TEST Engineering & Management

May-June 2020
ISSN: 0193-4120 Page No. 4963-4969

New Approach to Sentiment Polarity Detection by using ML Techniques

[1]Raghavendra Reddy, [2]Gopal K. Shyam

[1,2]School of C and IT, REVA University, Bangalore, India
[1]raghavendrareddy@reva.edu.in, [2]gopalkrishnashyam@reva.edu.in

Article Info
Volume 83
Page Number: 4963-4969
Publication Issue:
May - June 2020

Abstract

Sentiment analysis (SA) is a process of extracting the user's feelings, emotions and verifying whether a user-generated text expresses neutral, positive or negative opinion about a product, people, topic or an event. The development of internet based applications has directed enormous measure of customized surveys for different related data on the Web. These reviews can be collected from various sources such as social media, social network, Wiki, forums, blogs, news and websites. As a result of the growing number of customer reviews, finding appropriate customer reviews will play important rule in reducing information overload. Sentiment Analysis is considered as one of the useful tool for users to extract the required data, as well as to aggregate the collective sentiments of the reviews. Because of rapid development of social media and Internet technologies, sentiment analysis has turned into an essential opinion mining technique. There are three noteworthy systems being utilized for sentiment analysis; Machine learning, dictionary based, and rule-based methodology. Each individual method is having some limitations. So in order to overcome these limitations in this paper we proposes an integrated framework which combines the above methods to achieve better scalability and accuracy.

Keywords**:** *Sentiment Analysis, Machine Learning, Text Summarization, Review Analysis, Opinion Mining*

Article History
Article Received:*19 November 2019*
Revised: *27 January 2020*
Accepted: *24 February 2020*
Publication: *16 May 2020*

1. Introduction

Recently, Social networks has emerge as mainly substantial as easy-to-access, real-time, and affordable data sources in a range of fields [9-10]. And it has turn out to be common for changing ideas, sharing information, merchandising commercial enterprise and trade, strolling political and ideological campaigns, and promotion merchandise and services. The net has large supply of statistics alongside with a number user's perspectives, guidelines and so on. This will become extra vital now not solely to mine the useful resource however the remarks as nicely to improvise on the selection making primarily based on opinions [15-17]. An opinion is actually a superb or terrible view, attitude, or emotion about the entities. In general, sentiment analysis can be done at three different levels [14]:

- **Document Level**: Here whole document is considered in order to express a single opinion about a specific entity.
- **Sentence Level**: Here the given document is divided into number of sentences, and each sentence will express a unique opinion.
- **Feature/Aspect/Word/Term Level/ Finer-Grained Analysis**: Here the given sentence is divided into number of phrases, corresponding to different topics.

Sentiment Analysis Methods

Table. 1 indicates the difference between Sentiment Analysis Methods.

A. Machine learning Approach:

Figure 1 demonstrates the sentiment classification techniques. For the machine learning approach, a collection of function vectors are chosen and a series of tagged corpora is employed to put together a model. The

May-June 2020
ISSN: 0193-4120 Page No. 4963-4969

model can then be used to classify an untagged corpus of text. The resolution of points is indispensable to the success charge of the classification. Most commonly, a range of unigrams (single phrases from a record) or n-grams (two or extra phrases from a record in successive order) are selected as function vectors. The accuracy consequences for these algorithms substantially relies upon on the elements selected. In machine learning approaches we have different methods:

- **Support Vector Machine**: It supports high-dimensional input space. But it requires huge amount of training dataset as well as data collection process is difficult job.
- **Naïve Bayes**: It is suitable for only smaller dataset. But it assumes that all the features are conditionally independent.
- **Maximum Entropy**: It can handle larger dataset and it will overcome the limitation of Naïve Bayes Method.
- **KNN**: It is one of the efficient method. But it requires larger storage space as well as complexity is very high.
- **Multilingual Sentiment Analysis**: It is used to handle the sentimental analysis of different languages. But it requires training dataset for all the languages which we have considered for evaluation.
- **Feature Driven Sentiment Analysis**: it is suitable for larger projects. But complexity is very high.
- **Logistic Regression**: It is one of the efficient method and it will overcome the limitation of Naïve Bayes.

