SENTIMENT ANALYSIS OF INDIVIDUALS PRODUCT REVIEW USING MACHINE LEARNING

M V Keerthi[1], O. Chandana[2], P. Anjali[2], P. Revathi[2]

[1]Assistant Professor, [2]UG Scholar, [1,2]Department of Computer Science and Engineering

[1,2]Malla Reddy Engineering College for Women (A), Maisammaguda, Hyderabad, Telangana.

mvkeerthi29@gmail.com, chandanaoleti2626@gmail.com, peramrevathi27@gmail.com, pabbathireddyanjalireddy@gmail.com

ABSTRACT

Today, digital reviews play a pivotal role in enhancing global communications among consumers and influencing consumer buying patterns. E-commerce giants like Amazon, Flipkart, etc. provide a platform to consumers to share their experience and provide real insights about the performance of the product to future buyers. In order to extract valuable insights from a large set of reviews, classification of reviews into positive and negative sentiment is required. Sentiment Analysis is a computational study to extract subjective information from the text. In the proposed work, over 4,000,00 reviews have been classified into positive and negative sentiments using Sentiment Analysis. Out of the various classification models, Naïve Bayes, Support Vector Machine (SVM) and Decision Tree have been employed for classification of reviews. The evaluation of models is done using 10-Fold Cross Validation.

Keywords: Sentiment analysis, Reviews, Machine learning.

1. INTRODUCTION

With an ever-increasing demand of smart phones, the mobile phone market is expanding at an exponential pace. With such a boom in the smart-phone industry, there is a need to realize the holistic review of the brand and the model of phone. There are numerous brands present in the market, out of which some are dominant and occupy quite a big part of the industry. For instance, Samsung, Apple, etc. are names associated with brands which are famous throughout the world. Electronic commerce plays a vital role in increasing the sales of the mobile phones and influencing consumer buying patterns. Reviews available on such e-commerce platforms act as a guiding tool for the consumers to make informed decisions. Retail websites like Amazon.com offer different options to the reviewers for writing their reviews. For instance, the consumer can provide numerical rating from 1 to 5 or write comments about the product.

As there are innumerable products manufactured by many different brands, so providing relevant reviews to the consumers is the need of hour. Number of reviews associated with a product or a brand is increasing at an alarming rate, which is no less than handling the big data. Classifying the reviews on the basis of sentiment of customers into positive and negative sentiment provides sentiment orientation of the review, hence results in better judgement. Segregation of reviews on the basis of their sentiment can help future buyers to evaluate positive and negative feedback constructively and reach at better decisions as per their requirements. This evaluation acts as a testimony to the users who are looking to know the details and specifications of the smartphones; thereby increasing user credibility

In this research, unstructured data of Mobile Phone Reviews have been extracted from Amazon.com. It has been filtered to remove noisy data and has been pre-processed to evaluate sentiment of the reviews using supervised learning. The reviews have been classified using machine learning

classification models like Naïve Bayes, Support Vector Machine (SVM) and Decision Tree and have been cross validated to find the best classifier for this purpose.

2. LITERATURE SURVEY

1. TITLE: Big data consumer analytics and the transformation of marketing.

AUTHOR: S. Erevelles, N. Fukawa, and L. Swayne

Consumer analytics is at the epicenter of a Big Data revolution. Technology helps capture rich and plentiful data on consumer phenomena in real time. Thus, unprecedented volume, velocity, and variety of primary data, Big Data, are available from individual consumers. To better understand the impact of Big Data on various marketing activities, enabling firms to better exploit its benefits, a conceptual framework that builds on resource-based theory is proposed. Three resources—physical, human, and organizational capital—moderate the following: (1) the process of collecting and storing evidence of consumer activity as Big Data, (2) the process of extracting consumer insight from Big Data, and (3) the process of utilizing consumer insight to enhance dynamic/adaptive capabilities. Furthermore, unique resource requirements for firms to benefit from Big Data are discussed.

2. TITLE: Big data analytics.

AUTHOR: P. Russom et al.

Big Data Analytics offers a nearly endless source of business and informational insight, that can lead to operational improvement and new opportunities for companies to provide unrealized revenue across almost every industry. From use cases like customer personalization, to risk mitigation, to fraud detection, to internal operations analysis, and all the other new use cases arising near-daily, the Value hidden in company data has companies looking to create a cutting-edge analytics operation.

Discovering value within raw data poses many challenges for IT teams. Every company has different needs and different data assets. Business initiatives change quickly in an ever-accelerating marketplace, and keeping up with new directives can require agility and scalability. On top of that, a successful Big Data Analytics operation requires enormous computing resources, technological infrastructure, and highly skilled personnel.

All of these challenges can cause many operations to fail before they deliver value. In the past, a lack of computing power and access to automation made a true production-scale analytics operation beyond the reach of most companies: Big Data was too expensive, with too much hassle, and no clear ROI. With the rise of cloud computing and new technologies in compute resource management, Big Data tools are more accessible than ever before.

3. TITLE: Big data consumer analytics and the transformation of marketing.

AUTHOR: S. Erevelles, N. Fukawa, and L. Swayne

The data explosion over the last couple of years combined with technological advances has opened up a new way of analysing and gaining information from the data. The evolution of Big Data has led to the rise of Big Data Consumer Analytics wherein the large resources of data are used to extract information and then exploited to create value. This has transformed the way marketing is done and the interaction between the firm and its customers. In this research, we explore the properties of Big Data and how they can influence marketing campaigns by providing the right kind of information a marketer is looking for, that will help the firm to move ahead of its competitors. We shall also explore how the traditional knowledge-based view has limitations when implemented with Big Data and why it should be combined with an ignorance-based view to make the most of the data. The importance

and impact of using Big Data Consumer Analytics in the marketing strategies and how certain firms have created advantages for themselves with the help of the data has been discussed. Some of the challenges and limitations have been mentioned in the end.

4. TITLE: Predicting the semantic orientation of adjectives

AUTHOR: V. Hatzivassiloglou and K. R. McKeown,

We identify and validate from a large corpus constraints from conjunctions on the positive or negative semantic orientation of the conjoined adjectives. A log-linear regression model uses these constraints to predict whether conjoined adjectives are of same or different orientations, achieving 82% accuracy in this task when each conjunction is considered independently. Combining the constraints across many adjectives, a clustering algorithm separates the adjectives into groups of different orientations, and finally, adjectives are labeled positive or negative. Evaluations on real data and simulation experiments indicate high levels of performance: classification precision is more than 90% for adjectives that occur in a modest number of conjunctions in the corpus.

