JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

Structured application of Content Aware Support Vector Machines in Sentiment Analysis for Twitter Data

[1]Manikandan. B, [2]Dr. Chakaravarthi. S

[1]Research Scholar, [2]Professor
Department of Computer Science and Engineering,
Bharath Institute of Higher Education and Research, Chennai –73, India
E-Mail: bmanibala@gmail.com, chakra2603@gmail.com

Abstract:

One of the relatively new topic of potential research is the approach known as sentiment analysis in the area of micro-blogging. Albeit, prior research related works on the foregoing, the case differs for twitter. The impediment of applying the sentiment analysis in twitter is due to its limitation of only 140 characters in a single tweet. The Content Aware Support Vector Machines is a mobile application that yields satisfactory results. Today, most of the communication/dissemination of information happens using mobile applications. At times, the actual emotions with respect to the chat is not easily understood. Hence, this approach is expected to be very useful especially in understanding the emotion or sentiment in the chat.

Keywords:

Content Aware Support Vector Machines, Sentiment Analysis, Supervised Learning Technique, Twitter Analysis

1. INTRODUCTION

The popularity of Twitter over the recent years is unprecedented. As a consequence of such tremendous growth, many companies mine Twitter to analyse customers or users opinion about their products and services. It is true that, in the online analysis of products and services the way in which sentiments are conveyed has attracted a lot of attention for researchers. However, in microblogging there are very limited research. One of the key reason being the usage of informal language and very concise message. The most useful approach for sentiment analysis in other areas is the use of sentiment vocabulary and grammatical tagging. The question here is, whether the same approach would prove to be a success for sentiment analysis in Twitter. Therefore, this paper attempts to study in detail the foregoing question.

Besides the constrains of limited text in microblogging, it also encompasses almost the entirety of topic making it even more challenging. Hence, it requires an efficient method to easily recognise data to be utilized for training. Subsequently, it will not be difficult to build systems to mine Twitter sentiment notwithstanding its wide spectrum of topics it covers. Taking into account the need for an efficient method for building twitter data, this paper utilizes Twitter hashtags to categorize different types of tweets based on sentiments such as positive, negative, and neutral tweets. These can be subsequently used for training three way sentiment classifiers.

Today, technology has enabled people to express opinions/view freely. Hence, the information available online is fathomless. It is important to study the wide spectrum of opinions and one of the efficient ways to analyse the text is by deploying sentiment analysis method. It helps companies/organizations to understand customers feedback on their products, helps political parties to analyse the sentiments of the voters and for movie industries to understand the viewer's choice. All in all, the method is useful in making logical judgements and take corrective measures for future course of action.

"Bag of words" is a conventional techniques used in many sentiment analysis. The drawback with this method is that, it fails to recognize the language pattern / structure and subsequently categorize phrases with different meanings as one. This is because "Bag of words" accord importance to the collections of words rather than individual words. The aforementioned method also disregards word order which may lead to wrong classification. The succeeding section

discusses different methods of sentiment analyses. For better understanding of the existing methods and developments in the area of sentiment analysis and text classification, the Literature Survey summarizes different approaches used so far.

The computational study of different expressions such as opinions, sentiments and emotions in text encompassing a wide spectrum of natural language processing is known as Sentiment analysis. The main aim of it is to understand/learn opinions, attitudes and emotions of the viewers/customers/public for a product/service/topics. There has been a lot of research to gain better insights in the area of sentiment analysis. However, most of the study so far concentrates on categorizing reviews (text data) which are formal and large. The burgeoning of social media and simultaneous increase in the amount of information available has led researchers to make significant strides on sentiment analysis. Twitter is considered as one of the most popular microblogging site. It enables users to disseminate/share views/opinions/discuss on a wide spectrum of topics/products/services in a real time environment.

With ever increasing user base of 550 million since 2006, Twitter is one of the most promising and effective service. And extensive research for mining twitter data for information has become equally important. It puts on view all the tweets across the world in a real-time environment. The origin of Twitter concept could be insignificant with service only for providing personal status updates. However, today its growth and popularity is incomprehensible with tweets encompassing almost everything. Hence, we can see the impact of it in politics, lifestyle, products and services, movie industry etc. There is a basic structural difference between reviews and Tweets. The former is a condensed views/thoughts of the author identified by formal text patterns while the latter is comparatively informal and has text limitation of 140 characters. Tweets today play a significant role in the ever evolving market. It assists them in analysing user's experience with their products and services and most importantly it helps them in future decision making. Deploying an efficient technique for sentiment analysis on such data would have a far reaching impact. Methods such as machine learning, sentiment lexicons, hybrid approaches etc. have so far proven to be a success in extracting formal text. As mentioned earlier, tweets comprises of informal and limited text. Therefore, there is a lot to be accomplished in the area of extracting microblogging data. Upon further examination, it is found that the limit of just 140 character text per tweet puts a constrain on using appropriate words which consequently impedes the sentiment. Compared to other formal text form of media, tweets are often filled with incorrect spellings and colloquialisms. All of these often pose a challenge in sentiment analysis for tweet content. With enormous amount of data available from twitter and other microblogging websites encompassing a wide spectrum of topics, one can leverage the suggestive punctuations commonly used in microblogging language. It can provide room for sentiment extraction from tweets. The proposed work examines tweets using machine learning techniques integrating adapted polarity lexicon.

A detail understanding on sentimental analysis for twitter data is explained in Section 1. It also incorporates the impediments involved in examining such data. An extensive literature survey is given in Section 2 and section 3 includes the Proposed Algorithm. The outcome of the proposed method is discussed in section 4 and finally the culmination of the study is given in sections 5.

2. LITERATURE SURVEY

In this technology driven world, the opinion of the people in various fields especially products and services cannot be overlooked. To identify the contradictory tendencies of the viewers in Persian movies, the paper proposes an approach utilizing TF-IDF and transition point. Various classifiers are deployed to evaluate the scheme and prove the effectiveness of the latter [1].

To avoid any loss of information during various stages, the study proposes a comprehensive approach to sentiment examination. It involves deployment of various efficient systems in order to detect the negation. Use of specific pattern help in computing the sentiments and greatly enhances its ability to recognize negation [2].

