An Efficient Approach in Selection of Information-Gaining Features Using Sentiment Analysis

Arun Shalin LV[1], Dilli Babu.M[2], A.Manonmani[3], Y.Mary Reeja[4], Maria Sushma S[5], A.Rajesh Kumar[6]

[1]Assistant professor ,Department of Data science and cyber security Karunya institute of technology and sciences ,Coimbatore, Email: arunshalin7@gmail.com
[2]Assistant professor, Panimalar Engineering college, Department of information technology, Chennai, Email: deenshadilli@gmail.com
[3]Assistant professor (SG),EIE ,Saveetha Engineering College, Chennai, Email: manonmani@saveetha.ac.in
[4]Professor, Department of ECE, Saveetha School of Engineering, Saveetha University, Chennai, Email: maryreejay.sse@saveetha.com
[5]Assistant Professor, Dept of Electrical and Electronics Engineering ATME College of Engineering, Mysore, Email: s.mariasushma1@gmail.com
[6]Associate Professor, Computer Science and Engineering, N.S.N. College of Engineering and Technology, Karur, Email: arurajesh1980@gmail.com

Received: 11.04.2024 Revised : 19.05.2024 Accepted: 26.05.2024

ABSTRACT

Sentiment analysis, also known as opinion mining, has become increasingly important as the number of online review and social networking sites continues to expand rapidly. People's opinions on items, services, programs, and even politics whether they're positive or negative are heavily influenced by the feedback they receive from others who have used the same item. Big data refers to the massive amounts of data that can be analyzed because of this. The use of big data has expanded into every facet of the global economy. Feature selection is the process of selecting a subset of input variables that best separates input data while also reducing the impact of noise or unsuitable variables and producing efficiently higher prediction outcomes. Term Frequency-Inverse Document Frequency (TF-IDF) is used as a feature extraction method. The inverse document frequency (IDF) is used to standardize the term frequency for each word in the TF-IDF representation, hence decreasing the importance of a word's occurrence count. The focus here is on opinion mining using Information Gain (IG) based feature selection. Features' individual contributions to lowering the entropy can be used to derive IG. This reduces the processing time required by the learning algorithms while simultaneously improving classification accuracy by discarding superfluous characteristics from the initial feature set. The suggested method is evaluated using the Naive Bayes, Classification and Regression Tree (CART), and K-Nearest Neighbor (KNN) classifiers. It has been demonstrated that the proposed method yields optimal results.

Keywords: KNN, IDF, CART, IG, TF-IDF, social.

1. INTRODUCTION

Big data refers to the massive volumes of information present in fields such as remote sensing, healthcare records, and social media. These records might be completely unstructured, somewhat unstructured, or fully organized. Big data is a hot topic in computer science research, and the importance of sentiment analysis cannot be overstated [1]. Sentiment analysis for massive datasets is important for examining customer responses to products. Reviews posted on social media platforms like Twitter provide quick access to online data analysis. An issue like feedback with insufficient data might arise in the case of static sentiment analysis performed on a tiny data set. However, by collecting the tweets' underlying emotions, they are able to improve their offering. Data mining was the next formidable obstacle. The traditional relational database management system calls for a powerful computer setup. Using the Hadoop platform, which has an inbuilt structure like the map reduce utilized in query execution, yields quicker results. The end user will be able to make informed business decisions based on the analysis's findings, which will be presented to them in a graphical format [2].An NLP technique called "sentiment analysis" monitors how people feel about a product or topic generally.

Opinion mining is another name for sentiment analysis. Also included is the development of a system to collect and analyze the feedback provided in the form of blog entries, comments, tweets, or reviews. The sentiment analysis may be put to many different purposes. For instance, it may be able to judge the success of advertising campaigns or product launches, revealing the most well-liked items and helping to determine why they are so well received. Sentiment analysis is fraught with difficulties. The first is that a point of view might look good in one set of circumstances but bad in another. Second, Many other modes of expression are available. The conventional approaches to text processing rely on the idea that two texts with a similar meaning may be distinguished by a small number of subtle variations between them [3]. Sentiment analysis has traditionally focused on two tasks: document categorization and polarity determination. It's possible that this polarity is neutral, positive, or negative. Sentiment analysis is conducted on three distinct levels, namely (a) the document level, where the entire document is classified as positive, negative, or neutral; this is referred to as the document level sentiment classification. (b) At the sentence level, the sentences are categorized according to whether they are positive, negative, or neutral. Sentiment classification at the aspect and feature levels, also known as the aspect level sentiment classification, sorts phrases and documents into positive, negative, and neutral categories according to the prominence of certain features. [4].