5. TITLE: Recognizing contextual polarity in phrase-level sentiment analysis

AUTHOR: T. Wilson, J. Wiebe, and P. Hoffmann,

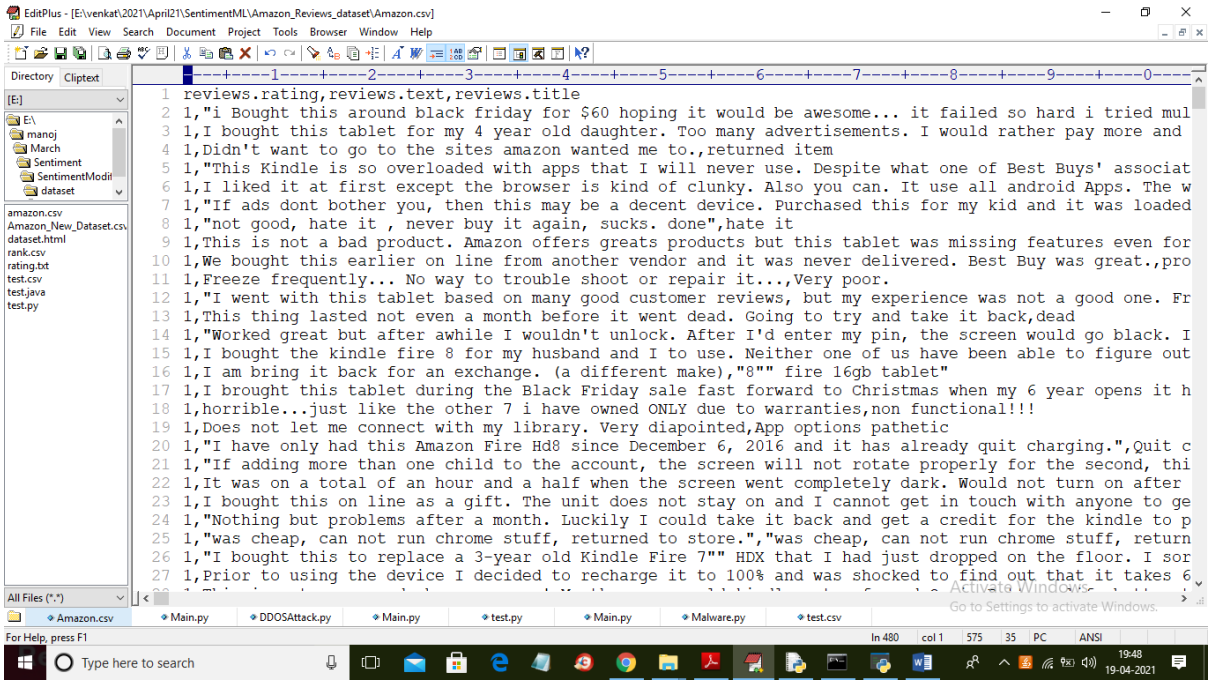
This paper presents a new approach to phrase-level sentiment analysis that first determines whether an expression is neutral or polar and then disambiguates the polarity of the polar expressions. With this approach, the system is able to automatically identify the contextual polarity for a large subset of sentiment expressions, achieving results that are significantly better than baseline.

3. PROPOSED SYSTEM

Sentiment analysis is not only confined to the English language but has been implemented for various languages. Sentiment analysis of Chinese text by implementing four feature selection methods and five classifiers viz. Centroid classifier, K-nearest neighbor, Window classifier, Naïve Bayes and SVM has been done [11]. Through this learning paradigm it was concluded that SVM outperforms all the other learning methods in terms of sentiment classification. Sentiment analysis on travel reviews using three machine learning models namely, Naïve Bayes, SVM and character based N-gram model has been performed in which SVM and N-gram approaches have better performance than Naïve Bayes [12]. It has been observed that in maximum number of cases SVM showcases best performance in comparison to other classification models.

4. RESULTS

In this project author is detecting sentiments from amazon reviews by using various machine learning algorithms such as SVM, Decision Tree and Naïve Bayes. In all 3 algorithms SVM is giving better accuracy and to train this algorithms author has used AMAZON reviews dataset and this dataset is saved inside ‘Amazon_Reviews_dataset’ folder. Below screen shot show example reviews from dataset



In above dataset first row contains column names and remaining rows contains dataset values and in above dataset first column contains sentiment values from 1 to 5 and its associated with each review and we will use above dataset to train all 3 machine learning algorithms.

To implement this project author has used following modules

1) Data Collection: Using this module we will upload AMAZON reviews dataset to application

2) Data Preprocessing: using this module we will read all reviews and then remove stop words, special symbols, punctuation and numeric data from all reviews and after applying Preprocessing we will extract features from all reviews.