Product reviews are generally not structured. Therefore, it is important to develop an efficient data mining technique for the analysis. In order to make the analysis more complete, the study proposes a model by incorporating emoticon. The experiment conducted with datasets of reviews for iPhone from Amazon prove that inclusion of smiley, emo tag etc. undoubtedly improves the performance [3].

Sentiment analysis is paramount for the progress of an organization today. The study attempts to conduct an extensive review on the various machine-learning algorithms used for the purpose. Through the review of all relevant papers it can be inferred that sentiment analysis is necessary before bringing out a new product in the market. It helps manufacturers to understand the trend and mind-set of the users and also the negative feedbacks helps in improving the product quality [4].

With the burgeoning of Amazon's presence in the food market it has become difficult to process its massive data using traditional techniques. The study deploys sentiment analysis using a processing technique known as apache spark. Three classification mechanisms are applied. Results prove that linear support vector classifier surpasses the performance of the other classifiers [5].

The study conducts a comparative examination on various methods involved in sentiment analysis. Irrespective of the methods used, sentiment analysis is proven to be the most successful technique of achieving higher accuracy. In order to counter each other's drawbacks and strengths, the proposed approach integrates lexicon-based and machine learning approach. The former for sub-processing reviews and latter in sentiment analysis. Besides online shopping, the approach can also be efficiently deployed for other domains [6].

The growth of e-commerce is unprecedented and it is understood that the reviews of the customer plays a vital role. The study aims to analyse the datasets of amazon products review using visualization techniques. It also categorizes positive and negative reviews and derive association. This information goes a long way in improving the products and services [7].

The study constitutes an extensive survey on SA practices and other aspects relevant to the field. It analysis existing works and represents the growing updates in sentiment analysis. Subsequent to the analysis of the papers, it can be inferred that there is much to be explored in the aspect of emotion classification and trait selection algorithms [8].

In order to process enormous datasets such as that of Amazon, the paper proposes an algorithm known as tunicate swarm algorithm. It has the ability to minimize the size of the feature set to 43% by retaining its accuracy. Over and above its ability to minimize feature size, it is also time efficient, scalable and outperforms other existing techniques. The analysis involves few benchmark such as time taken for calculation, size of feature etc. [9].

Together with a features extraction algorithm, the Spark NLP method is deployed in the study for sentiment analysis. It exhibits a high accuracy and precision. The Spark conditions are favorable for huge concurrent sentiment analysis with regards to scalability. The RDD approach used in the proposed model assures the availability and fault tolerance [10].

The paper examines various methods involved in text-based sentiment analysis. It measures the capability of various algorithms based on machine learning and categorizes them into supervised and unsupervised methods. It also considers various applications involved in semantic analysis in different domains [11].

Dissemination of information (both positive and negative) has been very common due to accessibility and availability of technology. In order to classify the reviews into positive and negative, the paper presents an approach for Persian movie reviews using convolutional neural networks and long-short-term memory algorithms. Results prove the latter algorithm to be more efficient with higher accuracy [12].

For identifying hate speech, the study gathers a complete data set comprising of Urdu Tweets. Various techniques are applied to enhance the performance of the classifier. It can be inferred from the results that sentiment analysis for the case in discussion can be efficiently resolved by getting rid of problems involving high dimensionality and class skew [13].

Today there are umpteen platforms to share view and information. Subsequently, it has transformed the way online business operates. To study the sentiment analysis at aspect level, the paper proposes a model based on machine learning. The special feature of the approach is that it allows admin to examine the review pairs. Precision, recall, F-score and accuracy are few variables that are effectively utilized in the study to evaluate the performance of the classifier [14].

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

There are many models utilized in overcoming the challenges involved in examining the product reviews and categorization of sentiment. It can be inferred that input dataset required in machine Learning and Deep Learning methods are really huge and are not cost efficient. Also, their performance has huge dependency on the problems and the input dataset [15].

Sentiment analysis is taking a central role in the improvement and growth of products. The paper focuses on the use of natural language processing algorithm. It helps in gathering accurate information (reviews) at a faster rate. For each feedback there is a final outcome which is subsequently captured and displayed as per the user's requirement [16].

The proposed method calculates sentiment scores and subsequently rates an alternate product using NS theory. To determine the efficiency of the method, real datasets are used and results prove that the proposed model is capable of tackling neutral data at both stages [17].

The study proposes a sentiment analysis model that recognizes/identifies the various aspects impacting sales. It analyses the trends of the customer that in turn determines the products popular amongst the customers and those that are not popular/lesser in demand. Classification of the sentiments is carried out using two different algorithms. The proposed model yields high accuracy [18].

3. PROPOSED ALGORITHM CONTENT AWARE SUPPORT VECTOR MACHINES

The algorithm used here is known as Content Aware Support Vector Machines which is a type of binary classification. It identifies correlationships between feature and a sentiment by carefully examining feature vectors with a labelled class. The following illustration would simplify the understanding of the process. Consider each vector as data point in vector space of dimension equivalent to feature vector size. Subsequently, a hyper plane is identified by SVM separating into two classes/functional margin. "Best "can most appropriately be defined as "a beneficial separation by the hyper plane". And the proximity to the training data point is important for the class. As the margin becomes bigger, the risks are assumed to be lesser. As soon as the system receives an unprocessed tweet, it extracts the feature vector 9 similar to how it extracted processed ones. Consequently, the learned model receives vector as an input. This operation identifies the position of the new data point on the hyper plane and then the class is delegated. The foregoing process has the potential to classify only up to two classes but sentiment classification comprises of three classes. To overcome such challenges, the proposed SVM utilizes binary classification which is a one-versus all solution. The process begins by testing the class to which an instance belong to and it gets terminated if the output is true. If not, it continues for the next class. Based on the result of whether the instance belong to that class or not, the third class is assigned.

