International Journal Of Multidisciplinary Research In Science, Engineering And Technology (IJMRSET)

Volume 1, Issue 1, April 2020

Text Mining of User Reviews to Predict Sentiment Using Machine Learning

R.Akila, S.Revathi

Dept. of CSE, BSA Crescent Institute of Science & Technology, Chennai, India

ABSTRACT: Online reviews are an important aspect for customers who buy the product and the organization who sell the product. A review can be about a product, a movie, a political campaign etc. A review can be given in simple words or in detail depending upon how much did the customer like the product. Organizations can track how well their product is fairing in the market and will it be able to sustain longer in the market through product reviews.Sentiment prediction or sentiment analysis is one way of telling how far are the reviews true , effective and will help the people. The objective of the work is to improve the prediction upon the review of the customer on a particular brand or product and bring out the feelings of customers via the reviews. The prediction which is being done is tried to be made more precise by increasing the efficiency and the working of the algorithm , as the prediction helps in expressing the feeling of the people in one word. Even for organizations, the prediction is found useful to improve their marketing and they can find out how much has their product affected the people in a positive or negative way through which they can improve upon. The paper is about sentiment analysis, to be precise it is described as examining the emotion of the people, who post their comments in various social media sites to express their feelings or sentiment. The sentiment analysis is done using various algorithms to provide an appropriate output. Sentiment analysis is the most well-known content grouping tool or method that investigates an approaching message or content and tells whether the hidden assumption is positive , negative our unbiased, based on the output .The algorithms logistic regression and TD-IDF are used and accuracy is compared between them for better performance. The input given is processed using two algorithms that is logistic regression and TFIDF vectorizer. One algorithm compares the independent variables and checks if they are correlated to the dependent variable and the other algorithm takes the given input and converts it into integers and extracts its features. Based on the prediction of two algorithm one is chosen to be the best than the other.

KEYWORDS: Topic Sentiment Analysis, TF-IDF Vectorization, Logistic regression

I.INTRODUCTION

Sentiment analysis is defined to be analyzing the sentiment based on the public comment so called the review to understand the intent of the writer on the thing they have described about. The comment can be upon anything, either it can be a movie, a product , a brand , a social cause etc. The review about the product or brand helps the organization to improve their product's quality or the way of marketing of their product or, as per the product's demand either to increase the sales or to target certain group of people etc. This sentimental analysis has its own applications and its output is either positive or neutral or negative based upon the review given as input. This is done using various machine learning algorithms.

Machine learning is subsidiary to artificial intelligence. To define artificial intelligence, artificial intelligence is the intelligence established by machines. Colloquially called machine intelligence. In other words it is said to be ,''the machine that understands the intent of the environment and respond them in action to achieve the goal''. Contributing to make the machines intelligent is the ultimate aim of artificial intelligence, the upcoming technology of computer science. Machine algorithm being a part of AI, it is defined as examining the given data, learning from it later anticipating the data based on already learnt data. To make predictions there are several algorithms namely decision tree, naïve bayes , random forest,KNN means and Support Vector Machine etc.

Linear regression and random forest algorithms are used to predict the sentiment.. The count vectorizer produces a library or vocabulary of words so that we can use that find new documents. The TF-IDF vectorizer is used with feature extraction. chi–square analysis is used to determine the frequency of the word and their precision score. This analysis helps in calculating precision and recall(f1 scores). linear regression is where both input and output is in numeric. For each difference of the input a coefficient is assigned. Based on the coefficient given, output is predicted. Here in linear regression it is assumed that the non target variables as independent variable and the target variable as dependent variable. The relationship between non target and target variable is explained using a graph, where as the

random forest algorithm is similar to decision tree but random forest contains many decision trees combining to form a forest. The inputs are given to different decision trees and the prediction from each tree is aggregated using summation function. Generally thousands of dataset of particular brand or product is taken as input. There are various parameters which add to characterize the achievement and validity of an Ecommerce store. In any case, one significant factor in raising the notoriety, standard and assessment of an Ecommerce store is Product Reviews As of here Amazon Alexa's review is taken as input.

