International Research Journal of Engineering and Technology (IRJET) e-ISSN: 2395-0056
Volume: 06 Issue: 07 | July 2019 www.irjet.net p-ISSN: 2395-0072

THE SENTIMENTAL ANALYSIS ON PRODUCT REVIEWS OF AMAZON DATA USING THE HYBRID APPROACH

Kajal 1*, Prince Verma2

[1]*M.Tech Scholar, Dept. of Computer Science Engineering, CT Institute of Engineering Management & Technology, Jalandhar, India*
[2]*H.O.D & Assistant Professor, Dept. of Computer Science Engineering, CT Institute of Engineering Management & Technology, Jalandhar, India*

Abstract - *The term sentiment specifies to the viewpoint or speculation of the person towards some particular domain like the topic, product information, etc. The sentiments are analyzed by the user towards any domain or elements by using this SA approach. Web or internet or social network is the best-known source to assemble sentiment information. Among popular social network portals or sites, Twitter has been the point of attraction to lot of researchers. Twitter is a big platform in which user post their messages or point of view related to any product or any other thing are said to tweets. In this paper, the study work is based on the sentiment analysis (SA) of product reviews of Amazon data by using the Twitter API .In the existing work, the SVM technique is applied that contains some issues and these issues can be resolved by applying the KNN technique for the sentiment analysis. We used six different training-testing compositions for performance analysis. Moreover, we present the performance contrast of the discussed techniques based on our recognized parameters.*

Key Words: **KNN (K-Nearest Neighbor), SA (Sentiment Analysis), Sentiments, SVM (Support vector machine), Twitter.**

I. INTRODUCTION

Twitter moreover another Social media (like facebook, WhatsApp etc), Social Networks, online posts and microblogging sites on the web became closely the largest destinations on the web for communication in between people or any other party to show their thoughts related to any products [1]-[2] or movies [3] allotment their regular understanding and convey their point of view or thoughts about actual time and upcoming events like political elections or sports[4] etc. The mining of such type of thoughtful data that can be extracted from social media is highly useful to conclude various fields [5]. Inappropriate, the growth of social media (SM) such as Twitter, Facebook, Snap-chat, Instagram, Wechat and also online reviews has powered interest in sentiment analysis [6].

In sentiment analysis, the Tweets, text, and speech that show opinions, thoughts, views, and emotions related to any product or service are extracted from the database sources. Arrangement of thoughts amid "positive", "negative" and "neutral" is completed through sentiment analysis way. Since getting any kind of products all user gain the knowledge about them through sentiment analysis way. Based on user's requirement the products or services can be served by the associations or organizations to their buyers or purchasers. The available accurate data is processed figure out or examined by the technique like textual information retrieval. The subjective attributes can be indicated by some textual contents or facts. The base of sentiment analysis is organized by the essence which mainly involves thoughts, views, opinions, and sentiments [7].

In broad, there are three levels where thoughts mining based on which sentiment analysis (SA) can be executed. They are given below:

a. Document-level sentiment analysis: In Document level, SA classifies the whole document and shows the polarity like positive or negative or neutral for any kind of service or product. It is the simplest type of sentiment analysis. Mostly under the document-level sentiment analysis [8], we can use two types of approaches that are shown below:

[1] Supervised learning
[2] Unsupervised learning

In the supervised learning approach that considers there is a definite no of groups or classes in which the whole documents should be classified and for individual class training data is accessible. In the Unsupervised learning approach, the analysis is with-in document conclusive the semantic direction thus of unique phrases. Mostly we can select the phrases by two main methods that are given below:

[1] A set of predefined part of speech
[2] A lexicon of sentient words [8]-[9].

b. Sentence-level sentiment analysis: To determine a particular sentence show a positive, negative or neutral thoughts or feelings for any kind of product or service, we can use Sentence-level sentiment analysis. This level of analysis is executed by two tasks that are subjective and objective. Among which the Subjective sentence is used for the classification is executed and the true information expressed by objective sentences [10]-[11].

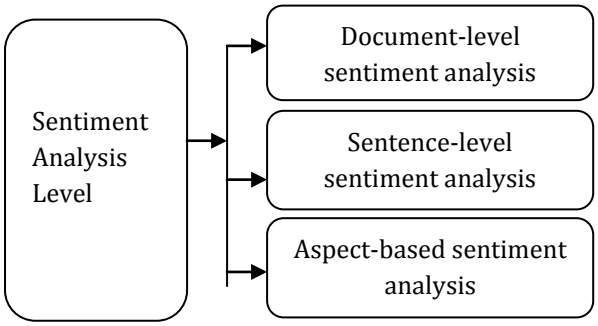

Fig 1: Sentiment Analysis Level

c. Aspect-based sentiment analysis: This approach work on a single entity.The research problem which targets on the recognition and extraction of every sentiment expressions within a provided document [11].