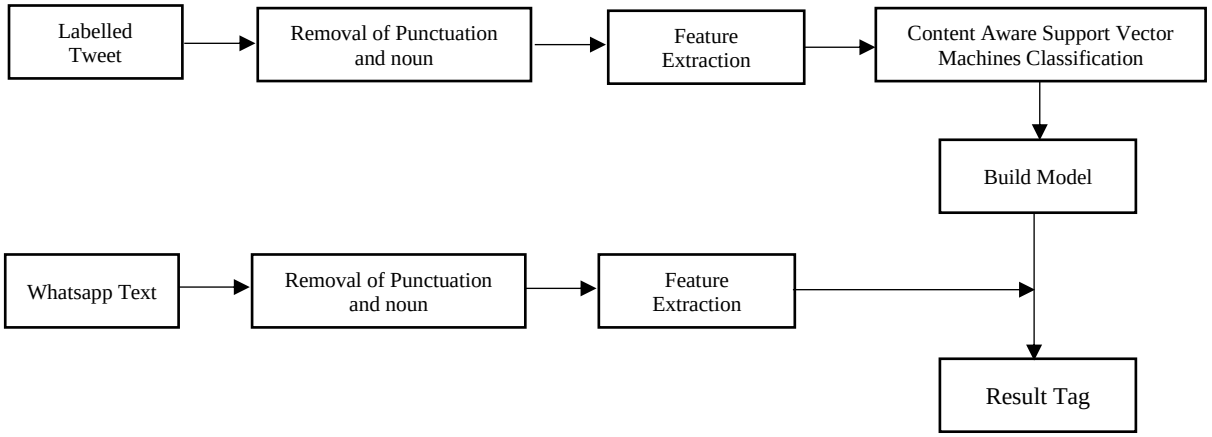


Fig.1: Content Aware Support Vector Machines Algorithm

Labelling of tweet is done by rearranging the input of Fig.1 architectural design. This is followed by training period wherein Noun and Punctuation are eliminated and tweet is broken up into stop words. They are then stemmed and each

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

word is named as token. Subsequently, features are extracted by applying the technique of semantic analysis. And sentiment analysis model is created utilizing CASVM. A comparative study is done between the feature and the model which helps in identifying positive and negative tags.

This model basically based on the existence of words in a document but does not necessarily consider the presence of certain word. For e.g., the word "indulging" is accorded more importance for ascertaining the polarity of the sentence in the earlier document. In the binary model, those words showing up in the document 1 get the value '1' and '0' is assigned to those not appearing in the document. Such interesting result inspires to attempt some other bag of words models and further enhance the process of sentiment analysis. The other bag of words model in the experiment is with term frequency-inverse document frequency scores. The distinction of this model is the presence of scores for every word instead of a vector of '0's and '1's. By multiplying TF and IDF for specific words, the score are evaluated. The equation below represents the score of any word in any document:

$$TFIDF(word, doc) = TF(word, doc) * IDF(word)$$

There are two arrangements that needs to be considered and calculated. From the collection of words, it is necessary to evaluate the frequency of inverse document followed by term frequency. This is applicable for each word in each document.

$$TF(word, doc) = Frequency of word \in the doc \ / No. of words \in the doc$$

$$I(word) = log e\ (1 + No. of docs \ / No. of docs with word)$$

Hence, the foregoing method enables the sentences in discussion to be changed to TF-IDF model. The process commences by forming an IDF dictionary comprising of 8 words that are regularly occurring. Subsequently, TF dictionary with TF values with equivalent words in each documents is worked out. **Table 1** represents the TFIDF model.

Table 1: TF-IDF model

Docs / Words	**the**	**feature**	**of**	**game**	**is**
D1	1	1	0	1	0
D2	0	1	1	0	0
D3	0	0	1	0	0
D4	0	0	0	0	1

There is a clear distinction between this model and that of binary bag of words model. Here, documents are not represented as vectors of '0's and '1's. It shows more clarity by assigning values within 0 and 1. It is understood that the TF-IDF model accords higher credit to the unusual words instead of considering all the words as equal (as in binary bag of words model). However, the competency of this model is challenged whenever it comes across sentences with negations. This is because the presence of negative words in a sentence greatly affects the polarity of word/sentence. So, to enhance the results of this model, the presence of negative words must be considered. The third model in the experiment deploys a negation strategy wherein words are negated depending on the earlier understanding of polar expressions. In this model, the succeeding words change with respect to the negation word. As soon as a negation word is noted and till punctuation is received, all the succeeding words are prefixed with a 'not_'. The practicability of this approach is questioned as negating all the words will lead to creation of unnecessary words in the compilation.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

Fig.1 clearly depicts the algorithm used for pre-processing in this model. In order to make the understanding less complex, the stop words are eliminated showing just the negation part. Hence, our approach utilizes input documents without punctuation. Subsequently, it connects the entire document wherein ACSVM technique is performed for each document.

Text Data Processing:

Input (KeyWord):

To obtain a less complex collection of twitter streaming API, tweepy is utilized which acquires raw tweets. It facilitates the easy accessibility via SampleStream and FilterStream. The former works in real time environment and supplies tweets in small and arbitrary form. However, the tweets delivered via the latter mode adhers to three specific filtering criteria as below:

- Based on a particular keyword in the tweets
- Based on user such as twitter user name
- Based on the geographic origin of the Tweets.

These criteria's can be specified by the developer as a combination or a single criteria. In this experiment, our focus would be on the SampleStream mode.

The collection of tweets have been made over a period of time instead of obtaining it at one time. This is to enhance the diversity of data and avoid mundane sentiments which can be triggered by some burning topics. For e.g. festive seasons such as Christmas and New Year is accompanied by so much hype all across the world. Hence, majority of tweets acquired during these seasons would apparently have positive sentiments referring to celebrations. Our approach of data collection spread over different periods of time/seasons would help overcome such issues. Accordingly, the first collection was initiated on 17th December 2015. Likewise, after few days the next sample was collected and moved on to mid of January the subsequent year. It culminated during the early part of second month of 2016.

Such arbitrary collection may have a mixture of both useful as well as unwanted data. It's a raw form of data referred to as python "dictionary" data type accompanied by pairs of key-value as furnished below:

- Popularity of the tweet
- User ID and screen name
- Authenticity
- hashtags
- Repeated or first time
- Registered user Language
- Location from where tweet originated(Geo-tag)
- Time stamp of the tweet

The information from the above key-value pairs is huge. Hence, during the experiment all unwanted data is eliminated and only the necessary ones are retained. The original content is saved in a different file. After the elimination and other processes are completed, it is now important to mark the dictionary key and language – the former as "text" and latter as "lang".

