A Sentiment Analysis Based Approach for Understanding the User Satisfaction on Android Application

Md. Mahfuzur Rahman[1],Sheikh Shah Mohammad Motiur Rahman[2], Shaikh Muhammad Allayear[3], Md. Fazlul Karim Patwary[4] and Md Tahsir Ahmed Munna[5]

[1] Division of Research,[2] Department of Software Engineering,[3] Department of Multimedia and Creative Technology
Daffodil International University,
Dhaka, Bangladesh
[4] Institute of Information Technology, Jahangirnagar University
Savar, Bangladesh
[5] National Research University Higher School of Economics
Moscow, Russia
mahfuzur.web@daffodilvarsity.edu.bd,motiur.swe@diu.edu.bd, drallayear.swe@diu.edu.bd, patwary@juniv.edu, tahsir.se@gmail.com

Abstract. The consistency of user satisfaction on mobile application has been more competitive because of the rapid growth of multi-featured applications. The analysis of user reviews or opinions can play a major role to understand the user's emotions or demands. Several approaches in different areas of sentiment analysis have been proposed recently. The main objective of this work to assist the developers in identifying the user's opinion on their apps whether positive or negative. A sentiment analysis based approach has been proposed in this paper. NLP based techniques Bags-of-Words, N-Gram and TF-IDF along with Machine Learning Classifiers namely KNN, Random Forest (RF), SVM, Decision Tree, Naive Byes has been used to determine and generate a well fitted model. It's been found that RF provides 87.1% accuracy, 91.4% precision, 81.8% recall, 86.3% F1-Score. 88.9% of accuracy, 90.8% of precision, 86.4% of recall and 88.5% of F1-Score are obtained from SVM.

Keywords: NLP, TF-IDF, Sentiment analysis, Machine Learning, Mobile Apps Review.

1 Introduction

Sentiment Analysis (SA) broadly involves data mining processes and techniques to extract insights however it is famous for Opinion Mining (OM) or emotion AI. Though it has different definition and uses but mostly it refers to the use of Natural Language Processing (NLP), text analysis, computational linguistic and biometrics. It systematically identifies, extracts, quantifies, and studies affective states and subjective information. People's sentiment and emotion take places towards certain entities. There are various categories people use their sentiment in various purposes.

Sophisticated categorization of a huge number of recent articles are provided by one survey. The related field to SA including emotion detection, building resources and transfer learning [1].

Machine learning approach and lexicon-based approach to evolve the problem of sentiment classification can be categorized. Machine learning and lexical need to enhance sentiment classification performance [2].

Application programming interfaces (APIs) are provided by many social media sites to prompt data collection and analysis by researchers and developers. For example, twitter has three different versions of APIs available such as the REST API, the Streaming API and the Search API. For collecting the status data and user information developers should use the REST API; with the Streaming API developers able to gather Twitter content in real-time, whereas the Search API allows developers to query specific Twitter content. Moreover, developers can mix those APIs to create their own applications. Hence, sentiment analysis seems having a strong fundament with the support of massive online data [3].

Subjective detection, sentiment prediction, aspect-based sentiment summarization, text summarization for opinions etc. are main field of sentiment research. Negative or positive opinions are predicting the polarity of text. The field of Natural Language Processing (NLP) is discovered the concept of sentiment analysis [4].

Classifying an opinionated document express as a positive or negative opinion and classifying a sentence or a clause of the sentence as subjective or objective has been studied in one paper [5]. The reviewers are determined the product by positive or negative review. Positive or negative opinion is called sentence-level sentiment classification.