B. Rule Based Approach:

It is utilized by characterizing different principles for getting the opinion, generated by tokenizing each sentence in every report and afterward testing each token, or word, for its presence [18]. In a given record, if there is a word which is classified as a positive sentiment then it is ranked with +1. Generally each post begins with an unbiased score of zero, and was viewed as positive, if the ultimate polarity score was more prominent than zero. And if the polarity score is less than zero then it is marked with negative. After the yield of rule based methodology it will take a look at or ask whether or not the yield is right or not. In the event that the sentence carries any phrase which is absent in the database which may assist in the investigation of movie review, then at that point such phrases are to be added to the database.

C. Lexical Based Approach:

It utilizes a dictionary which is having positive and negative words to decide the sentiment polarity. Preprocessing is considered as first step in lexical approach which is used as a part of text to be analyzed [19]. The initial polarity score is assigned to zero. Polarity score will be updated based on lexicon is present in the dictionary of words or not. And if it is present then is it positive or negative. Based on this score values will be updated. Based on the final sum of polarity score it will classify the given reviews as positive or negative. In this methodology, a dictionary is built to save the polarity scores of lexicons. For computing polarity rating of every phrase of the text, if it is exist in the dictionary, then it is considered to have 'overall polarity score'. For instance, if a lexicon suits a phrase marked as positive within the dictionary, then the complete polarity rating of the textual content is increased. If the normal polarity rating of a textual content is positive, then that textual content is categorized as positive, in any other case it is labeled as negative. Lexical methodology are reported to possess extensively high accuracy. Significant disadvantage of Lexicon based methodology is that, the power of the sentiment classification relies upon the dimension of the lexicon. As the dimension of the lexicon increases this method turns into greater erroneous and tedious.

D. Hybrid based approach:

It uses the most common features of lexicon methodology and ML mechanism for classification [20].

Naïve Baye's technique is based on Baye's probability rule with an assumption of feature independence in an input sample[11]. It works by building a probability distribution model for each lexical element. The classification is achieved by predicting the label that best represents the probability distribution of a test document. The classifier is trained with training samples comprising of reviews with labels of polarity like positive, negative or neutral. Naïve Baye's classifier is optimal solution for classification task when the training data is not huge. As the features increase, the sensitivity of the algorithm decreases. The advantages of Naive Baye's technique are that it is relatively simple and efficient in classification accuracy.

$$P(y| x) = [P(x | y) x P(y)] / [P(x)] \ldots\ldots\ldots\ldots\ldots\ldots (1)$$

Where $x=(x_1, x_2, x_3, \ldots\ldots\ldots, x_n)$ is set of input feature vectors, y is the target class and the equation outputs the maximum probability of the input feature vectors representing the class y.

Logistic Regression, a supervised machine learning model which maps the input data related to different categories into discrete output classes [12]. The logistic regression is based on sigmoid function. The threshold is applied initially to the regression output in order to restrict the output to the range [0, 1] in linear regression. This constitutes the sigmoid function. It is a regression model that is mainly used for classifying a sample input to its class. Logistic regression (LR) is classified into binary, multinomial and ordinal LR algorithms based on the number of output classes and order of the target categories.

May-June 2020
ISSN: 0193-4120 Page No. 4963-4969

Table 1: comparison of Sentiment Analysis Methods

Sl. No	Sentiment Analysis Approach	Classification Methods	Advantages	Disadvantages
1	Machine Learning	Supervised and Unsupervised Machine learning Approach	-Usage of Dictionary is not required -High classification accuracy	-Trained classifier algorithm for one domain may or may not work for other domains.
2	Rule Based Approach	Supervised and Unsupervised Machine learning Approach	-Sentence level analysis performs better than document and aspect level analysis	-Performance and accuracy of the system is completely depending on the parameters considered for the classifier.
3	Lexical Based Approach	Unsupervised Machine learning Approach	-Labelled data is not required	- The size of the dictionary defines strength of the sentiment classification.