TWEETS RETRIEVAL:

Tweets Retrieval: The manual process of labelling is both tedious and expensive. Hence, the process of filtering continues such that diversity is retained. Furnished below are few criteria used for filtering:

- Tweets with "RT"(Retweets) are eliminated

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

- Very concise tweets having less than 20 characters are eliminated
- Tweets containing less than 15% of 2,000 listed common English words are discarded
- Tweets with almost 90% similarity are also discarded.

Once the filtering process is complete, the remaining tweets for labelling are about 30% of the collected tweets accounting for 10,173.

Data Pre-processing: It comprises of the following three steps explained one by one:

Tokenization: It involves disintegration of text into significant components such as words, symbols etc. The components are known as TOKENS and they can be segregated by characters such as whitespace or punctuations. This step is important in order to identify them as individual parts in a tweet. E.g. Emoticons and abbreviations are identified as individual tokens.

Normalization:

It involves identification of abbreviations, identification of all-caps short sentences and repetitions. The All-caps words are converted to lower case, abbreviations are restored to their actual meaning and repetitions are also replaced by proper single word. Over and above, the normalization process also tracks the existence of other tokens such as #hashtags, user-tags, URLs etc. The whole exercise is to enhance the process of normalization which will further strengthen the POS tagger performance to a considerable extend. It will also successfully sum up the pre-processing step.

Part-of-speech:

In this process, tags are allocated to each word in the sentence. We are aware that a grammatically correct sentence has parts of speech component. Here, the words are allocated their part such as noun, pronoun, verb, adjective, adverb, etc. Once the parts are allocated, it is easier to track the count of the number of parts of speech that exists in a particular tweet.

In order to demonstrate the integration of an algorithm to an application, it is paramount to construct a sentiment analysis of Twitter data. As discussed earlier, twitter covers a wide spectrum of topics. Hence, it is important to first select the topic for the analysis. Once a topic is selected, the tweets are collected using the keywords of the selected topic and sentiment analysis can be conveniently performed on those tweets. At the end of the exercise, we will understand the general perception/mood of the people with respect to that topic/issue. The negative or positive views of the people can be ascertained.

4. RESULTS AND DISCUSSION

The program used in this application is known as Python Script. A total of 45000 data set is considered for training and 40000 data set for testing (Table 2).

Table 2: Used Data set details

Training Data	45000
Negative	23754
Positive	21246

Step 1: Gather Tweets

The introductory step involves choosing a topic for analysis. Any phrase can be used to define input inside the sentiment-analysis.js. For e.g. The word used here is anticipated to retrieve positive results.

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

```
var algorithmia = require("algorithmia");

var client = algorithmia(process.env.ALGORITHMIA_API_KEY);

var input = "happy"; var no_retweets = [];

console.log("Analyzing tweets with phrase: " + input);

client.algo("/diego/RetrieveTweetsWithKeyword/0.1.2").pipe(input).

then(function(output)

        { if (output.error)

{ console.log(output.error);

} else { var tweets = []; var tweets = output.result;

for (var i = 0; i < output.result.length; i++)

                { // Remove retweets. All retweets contain "RT" in the string.

if (tweets[i].indexOf('RT') == -1) { no_retweets.push (tweets[i]);

} } }

// The subsequent step would encompass this function.

analyze_tweets(no_retweets); });
```

To progress the selected topic to the algorithm, the API of the algorithm is utilized. For each call, the number of tweets is limited to 500. It retrieves texts that contains the keywords and captures those tweets with the phrase. Another important step is the elimination of repetitions to avoid duplicate data. By attaching RT as a prefix to the repeated tweets, it is convenient to identify and eliminate those tweets from our data set.

Step 2:

Following the gathering and elimination process, the data set is now ready for Sentiment Analysis. It is performed on each tweet and finally an average score for all the tweets together are calculated.

```
var analyze_tweets = function(no_retweets)

 { var total_score = 0; var score_count = 0; var final_score = 0; // perform sentiment analysis on each tweet  followed by average score calculation.

for (var j = 0; j < no_retweets.length; j++)

{ client.algo("nlp/SentimentAnalysis/0.1.1").pipe(no_retweets[j]).then(function(output)

 {if (output.error)

{console.log (output.error); }

 else { console.log(output.result); score_count = score_count + 1;

total_score = total_score + output.result; } // Calculate average score.

if (score_count == no_retweets.length)
```

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

{ final_score = total_score / score_count; console.log('final score: ' + final_score); } }) } }

Each tweet is repeated through no-retweets in order to transmit as input to the Sentiment Analysis algorithm. The output is added to the total_score variable along with the end result from that API call. The number of tweets that goes through variable score_count is kept on track. Finally, when the number of tweets reaches the score for analysis, the final score is evaluated by averaging the total_score. The final result obtained is very decisive as it signifies different sentiments. It ranges from [0-4] and denotes sentiments in the following order: very negative, negative, neutral, positive, and very positive sentiment.

Classified Tweets:

Depending on the sentiments, the tweets are categorized as positive, negative and neutral. In order to assist the labellers, the following specifications are provided:

Positive: It encompasses a wide range of interpretation. Any tweet that has a viewpoint which is appreciative, contented, thrilled or cheerful in nature is considered positive tweet. Also, a tweet which is more of positive words than other sentiment can also be counted as positive tweet.

Negative: Any sorrowful, gloomy-ridden, despondent opinions expressed in tweet is considered to be of negative in nature. The content may not necessarily by full of such negative sentiments. However, if the pessimistic expression dominates the tweet then it is also categorized as a negative tweet.

Neutral: Any impartial, unbiased, just and even-handed opinion wherein the user has no personal connection with the particular issue/service/product is considered neutral. In most of the circumstances, such tweets are purely for the dissemination of information E.g. We can categorize commercials, promotions or even announcements under such umbrella. They don't intent to create polarization amongst anyone.

Over and above the aforementioned criteria's, tweets of other language besides English are abandoned and not considered in the training data. Having briefly explained the methods applied for text formatting, the subsequent part would incorporate the different features explored in the study. It is clearly indicated below that any variable that assists our classifier in bringing out a distinction between different classes is counted as a feature. In our system, two major classifications are discussed namely objectivity / subjectivity and positivity / negativity. Before we delve deeper in understanding the two classification, it can be briefly explained that the former distinguishes between different personal perspective and decisions based on facts. The latter distinguishes different sentiments of which the two extreme sides of the spectrum would be positive and negative.