Due to the rapid growth of online shop, most of the people are getting attracted to buy products which are needed through online. On that case, customer reviews on the products getting importance for online merchants. So that, they can measure the satisfaction rate of their customers and can make decision [6]. The main contributions of this paper are listed as follows:

1. A model has been proposed to make a decision and be financially profitable for android apps developers using sentence level sentiment analysis.
2. More focus on android apps reviews which will help developers to know about the real scenario of the developed application in market.
3. Multiple Machine learning classifiers has been evaluated to find out the classifier which can be fitted well in the proposed model.
4. The result of the evaluation provides a strong basis for building effective tools for Sentiment Analysis on the reviews from mobile application users.
5. The effectiveness of classification algorithm is evaluated in terms of accuracy score, precision, recall and F1-score.

To Analyzing overall sentiment of United States people to determine of the 2016 United States presidential election, there are many tweets collecting from United States people. These tweets are collecting from not only United States people but also others foreign countries [7].

The organization of this paper is structured as follows: section 2 describes the related works done in this relevant field. The proposed model has been constructed and described in section 3. The experiments, result analysis and performance evaluation has been briefly discussed in section 4 explains Result and discussion. Section 5 concludes the paper with a possible future step.

2 Literature Review

Recently sentiment analysis is one of the big areas for researchers because increasing social media networks, blogger sites, different types of forum, feedback of products. In this section, provide an overview of previous studies regarding sentiment analysis.

Monisha Kanakaraj [8] proposed an idea to increase the accuracy of classification by semantics and Word Sense Disambiguation (WSD) in NLP (Natural Language Processing Techniques). Ensemble classification is analyzed to mined text information in sentiment analysis. Ensemble classifiers is outperformed traditional machine learning classifiers by 3-5%.

Purchasing online products are preferred maximum customer. Devendra Kamalpurkar [9] proposed a technique, classifying negative or positive reviews using machine learning based approach based on feature. Based on user's feedback, users choose to purchase Products.

Using two product's datasets namely Nokia Lumia 1020 and Apple iPhone 4s, Venkata Rajeev P [10] determine analysis combination on four parameters: star rating average, polarity rating, reviews per month and helpful score. Apple iPhone 4s is better than Nokia Lumia 1020 to choosing the product. Star rating average 3.98, polarity rating 0.51, Reviews er month 39, helpful score 78% and grand score 8.03.

Mining products based on feature by customer's negative or positive opinion. Weishu Hu [11] proposed a SentiWordNet-based algorithm and divides the opinion analysis tasks into three steps: identifying opinion sentences and their polarity, mining the features that are customers' opinion and removing the incorrect features.

Mingquing Hu [6] gathered the customers' comments on electronic products from Amazon.com and Clnet.com. They have considered the reviews of five products during their experiment such as Digital Cameras, DVD player, mp3 player and cellular phone in quantity 2,1,1,1 and 1 respectively. During their work, they focused on only the features of the product mentioned by the reviewer. After that, they applied different techniques to mine the features. As a result, the obtained maximum 80% of recall and 72% of precision on average.

Individual performance is improved by the different types of sentiment approaches and consists of three combinations: ranking algorithms for scoring sentiment features as bi-grams and skip-grams extracted from annotated corpora, a polarity classifier based on a deep learning algorithm and a semi-supervised system founded on the combination of sentiment resources. TASS competition is evaluated based on general corpus average [12]

With statistical significance test, five text categorization methods are reported by a controlled study. Five methods Support Vector Machine (SVM), K-Nearest Neighbour(KNN), Neural Network (NNet), LLSF (Linear Least Squares Fit) and Naive Byes (NB). Dealing with a skewed category and performance as function of the training-set category frequency are focused on robustness of these methods. Yang [13] claimed that when the number of positive training instances per category are small, SVM, KNN and LLSF significantly outperform.

Nicholas [14] investigated cross-domain data to detect polarity of sentiment analysis which data collected from YouTube. It represents bag-of-words feature.

Most significant challenges in sentiment evaluation are focused. It explains techniques and access of sentiment analysis objection. It represents challenges of sentiment [15].

Murty [16] explained cross-domain text classification techniques impotences and strengths. basic knowledge right from the beginning of text classification methods are provided by cross-domain text classification algorithms.