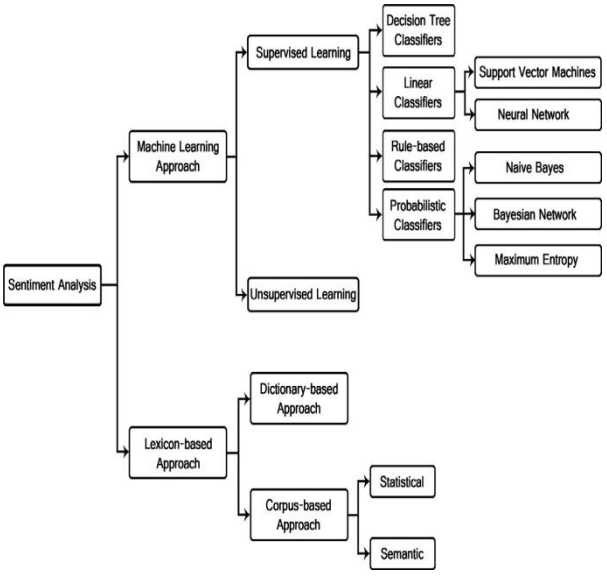

Figure 1: Sentiment classification techniques.

2. Literature Survey

Literature survey which is discussed in the table 2 will helps us to identify some of the possible research challenges that need to be addressed.

Table 2: Summary of Literature survey

Ref. No.	Methodology	Advantages	Drawbacks/ Feature Enhancements	Evaluation Matrices (%)	
				Accuracy	Precision(P), Recall (R), F-Measure(F)
[1]	novel Genetic Algorithm (GA)	Improves scalability & efficiency. It reduces feature size	Extended to cyber-intelligence and other applications	80	P-85 R-89 F-78
[2]	Random Forest	better prediction	It can be extended to other applications	85.95	P-89.2, R-89 F-89
[3]	CNN	good performance	Limited dataset	83.62	P-88, R-87.32 F-87.66
[4]	word embedding clustering-based deep hypergraph model +NN	Improves classification accuracy.	Multi-modal features and task specific embedding learning can be employed.	85	N/A

May-June 2020
ISSN: 0193-4120 Page No. 4963-4969

[5]	CNN	Classification accuracy is improved	Scope to improve the accuracy	71	N/A
[6]	SVM, NB, ME, Stochastic Gradient Descent (SGD)	Good Accuracy	different feature selection mechanism may be identified	85	N/A
[7]	SVM,BOW and TF-IDF approach	Helpful in understanding the human affection.	-needs to identify different emotional context and use appropriate interactive annotation tool.	74.6	F:68% P:86.5%
[8]	SVM, NB, DT , emotion space model (ESM)	ESM is not parameter sensitive. Higher classification accuracy.	Large-scale dataset is needed.	SVM:78.31, NB:63.28, DT:79.21	N/A