Given below are the features examined for objective/subjective classification:

Number of exclamation marks in a tweet
Number of question marks in a tweet
Presence of exclamation marks in a tweet
Presence of question marks in a tweet
Presence of url in a tweet
Presence of emoticons in a tweet
Unigram word models calculated using Naive Bayes
Prior polarity of words through online lexicon MPQA
Number of digits in a tweet
Number of capitalized words in a tweet
Number of capitalized characters in a tweet
Number of punctuation marks / symbols in a tweet
Ratio of non-dictionary words to the total number of words in the tweet
Length of the tweet Number of adjectives in a tweet
Number of comparative adjectives in a tweet
Number of superlative adjectives in a tweet
Number of base-form verbs in a tweet
Number of past tense verbs in a tweet
Number of present participle verbs in a tweet
Number of past participle verbs in a tweet

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

Given below are the features examined for positive / negative classification

Overall emoticon score (where 1 is added to the score in case of positive emoticon, and 1 is subtracted in case of negative emoticon)
Overall score from online polarity lexicon MPQA (where presence of strong positive word in the tweet increases the score by 1.0 and the presence of weak negative word would decrease the score by 0.5)
Unigram word models calculated using Naive Bayes
Number of total emoticons in the tweet
Number of positive emoticons in a tweet
Number of negative emoticons in a tweet
Number of positive words from MPQA lexicon in tweet
Number of negative words from MPQA lexicon in tweet
Number of base-form verbs in a tweet
Number of past tense verbs in a tweet
Number of present participle verbs in a tweet

Number of non-3rd person singular present verbs in a tweet
Number of plural nouns in a tweet
Number of singular proper nouns in a tweet
Number of cardinal numbers in a tweet
Number of prepositions or coordinating conjunctions in a tweet
Number of adverbs in a tweet
Number of wh-adverbs in a tweet
Number of verbs of all forms in a tweet

Proposed Algorithm:

Step 1: Process along the document compilation
Step 2: Set new_word_list to empty
Step 3: Document conversion to lower case and single character elimination

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

Step 4: Document is broken up into words
Step 5: Loop through the list of words
Step 6: Identify words with negation
Step 7: If identified
Set temp_word to "not"
Step 8: End
Step 9: Else if temp_word is "not"

Add "not_" before the word

End

Step 10: Append the work to the new_word_list
Step 11: Set temp_word to Empty

End

Step 12: A new document is formed by integrating the words in the list
Step 13: End

Feature Selection:

Input: A finite list *A* = (*a1, a2, a3…an)* of tokens and a labelled sentiment T

Output: a list of optimal features

Let the initial arbitrarily seeded population be denoted by *P* and number of generations denoted by k *numGenerations k*

count 0

while *count* <*numGenerations* **do**

ProduceNextGeneration(*P*; *A*; *T*)

end

return *P*0

CASVM Model:

$$W = \{w_1, w_2, w_3, \ldots, w_n\} \quad (1)$$

Then, choose a set S, such that;

$$S \subset W \quad \wedge \quad \sum_{i=1}^{n} S_i = \tau \quad (or\ closest) \quad (2)$$

where S_i is the sentiment score of *i-th* token in subset S. Then, we introduce a vector $\vec{x}$ as a feasible solution.

$$Feasible\ Solution : \vec{x} = (x_1, x_2, x_3, \ldots, x_n) \quad (3)$$

where $x_i \in \{0, 1\}$

$$\sum^{n} w_i x_i = \tau \quad (or \quad closest) \quad (4)$$

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

This section discusses in detail the results from the proposed approach. However, it is important to first discuss and understand the type of dataset incorporated in the experiments followed by the resulted yielded.

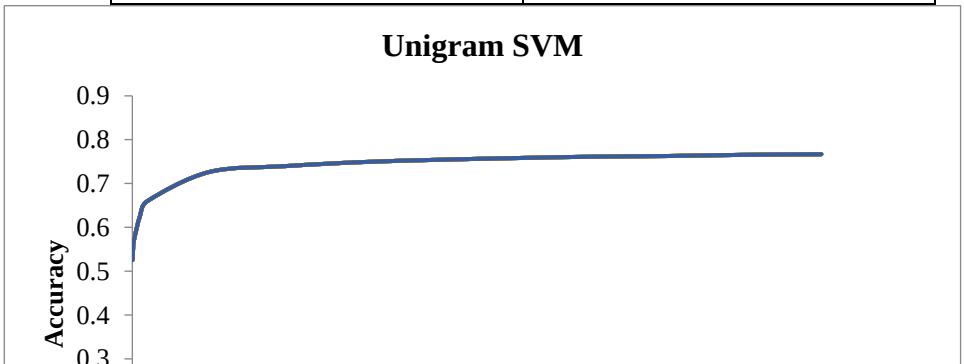
Dataset: Twitter dataset from https://data.world/data-society/twitter-user-data forms the core of the text compilation in this experiment.

There are three data sets and we present the results obtained from each of the dataset. We believe that this presentation will provide a better vantage point in understanding the implementation of the proposed approach. The final score as well as the precision measure for the results were calculated based on (F1pos + F1neg)/2 measure.

The proposed approach showed enhanced performance by surpassing the standard in all cases. And notwithstanding the presence of higher positive texts leading to unbalance in the text compilation, the results have been highly satisfactory. It can be deduced that it will be possible to identify more precise features through in-depth analysis on the use of centrality measures and graph techniques. This can be very promising for utilization in supervised learning for Sentiment Analysis.

Table 3: Unigram SVM

Data Set	Accuracy
10	0.525450571021
50	0.550521948608
100	0.569726980728
500	0.6261375803
1000	0.660421127766
5000	0.726222341185
10000	0.739806388294
15000	0.748973947181
20000	0.75426034975
25000	0.758096895075
30000	0.76130888651
35000	0.762847965739
40000	0.76556923626
45000	0.766862955032



JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

Fig. 2: Unigram SVM

The Accuracy of Unigram Data Set using SVM Algorithm is represented in Table 3 and plotted in the graph (Fig.2) with the number of Data Set arranged along X-axis and Accuracy shown along Y – Axis. Here, the abrupt increase in accuracy value is observed as the data set value goes below 10000. Subsequently, gradual increase is observed parallel to X-axis. It can be deduced that, higher data set value triggers a gradual change in accuracy level with negligible level of inaccuracy.