Murty [17] Propond a new algorithm named LS-SVM which is efficient scheme for documents classification. LS-SVM is enhancing accuracy and retrenchment dimensionality of large text data with Singular Value Decomposition.

3 Proposed Model

In this section, the proposed model has been presented and described in details.

3.1 Data Collection

Reviews are collected from google play store using Web based Crawler. Then collected positive reviews (PR) are labeled as 1 and for negative reviews (NR) are labeled as 0.

3.2 Pre Processing

For balancing the dataset, in this step the total number of PR and total number of NR have been checked. After that, a balanced dataset has been generated from equal number of PR and NR. The following tasks have been also done as a part of pre-processing of the data:

- Steaming: Stemming is the process of reducing a word to its word stem that affixes to suffixes and prefixes or to the roots of words known as a lemma. Stemming is important in natural language understanding and NLP.
- Remove stopwords: The process has been carried out by removing frequently used stopwords (preposition, irrelevant words, special character etc).

3.3 Bag-Of-Words (BoW) Model Certain

The bag-of-words model is widely used in feature extraction procedures of Natural Language Processing (NLP) for a sentence or a document. Bag-Of-Words Model has been generated which includes following steps.

- Words Vocabulary: If there is a word occurred 20 times in the corpus, it will be counted only once in the word's vocabulary. During defining the words vocabulary, every word will ensure the uniqueness of that presence.
- Sentence Vector and Manage Vocabulary: Each word from vocabulary will be scored in binary which is known as word scoring in NLP. This process has been conducted by applying n-gram to minimize the large size of vocabulary. The output of this process is (for one sentence):
- Feature Extraction: TF-IDF has been used to extract the features from the vectorized vocabulary. TF-IDF stands for Terms Frequency Inverse and Document Frequency which can define the importance of words corpus datasets

3.4 Extracted Feature Vector

As an output of BoW, a vector containing the extracted and minimized features has been found in this stage.

3.5 Splitting

In this stage, the dataset has been split in ratio of 67% for train the model and the rest 33% for test.

3.6 Machine Learning Algorithm

In this stage, machine learning algorithms can be applied for training and test. In this research work, K-Nearest Neighbors (KNN), Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT), Naive Byes (NB) classification techniques have been applied to classify the sentiment and evaluate the performance of the classifiers. As a result, a decision can be made whether which algorithm performs better with proposed model.