II.LITERATURE SURVEY

Anh-Dung-Vo et.al[1] studied and devised methods on how to bring out the feelings of the customers. The areas of aspect and the opinions on a product is very important when it comes to product reviews and extracting the customer feelings through this. The dataset of camera reviews is used for this project and the first step in the process is knowledge extraction through natural language processing tools such as NER, Dependency parser and the second step is to apply the gained knowledge to analyze or understand the new data. The dataset used has to be considered for many loopholes such as dependency errors, emotions and other features. The window extraction process features the extraction of raw knowledge which is a set of vocabulary for the model and this step is followed projection where a particular range of data is tested. The results are quite optimal with average F1 scores.

Deokgun Park et.al[2] devised a solution for text analysis which in today's world is of most importance when it comes to analyzing big chunks of data. The deep understanding of large chunks of data comes when the words is divided into lexicons which provide high quality information about the data using NLP techniques and for analysis of text. Word embeddings are done on words to extract semantic relations in a vector format. The technique here is to lock the semantic correlated word to the word that is embedded in a vector space. Topic modeling is one of the models used here which uses LDA to improvise the critical word search in a particular document. Concept vector technique proposed in this paper is where more diverse concepts are taken and lexicons are built over them and how the lexicons are used in terms of the context in a document.

Azin Ashkan et al[3] studied and devised techniques to classify queries. Knowing the intent behind the user queries helps improve user search results. Understanding the user intent becomes important for organizations to improve user satisfaction. This paper focuses on ad-click through and content of search engines for detecting query intent. Online commercial intention is where the user wants to utilize a commercial service. A commercial query is taken and all products that are to be purchased in future are taken. Query intent detection is based on SERP(search engine result pages) which are a result of query or keyword by the user. Using click through, the click through rate depends on the factors such as nature of information needed for the user etc. The query the user intends to search is elaborately divided into user defined, navigational, informational and through which the user behavior is also monitored.

Samantha Aculick et.al[4] investigated many algorithms for text analysis and text mining. The natural language processing has many methods to identify the meaning and intention behind texts and words. The n grams classifier is the first method which identifies intents in a sentence or posts by splitting the sentences and giving the n gram weights to each word where the intents are compared with the weights and the given predefined vocabulary for classifying the intention of the sentence or post. The next algorithm investigated is the part of speech also called as POS tagging is a method which uses distinctive patterns with vocabularies that are used to find more expressions in a text. The pos tags contain all the English vocabulary in them so that they can split up the sentences .according to the pattern and match them up with other sentences to see their co-relation. The support vector machine is machine learning algorithm which is useful in training and testing the models which takes the input from the previous two algorithms implemented. Support vector is a technique where a hyper-plane divides two variables by calculating the distance between them and classifies them accordingly.

Joseph Lilleberg et.al[5] taught of a new approach to classify and take out the features in the words by implementing word2vec algorithm with TF-IDF where the latter algorithm takes in a more precise set of vocabulary that helps to classify text more precisely. The word2vec is used for document classification and categorization of data. The word2vec is implemented mainly by using the bag of words and skip gram model. The bag of words predicts the words based on the features extracted and the skip gram model learns and predicts the new words by the current word it is featured with. The approach here combines the TDIDF weight score with the sums of the word2vec word count that would be in a document classification. This methodology uses vector to represent the words and to outweigh the results

of TDIDF with or without stop words word2vec is trained and tested on the support vector classifier model and higher accuracy is achieved.

Mridula. et.al[6] have proposed a method for opinion and sentimental analysis of tweets whose results are highly beneficial for predicting an event to occur. The paper focuses on various aspects and algorithms such as pos tagger are used and surveys on what model can be best chosen for sentimental analysis. The articles are analyzed using various algorithms. The articles categorized are summarized in this paper. The many sentimental analysis classifiers are analyzed for their various methodologies and an appropriate mechanism is chosen. The classification and prediction algorithms such as the Naïve-Bayes, SVM classifier are used and the models are trained and tested. Initially the tweets on which the algorithms work are being feature extracted using the TF-IDF algorithm which tells the features that are highly essential to a document or in a dataset of tweets to predict the emotions of the user. It can tell how important a word is in a particular phrase. The results predicted suggest that the feature extraction can be extended to bigram or trigram model for more precision and other NLP techniques such as POS tagger and bag of words model can be used.