Table 4: SVM Bigram

Data Set	**Accuracy**
10	0.500223054961
50	0.574232690935
100	0.56437366167
500	0.632293897216
1000	0.657989828694
5000	0.725486259814
10000	0.746609564597
15000	0.756468593862
20000	0.761487330478
25000	0.767375981442
30000	0.771011777302
35000	0.77210474661
40000	0.775941291934
45000	0.777324232691

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

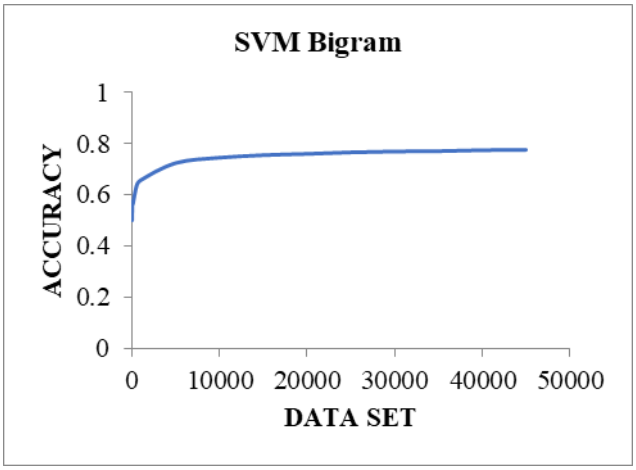


Fig. 3: SVM Bigram

The Accuracy of Bigram Data Set using SVM Algorithm is represented in Table 4 and plotted in the graph (Fig.3) with the number of Data Set arranged along X-axis and Accuracy shown along Y – Axis. Here, the abrupt increase in accuracy value is observed as the data set value goes below 10000. Subsequently, gradual increase is observed parallel to X-axis. It can be deduced that, higher data set value triggers a gradual change in accuracy level with negligible level of inaccuracy.

Table 5: Difference in accuracy value between Unigram and Bigram -SVM

Algorithm	**Accuracy**
SVM Model Unigram	76.63
SVM Model Bigram	77.73

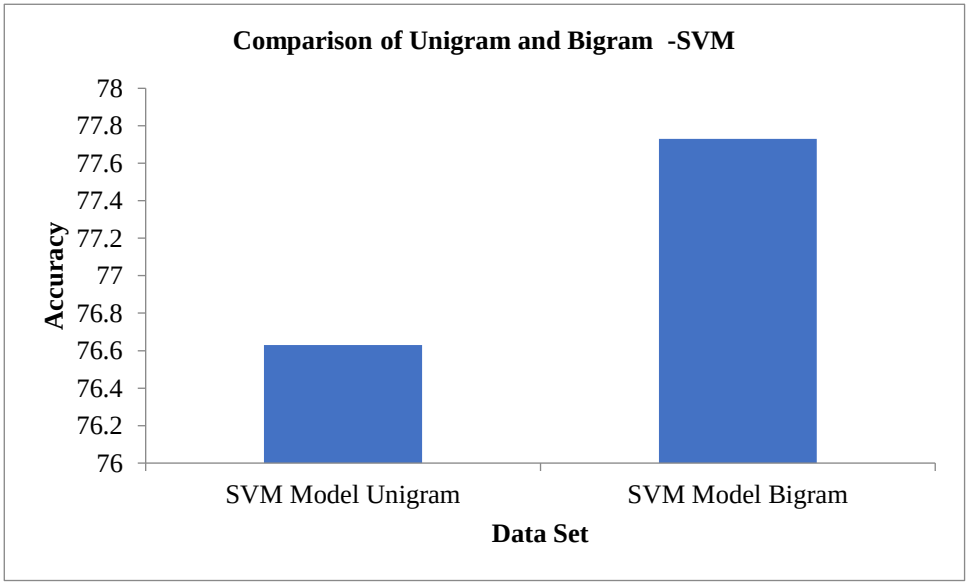


Fig. 4: Graph showing the difference in accuracy value between Unigram and Bigram -SVM

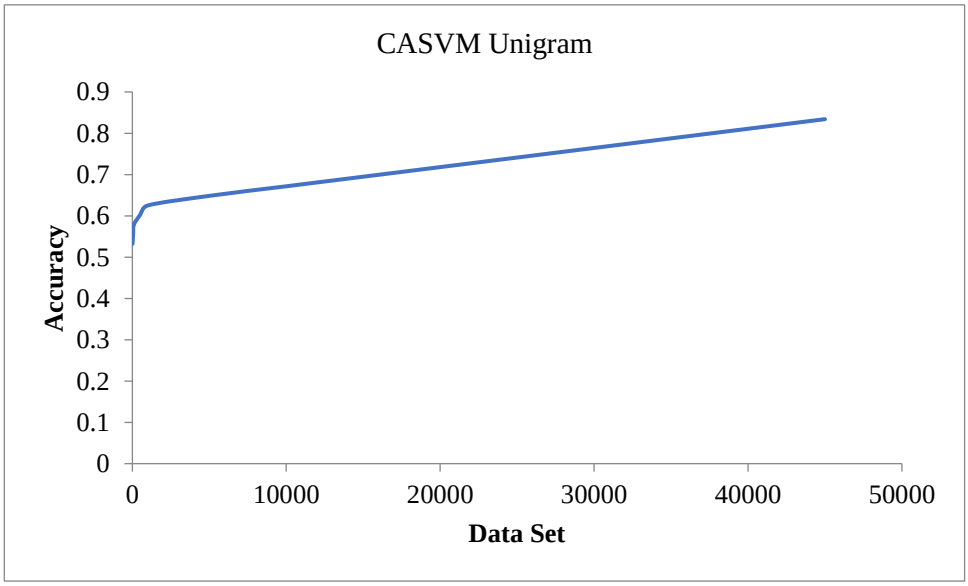
JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

The difference in accuracy value between Unigram and Bigram Data Set (SVM) is represented in Table 5 and plotted in Fig.4 with Data Set label plotted along X-axis and Accuracy plotted along Y – Axis. The two representation clearly indicates that the accuracy of bigram is slightly higher (by 1.1%). The use of trigram and beyond is believed to yield better results in terms of accuracy.