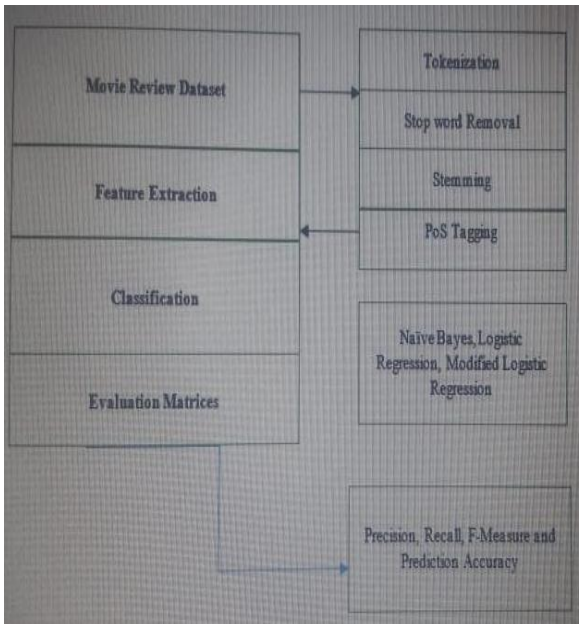

Figure 2: Work flow diagram

There is a requirement to automate the method of sentiment analysis so that it becomes very easy to finding out public's opinions without perusing a huge number of tweets physically. Here we proposed a simple yet efficient model, called MLR for improving sentiment classification accuracy. An analysis is carried out to find out about how to process the dataset and choose specific features to improve prediction performance.

The structure of the paper is organized as follows: Section 2 gives the literature survey associated with this work. Section 3 indicates the proposed work, Section 4 provides the experimental outcomes and discussion. Section 5 indicates conclusion and future work.

The significant commitment of this paper is:

- Two distinctive machine learning techniques such as Logistic Regression and Naive Bayes classifiers are considered for classification of movie reviews using Bigram approach.
- Proposed Modified Logistic Regression (MLR) algorithm for prediction in quantitative and qualitative aspects.
- The performance of the above algorithms are assessed by utilizing Accuracy, F-measure, Precision and Recall. The outcomes generated from this paper shows the higher accuracy as compare to the other authors.

3. Methodology

Problem Definition: Assume that we have collected a large number of documents say 'D' and set of polarity classes C_i, such that C_i={Positive, Negative, Neutral}. The main objective here is we need to find the unknown target function $Q:DC_i$, which describes the polarity of documents according to a golden standard. Here, each document d_i in D where D = {d_1, d_2, d_3, ...d_n} are classified into class C where C = {Positive, Negative}. In multi-class sentiment classification, the d_i is classified into class C where C = {StrongPositive, Positive, Neutral, Negative, Strong Negative}. Here we have used binary classification approach.

3.1 Dataset

Different steps that need to be followed as a part of sentiment analysis is demonstrated in figure 2. Sentiment analysis process begins with preparing dataset containing reviews. By using Natural Language processing (NLP) techniques we need to be pre-processed the dataset as it may contains some noisy. Then the relevant aspects which is required for sentiment analysis needs to be collected and lastly the classifier is trained and tested on dataset.

There are a lot of sources from where one can extract reviews of different movies or products. For our work we have considered Internet Movie Database (IMDb) dataset. And it is having 50,000 reviews. Among that 12500 are positive test reviews, and 12500 positive train reviews. Similarly, there are 12500 negative labeled test reviews, 12500 negative labeled train reviews. Count-Vectorizer and TF-IDF approaches are used to change the textual content into a numerical vector, which is then regarded as input to the proposed approach.

May-June 2020
ISSN: 0193-4120 Page No. 4963-4969

3.2 Preprocessing

Once the dataset is collected often needs to be preprocessed before starting classification and analysis process. Some popular preprocessing steps are:

a) Case Conversion: Here all the letters should be changed over into either lower case or capitalized letters in order to keep away from the ambiguity among "Text" and "text" for additional processing.

b) Tokenization: is utilized to interrupt a sentence into words, phrases, symbols or different significant tokens with the aid of eliminating punctuation marks.

c) Stop-words Removal: The most regularly utilized words like an, a, the, has, have and so forth which convey no significance ought to be expelled from the information content.

d) Stemming: Stemming is the procedure to convey a phrase into its root structure, while disregarding different PoS of the word.

e) Spelling Correction: Spelling revision are often performed by utilizing automated choice of increasingly likely word.

3.3 Feature Extraction

In this phase, the movie features are extracted from each sentence. For discovering the polarity of textual content document, it's necessary to know the sentiment rating with its utilization as well as their relationship with all the close by words. We have used term presence and frequency and parts of speech tagging.