Every industry and activity in the global economy now relies heavily on big data. A typical field, like the stock market, will provide a flood of data, such as bids, buys, and puts, every single second. The local and international news, natural disasters, and government reports are just few of the areas that are affected by this type of data. Given the sheer volume and complexity of the data that has to be properly categorized and presented to users for their comfort and ease of access, it is next to impossible to ensure that users receive the correct information [5]. Big data classification according to format, processing method, analysis to be applied, data sources, data load, processing, analysis, and storage are major obstacles that must be overcome in order to address the aforementioned problems. Classification makes use of a wide variety of methods. Here, we propose the use of K-Nearest Neighbor (KNN) classifiers, a Classification and Regression Tree (CART), Naive Bayes, a feature selection technique based on Information Gain (IG), and a feature extraction method based on Term Frequency-Inverse Document Frequency (TF-IDF).

2. BACKGROUND WORK

Since NLP techniques are needed for the automatic identification of the features that are being analyzed, feature extraction is the greatest challenge in sentiment analysis. When you add in the fact that the views on the stores themselves are features in an area that is not pre-specified, you have a far more difficult task on your hands [6]. Aside from what would seem like the most important factor (the price of the goods), additional factors include how easy it is to use the site, how much it costs to ship, how long it takes to ship, how reliable the business is, how well the fruit is packaged, and how well customers are cared for. The TFIDF statistic, which calculates weights for each phrase based on their relative relevance, will be used as the weighted frequency statistic. The term's usefulness for a given text relies on how many words it contains. This is why a TFIDF denominator needs to be recalculated every time a document's word count changes. The amount of records containing the word (the words that occur on the records that are down-weighted) has also been changed. All of the documents in this corpus will now be represented as a Bag of Words (BoW), which is both a simple and effective approach to text categorization. In the process of emotion categorization, the BOW employs a complex mechanism for comprehending the texts. The Part of Speech (POS) tagging: the POS tagging is a linguistic approach that has been used since 1960 and has recently garnered a lot of attention from NLP researchers interested in the extraction of product features. Common examples are nouns and noun phrases. Each word in a document will be given a unique tag by the POS tagger, as well as an indication of its morphological category, such as noun, verb, or adjective. The POS taggers will work well if the feature is extracted explicitly [7].

Positive, negative, and neutral textual evaluations on the same topic have been categorized using document-level attributes. The document's contents will often dictate the tone's general polarity. The potential positivity, negativity, or neutrality of an utterance is identified using elements at the phrase level. Finally, the name, location, and address have all been extracted using the different entity characteristics. Both single-word and multi-word applications of the TF-IDF weighting system have been studied for document sentiment categorization [8]. Using the frequency-based feature extraction technique, we have identified a list of nouns and a subset of noun phrases for each text. To do this, we append to the text any words labeled with the POS of the "N" (called the noun) and any noun set labeled with the POS of the "N to N" (thought to be noun phrases). The number of nouns acquired in the previous stage will be counted and compared to sum of words contained in said lists in the next step. This is achieved by constructing a new set in which all the terms retrieved in the first phase are included, and then specifying their frequency. Important characteristics will be retrieved in the following stage, and they will be the nouns whose frequency is greater than a threshold. This threshold for frequency can be set at any value based on past experience. Finally, the following concept is utilized to extract characteristics that occur less frequently than the frequency criterion. All the aspects are described using these opinion words [9].