Mandar Deshpande et.al[7] have applied natural language processing techniques to data, to get the emotion factor out of it in an artificial intelligence approach. The area of emotion intelligence is still blooming and is at its peak of usage now. The artificial intelligence here doesn't limit to only text but also the images ,audio and other stuff which it can detect. The preprocessed data is categorically labeled into negative and neutral tweets based on five types of emotions that we see in day to day life. The emotional classification is always done on dense data which deals with coarse level analysis on the attribute. The analysis here is done on sentence level. The emotions are manually labeled and then they are trained using the supervised classification algorithms and the results predicted suggest that these classification models cannot be as accurate as manual labeling as they are restricted to trained set of variables.

Mita K. Dalal et.al[8] have explained the methodology for automatic text classification which usually refers to categorizing data automatically based on their feature semantics and content. The data is automatically labeled as part of a class which is predefined or set by the machine learning algorithm which is based on supervised learning. The normal method of training and testing a model using the classifier by extracting the features is followed. As automatic classification has lot of issues such as unstructured text , dealing with the dense attributes and finalizing a proper classifier for the dataset chosen and to address this models such as multinomial models are suggested where the document vector or array stores the occurrence of each term and their frequency. The pre processing errors can be rectified up to an extent by the multiword model. In this model the synonymous words and their occurrence has been tried to overcome by the LSI model and multiword model. The feature selection done on the data is about implementing the classifiers such as SVM but they have drawbacks such as they cannot be carried over to multilevel classification.

M. Krendzelak et.al[9] have presented the most common ways to categorize text. Since the text classifiers are domain oriented they need to be fine tuned so that they can fit in with any model and can classify text more correctly and make them generic to categorize. The approach here is to hierarchically structure the classifiers to improve their performance tough this approach hasn't had that much significance due to unsatisfactory results. Since the method suggested is hierarchical the first step in that is using the expert systems which is manually classifying and labeling that had a high accuracy score. The components trained with hierarchical structures tend to choose only the best classification stems out of the whole set of classifiers. One such structure can be the neural network where many variables of input are trained to give a single real output of the data(image, text, audio).

Wen Pu et.al[10] have proposed an approach to use local n-gram model to analyze text where the repetition of words is also considered while classifying the text or words. The words with repetition are retrieved and are distributed to the data and this helps to identify which data or document is in great similarity with the local data. The first step in this approach is to classify the words according to their weights such as bigram or trigram and then all the bag of words can be converted into TDIDF scores which takes the td score and the reciprocal of the TD(IDF) score and multiplies to give the total weight or the vector space of the word. The main idea of this is to match the similarities between two posts or documents or tweets by using the feature extraction technique instead of just using the conventional dictionary and vocabulary to match them. After comparing the results it can be said that the local bag model is able to extract more semantic features and we can see the pattern of words that are repeated more frequently in a more clearer manner.

Asif Ekbal et.al[11] have proposed a paper that intends to detect the plagiarized content based on text similarity in the documents to be checked. This approach uses the n gram model and the vector space model for tokenizing , using the pos tagger and other NLTK tools to classify the documents which are plagiarized and also find the source of the document from which the content has been plagiarized. Lemmas are used in place of tokens for

generating the dictionary of words and each document is converted into a vector space which is then transformed into unigram or bigram model. The similarity content or coefficient is calculated by using the vector n-gram which contains the lemmas sorted and by comparing the n-gram count which is got by using a merge algorithm between the source and the plagiarized document. The result got is quite good with little improvements needed in content detection.