3.4 Classification

The major disadvantage of the logistic regression is when the training data size is little comparative with the quantity of features, than it can help decrease over fitting and end in a more generalized model. And it is associated with 2-class problem i.e algorithm fails when it contrasts and classify the reviews with two independent variables or this algorithm fails whenever we are classifying words based on the comparable meaning.

Modified Logistic Regression (MLR)

To address two-class problem which is present in existing logistic regression techniques, we propose a Modified Logistic Regression (MLR) Technique which is defined as Bilinear Model that is the combination of both joint distribution and the input to output mapping and then predicts the target function. This proposed MLR divides the given input dataset and classifies the reviews based on number of occurrences of bag of words and stop words based on the target variable by correlating the variable, which improves prediction accuracy of the input reviews. This proposed MLR technique also takes the unlabelled dataset as input; this helps by modifying the existing logistic regression technique which helps in overcoming 2-class problem and made working for two independent variable classification and prediction.

In the proposed method classification of reviews is finished by utilizing MLR approach. And it incorporates Bilinear Model which expect that there is a limited arrangement of classes into which the reviews should be classified and training data is accessible for each. The classification procedure is carried out with the aid of combing each joint distribution and the input to output mapping procedures, means the chosen characteristic for classifying the review will be in contrast with comparable phrases as well as the phrases with similar meaning, which can be accomplished the use of the integration of advanced part of speech are going to be classified as similar group of review. This will be done by utilizing following various steps:

- **Support Count for Splitting the Input dataset:**

Before choosing feature for classification we have to set the support count for splitting the given dataset, in our work the support count is set depending on the quantity of opinions viewed for analysis and by splitting the given dataset we will process the data quicker or we will do multiprocessing.

Vect = CountVectorizer (min_df=5)……..(1)

The above equation specifies for the vectorizer we have set the splitting count as 5 based on the input dataset and this will split the dataset according to the count and then, modified logistic regression can be applied to classify the reviews.

- **Classifying based on unlabeled Dataset**

This module will considers unlabelled dataset for analyses and it will be classified dependent on the kind of opinions which is identical to logistic regression approach, but here we incorporated POS module for classifying the opinions dependent on the two independent variables with similar meaning are often classified as similar group which enhance the accuracy of the current logistic regression approach.

ngram_range= (2, 2)) …… (2)

The Modified logistic regression is based on a bilinear condition module with two autonomous input parameters as in linear regression to foresee the likelihood of the input belonging to a unique class. A viable output that represents a class. Utilizing bilinear function, the output values can vary between 0 and 1.

4. Results

The existing algorithms fails while it contrasts and classify the opinions with two autonomous variables or these algorithms fails once we choose to classify based on the phrases which have comparable meaning. In the proposed model we incorporated PoS (Part of Speech) module for classifying the opinions dependent on the two autonomous variables with comparable meaning are often classified as comparable group which improve the accuracy of the existing logistic regression algorithm. Which is not present in normal logistic regression.

May-June 2020
ISSN: 0193-4120 Page No. 4963-4969

The movie dataset is taken from standard website and then classification is done based on the prediction attribute in the data set while classifying we considered splitting attribute and support count modules which is main drawback in the existing logistic Regression. Therefore with the proposed technique we can able to achieve 88% of the accuracy for 25000 instances of reviews where in existing Logistic Regression technique 83% accuracy and in Naive Bayes technique 80% accuracy is achieved, because the algorithms fails in classifying with 2 class and word with similar meaning is treated as separate for classification.

In this paper different parameters are considered for the evaluation of the proposed scheme like precision, recall, F-Measure and prediction of the movie reviews. Table 3 depicts the result of Navie Bayes, Logistic Regression, and Modified Logistic Regression classifiers.