Feature extraction refers to the process by which feature subsets are extracted from the data in preparation for the use of a learning algorithm. The optimal feature subset will have a small number of extremely accurate dimensions and will leave out the extraneous ones. Additionally, the curse of dimensionality can be avoided at this first processing phase. This is essential text processing for the purpose of sentiment analysis, since it transforms a text fragment into the feature vector. Making a good choice of features is essential for achieving accuracy in the learning job [10]. Dependency grammar-based feature extraction: is a bootstrapping technique that gets going with a minimal set of words. However, the retrieved features would be employed in the succeeding cycle of feature extraction, thus depleting the lexicon in the process. The approach is grounded in the rules that may be found in language-dependent relationships. For instance, in the statement "this phone has an appearance that is good," the characteristic "appearance" is extracted since it is understood that the word "good" is an opinion term.

3. METHODOLOGY

Data mining's approach to dimensionality reduction will be feature selection, which selects subsets of original features according to predetermined criteria. Data mining techniques can be sped up, mining settings may be optimized, and unnecessary or noisy data can be filtered out of the final output, among other benefits. In this part, we covered the TF-IDF feature extraction approach, the IG feature selection method, Naive Bayes, the KNN, and the CART.

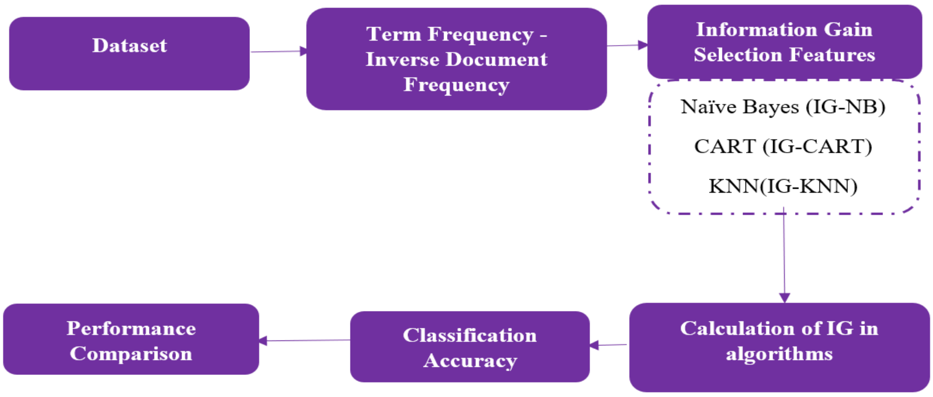

Dataset
Term Frequency - Inverse Document Frequency
Information Gain Selection Features
Naïve Bayes (IG-NB)
CART (IG-CART)
KNN(IG-KNN)
Calculation of IG in algorithms
Classification Accuracy
Performance Comparison

Fig 1. Process of proposed methodology

3.1 Dataset

The Amazon book reviews provided new information for the development of the emotion domain. The reviewers' names, locations, review dates, review titles, evaluated products, and review descriptions were all included alongside numerical ratings ranging from 0 (not reviewed) to 5 (highly recommended) for each book. Reviews with fewer than three stars were considered unfavourable, while those with three or more were considered good. Since the remaining polarity was ambiguous, the remainder had to go. After all the data was sorted, we were left with a dataset that was rather evenly split between positive and negative ratings (800 samples in total).

3.2 Term Frequency - Inverse Document Frequency (TF - IDF)

The phrase Frequency (TF) measures the actual frequency with which a given phrase appears inside a given document. In addition, the percentage of occurrences of each term in the text is displayed. If a paper has 100 words and the word "dog" appears twice, its frequency would be 2/100=0.02. The frequency with which a given term t (the word or the token) appears in document d may be properly defined as TF. Additionally, the length of each document may differ, with a phrase appearing more frequently in larger papers than shorter ones (Kumar &Abirami, 2015). To achieve this normalization, we will divide the the document's total word count by the frequency with which specific words appear.

$$Normalized\ TF = \frac{tf(t,d)}{n_d}$$

In which, tf(t,d) = the term frequency
nd = sum of all the words in doc d.
If you provide DF a word t, it will return the frequency with which particular word appears in documents. Inverse document frequency (IDF) is a measure of the significance of a phrase that is the polar opposite of DF. After the frequency of each phrase is determined, they will all be of equal weight (Basari et al., 2013). It is also common knowledge that words like "the," "of," and "that" may show up rather frequently in writing.
While the TF-IDF is most often employed for document ranking through the relevance of various text mining tasks like search engine rankings, it can also be utilized for the categorization of texts in the Naive Bayes. If there are 100 words in a text and the word "dog" appears twice, the TF for the word "dog" would be 2/100=0.02. If there are 10 million documents and dog appears in 1000 of them, then the information distribution function (IDF) is log (10,000,000/1000) = 4. Thus, the TF-IDF weight will equal 0.02*4=0.08, the product of two small numbers.