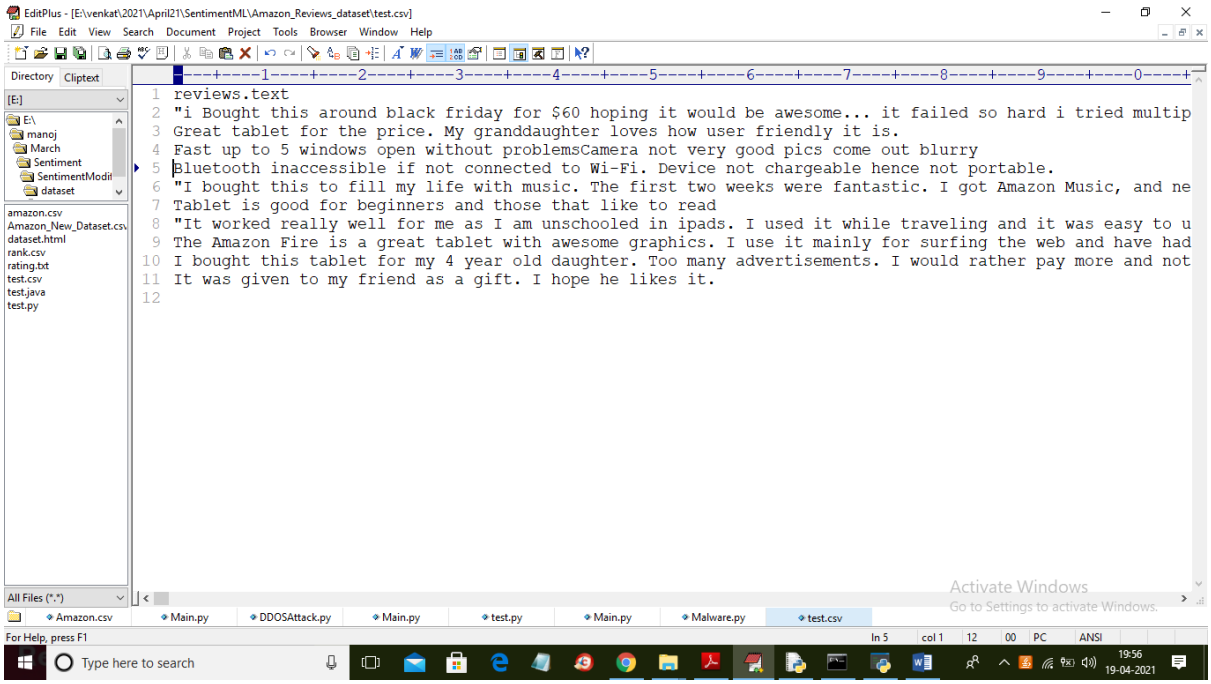
3) Features Extraction: here we will apply TF-IDF (term frequency Inverse Document Frequency) algorithm to convert string reviews into numeric vector. Each word count will be put in vector in place of words.

4) Run SVM Algorithm: We will apply SVM algorithm on TF-IDF vector to train SVM algorithm and then we apply test data on SVM trained model to calculate SVM prediction accuracy

5) Run Naïve Bayes Algorithm: We will apply Naïve Bayes algorithm on TF-IDF vector to train Naïve Bayes algorithm and then we apply test data on Naïve Bayes trained model to calculate Naïve Bayes prediction accuracy

6) Run Decision Tree Algorithm: We will apply Decision Tree algorithm on TF-IDF vector to train Decision Tree algorithm and then we apply test data on Decision Tree trained model to calculate Decision Tree prediction accuracy

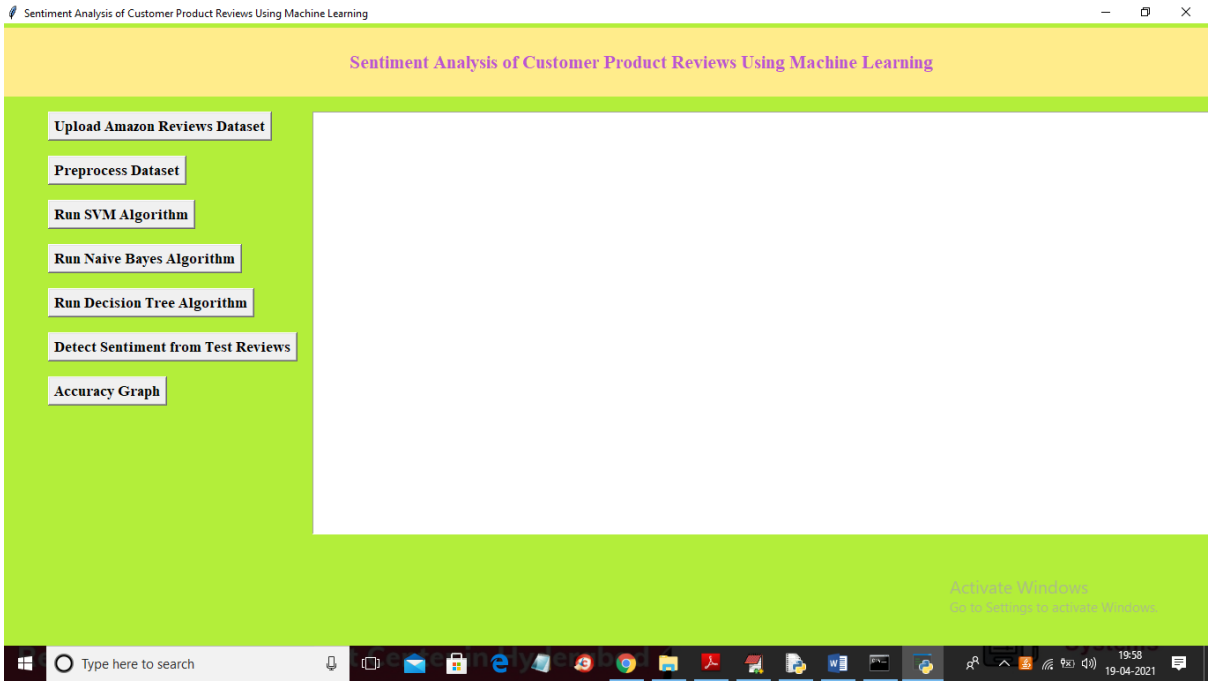
7) Detect Sentiment from Test Reviews: Using this module we will upload test reviews and then ML algorithm will predict sentiment for each review and in below test reviews dataset we can see there is no sentiment value and ML will predict sentiment for each test value