Jan Martinovic et.al[12] have discussed on the techniques of how information retrieved form a database can help to answer the queries of a user handling the information. In order for the document to be relevant to the user queries techniques such as compression and clustering need to be done on the information retrieved. Word compression is one of the techniques that is incorporated here. Clustering or grouping is done on the documents to find the similarity index by calculating the distance between the documents or the vector space of the documents. The topics are searched thematically for the query lists that are given to the documents. The count of words in a particular document is extracted by using specific algorithm which is used for hierarchical clustering and topic modeling. The results are got by using a two technique model which uses both topic evaluation and word compression techniques for classification of huge corpus of newsfeeds and newsletters used as a dataset. The impact of clustering the document is also discussed here and its effects on the compression techniques are highlighted.

Kazem Taghva et.al[13] have showcased how features can determine or have an impact on the classification of texts. Bayes algorithm is implemented here. The text in each document is converted into a vector array and each document is transformed to vector space for classification. The word selection technique here is very important as it decides whether a document belongs to a particular category. The confusion matrix is built and the dataset is trained and tested on the feature selected vocabulary. The single labeled data is taken and some categories are taken out of it and the multi labeled categories are omitted here. The features and attributes like size and style of the font, the letters are proved to have a significant impact on the documents.

Hang Cui et.al[14] have proposed a technique for query development dependent on interactions of the user or client recorded in client logs. The focal thought is to extricate connections between's query terms and record terms by dissecting user logs. Queries to web search tools on the Web are normally short. They don't give adequate data to a successful determination of significant reports. Since the data contains the URL added along with it ,it can be used to extract client sessions which he/she has used to surf the query. The sessions contain the document terms linked to them and the queries are now correlated or matched with these terms . The result is that for long queries their corresponding and relevant short query is shown.

Sanduo Zhou et.al[15] have introduced the developments that have taken place for question classification and the methodologies have been studied. The queries are classified according to the user behavior and also the query logs. The techniques of machine learning which is either learned or unsupervised is trained and tested and the results are compared. Day to day the amount of information getting generated is growing as well as the query logs and to handle this, Big data techniques are proposed to be incorporated on query logs to perform analysis and visualize the data.

III.METHODOLOGY

Products and purchases made around the world are done by people by seeing the product reviews and the popularity of the product in the market. Reviews are always subjective in nature and depend upon the product type on which they are being rated. To have a good overview of the review and its clear meaning reviews need to fine tuned and their precise meaning has to be taken out which is a big problem for all sentimental analysis projects. Reviews benefit a large amount of people from the users to the organizations that market them and so they need to be given high importance. One significant assignment for the Ecommerce store is to keep up its notoriety in the online market. Naturally, it requires a great deal of exertion to pick up that notoriety yet very little to lose it. Product Reviews are the most ideal approach to keep up their series of winning streak in the market. Item Reviews and inputs have changed the diversion for online market since web has turned into a very family thing. The proposed system takes the data of user reviews on Amazon Alexa which has ratings as well as the verified reviews on the product. The algorithms of Logistic Regression and TF-IDF vectorizer are applied and their accuracies are compared. Since the project mainly deals with information retrieval and text mining the TF-IDF algorithm is used for feature extraction with the help of count vectorizer which builds up a vocabulary of words that can be encoded into a another review dataset to get the vocabulary. After feature extraction the random forest classifier is used for prediction and fine tuning and gives the F1 scores for the review which are actually on the higher side for positive labels. Chi square graph is plotted with the words with highest occurrences or the feature extracted words and the sentiment of the review is predicted as positive and negative labels.

Advantages of Proposed System

The advantages of the proposed system covers the disadvantages that are in the existing system, they are

- Feature extraction is done as the main process.
- Higher F1 scores for positive labels.
- Improved accuracy for the NLP method used.
- Algorithms are compared for their efficiency.