3.3. IG based Selection of Features

The IG will be used to evaluate the quality of a word in the event of a machine learning-based categorization. Features' contributions to a decrease in the global entropy can be used to compute the IG (Sharma &Dey, 2012). For a given instance or tuple in partition D, the entropy is assumed to be the information required to label the instance with a certain class.

$$Info(D) = -\sum_{i=1}^{m} (P_i) \log_2 (P_i)$$

Where m is the total number of classes (m=2 for a binary classification) and Pi is the estimated probability that a random instance found in partition D belongs to class Ci, calculated as |Ci, D| / |D| (where |D| is the total number of instances and |Ci, D| is the proportion of those instances that fall into class Ci). The reason that this uses a compact encoding scheme is explained by the existence of a log function with a base of 2. If a feature attribute Aa1,...,av requires partitioning (or classifying) an instance in D, then D will be divided into v partitions set D1,D2,....,Dv. A categorization requires the actual data size, expressed in bits. The IG has ranked the features, and then picked the ones with the highest IG score. As a result, the entropy of the system can be decreased by proper classification of the examples used to rank the characteristics.

$$Info_A(D) = -\sum_{j=1}^{v} \frac{|D_i|}{|D|} \times Info(D_j)$$

4. IMPLEMENTATION RESULTS

4.1 Naïve Bayes-Information Gain

In this scenario, we will make use of a probabilistic classifier that is founded on the Bayes theorem in addition to a number of other robust and independent assumptions. Given the value of the class variable, this is the most basic form of the Bayesian network, and it assumes that all characteristics can be considered independent. Each characteristic in this category was thought to be conditionally independent of the others, leading to the feature's name, conditional independence. The method is then used to a subset of issues, primarily those whose formulation involves grouping the item into a distinct category. In this way, it examines every occurrence of the positive and negative series to determine if the word probability is being used positively or negatively:

$$P\ (\text{sentiment} \mid \text{sentence}) = P\ (\text{sentiment}).\ P\ (\text{sentence})\ /\ P\ (\text{sentence})$$

Naive Bayes has been trained on a single set of annotated characteristics (the labels being either "positive" or "negative") to predict the likelihood of the word appearing in a positive or negative context. In the realm of statistics, the accuracy of a management system is defined by how near its measurements get to the actual value of a given quantity. Recall refers to the proportion of relevant instances recovered, whereas precision measures how well those instances match the query. Below is a description of how this Naive Bayes algorithm works in practice.
First, let's say we have a dataset in which the letters D indicate the classes. Each tuple will be represented by the n-dimensional element vector X = (x1, x2,....,xn).

The real number of classes, C1, C2, C3,..., Cm, shall be taken into account. When the Naive Bayesian classifier assigns an unknown tuple X to class Ci, it also predicts that X will be a member of a class with a greater posterior probability that is conditioned on X. The Bayes Theorem will be used to calculate the aforementioned posterior probabilities if and only if P(CiX) > P(CjX) for 1 j m, and if I j.

4.2 IG-Classification and Regression Tree

The CART process generates a series of trimmed trees in a nested structure. During the pruning process, a "honest" or "right sized" tree is determined by assessing the accuracy of the tree's predictions. The CART does not provide any training information based on intrinsic performance characteristics since these attributes cannot be trusted. Instead, the performance of the tree was tested using cross-validation and by creating independent test data; tree selection will commence after the test-data-led assessment is completed. For example, in the above equation, nodes are separated based on the Gini impurity criterion:

$$l(t) = \sum_{i=j} p(i \mid t) p(j \mid t)$$

For a CART representation of input X, there exists a binary tree that yields the output. The Y, a random variable with a stochastic correlation to the input, will be approximated in this way. The following deterministic mapping illustrates this: Each leaf node in the tree will be tagged with the binary function of the input X. Each such external node will be linked to this for a specific Y-labeled output. The X input will be compared to the output of a binary function that takes its initial value at the root of the tree. If the result is zero, then the left fork is taken. In case of a "1" result, the correct option will be selected. The process will iterate until the external leaf or node is reached, at which point the Y will be linked to the output label. Expected losses range from Y to, and the tree finds their minimum value.