This paper contains 5 sections that describe the sentiment analysis of Twitter data. Section I contains the Introduction about the sentiment analysis, its levels. Section II presents various literature surveys in the field of sentiment or opinion mining. Section III represented the sentiment analysis research methodology and Section IV elaborates experimental results and discussion of work. The Last section V defines the conclusion and future scope of sentiment analysis.

II. LITERATURE REVIEW

From the last set of years, many articles, papers, and books have been written on sentimental analysis. At a similar time, some researcher's center of attention more on exact burden like finding the subjectivity expression, subjectivity clues, topics, and sentiments of words and extracting sources of opinions, while others target is on assigning sentiments to the whole document. All analysts of sentiment analysis have become distinct approaches to naturally predict the expression, sentiments of words or a document. The data set for sentimental analysis considered are a movie, product review or social media data from the source of the internet. Because of the importance of the sentiments analysis, researchers have developed different techniques to enhance the efficiency of the analysis. They use pattern-based approaches, NLP (Natural Language Processing), lexicon based analysis as well as machine learning techniques, etc.

Huma Parveen, et.al (2016) proposed the Hadoop Framework that was used for processing the data set of movie accessible on the twitter in a variety of feedback, reviews, and comments [12]. The outcomes of SA on twitter data displayed in the type of different sections similar to positive, negative and neutral sentiments. The pre-processing of data was done to remove noise. This type of analysis would help any organization to increase their business productivity. This method also provided the fast downloading approach for proficient twitter trend analysis. The concluded results indicated that in the future, sentiment analysis will be done on other social networking sites also.

Ankit, et.al (2018) proposed that Ensemble Classifier was used to improve the precision and performance of the sentiment classification technique [13]. This classifier combined the base learning classifier to form a solitary classifier. The obtained results indicated that the ensemble classifier gave better performance than a stand-alone classifier. In sentiment classification technique, the role of data pre-processing and feature representation was also explored. The ensemble classifier system was developed by using various base learners like Naïve Bayes classifier, SVMs, random forest classifier, and logistic regression. This approach was very helpful for companies to monitor consumer opinions regarding their products. In the future, the main center of attention will be on the study of natural tweets since the different tweet sentiments.

Sushree Das, et.al (2018) stated that a stream-based setting by using the incremental active learning approach gave the capability to the algorithm for choosing new training data from a data stream for hand-labeling [14]. Stream-based active learning in the financial domain could be helpful to both sentiment analysis and the active learning research area. With the use of RNNs Long Short –Term Memory, this experiment also proved helpful for feasibility study through batch processing. To analyze the sentiments and recent stock trends, a hybrid model could also be developed. This model would improve the reliability of the prediction. In the future for analyzing the stock data, addition also applies the machine learning algorithms. Some additional methods of data ingestion like data ingestion through the Apache Flume or NodeIJS can also be used in the future.

Symeon Symeonidis, et.al (2018) proposed that various pre-processing techniques evaluated on their ensuring the accuracy of the classification and the number of features they developed [15]. The obtained results indicated that some methods like lemmatization and removing numbers also replacing contractions improved precision while other techniques like removing punctuations did not. To explore the interactions between the methods when they were employed in a pipeline manner, an ablation and combination study was done. The outcomes of these techniques indicated the importance of techniques like replacing numbers and replacing repetitions of punctuations.

Neethu M S, et.al (2013), in this paper, they analyze the twitter data related to Electronic products with Machine Learning approach [16]. They existent a new Feature-Vector for classification of the tweets and extricate peoples' opinions about Electronic products.Thus created a Feature-Vector from 8 relevant features. Naïve-Bayes and SVM classifiers are implemented using built-in functions of Matlab. Max-Entropy classifier is implemented using

Maximum-Entropy software. All the used classifiers have almost equal performance.