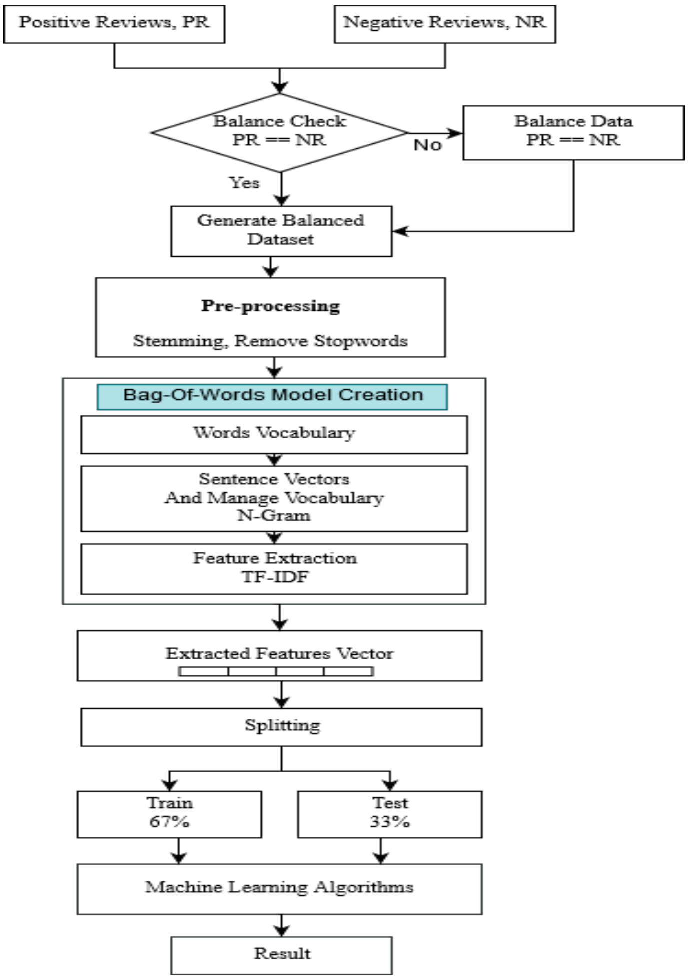

Fig. 1. Architecture of Proposed Model.

4 Environment Setup and Experiment

The experiments and evaluation of the classifiers in proposed approach has been discussed in this section.

4.1 Environmental Setup

A desktop computer in configuration with Intel Core i5 Processor, 8 GB DDR3 RAM has been used during the experiment. Windows 10 64 bit was the installed Operating System. The implementation of the model in coding has been done using python 3.5

programming language along with the packages of scikit-learn, pandas, scipy, numpy etc.

4.2 Dataset Used

"Mobile App Review" dataset collected by a web Crawler has been used [18]**.** The dataset contains 20k reviews where 10k data are positive and 10k data are negative. Every app has about 40 reviews. The reviews are classified into positive and negative reviews and ignored neutral reviews. For positive review and negative review node value are 1 and 0 respectively.

4.3 Evaluation Parameter

There are some parameters or metrics [19] are used to evaluate the performance of classifiers.

Table 1. Evaluation Parameters.

Parameters	**Definition**	**Formula**
Precision	The measurement exact value of the result provided by the classifier is defined as precision.	$\frac{TP}{TP+FP}$
Recall	The thoroughness of the result provided by classifier measured by recall.	$\frac{TP}{TP+FN}$
F1-Score	The harmonic mean of exactness and completeness is known as F1-Score as well as F-Score.	$\frac{2*Precision*Recall}{Precision+Recall}$
Accuracy	The ratio of accurately classified data to total number of data is defined as accuracy.	$\frac{TP+TN}{TP+TN+FP+FN}$

TP (True Positive): The reviews which are positive and also classified as positive; FP (False Positive): False Positive represent the reviews are positive but classifier classified as negative; TN (True Negative): True Negative represents the reviews are negative and also classifier classified as negative; FN (False Negative): False Negative represent the reviews are negative but classifier classified as negative;

4.4 Result Analysis and Evaluation

Table 2 represents the comparison of the performance of proposed model along with KNN, Random Forest, SVM, Decision Tree and Naive Byes classifiers. KNN provides 62.4% of accuracy, 70.2% of precision, 42.9% of recall and 53.3% f1-score. 87.1%, 91.4%, 81.8%, 86.3% of accuracy, precision, recall and f1-score have been found respectively for random forest. SVM provides 88.9%, 90.8%, 86.4%, 88.5% of accuracy, precision, recall and f1-score respectfully. 85.4% accuracy, 87.3% precision, 82.2% recall and 84.7% f1-score have been obtained from decision tree classifiers. The results obtained from Naive Bayes are 60.3% accuracy, 78.4% precision, 28.4% recall and 41.7% f1-score. Maximum precision score obtains 91.4 % by Random Forest classifiers. SVM is also given better performance for Recall and F score 86.4% and 88.5% respectively. In case of accuracy, SVM has been performed better with the proposed.

Table 2. Comparison of Experimented Classifiers.

