 1 shows the result of applying the Sentiment VADER on the dataset. The labeled data is counted for the number of positive, negative and neutral values which is shown in the Table 1. Figure 4 shows the graphically representation of Sentiment VADER result when applied on the dataset. Comments that labeled neutral are removed making a binary classification for predicting the sentiment as either positive or negative. The dataset with positive and negative comment is used to train the Naive Bayes classifier. From Table 1 it is clear that the positive comments are more than the negative comments in the training dataset. The accuracy of predicting positive comments is 62.99% which is more than 50% causing an imbalance. For a binary classification the accuracy of predicting a class has to be 50%. Since accuracy of predicting positive comments is 62.99% creating imbalance SMOTE algorithm is used to restore the balance. SMOTE algorithm applied on the dataset creates a balance for predicting the sentiment class as either positive or negative.

Table 1. Results of sentiment VADER

Sentiment Values	Comment Count	Percentage (%)
Positive (Value 1)	1631	37.29
Negative (Value -1)	955	21..83
Neutral (Value 0)	1787	40.86

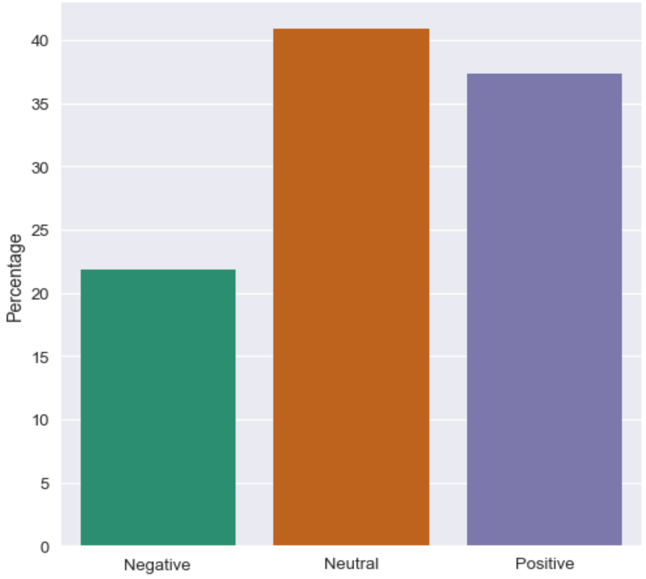

Figure 4. Graphical representation of sentiment VADER result

Naive Bayes classifier is trained with the comments having the labels. The classifier is then applied on the testing dataset and accuracy of the classifier is found to be 79.78% with the F1 score of 83.72%. Confusion matrix also called error matrix visually shows the performance of the classifier for the test data whose actual values are known. It provides the number of correct and incorrect class prediction. Table 2 shows the confusion matrix associated with the Naive Bayes classifier applies on the test data. Out of 410 negative comments in the test dataset 305 were correctly predicted as negative class and 105 were predicted positive but was negative. With 688 positive comments in the test dataset 571 were predicted as positive and 117 predicted negative but was positive.

Table 2. Confusion matrix

	Predicted Negative	Predicted Positive
Actual Negative	305(True Negative)	105(False Positive)
Actual Positive	117(False Negative)	571(True Positive)

6. CONCLUSION

Hybrid approach for performing sentiment analysis of LG G7 mobile unboxing video was implemented. From the extracted dataset of LG G7 mobile unboxing video only the comments were concentrated to perform sentiment analysis. The objective of considering comments for studying sentiment was that it expresses actual opinion than likes/dislikes. Rule based lexicon approach Sentiment VADER was applied to label the comments. In order to perform binary classification neutral comments were removed and the Naive Bayes classifier was trained with 70% of the data. The classifier was then tested on the 30% unseen data and an accuracy of 79.78% and F1 Score of 83.72% was achieved. The confusion matrix obtained by the performance of the classifier shows that the classifier has performed significantly well in predicting the sentiment of the comments. It is also observed that the lexicon approach Sentiment VADER used for the social media text has a good impact on the Naive Bayes classifier in predicting the sentiment.

REFERENCES

[1] Liu B, “Sentiment analysis and opinion mining,” in *Synthesis Lectures on Human Language Technologies: Morgan & Claypool Publishers*, May 2012.

[2] Hanif Bhuiyan, Jinat Ara, Rajon Bardhan, *et al.*, “Retrieving YouTube Video by Sentiment Analysis on User Comment,” in *IEEE International Conference on Signal and Image Processing Applications*, Kuching, Malaysia, Sep 2017.

[3] Fiktor Imanuel Tanesab, Irwan S embiring, Hindriyanto Dwi Purnomo, “Sentiment Analysis Model Based OnYoutube Comment Using Support Vector Machine,” in *International Journal of Computer Science and Software Engineering (IJCSSE)*, vol. 6, no. 8, pp. 180-185, 2017.