Rasika Wagh, et.al (2018) reviewed that the sentiment analysis involved the area of data mining and NLP [17]. The sentiment analysis research of Twitter data could be done in many ways. Various types and techniques of sentiment analysis were used to done extraction of sentiments from tweets after which a comparative study is done. In classify to construct opinion on sentiment, the analyzation of twitter data was done from a different point of view. The study of the literature indicated that when the semantic analysis Word-Net was followed up by the machine learning techniques like SVM, Naïve-Bayes and maximum entropy then the accuracy improved. The accuracy could also be increased by up to 4-5% by using the hybrid approach.

M.Trupthi, et.al (2017) proposed an interactive automatic system that used Hadoop to analyze the sentiment of the tweets posted in social media [18]. In the process of Sentiment analysis, genuine tweets were used. The main motive of research was to perform real-time sentiment analysis on tweeter data and to give time-based prediction to the user. The main features of the system were a training module which was done by using Hadoop and Map-Reduce. The classification method was based on Naïve Bayes, Time-Variant Analytics and the Continuous Learning system. The proposed system had some limitations like the use of Uni-gram Naïve, adaptability to only English Language and the lack of actual intended meaning. But these problems could be removed by making certain changes like the use of n-gram classification instead of Uni-gram, to build more adaptive language systems, etc. In the future, this system will be helpful to people and industries based on sentiment analysis like Sales Marketing, Product Marketing, etc.

III. SENTIMENTAL ANALYSIS RESEARCH METHODOLOGY

This section presents the steps of the sentiment analysis of twitter data shown in fig 2. The method classifies a Twitter data set of English language into positive or negative or neutral sentiments.

A. Standard Product Dataset

We gathered our dataset by consulting the Twitter API. Two types of datasets are generated manually here amongst which one is used for training and another is used for testing. X: Y is the relation present within the training set. Here the X is represented by the score of probable opinion word and Y is the representation of whether the score is positive or negative. By gathering reviews from the e-commerce sites, the testing set is generated. A review of whether the testing set is positive or negative is manually tagged. The reviews will be separated based on positive and negative sentiments they include once the training is completed after that the system is tested, with the help of reviews. The accuracy of the system can be determined based on an output that is generated by the system.

B. Preprocessing the data

In this segment, the input data is pre-processed before extracting the features. Dataset used in EXCEL or CSV file is used and comments are expressed in ENGLISH language. Stemming [22], error correction and stop word removal are the three main preprocessing techniques which are performed here [20].

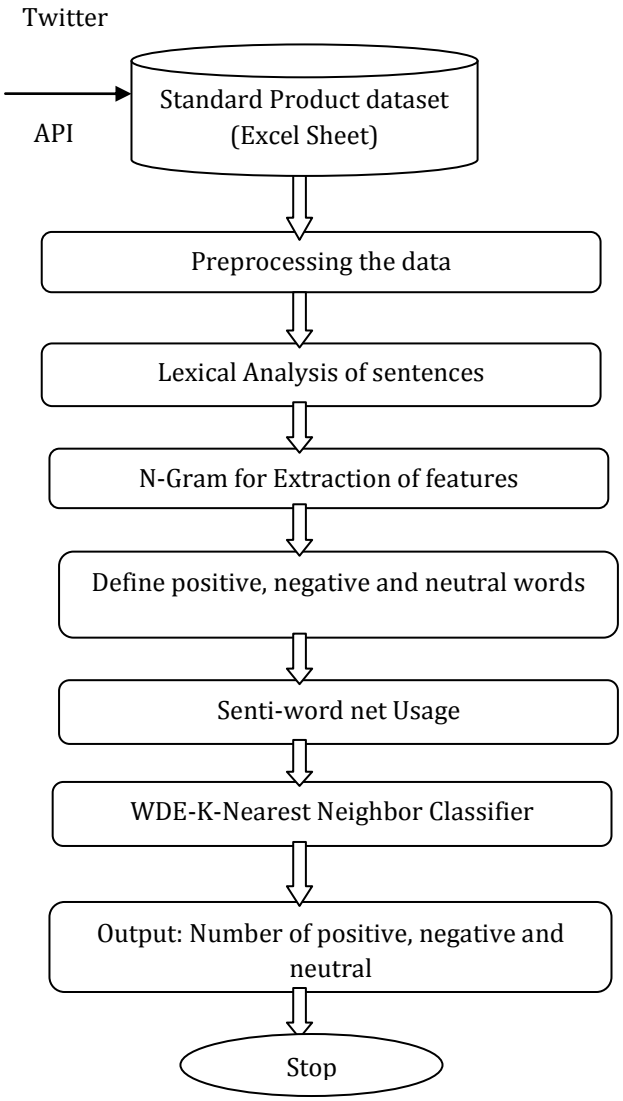

Fig 2: Steps of Proposed Sentimental Analysis Flowchart

C. Lexical Analysis of Sentences

In the dataset, Sentence which includes either a positive or negative sentiments. To minimize the complete size of the review, such sentences can be removed which might not include any sentiments within them. The regular expressions involved within python do not recognize these sentences of questions [19].