Classifiers	Accuracy	Precision	Recall	F-1 Score
KNN	62.4%	70.2%	42.9%	53.3%
RF	87.1%	**91.4%**	81.8%	86.3%
SVM	**88.9%**	90.8%	**86.4%**	**88.5%**
DT	85.4%	87.3%	82.2%	84.7%
NB	60.3%	78.4%	28.4%	41.7%

Figure 2 depicts the comparison chart of the performance of precision from proposed model obtained from KNN, Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT) and Naive Byes (NB) classifiers. The highest Precision score 91.4% by Random Forest classifier.

Figure 3 depicts the comparison chart of the performance of Recall from proposed model obtained from KNN, Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT) and Naive Byes (NB) classifiers. The highest Recall is given by Support Vector Machine and score 86.4%.

Figure 4 depicts the comparison chart of the performance of F-1 SCore from proposed model obtained from KNN, Random Forest (RF), Support Vector Machine (SVM), Decision Tree (DT) and Naive Byes (NB) classifiers. Support Vector Machine is the highest F-1 score.

The final accuracy obtained from different classifiers have been depicted in figure 5. It’s been found that Support Vector Machine (SVM) provide better performance in terms of accuracy and proved as a strong candidate to be used in the proposed model. SVM provides 88.90% of accuracy.