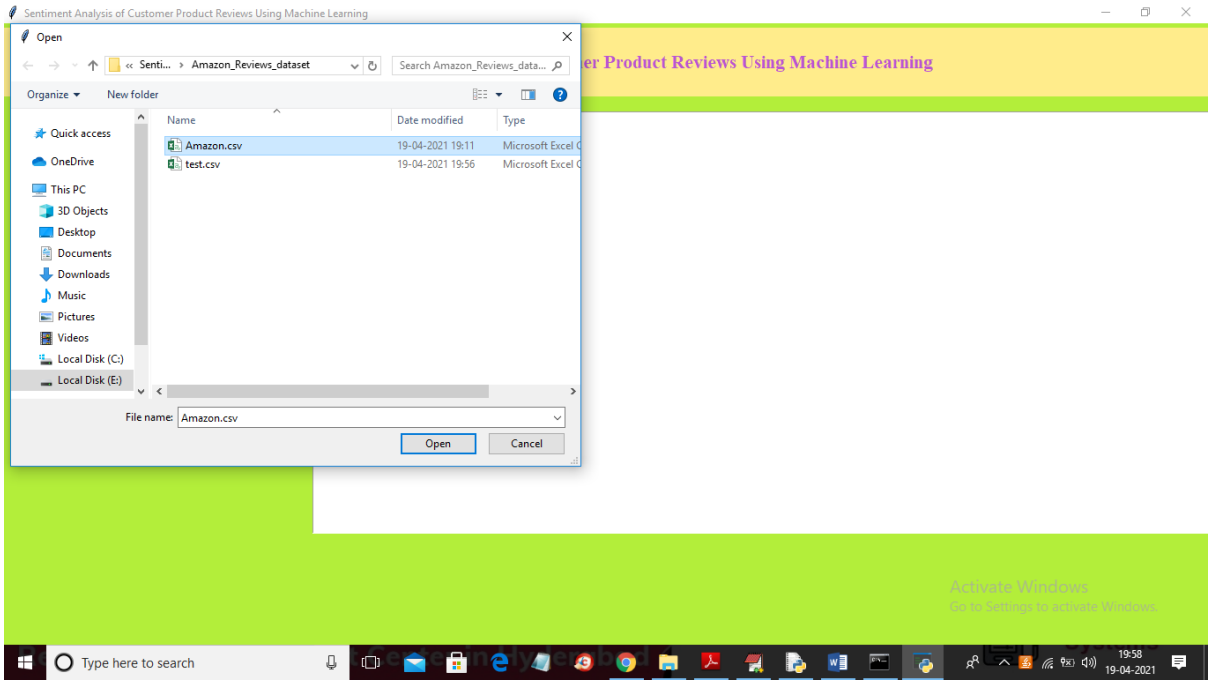
In above test data we have only test reviews and by applying ML trained model on above test data we can predict sentiment label.

Screen Shots

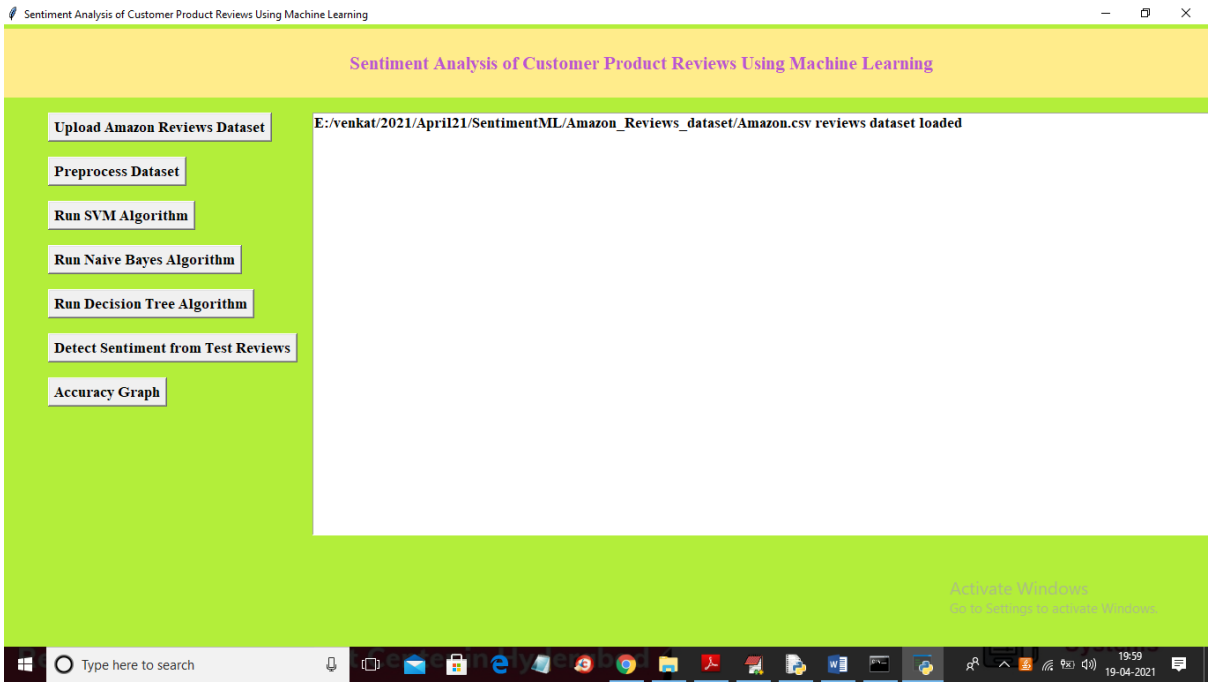
To run project double click on 'run.bat' file to get below screen



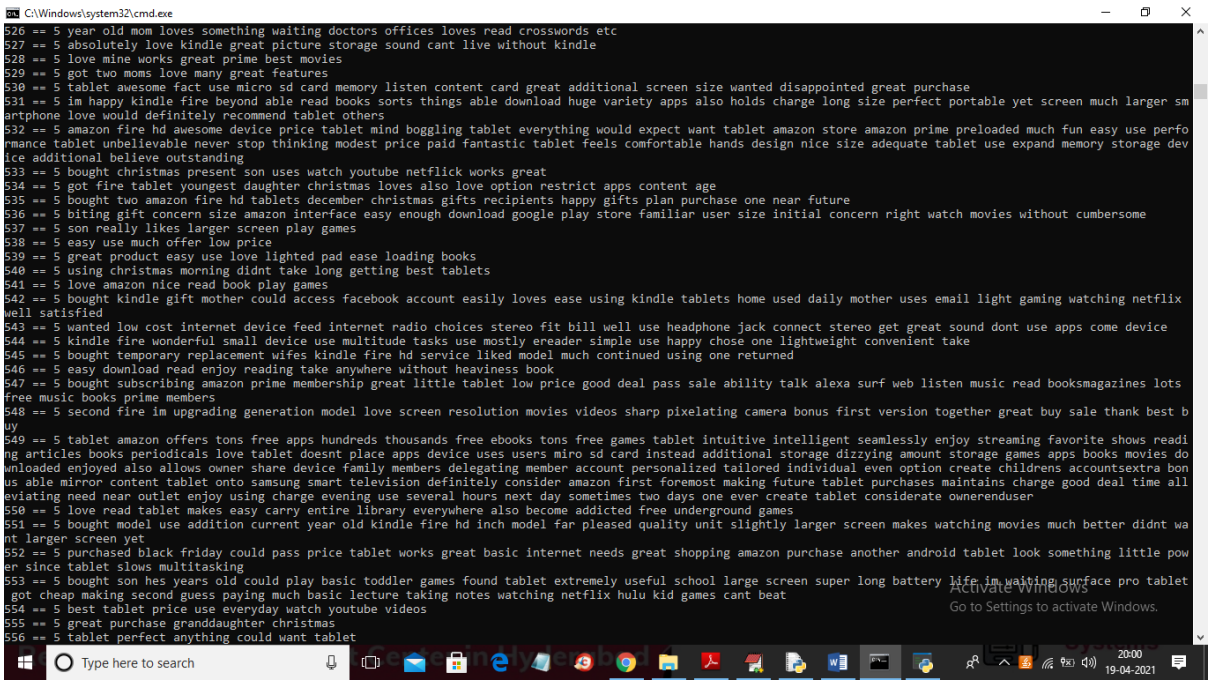
In above screen click on 'Upload Amazon Reviews Dataset' button to upload dataset