SYSTEM DESIGN

The system design or flow for predicting the reviews is as follows:

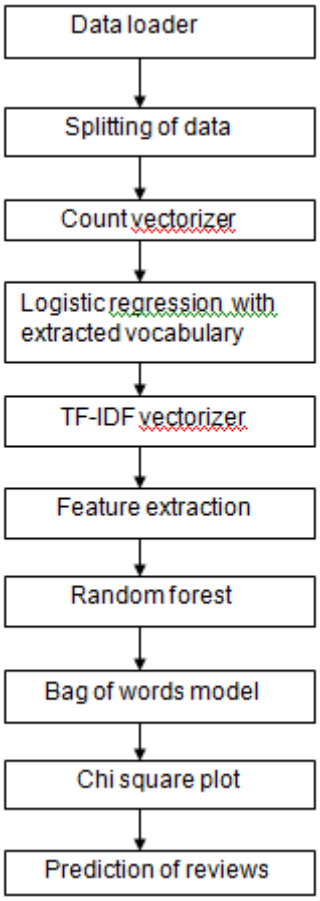

Figure 1.1 System Architecture

The Figure 1.1 shows the system architecture and how the process is carried out. The data is first split according to the ratings that are given. The ratings are given from 1 to 5 and it is classified into binary type as 1 for ratings more than 3 and 0 for ratings 3 and less than 3. The count vectorizer extracts the vocabulary of words and tells the word count of each word extracted. The words extracted can be encoded or transformed to other vocabulary for a different dataset or a document. The logistic regression model is fit into the data set for calculating the accuracy of the algorithm. The result of the extracted vocabulary from the count vectorizer is given as input to the logistic classifier that has in built packages to calculate the relation between the variables and give the accuracy. The TF-IDF vectorizer is used and feature extraction is done. The weights are calculated separately for Term Frequency and Inverse Document Frequency and their product is taken as weight of the word. F1 scores are predicted using the random forest classifier by using bag of words or n-gram model. Chi square plot is used to plot the graph between the label and their occurrence.

IV.IMPLEMENTATION

The algorithms used in the project are linear regression , TF-IDF vectorizer, random forest classifier.The reviews are first processed by the count vectorizer which takes in the word count and extracts the vocabulary of the

data given. The count vectorizer mechanism processes each and every word in the document or review dataset and tells the user how important a word is in the document. The vocabulary that is built from the dataset can be applied to get the vocabulary of other reviews or documents.

Logistic regression is applied next on the vocabulary that is got from the count vectorizer. It checks the relation between the independent and the dependent variables where the regression can be simple or multi linear in nature depending upon the independent variables for the dataset. The dependent variable or binary variable is the positively related column which contains 1 and 0 for positive and negative labels and the independent variable is verified reviews column which are the ones to be predicted as the output.

The TF-IDF vectorizer algorithm is a machine learning algorithm which takes the most repeated words in the data set and applies feature extraction. Each token is converted into a matrix format in the feature extracted index where each word is present in the dictionary. The values of the weights of the TF-IDF are returned in float and the highest count corresponds to the rarest word in the dataset. The formula of the TFIDF vectorizer and how the weight of the words are calculated for feature extraction.

The random forest classifier is a forest of trees or decision trees that categorize a dependent variable in tree format. This algorithm is used for prediction and fine tuning of the TF-IDF vectorizer for improving the accuracy. It gives the classification report of the negative and positive percentage of labels and their F1 scores. The N-gram model is applied and the trigrams are separated from the other extracted features.The chi square model is used and the chi coefficient is found out by plotting the graph. The highly recurrent trigram tokens are plotted and their occurrence in the whole review is numbered and seen for which word has highest appearance. The reviews labels are predicted after training and testing the model after fitting the random forest classifier by giving the corresponding testing set label number for the trained set and the outcome is either 1 or 0(Pos or Neg).

COUNT VECTORIZER

The count vectorizer tells the word count in a document or a review. The words are converted into the arrays and is given a count in integer data type.

Steps Involved

The process in count vectorizer that tells the occurrence of word is as follows:

Step 1: The data collected from the Amazon site is pre-processed and the empty rows are deleted from the dataset and the ratings are split into >3 as positive and <3 as negative.
Step 2: Count vectorizer tells the word count of each and every word in the data set.
Step 3: The words are tokenized and are stored in numpy arrays in vectorized form.
Step 4: For each word the vocabulary is extracted and is given an integer value.
Step 5: The word count is stored in the arrays and are returned when called after training the dataset.
Step 6: The arrays display the word count of each trained word and displays 0 if the word is not there in the vocabulary.