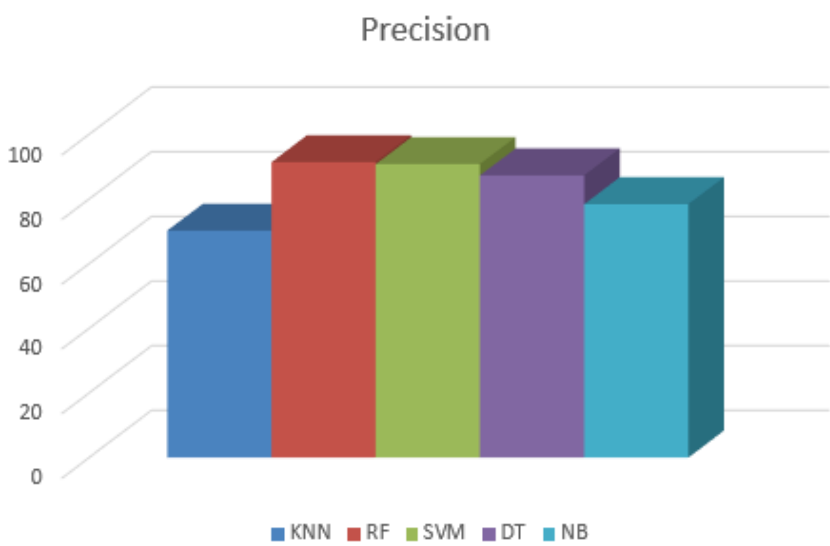

Fig. 2. Comparison of classifier performance for Precision

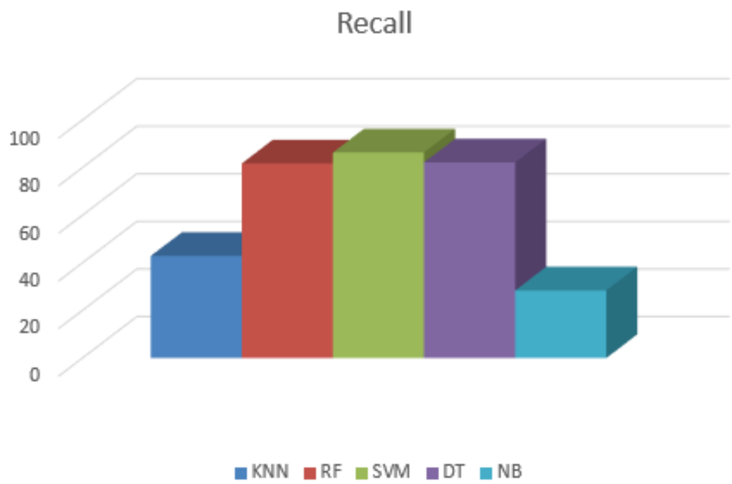

Fig. 3. Comparison of classifier performance for Recall

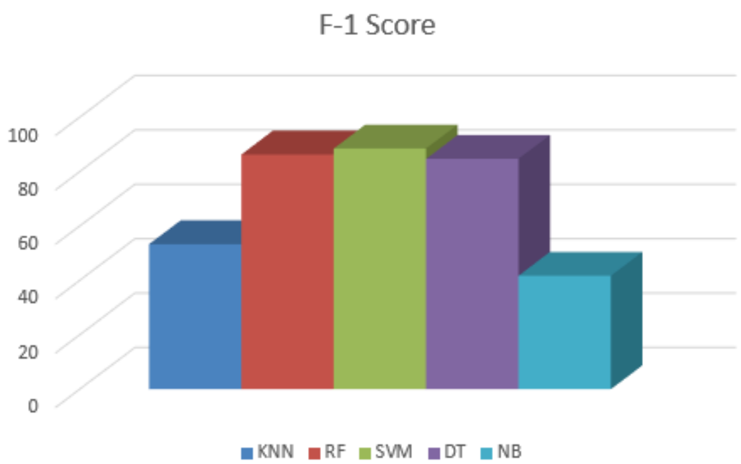

Fig. 4. Comparison of classifier performance for F-1 Score

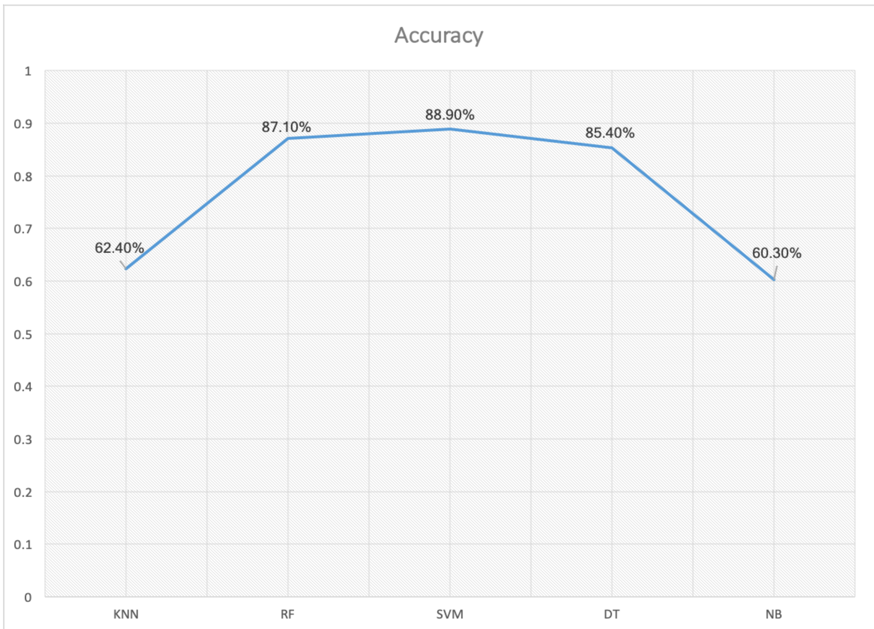

Fig. 5. Accuracy Comparison of the classifiers on the proposed model

To recapitulate, it's been claimed that the proposed approach can be a strong basis of sentiment analysis with SVM classifier.

5 Conclusion

Sentiment analysis of mobile application users has been explained as well as has been proposed a model. The proposed model has been broadly discussed and evaluated. Sentence level sentiment classification has been focused on this research work. Multiple machine learning classifiers also being assessed on the model. From table II and figure 2, it can be concluded that the model can be a strong basement for building effective tools for Sentiment Analysis. Maximum accuracy has been found from SVM which is 88.9%. Even in case of recall and f1-score, it's been found that SVM performs better with the proposed model. But in case of precision, Random forest classifiers provide maximum precision which is 91.4%.