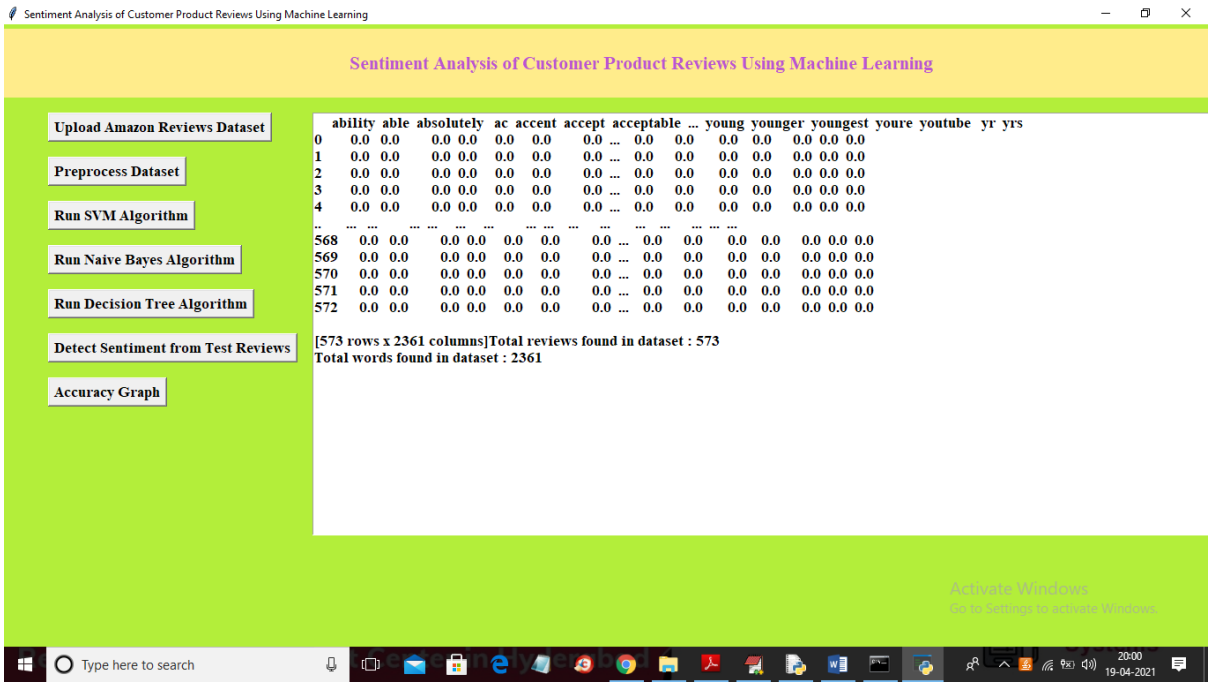
In above screen we are selecting and uploading 'Amazon.csv' file and then click on 'Open' button to load dataset and to get below screen



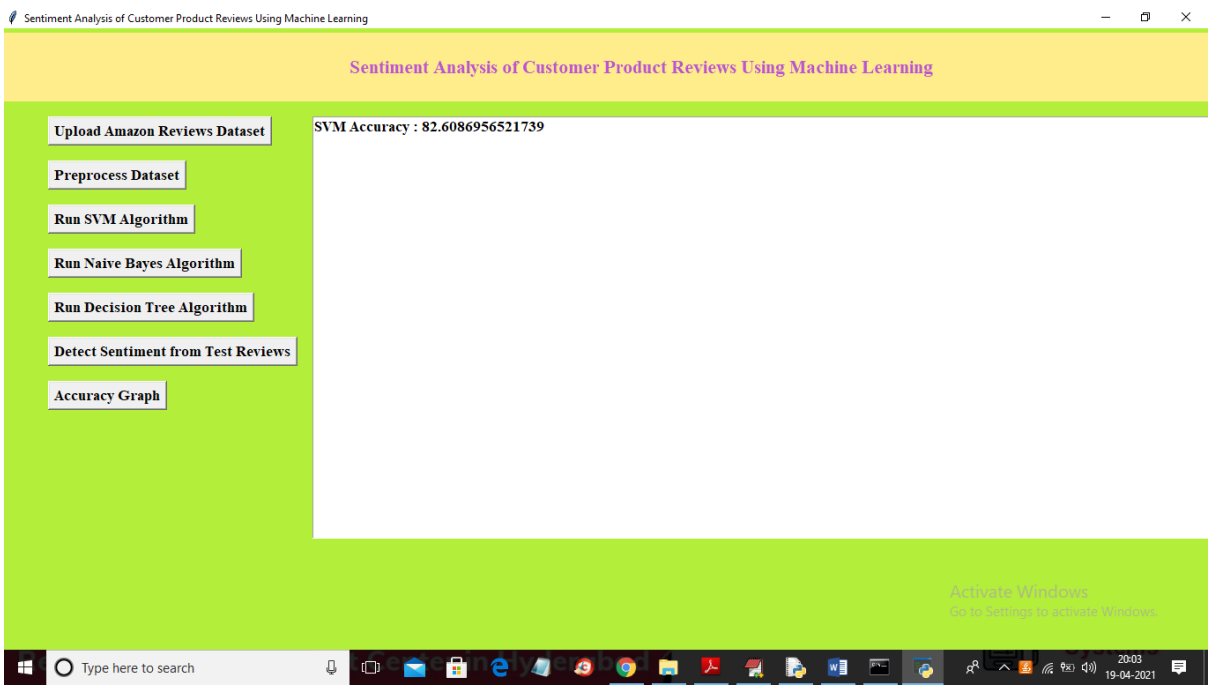
In above screen dataset loaded and now click on 'Preprocess Dataset' button to read all reviews from dataset and then apply Preprocess steps to get below screen