4.3 IG-K-Nearest Neighbour

In addition, the KNN will rank almost all reviews in the training set in order to categorize the review that comes first in the classification. According to research by Alsaffar and Omar (2015), the test review's grade is determined by the average class weights of its reviewers. The following equation depicts a weighted sum used in the KNN classification:

$$score(d,t_i) = \sum_{d_j = KNN(d)} sim(d,d_j)\delta(d_j,c_i)$$

Wherein (d) stands for a collection of K-Nearest Neighbors for examination. The (dj,ci) can be either 1 or 0 depending on whether or not dj is a member of ci. If the test review d belongs to the group with the highest total weight, then. Here's how the KNN classifier does its thing:

1. Initialize the K value.
2. Calculate the distance between the input data and the training datas.
3. Sort the actual distances.
4. Take the highest value of KNN.
5. Apply a simple majority.
6. Make a prediction about the class label using the input data and some additional neighbours.

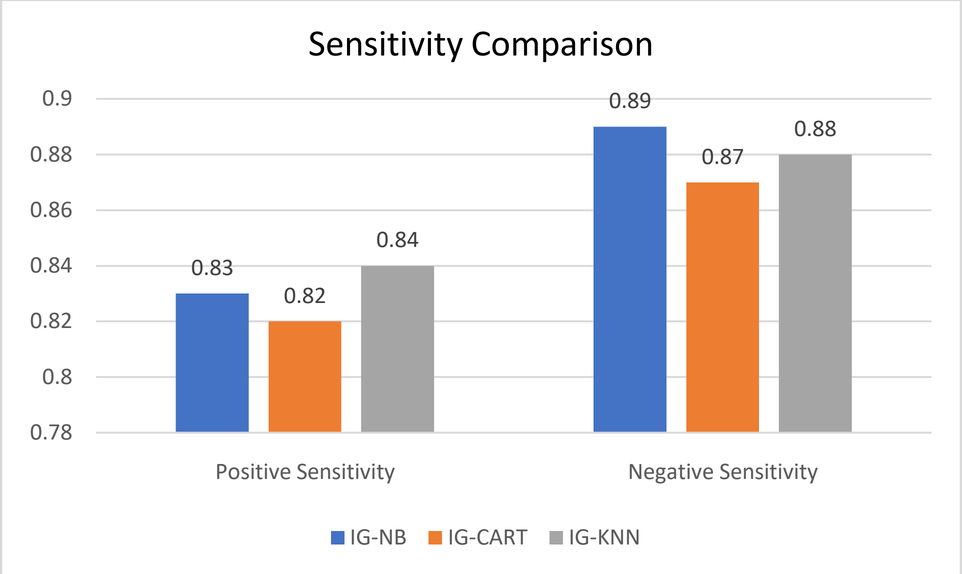
In this case, we have three distinct categories identified as X, Y, and Z. A label for a certain category will be found for the data set P. In this example, K=5. The Euclidean distance between each pair of objects will be calculated. There are a total of four neighbors in the case of the X class label. Z has a single tuple in its class label. Thus, a sample P is placed in the class X as its primary class.