Table 3: Average Evaluation measure of three Classifiers

Classifiers	Evaluation Measures		
	Precision	Recall	F-Measure
NB	68.72%	65.2%	70.8%
LR	75.2%	73.2%	73.2%
MLR	86%	78%	76.5%

Figure 3 describes the accuracy of prediction i.e. prediction percentage for movie based reviews. Here, x-axis represents the prediction accuracy in percentage, and y-axis represents the number of reviews considered for analysis. Initially, we considered 10,000 reviews as input dataset for training and testing the dataset, respectively. We have achieved around 84% accuracy from the proposed MLR technique, and 78% of accuracy from exiting algorithms. Then, 15,000 reviews are considered as input dataset for training and testing the dataset, respectively. An accuracy of around 86% from the proposed MLR technique, and 80% of accuracy from exiting algorithms is achieved. Finally 25,000 reviews are considered as input dataset for training and testing dataset, respectively. An accuracy of around 88% was achieved from the proposed MLR technique and 82% of accuracy from exiting algorithms.

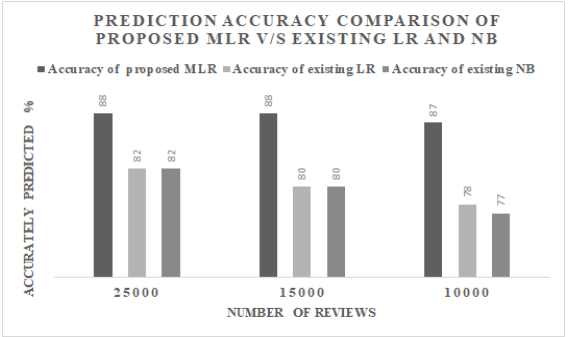

Figure 3: Accuracy comparison of proposed MLR v/s existing LR and NB.

We observe that the proposed MLR technique proves more accurate when compared to exiting techniques even after varying the size of the dataset. The existing algorithms fails when it contrasts and classifies the opinions with multiple autonomous variables or these algorithms fails when we choose to classify dependent on the phrases which have comparable meaning. The integration of part of speech module for classifying the opinions dependent on the multiple independent variables with comparable meaning enhance the accuracy of the current LR algorithm.

5. Conclusion and Future work

This paper concludes the analysis and classification for various movie reviews taken from different Movie based applications. This paper applies various methods to analyze the reviews on the movies like data preprocessing, clustering and classification. This paper presents modified logistic regression based classification technique to analyze the reviews taken from various Movie based applications. The details of various movies are taken from Movie based application. Then classification is done based on the prediction attribute in the data set while classifying we considered the total number of comments and rating of the users on the movies. Therefore with the proposed technique we can able to achieve 88% of the accuracy for 25000 instances of reviews. As shown in the comparison of accuracy of proposed techniques is more even varying the size of the dataset. In future this classification results can be implemented directly on the Movie based applications using web crawling method and multiple movie reviews can be analyzed. This work can also be extended with various machine learning algorithms to improve the accuracy of sentimental analysis.

Reference

[1] Farkhund Iqbal, Jahanzeb Maqbool, Benjamin C. M. Fung, Rabia Batool, Asad Masood Khattak, Saiqa Aleem and Patrick C. K. Hunga. “A Hybrid Framework for Sentiment Analysis Using Genetic Algorithm Based Feature Reduction”, IEEE Access, vol. 7, pp. 14637-14652, 2019.

[2] Sahu, Tirath Prasad and Sanjeev Ahuja. “Sentiment analysis of movie reviews: A study on feature selection & classification algorithms.” 2016 International Conference on Microelectronics, Computing and Communications (MicroCom), pp. 1-6, 2016.

[3] Z. Jianqiang, G. Xiaolin and Z. Xuejun, "Deep Convolution Neural Networks for Twitter Sentiment Analysis," in IEEE Access, vol. 6, pp. 23253-23260, 2018.