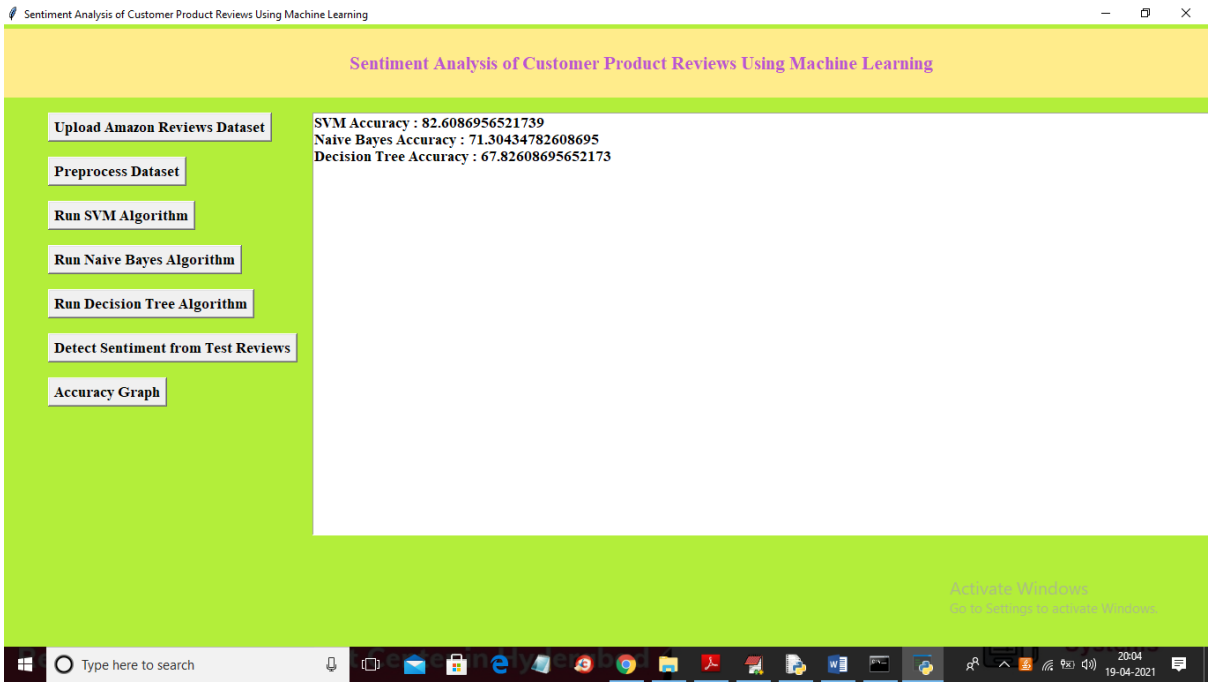
In above black console we can see application read all reviews from dataset and then generate below TF-IDF vector



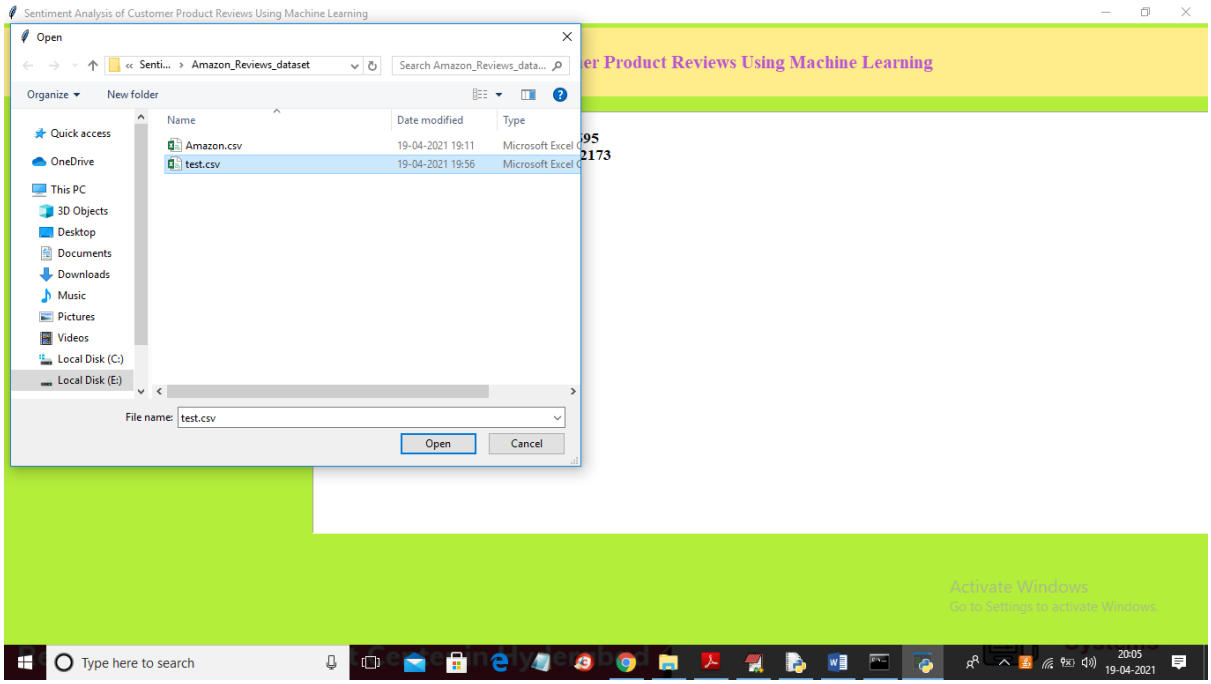
In above screen in text area we can see application extract all words from reviews and then put in top line of above test area and in remaining rows if that word appear then it put average count value of that word and if word not appear then 0 will put. In above screen vector generated and I am showing few records from that vector. In that vector total reviews are 573 and all reviews contains total 2361 unique words. Now vector is ready and now click on 'Run SVM Algorithm' button to train SVM with above vector