Table 6: CASVM Model Unigram

Data Set	Accuracy
10	0.532305
50	0.555515
100	0.578725
500	0.601935
1000	0.625145
5000	0.648355
10000	0.671565
15000	0.694775
20000	0.717985
25000	0.741195
30000	0.764405
35000	0.787615
40000	0.810825
45000	0.834035

The Accuracy of Unigram Data Set using CASVM Model is represented in Table 6 and plotted in the graph (Fig.5) with the number of Data Set arranged along X-axis and Accuracy shown along Y – Axis. Here, the abrupt increase in accuracy value is observed as the data set value goes below 10000. Subsequently, gradual increase (slanting) is observed. It can be deduced that, higher data set value triggers a gradual change in accuracy level with negligible level of inaccuracy.



JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

Fig. 5: CASVM Model Unigram

Table 7: CASVM Bigram

Data Set	Accuracy
10	0.523905
50	0.550815
100	0.577725
500	0.604635
1000	0.631545
5000	0.658455
10000	0.685365
15000	0.712275
20000	0.739185
25000	0.766095
30000	0.793005
35000	0.819915
40000	0.846825
45000	0.873735

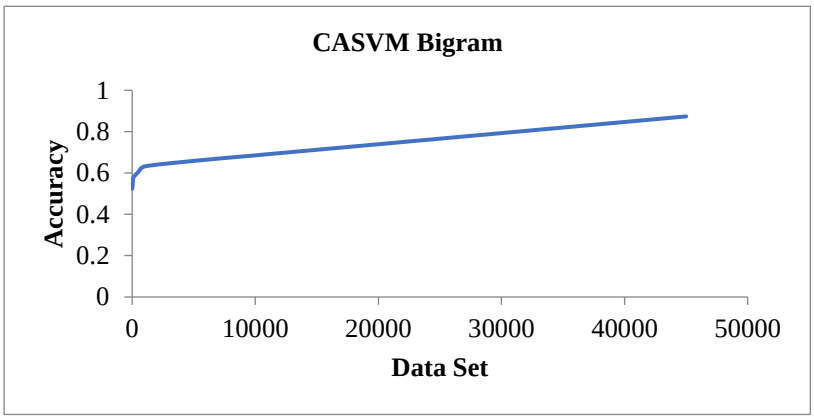


Fig. 6: CASVM Bigram

The Accuracy of Bigram Data Set using CASVM Model is represented in Table 7 and plotted in the graph (Fig.6) with the number of Data Set arranged along X-axis and Accuracy shown along Y – Axis. Here, the abrupt increase in accuracy value is observed as the data set value goes below 10000. Subsequently, gradual increase (slanting) is observed. It can be deduced that, higher data set value triggers a gradual change in accuracy level with negligible level of inaccuracy.

Table 8: Differences in accuracy value between Unigram and Bigram using CASVM

Algorithm	Accuracy
CASVM Model Unigram	83.4035

JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

CASVM Model Bigram	87.3735

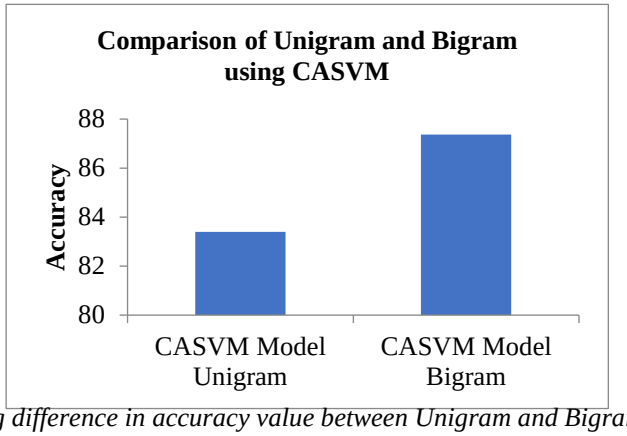
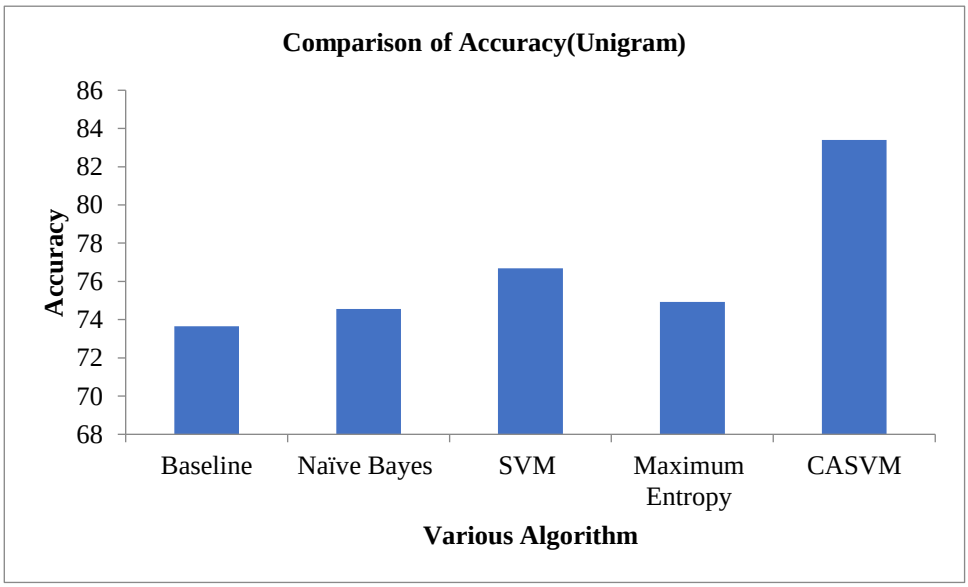


Fig. 7: Graph showing difference in accuracy value between Unigram and Bigram using CASVM

The difference in accuracy value between Unigram and Bigram Data Set (CASVM Model) is represented in Table 8 and plotted in Fig.7 with Data Set label plotted along X-axis and Accuracy plotted along Y – Axis. The two representation clearly indicates that the accuracy of bigram is slightly higher (by 3.97%). The use of trigram and beyond is believed to yield better results in terms of accuracy.