As a future step, the proposed model will be enhanced based on "Ensemble Feature Selection Scheme" and will be investigated.

References

1. Medhat, W., Hassan, A., & Korashy, H. (2014). Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), 1093-1113.
2. Ghosh, M., & Sanyal, G. (2018). Performance Assessment of Multiple Classifiers Based on Ensemble Feature Selection Scheme for Sentiment Analysis. Applied Computational Intelligence and Soft Computing, 2018.
3. Twitter (2014) Twitter apis. https://dev.twitter.com/start, [Accessed: 10- January- 2019]
4. Saifee, V., & Jay, T. Applications and Challenges for Sentiment Analysis: A Survey. International Journal of Engineering Research & Technology (IJERT), 2.
5. Liu, B. (2010). Sentiment Analysis and Subjectivity. Handbook of natural language processing, 2, 627-666.
6. Hu, M., & Liu, B. (2004, July). Mining opinion features in customer reviews. In AAAI (Vol. 4, No. 4, pp. 755-760).
7. Agrawal, A., & Hamling, T. (2019). Sentiment analysis of tweets to gain insights into the 2016 US election. Columbia Undergraduate Science Journal, 11(2017).
8. Kanakaraj, M., & Guddeti, R. M. R. (2015, February). Performance analysis of Ensemble methods on Twitter sentiment analysis using NLP techniques. In Semantic Computing (ICSC), 2015 IEEE International Conference on (pp. 169-170). IEEE.
9. Devendra Kamalpurkar, Ninad Bagwe, R. Harikrishnan, Salil Shahane "Sentiment Analysis of product Reviews" International Journal of Engineering Science and Research Technology Conference on.(IEEE ICSC 2015)
10. Hu, W., Gong, Z., & Guo, J. (2010, November). Mining product features from online reviews. In e-Business Engineering (ICEBE), 2010 IEEE 7th International Conference on (pp. 24-29). IEEE.
11. Hu, M., & Liu, B. (2004, July). Mining opinion features in customer reviews. In AAAI (Vol. 4, No. 4, pp. 755-760).
12. Martınez-Cámara, E., Gutiérrez-Vázquez, Y., Fernández, J., Montejo-Ráez, A., & Munoz-Guillena, R. (2015). Ensemble classifier for Twitter Sentiment Analysis.
13. Yang, Y., & Liu, X. (1999, August). A re-examination of text categorization methods. In Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval (pp. 42-49). ACM.
14. Cummins, N., Amiriparian, S., Ottl, S., Gerczuk, M., Schmitt, M., & Schuller, B. (2018, April). Multimodal Bag-of-Words for cross domains sentiment analysis. In 2018 IEEE

International Conference on Acoustics, Speech and Signal Processing (ICASSP) (pp. 4954-4958). IEEE.

15. Hussein, D. M. E. D. M. (2018). A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences, 30(4), 330-338.
16. Murty, M. R., Murthy, J. V. R., Reddy, P. P., & Satapathy, S. C. (2012, March). A survey of cross-domain text categorization techniques. In *2012 1st International Conference on Recent Advances in Information Technology (RAIT)* (pp. 499-504). IEEE.
17. Murty, M. R., Murthy, J. V. R., & PVGD, P. R. (2011). Text Document Classification based-on Least Square Support Vector Machines with Singular Value Decomposition. *International Journal of Computer Applications*, *27*(7), 21-26.
18. "Android App Review Dataset", https://github.com/amitt001/Android-App-Reviews-Dataset [Accessed: 10- January- 2019]
19. Rahman, S. S. M. M., Rahman, M. H., Sarker, K., Rahman, M. S., Ahsan, N., & Sarker, M. M. (2018, July). Supervised Ensemble Machine Learning Aided Performance Evaluation of Sentiment Classification. In Journal of Physics: Conference Series (Vol. 1060, No. 1, p. 012036). IOP