[4] Youngsub Han, Kwangmi Ko Kim, “Sentiment Analysis on Social Media Using Morphological Sentence Pattern Model,” in *IEEE 15th International Conference on Software Engineering Research, Management and Applications (SERA)*, London, UK, Jun 2017.

[5] Shanta Rangaswamy, *et al.*, “Metadata Extraction and Classification of YouTube Videos Using Sentiment Analysis,” in *IEEE International Carnahan Conference on Security Technology (ICCST)*, Orlando, FL, USA, Oct 2016.

[6] Vipul Kumar Chauhan, Ashish Bansal, Dr. Amita Goel, “Twitter Sentiment Analysis Using Vader,” in *International Journal of Advance Research, Ideas and Innovations in Technology (IJARIIT)*, vol. 4, no. 1, pp. 485-489, 2018.

[7] Meghana Ashok, *et al.*, “A Personalized Recommender System using Machine Learning based Sentiment Analysis over Social Data,” in *IEEE Students' Conference on Electrical, Electronics and Computer Science*, Bhopal, India, Mar 2016.

[8] Smitashree Choudhury, John G. Breslin, “User Sentiment Detection: A YouTube Use Case,” In: *21st National Conference on Artificial Intelligence and Cognitive Science*, Galway, Ireland, Aug - Sep 2010.

[9] Amar Krishna, Joseph Zambreno, Sandeep Krishnan, “Polarity Trend Analysis of Public Sentiment on YouTube,” in *COMAD '13 Proceedings of the 19th International Conference on Management of Data*, Ahmedabad, India, Dec 2013, pp. 125-128.

[10] AsadUllahRafiq Khan, Madiha Khan, Mohammad Badruddin Khan, “Naive Multi-label classification of YouTube comments using comparative opinion mining,” *Procedia Computer Science, Elsevier*. vol. 82, pp. 57-64, 2016.

[11] Rishanki Jain, “Sentiment Analysis on YouTube Movie Trailer comments to determine the impact on Box-Office Earning,” *Oklahoma State University*, 2018. Available: https://www.sas.com/content/dam/SAS/support/en/sas-global-forum-proceedings/2018/2719-2018.pdf.

[12] Siersdorfer, Sergiu Chelaru, Jose San Pedro, “How Useful are Your Comments? Analyzing and Predicting YouTube Comments and Comment Ratings,” in *Proceedings of the 17th International Conference on World Wide Web*, Raleigh, North Carolina, USA, Apr 2010, pp. 891-900, 2010.

[13] Lakshmish Kaushik, Abhijeet Sangwan, John H.L. Hansen, “Automatic Sentiment Extraction from YouTube Videos,” in *IEEE Workshop on Automatic Speech Recognition and Understanding (ASRU)*, Olomouc, Czech Republic, Dec 2013, pp. 239–244.

[14] AliakseiSeveryn, *et al.*, “Opinion Mining on YouTube,” in *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Baltimore, Maryland, USA, Jun 2014, pp. 1252–1261.

[15] Peter D. Turney, “Thumbs Up or Thumbs Down? Semantic Orientation Applied to Unsupervised Classification of Reviews,” in *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL)*, Philadelphia, Jul 2002, pp. 417-424.

[16] Bo Pang and Lillian Lee, Shivakumar Vaithyanathan, “Thumbs up? Sentiment Classification using Machine Learning Techniques,” in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Philadelphia, Jul 2002, pp. 79–86.

[17] Hutto C, Gilbert E, “VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text,” in *8th international AAAI conference on weblogs and social media (ICWSM)*, 2014.

[18] Chauhan Vipul Kumar, *et al.*, “Twitter Sentiment Analysis Using Vader,” in *I.J Ideas and Innovations in Technology*, pp. 485-489, 2018.

[19] Lopamudra Dey, *et al.*, “Sentiment Analysis of Review Datasets Using Naïve Bayes‘and K-NN Classifier,” in *I.J. Information Engineering and Electronic Business*, pp 54-62, Jul 2016.

[20] Norman Jasmine, *et al.*, “A Naive-Bayes Strategy for Sentiment Analysis on Demonetization and Indian Budget,” in *I.J Pure and Applied Mathematics*, pp. 23-31, 2017.

[21] Bhanap and Kawthekar, “Sentiment Analysis Of Mobile Datasets Using Naïve Bayes Algorithm,” in *I.J Advanced Research in Computer Science*, pp 785-787, Mar-Apr 2018.

BIOGRAPHIES OF AUTHORS

Chaithra V D is PG Scholar at CHRIST (Deemed to be University), Bengaluru. She received a B.Sc in Mathematics, Electronics and Computer Science from Jyoti Nivas College (Autonomous) in 2016, Bengaluru. Her research interests are Big data and Data Analytics.