Table 9: Difference in accuracy by applying various Algorithm (For Unigram)

Method	**Accuracy(Unigram)**
Baseline	73.65
Naïve Bayes	74.56
SVM	76.68
Maximum Entropy	74.93
CASVM	83.4035



JOURNAL OF ALGEBRAIC STATISTICS
Volume 13, No. 1, 2022, p. 593-611
https://publishoa.com
ISSN: 1309-3452

Fig.8: Graph showing the difference in accuracy by applying various Algorithm (For Unigram)

The difference in accuracy value using various techniques for Unigram is represented in Table 9 and plotted in Fig.8. The Data Set label plotted along X-axis and Accuracy plotted along Y – Axis. Compared to all other methods deployed in the experiment, it is observed that the accuracy of CASVM is comparatively higher with 83.40% accuracy while all other methods are below 80%. Such results are very encouraging and motivates to delve deeper for enhanced result.

5. CONCLUSION

The proposed approach showed enhanced performance by surpassing the standard in all cases. The use of supervised learning method representing different elements using graph further proved to be very efficient for Twitter dataset. Continuous and extensive study is still required especially in those cases where the text are too small and use of colloquialisms are common (e.g. Datasets from live journal and sms). There are two distinct feature of this approach – one is the use of graph based representation and the other is the use of centrality measures. Unlike the use of conventional features, it helps to identify words that has sentimental relationships. This area of study is very promising and we aspire to delve deeper and overcome the challenges associated with the research area. Future work would focus on experimenting with other graph representations for texts. Supervised / unsupervised classification algorithms would be explored for such texts with sentence from regional languages as well as abbreviations that are not yet approved.

REFERENCES:

[1]. Dashtipour, K., Gogate, M., Gelbukh, A., & Hussain, A. (2021, June). Adopting Transition Point Technique for Persian Sentiment Analysis. In *ICOTEN*.

[2]. Mukherjee, P., Badr, Y., Doppalapudi, S., Srinivasan, S. M., Sangwan, R. S., & Sharma, R. (2021). Effect of Negation in Sentences on Sentiment Analysis and Polarity Detection. *Procedia Computer Science*, *185*, 370-379.

[3]. Kakudi Habiba, A. Emoticon Aware Aspect Based Sentiment Analysis of Online Product Review. Dutse Journal of Pure and Applied Sciences (DUJOPAS), Vol. 7 No. 2a June 2021, 166-179.

[4]. Paruchuri, H. A. R. I. S. H., Vadlamudi, S. I. D. D. H. A. R. T. H. A., Ahmed, A. A. A., Eid, W. E. S. A. M., & Donepudi, P. K. (2021). Product Reviews Sentiment Analysis using Machine Learning: A Systematic Literature Review. *Turkish Journal of Physiotherapy and Rehabilitation*, *23*(2), 2362-2368.

[5]. Ahmed, H. M., Javed Awan, M., Khan, N. S., Yasin, A., & Faisal Shehzad, H. M. (2021). Sentiment Analysis of Online Food Reviews using Big Data Analytics. *Hafiz Muhammad Ahmed, Mazhar Javed Awan, Nabeel Sabir Khan, Awais Yasin, Hafiz Muhammad Faisal Shehzad (2021) Sentiment Analysis of Online Food Reviews using Big Data Analytics. Elementary Education Online*, *20*(2), 827-836.

[6]. Kamaruddin, N. (2021). Comparative Study on Sentiment Analysis Approach for Online Shopping Review. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(3), 1358-1370.

[7]. Rashid, A., & Huang, C. Y. (2021). Sentiment Analysis on Consumer Reviews of Amazon Products. *International Journal of Computer Theory and Engineering*, *13*(2).

[8]. Ilavendhan, A. (2021). An Empirical Analysis on Various Techniques Used to Detect the Polarity of Customer Satisfaction in Sentiment Analysis. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(3), 4376-4385.

[9]. Daniel, D. A. J., & Meena, M. J. (2021). A Novel Sentiment Analysis for Amazon Data with TSA based Feature Selection. *Scalable Computing: Practice and Experience*, *22*(1), 53-66.

[10]. Jha, B. K., Sivasankari, G. G., & Venugopal, K. R. (2021). Sentiment Analysis for E-Commerce Products Using Natural Language Processing. *Annals of the Romanian Society for Cell Biology*, 166-175.

[11]. Zad, S., Heidari, M., Jones, J. H., & Uzuner, O. (2021, May). A survey on concept-level sentiment analysis techniques of textual data. In *2021 IEEE World AI IoT Congress (AIIoT)* (pp. 0285-0291). IEEE.

[12]. Dashtipour, K., Gogate, M., Adeel, A., Larijani, H., & Hussain, A. (2021). Sentiment analysis of persian movie reviews using deep learning. *Entropy*, *23*(5), 596.

[13]. Ali, M. Z., Ehsan-ul-Haq, A., Rauf, S., Javed, K., & Hussain, S. (2021). Improving Hate Speech Detection of Urdu Tweets using Sentiment Analysis. *IEEE Access*.

[14]. Bhoite, S. S., & Londhe, S. K. (2021). Aspect Based Online Sentiment Analysis Product Review and Feature Using Machine Learning. *International Research Journal on Advanced Science Hub*, *3*, 54-59.

[15]. Kalaivani, M. S., & Jayalakshmi, S. (2021). Sentiment analysis on micro-blog data using machine learning techniques-A Review. In *IOP Conference Series: Materials Science and Engineering* (Vol. 1049, No. 1, p. 012012). IOP Publishing.

[16]. Chitra, P., Karthik, T. S., Nithya, S., Poornima, J. J., Rao, J. S., Upadhyaya, M., ... & Manjunath, T. C. (2021). Sentiment analysis of product feedback using natural language processing. *Materials Today: Proceedings*.

[17]. Awajan, I., Mohamad, M., & Al-Quran, A. (2021). Sentiment Analysis Technique and Neutrosophic Set Theory for Mining and Ranking Big Data From Online Reviews. *IEEE Access*, *9*, 47338-47353.

[18]. Sindhu, C., Rajkakati, D., & Shelukar, C. (2021). Context-Based Sentiment Analysis on Amazon Product Customer Feedback Data. In *Artificial Intelligence Techniques for Advanced Computing Applications* (pp. 515-527). Springer, Singapore.