D. N-Gram for Extraction of Features

The main problem arises inside the SA though extractive the features from data. The N-gram algorithm is applied which can extract the features and also post tag the sentences.

E. Define Positive, Negative and Neutral Words

The grammatical dependencies present the words in the sentences will be gathered by the parser and given as output [18].To identify the opinion word for features that have been gathered from the last step, the dependencies have to be looked upon in further steps [19]. There is also a need to contain the direct dependencies and transitive dependencies within this step [20].

F .SentiWordNet Usage

Sentiwordnet is generated especially within the opinion mining applications. There are 3 main relevant polarities present for each word within the Sentiwordnet which are positivity, negativity and neutral [17]-[20].

G. WDE-K-Nearest Neighbor Classifier

In organize to use a classifier within this approach, WDE-KNN is preferred. Since sentiment analysis (SA) is a binary classification and there are huge datasets that can be executed, WDE-KNN is chosen here. The generated training set is useful for training the classifier here. There is X: Y relation provided within the training set. In which the x is represented by the score of an opinion word and it is represented by score whether the word is positive or negative [17]-[20]-[21]. A score of the view word related to a feature within the review is specified as input to the KNN classifier.

H. Output

The output comes in the form of extraction of features wise opinions i.e. Number of positive, negative and neutral. The accuracy of the system can be decided on the base of output.

IV. EXPERIMENT RESULT & DISCUSSION

In this research work is related to sentiment analysis of twitter dataset. This section presents the environment setups; the data is collected over the twitter by using Twitter API and the conducted experiments for studying the method performance.

1). Accuracy: For Accuracy analysis, Table 1 shows the overall result and the graphical representation demonstrated in Chart 1. The classifiers used were SVM and KNN for classifications. The classifiers were trained and tested several times on the same dataset. KNN classifier achieved high accuracy 84.63 %(in case of 40-60 ratio) i.e. 40 percent of the data was used for training and 60 percent of the data was used for testing , 85.32% (in case of 50-50 ratio)i.e. 50 percent of the data was used for training and 50 percent of the data was used for testing, 85.92%(in the case of 60-40) i.e. 60 percent of the data was used for training and 40 percent of the data was used for testing, 91.08%(in case of 70-30) i.e. 70 percent of the data was used for training and 30 percent of the data was used for testing, 91.15 %(in case of 80-20) i.e. 80 percent of the data was used for training and 20 percent of the data was used for testing and 96.43(in case of 90-10) i.e. 90 percent of the data was used for training and 10 percent of the data was used for testing. This experimental result shows that the KNN approach achieved the highest accuracy for the classes like positive, negative and neutral than the SVM approach, as elaborated in Chart 1.

Table 1: Accuracy Analysis

Training ,Test Ratio	SVM classifier (%)	KNN classifier (%)
40-60	80.63	84.63
50-50	81.17	85.32
60-40	82.59	85.95
70-30	84.17	91.08
80-20	85.88	91.15
90-10	85.92	96.43

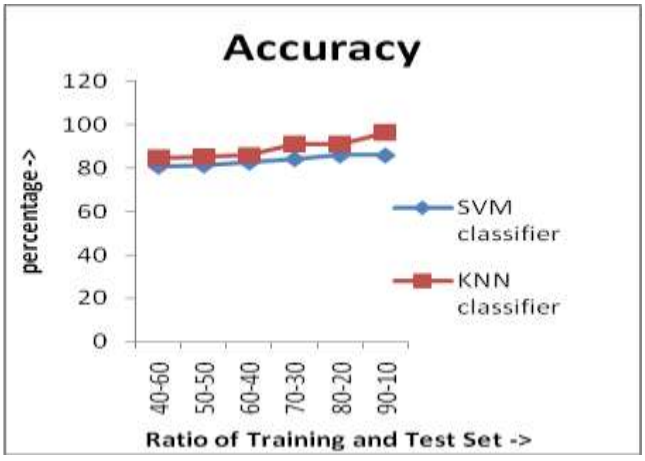

Chart 1: Accuracy Analysis

2).Execution Time: For Execution Time, Table 2 shows that overall results for the different number of training and test ratio. The graphical representation of all these results presented in Chart 2. In case of (40-60 ratio) i.e. 40 percent of the data was used for training and 60 percent of the data was used for testing and the experiment result show that KNN approach required only 3.08625 seconds which is comparatively low as compared to the SVM approach. In case of (50-50 ratio) i.e. 50 percent of the data was used for training and 50 percent of the data was used for testing and the experiment result show that KNN approach required only 3.49785 seconds which is comparatively low as