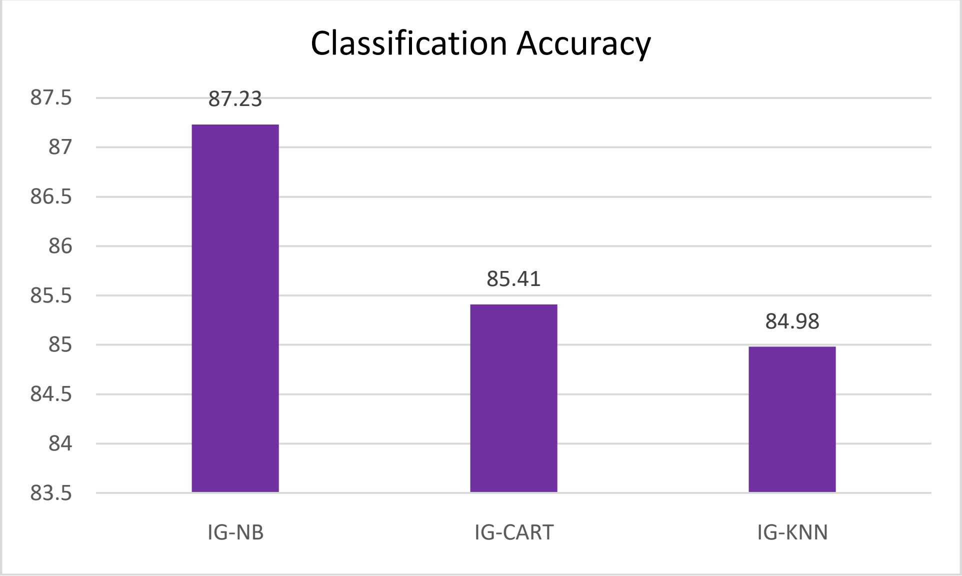
4.4 Results Comparison

The Google Map Reduce framework is widely used as an environment implementation of big data. Both a Map and a Reduce operation are included in this model. Users define calculations in the former, while results are aggregated in the latter. The Apache Hadoop framework is one such paradigm; it is an open-source Java program that facilitates the processing and administration of massive information. It builds upon a Hadoop DFS to distribute data files across several nodes for accelerated transfer speeds. The Apache Hadoop platform serves as the foundation for the proposed system. Three different IG algorithms—the IG-KNN, IG-CART, and IG-Naive Bayes—are employed here. Table following displays the findings summary. The figures below illustrate the classification accuracy, the positive and negative sensitivity, and the positive and negative predictive values.

Table 1. Results Summary

Features	**IG-NB**	**IG-CART**	**IG-KNN**
Classification accuracy	87.23	85.41	84.98
Positive Sensitivity	0.83	0.82	0.84
Negative Sensitivity	0.89	0.87	0.88
Predictive Value for positive	0.89	0.87	0.85
Predictive Value for negative	0.83	0.82	0.84





It is clear from the preceding illustration that IG-Naive Bayes outperforms IG-CART and IG-KNN in terms of classification accuracy. The Naive Bayes classifier has various advantages, including the ability to produce competitive results at a faster rate than more complicated methods. Although it assumes unrealistically that features have been independent of each other, the fact that it is approachable makes it quite popular. Aside from text classifications, this Naive Bayes classifier has shown to be quite active in the fields of practical application such as medical diagnostics and system performance monitoring.

5. CONCLUSION

The primary goal of this study is to facilitate faster processing of data sets through the use of sentiment analysis. It also involves a bias against differing points of view. Opinionated text contains more dimensions, which means it might affect classifier accuracy. As a common property utilized in combination with the filtering method, the IG was also proposed in this study. The IG is calculated for each feature and then the features are ranked from best to worst. Classifiers used include Naive Bayes, KNN, and CART. Both the Naive Bayes and the CART are effective in weeding out superfluous details and variables, respectively. The KNN performs well in all applications with multiple class labels, and it is highly resilient on noisy training data. The findings demonstrated that the IG-Naive Bayes had a greater accuracy of classification than the IG-CART and the IG-KNN by roughly 0.87% and 1.02%, respectively.

REFERENCES

[1] Ramesh, R, Divya, G, Divya, D &Merin, KK 2015, 'Big Data Sentiment Analysis using Hadoop IJIRST', International Journal for Innovative Research in Science & Technology, vol. 1, no. 11

[2] Ramanujam, RS, Nancyamala, R, Nivedha, J &Kokila, J 2015, 'Sentiment analysis using big data. In Computation of Power, Energy Information and Commuincation (ICCPEIC)', 2015 International Conference on IEEE, pp. 0480-0484

[3] JeevanandamJotheeswaran, DR &Kumaraswamy, YS 2013, 'Opinion mining using decision tree based feature selection through Manhattan hierarchical cluster measure', Journal of Theoretical and Applied Information Technology, vol. 58, no. 1, pp. 72-80.