V.RESULT

The accuracy scores in Figure 1.2 are got by comparing the two algorithms are of 75% and 54% of Logistic regression and TF-IDF vectorizer respectively. The X axis consists of the features extracted count and the Y axis consists of the accuracy score in percentage. The features extracted are of 10000, 20000 and 30000 in number which have the accuracy of 88%, 87% and 87% respectively which is got after fine tuning the TF-IDF algorithm by random forest classifier.

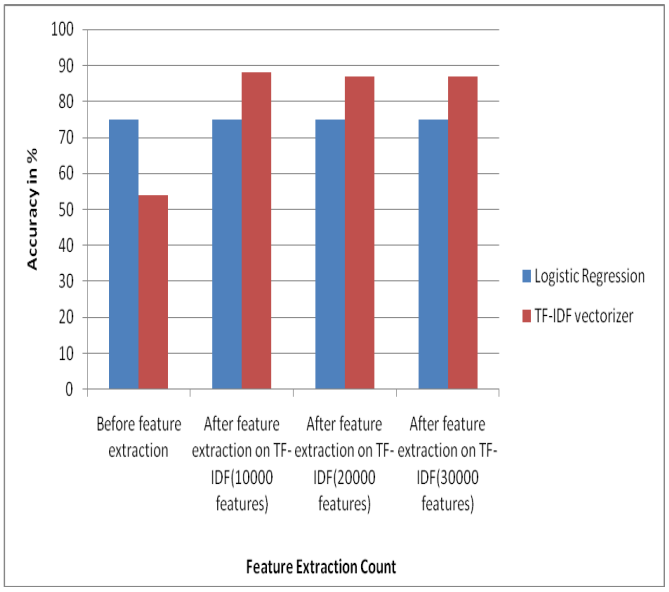

Figure 1.2 Algoriths Performance Comparison

Since the feature extraction is not possible with the logistic regression classifier, its value remains the same throughout the prediction process. The method of feature extraction can be applied only to TF-IDF algorithm and so features are extracted with has more efficiency in terms of accuracy than the logistic regression , which is done after fine tuning with the random forest classifier.

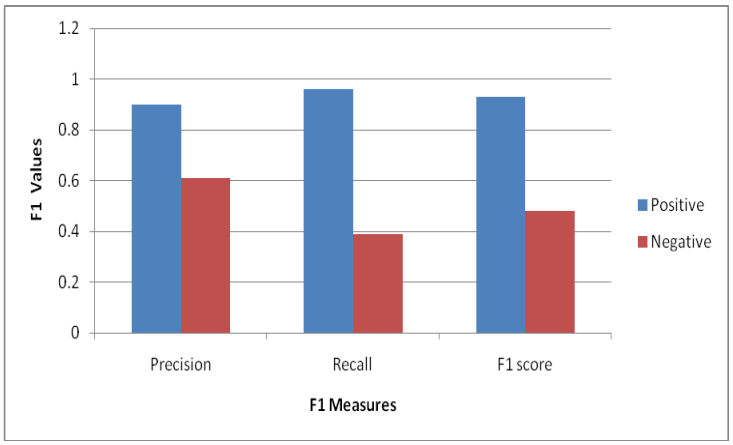

Figure 1.3 F1 scores of Positive and Negtive reviews

The Figure 1.3 here depicts the F1 scores of the reviews given. The X axis is the scores that are predicted by the random forest classifier and the Y axis depicts the F1 measures. The F1 scores of the negative and positive labels are given in the classification report and the positive labels have high F1 score 93% and negative has score of 48%. The F1 scores depicted here are got from both the logistic regression and the TF-IDF vectorizer. The positive labels have got a much more greater F1 score than the negative labels and thus the algorithms of logistic regression and TF-IDF vectorizer used is able to predict the positive reviews with more precision than the negative ones.