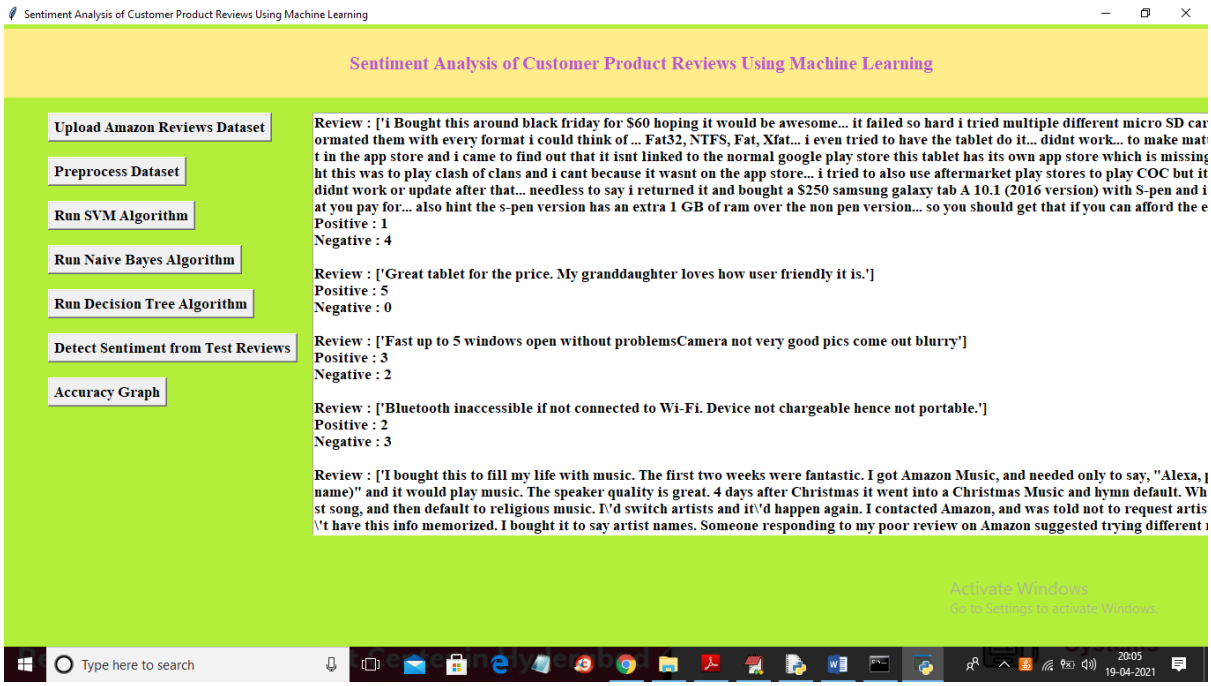
In above screen with SVM we got 82% accuracy and now click on Naïve Bayes and Decision tree button to get their accuracy



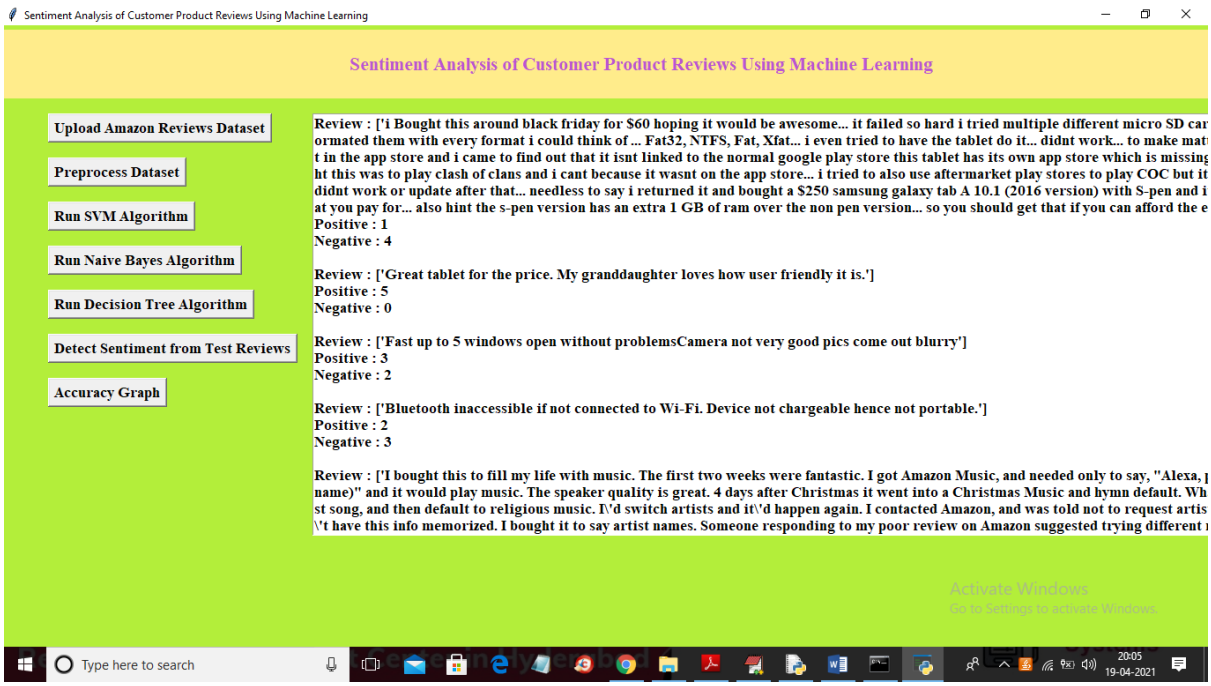
In above screen with all 3 algorithms SVM gave better prediction accuracy and now click on 'Detect Sentiment from Test Reviews' button to upload test reviews