[4] Sharma, A &Dey, S 2012, 'A comparative study of feature selection and machine learning techniques for sentiment analysis', In Proceedings of the 2012 ACM Research in Applied Computation Symposium, ACM, pp. 1-7

[5] Koturwar, P, Girase, S &Mukhopadhyay, D 2015, 'A survey of classification techniques in the area of big data', arXiv preprint arXiv:1503.07477

[6] Asghar, Muhammad Zubair, Aurangzeb Khan, Shakeel Ahmad &FazalMasudKundi 2014, 'A review of feature extraction in sentiment analysis', Journal of Basic and Applied Scientific Research, vol. 4, no. 3, pp. 181-186

[7] Kumar, JA &Abirami, S 2015, 'An Experimental Study Of Feature Extraction Techniques In Opinion Mining', International Journal on Soft Computing, Artificial Intelligence and Applications (IJSCAI), vol. 4, no. 1

[8] Kumar, PS &Kathiravan, AV 2015, 'An Efficient Product Hybrid Feature Classification on Opinion Mining using Ant Optimization Rule', International Journal of Computational Intelligence and Informatics, vol. 4, no. 4

[9] Golpar-Rabooki, E, Zarghamifar, S &Rezaeenour, J 2015, 'Feature extraction in opinion mining through Persian reviews', Journal of AI and Data Mining, vol. 3, no. 2, pp. 169-179.

[10] Jeong, H, Shin, D & Choi, J 2011, 'Ferom: Feature extraction and refinement for opinion mining', Etri Journal, vol. 33, no. 5, pp. 720- 730.

[11] Mandal, S & Gupta, S 2016, 'A Lexicon-based text classification model to analyse and predict sentiments from online reviews', In Computer, Electrical & Communication Engineering (ICCECE), 2016 International Conference on IEEE, pp. 1-7. 91.

[12] Manek, AS, Shenoy, PD, Mohan, MC &Venugopal, KR 2017, 'Aspect term extraction for sentiment analysis in large movie reviews using Gini Index feature selection method and SVM classifier', World wide web, vol. 20, no. 2, pp. 135-154

[13] Azmin, S &Dhar, K 2019, 'Emotion Detection from Bangla Text Corpus Using Naïve Bayes Classifier', in 4th International Conference on Electrical Information and Communication Technology, EICT no. December, pp. 1-5.

[14] Bagi, R, Dutta, T & Gupta, HP 2020, 'Cluttered TextSpotter: An End-to-End Trainable Light-Weight Scene Text Spotter for Cluttered Environment', IEEE Access, vol. 8, pp. 111433-111447,

[15] G. Xu, Y. Meng, Z. Chen, X. Qiu, C. Wang, and H. Yao, "Research on Topic Detection and Tracking for Online News Texts," IEEE Access, vol. 7, pp. 58407–58418, 2019.

[16] Gao, Y & Zhu, W 2016, 'Detecting affective states from text based on a multi-component emotion model', Comput. Speech Lang., vol. 36, pp. 42-57, 2016

[17] Ghosh, S, Hiware, K, Ganguly, N, Mitra, B & De, P 2018, 'Emotion detection from touch interactions during text entry on smartphones', Int. J. Hum. Comput. Stud., vol. 130, pp. 47-57,

[18] Park, D, Kim, S, Lee, J, Choo, J, Diakopoulos, N &Elmqvist, N 2018, 'ConceptVector: Text Visual Analytics via Interactive Lexicon Building Using Word Embedding', IEEE Trans. Vis. Comput. Graph., vol. 24, no. 1, pp. 361-370.

[19] Park, SH, Bae, BC & Cheong, YG 2020, 'Emotion recognition from text stories using an emotion embedding model', in Proceedings IEEE International Conference on Big Data and Smart Computing, BigComp, pp. 579-583.

[20] S. B. Sadkhan and A. D. Radhi, "Fuzzy Logic used in Textual Emotion Detection," in 2017 2nd Al-Sadiq International Conference on Multidisciplinary in IT and Communication Science and Applications, AIC-MITCSA 2017, 2017, pp. 242–245.