compared to the SVM approach. In the case of (60-40 ratio) i.e. 60 percent of the data was used for training and 40 percent of the data was used for testing and the experiment results prove that the KNN approach required only 3.27577 seconds which is comparatively low as compared to the approach SVM.

In the case of (70-30 ratio) i.e. 70 percent of the data was used for training and 30 percent of the data was used for testing and the experiment results prove that the KNN approach required only 3.83721 seconds which is comparatively low as compared to the SVM approach. In the case of (80-20 ratio) i.e. 80 percent of the data was used for training and 20 percent of the data was used for testing and the experiment results prove that the KNN approach required only 3.44296 seconds which is comparatively low as compared to the SVM approach. In the case of (90-10 ratio) i.e. 90 percent of the data was used for training and 10 percent of the data was used for testing and the experiment results prove that the KNN approach required only 3.37808 seconds which is comparatively low as compared to the SVM approach.

Table 2: Execution Time

Training ,Test Ratio	SVM classifier (Time in seconds)	KNN classifier (Time in seconds)
40-60	7.07257	3.08625
50-50	8.27024	3.49785
60-40	9.22069	3.27577
70-30	9.76641	3.83721
80-20	10.5381	3.44296
90-10	11.26555	3.37808

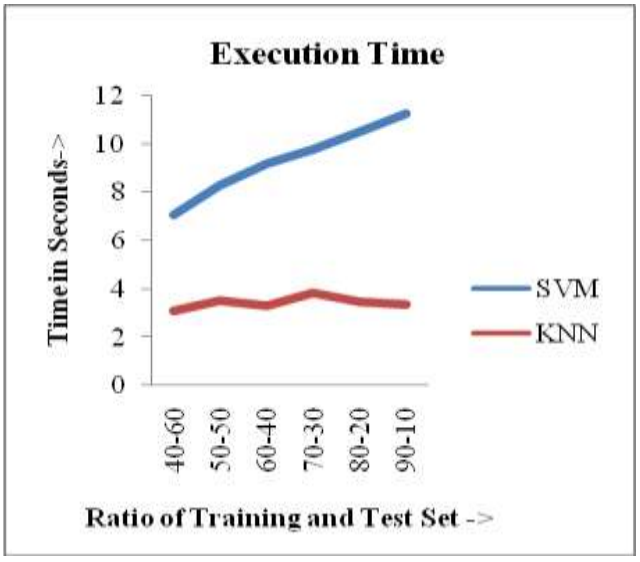

Chart 2: Execution Time Analysis

This experiment result proved that the KNN approach required low time (in seconds) for execution as compared to the SVM approach, as presented in Chart 2.

V. CONCLUSIONS AND FUTURE SCOPE

Sentiment analysis of tweets is an important area of research. The sentiment analysis of Twitter data, we can analyze user behavior. Through classification and analysis of sentiments on Twitter, we can get perceptive of people's attitudes about the product. Firstly N-gram approach is applied for sentiment analysis of Twitter data. By using N-gram we can extract the features then analyzed the input data. Furthermore, We applying a classification technique for data classification. In this work, we can study and comparative analysis the techniques used are SVM and KNN. The classification of KNN performs better as compared to the classification of SVM because KNN uses multiple hyper-planes for data classification. The SVM and KNN techniques are implemented in Python and Experiment results analysis shows that the KNN approach achieved a high accuracy score over the SVM approach. The execution time of the KNN approach is low as compared to SVM. In the future, we would similar to analyze and compare sentiment analysis with other domains.

REFERENCES

[1] B. O'Connor, R. Balasubramanyan, BR. Routledge, and N. Smith, ''from tweets to polls: Linking text sentiment to public opinion time series,'' in Proc. International. AAAI Confer. Weblogs Social Media, pp. 26-33, May 2010.