VI. CONCLUSION AND FUTURE WORK

The aim of the work is achieved as the sentiment of the user is analyzed and the accuracy is also improved. The feature extraction of words has helped to increase the efficiency of the algorithm in terms of F1 scores by the use

of machine learning and Natural Language Processing tools. The coefficients or the feature extracted labels have been depicted clearly in the chi square plot and the algorithm TF-IDF implemented has shown good results when fine tuned using the random forest classifier.

There is still scope for improvement here as the data set implemented here is a small dataset featuring the reviews in just thousands and in the future a large data set can be taken and implemented on to compare and see the results. The scores of negative reviews aren't so good when compared to positive reviews and so the negative prediction of reviews are to be improved. To dive deeper and to give clearer output, intent analysis can be done to know the intention of the user such as emotional classification that uses advanced deep learning algorithms. The algorithms linear regression and TF-IDF isn't able to predict the negative reviews as efficiently as the positive reviews so this problem needs to be address by taking a large dataset that has much wider perspective of opinions pointed out.

REFERENCES

[1] Anh-Dung vo , Quang-Phuoc Nguyen, and Cheol-Young Ock, "Opinion–Aspect Relations in Cognizing Customer Feelings via Reviews", IEEE, Vol 6, 5415–5426, 2018.
[2] Park, D., Kim, S., Lee, J., Choo, J., Diakopoulos, N and Elmqvist, N. "ConceptVector: Text Visual Analytics via Interactive Lexicon Building Using Word Embedding", IEEE Transactions on Visualization and Computer Graphics, Vol 24(1), 361–370, 2017.
[3] Ashkan, A., Clarke, C. L. A., Agichtein, E., and Guo, Q, "Classifying and Characterizing Query Intent", Advances in Information Retrieval, 578–586, 2009.
[4] Samantha Akulick and El Sayed Mahmoud, "Intent Detection through Text Mining and Analysis", Future Technologies Conference (FTC), 2017.
[5] Lilleberg, J., Zhu, Y., and Zhang, Y, " Support vector machines and Word2vec for text classification with semantic features", IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI*CC), 2015.
[6] Mridula, A., and Kavitha, C. R., "Opinion Mining and Sentiment Study of Tweets Polarity Using Machine Learning" , Second International Conference on Inventive Communication and Computational Technologies (ICICCT), 2018.
[7] Deshpande, M., and Rao, V., "Depression detection using emotion artificial intelligence", International Conference on Intelligent Sustainable Systems (ICISS), 2017.
[8] Mita K. Dalal., and Mukesh A., " Automatic Text Classification: A Technical Review", International Journal of Computer Applications Volume 28– No.2, 2011.
[9] Krendzelak, M.,and Jakab, F., " Text categorization with machine learning and hierarchical structures", 13th International Conference on Emerging eLearning Technologies and Applications (ICETA), 2015.
[10] Pu, W., Liu, N., Yan, S., Yan, J., Xie, K., and Chen, Z., " Local Word Bag Model for Text Categorization", Seventh IEEE International Conference on Data Mining, 2007.
[11] Ekbal, A., Saha, S., and Choudhary, G., "Plagiarism detection in text using Vector Space Model", 12th International Conference on Hybrid Intelligent Systems (HIS), 2012.
[12] Martinovic, J., and Dvorsky, J., " Document Classification Based on the Topic Evaluation and Its Usage in Data Compression", IEEE/WIC/ACM International Conferences on Web Intelligence and Intelligent Agent Technology, 2007.
[13] Taghva, K., and Vergara, J., "Feature Selection for Document Type Classification", Fifth International Conference on Information Technology: New Generations, 2008.
[14] Hang Cui, Ji-Rong Wen, Jian-Yun Nie, and Wei-Ying Ma., " Query expansion by mining user logs", IEEE Transactions on Knowledge and Data Engineering, 15(4), 829–839, 2003.
[15] Zhou, S., Cheng, K., and Men, L., "The Survey of Large-Scale Query Classification", AIP Conference Proceedings, 2017.