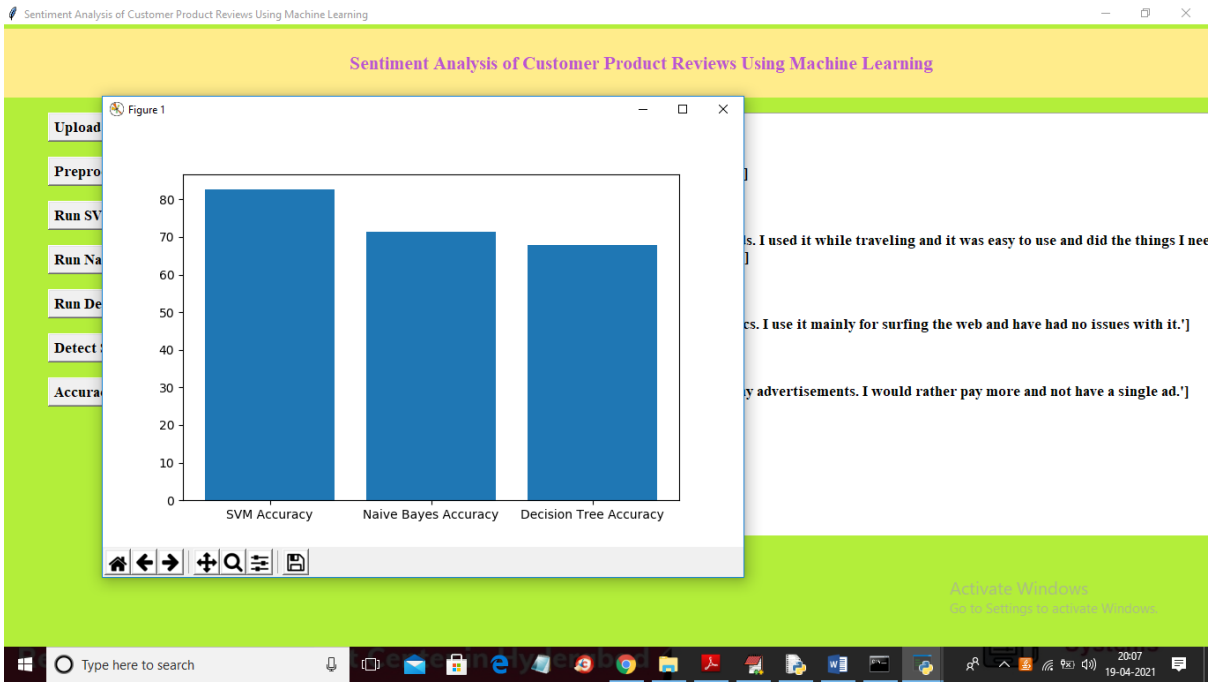
In above screen selecting and uploading 'test.csv' file and then click on 'Open' button to get below prediction result



In above screen first I am displaying reviews from uploaded test file and then predicting positive and negative sentiment for each review and you can scroll down above text area to get all outputs



In above screen we can see sentiment prediction result for all reviews and now click on 'Accuracy Graph' button to get below graph



In above graph x-axis represents algorithm name and y-axis represents accuracy of those algorithms and in all 3 algorithms SVM got higher accuracy

5. CONCLUSION

An evolutionary shift from offline markets to digital markets has increased the dependency of customers on online reviews to a great extent. Online reviews have become a platform for building trust and influencing consumer buying patterns. With such dependency there is a need to handle such large volume of reviews and present credible reviews before the consumer. Our research is aiming to achieve this by conducting sentiment analysis of mobile phone reviews and classifying the reviews

into positive and negative sentiment. After balancing the data with almost equal ratio of positive and negative reviews, three classification models have been used to classify reviews. Out of the three classifiers, i.e., Naïve Bayes, SVM and Decision Tree, predictive accuracy of SVM is found to be the best. The accuracy results have been cross validated and the highest value of accuracy achieved was 81.75% for SVM among the three models.

Future Scope

In future, the work can be extended to perform multiclass classification of reviews which will provide delineated nature of review to the consumer, hence better judgement of the product. It can also be used to predict rating of a product from the review. This will provide users with reliable rating because sometimes the rating received by the product and the sentiment of the review do not provide justice to each other. The proposed extension of work will be very beneficial for the e-commerce industry as it will augment user satisfaction and trust.

REFERENCES

[1] S. Erevelles, N. Fukawa, and L. Swayne, "Big data consumer analytics and the transformation of marketing," Journal of Business Research, vol. 69, no. 2, pp. 897–904, 2016.
[2] P. Russom et al., "Big data analytics," TDWI best practices report, fourth quarter, pp. 1–35, 2011.
[3] S. Erevelles, N. Fukawa, and L. Swayne, "Big data consumer analytics and the transformation of marketing," Journal of Business Research, vol. 69, no. 2, pp. 897–904, 2016.
[4] V. Hatzivassiloglou and K. R. McKeown, "Predicting the semantic orientation of adjectives," in Proceedings of the eighth conference on European chapter of the Association for Computational Linguistics. Association for Computational Linguistics, 1997, pp. 174–181.
[5] T. Wilson, J. Wiebe, and P. Hoffmann, "Recognizing contextual polarity in phrase-level sentiment analysis," in Proceedings of the conference on human language technology and empirical methods in natural language processing. Association for Computational Linguistics, 2005, pp. 347–354.
[6] B. Pang and L. Lee, "A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts," in Proceedings of the 42nd annual meeting on Association for Computational Linguistics. Association for Computational Linguistics, 2004, p. 271.
[7] A. Pak and P. Paroubek, "Twitter as a corpus for sentiment analysis and opinion mining." in LREc, vol. 10, no. 2010, 2010.
[8] M. WAHYUDI and D. A. KRISTIYANTI, "Sentiment analysis of smartphone product review using support vector machine algorithm-based particle swarm optimization." Journal of Theoretical & Applied Information Technology, vol. 91, no. 1
[9] D. N. Devi, C. K. Kumar, and S. Prasad, "A feature based approach for sentiment analysis by using support vector machine," in Advanced Computing (IACC), 2016 IEEE 6th International Conference on. IEEE, 2016, pp. 3–8.
[10] V. Narayanan, I. Arora, and A. Bhatia, "Fast and accurate sentiment classification using an enhanced naive bayes model," in International Conference on Intelligent Data Engineering and Automated Learning. Springer, 2013, pp. 194–201.