[4] X. Yuan, M. Sun, Z. Chen, J. Gao and P. Li, "Semantic Clustering-Based Deep Hypergraph Model for Online Reviews Semantic Classification in Cyber-Physical-Social

May-June 2020
ISSN: 0193-4120 Page No. 4963-4969

Systems," in IEEE Access, vol. 6, pp. 17942-17951, 2018.

[5] J. Zhang and C. Chow, "MOCA: Multi-Objective, Collaborative, and Attentive Sentiment Analysis," in IEEE Access, vol. 7, pp. 10927-10936, 2019.

[6] Tripathy, A., Agrawal, A., & Rath, S. K, "Classification of sentiment reviews using n-gram machine learning approach", Expert Systems with Applications, Elsevier, vol 57, pp 117–126, 2016.

[7] Zhang, D., Xu, H., Su, Z., & Xu, Y, " Chinese comments sentiment classification based on word2vec and svm perf", Expert Systems with Applications, Elsevier, vol 42, pp 1857-1863, 2015.

[8] Niu T., Zhu S., Pang L., El Saddik A., "Sentiment Analysis on Multi-View Social Data", International Conference on Multimedia Modeling Multi Media Modeling. MMM 2016. Lecture Notes in Computer Science, Springer, vol 9517, pp 15-27, 2016.

[9] Luo, B., Zeng, J., & Duan, J, "Emotion space model for classifying opinions in stock message board", Expert Systems with Applications, Elsevier, vol 44, pp 138–146, 2016.

[10] Liu, S. M., & Chen, J.-H, "A multi-label classification based approach for sentiment classification", Expert Systems with Applications, Elsevier, vol 42, pp 1083–1093, 2015.

[11] Tang, B. Qin, F. Wei, L. Dong, T. Liu, M. Zhou, "A joint segmentation and classification framework for sentence level sentiment classification", IEEE/ACM Transactions on Audio, Speech, and Language Processing, pp 1750–1761, 2016.

[12] Hichem Rahab, Abdelhafid Zitouni, Mahieddine Djoudi, "SANA: Sentiment analysis on newspapers comments in Algeria", Journal of King Saud University - Computer and Information Sciences, Elsevier, 2019

[13] Ravi, K., & Ravi, V," A survey on opinion mining and sentiment analysis: Tasks, approaches and applications", Knowledge-Based Systems, vol 89, pp 14–46, 2015

[14] Liu, Y., Bi, J.-W., & Fan, Z., "Multi-class sentiment classification: The experimental comparisons of feature selection and machine learning algorithms", Expert Systems with Applications, vol 80, pp 323–339, 2017

[15] Serrano-Guerrero, J., Olivas, J. A., Romero, F. P., & Herrera-Viedma, E, "Sentiment analysis: a review and comparative analysis of web services", Information Sciences, vol 311, pp 18-38, 2015.

[16] Liu, Y., Bi, J.W., & Fan, Z.P, "Ranking products through online reviews: A method based on sentiment analysis technique and intuitionistic fuzzy set theory", Information Fusion, vol 36, pp 149-161, 2017.

[17] Lee, A. J., Yang, F. C., Chen, C. H., Wang, C. S., & Sun, C. Y, "Mining perceptual maps from consumer reviews", Decision Support Systems, vol 82, pp 12-25, 2016.

[18] Chatterjee, A., Gupta, U., Chinnakotla, M. K., Srikanth, R., Galley, M., & Agrawal, P, "Understanding Emotions in Text Using Deep Learning and Big Data", Computers in Human Behavior, vol 93,pp 309–317, 2019

[19] S. Poria, E. Cambria, R. Bajpai, and A. Hussain, "A review of affective computing: From unimodal analysis to multimodal fusion," Inf. Fusion, vol. 37, pp. 98–125, 2017

[20] N. Majumder, D. Hazarika, A. Gelbukh, E. Cambria, and S. Poria, "Multimodal sentiment analysis using hierarchical fusion with context modeling," Knowl. Based Syst., vol. 161, pp. 124–133, 2018.