[2] M. Anthony Cabanlit and K. Junshean, '' Optimizing N-Gram based text feature selection in sentiment analysis for commercial products in Twitter through polarity lexicons,'' in Proc. 5th Int. Conference. Inform, Intell. Syst. Application., pp. 94-97, Jul. 2014.

[3] Umesh Rao Hodeghatta, ''Sentiment analysis of Hollywood movies on Twitter,'' in Proc. IEEE/ ACM ASONAM, pp. 1401-1404, Aug. 2013.

[4] Juan M. Soler, Fernando Cuartero, and Manuel Roblizo, "Twitter as a tool for predicting elections results," in Proc. IEEE/ACM Int. Conf. Adv. Social Netw. Anal. Mining (ASONAM), 2012, pp. 1194–1200.

[5] R. Barahate Sachin and M. Shelake Vijay, "A Survey and Future Vision of Data mining in Educational Field", in Proc. 2nd Int. Conf. on Advanced Computing & Communication Technology, pp. 96-100, 2012.

[6] Liu B. "Sentiment analysis and opinion mining." San Rafael, Computer Application: Morgan & Claypool, 2012.

[7] R. Parikh and M.Movassate, "Sentiment Analysis of the User- Generated Twitter Update using Various Classification Technique", Final Report CS224N , 2009.

[8] R. Feldman, "Techniques and Applications for Sentiment Analysis", Communications of the Association for Computing Machinery, pp 8289, Volume. 56, Issue-4, 2013 .

[9]E.Divya,"Real Time Sentiment Classification Using Unsupervised Reviews" Inter. Jour. of Sci. & Engi..Research (IJSER), Vol 5(3), March-2014.

[10] R. Kaur, P. Verma "Sentiment Analysis of Movie Reviews: A Study of Machine Learning Algorithm with Various Feature Selection Methods", (IJCSE), Vol-5, Iss-9, 2017.

[11] N. Nehra, "A Survey on Sentiment Analysis Of Movie Reviews", (IJIRT) , pp 36-40,Volume1, Iss.7, 2014.

[12] H. Parveen, P. S. Pandey, "Sentiment Analysis on Twitter Data-set using NB Algorithm", Senti WordNet Dictnary,2016.

[13] Ankit, Nabizath Saleena, "An Ensemble Classification System for Twitter Sentiment Analysis" , Int. Conf. on Comp. Int. and Data Sci. (ICCIDS),2018.

[14] S. Das, R. Kumar Behera, M. Kumar, S. Kumar Rath, "Real Time Sentiment Analysis of Twitter Streaming Data For Stock Prediction", Int. Conf. on Comp. Int. and Data Sci. , 2018.

[15] Sy. Symeonidis, Di. Effrosynidis, Avi. Arampatzis, "A comparative evaluation of pre-processing techniques and their interactions for twitter SA", Expert Sys. With App. (ESA), 2018.

[16] Neethu, R. Rajasree and M. S., "Sentiment analysis in twitter using machine learning techniques." Comp., Comm. and Net. Technologies (ICCCNT), IV Int. Conf. on IEEE, 2013.

[17] R. Wagh, P. Punde, "Survey on Sentiment Analysis using Twitter Dataset", Proc. of the 2nd Inter. conf. on Electronics, Communication and Aerospace Tech. (ICECA) ,2018.

[18] M.Trupthi, G.Narasimha, S. Pabboju, "Sentiment Analysis on Twitter using Streaming API", 7th Inter. Advance Comp. Confer., IEEE, 2017.

[19] Z. Rezaei, M. Jalali, "Sentiment Analysis on Twitter using McDiarmid Tree Algorithm", 7th Inter. Confer. on Computer and Knowledge Eng. (ICCKE), October 26-27 ,2017.

[20] Kajal, Prince Verma "Hybrid approach for sentimental analysis of twitter data", Inter. Jour. of Comp. Sci. and Eng.., Volume-7 (6), Jun 2019.

[21] Martín-V. M T, Rushdi S. M, Urena-Lopez L A, Montejo-Raez A, "Experiments with SVM to classify opinions in different domains", Expert Sys. With Applications (ESA), pp:14799- 14804, 2011.

[22] H. Kaur, Prabhjeet Kaur, "Dimensionality Reduction In Sentiment Analysis Using Colony-Support Vector Machine", Inter. Journal of Innovative Tech. and Exploring Eng. (IJITEE), Vol-8 , Iss-8, 2019