IOSR Journal of Engineering (IOSRJEN) **www.iosrjen.org**
ISSN (e): 2250-3021, ISSN (p): 2278-8719
PP 17-25

An Efficent Sentimental Analysis for Twitter Using Neural Network based on Rmsprop

S. Sathish Kumar[1], Dr. Aruchamy Rajini M.Sc., M.Phil., Ph.D[2]

[1]Research Scholar, Dept. Of. Computer Science, Hindustan College of Arts and Science Coimbatore, Tamil Nadu, India

[2]Assistant Professor, Department of computer Science, NGM College, Pollachi-624001, Tamil Nadu, India

Abstract: *With the advancement of web technology there is a tremendous volume of information present in the web for internet users. Nowadays millions of people are utilizing social network sites like Facebook, Twitter, Google Plus, etc. to express their emotions, opinion, share their perspectives and to interface with different sources. Social media generating a large volume of sentiment rich data in the form of tweets, status updates, comments, reviews, etc. Sentiment Analysis is significant in all fields especially in business to understand the conversations and discussions to identify the negative sentiments and turn poor experiences into good ones. This analysis classifies the type of users by analysis of their posted data on the social web sites. In some cases,peoples express opinions in complex ways and it is complicated to identify the opinions. It can be overcome by developing a new method for tweet classification. In this paper, thefeature set of tweet datasets is integrated into a classifier for training and predicting sentiment classification labels. Convolutional Neural Network is implemented to classifying tweet as negative or positive and Neutral sentiment. Despite the sentiment analysis-based text information, the analysis of image content is challenging task. In this work predicting the sentiments by text-based CNN network is developed. In which the image undergoes pre-processing and feature extraction process for efficient classification of the tweets. The performance evaluation of proposed method is simulated in MATLAB Tool which achieves the efficient classification accuracy.*

Keywords: *Sentiment Analysis, Tweets comments, Preprocessing, Feature Extraction and Convolutional Neural Network*

I. INTRODUCTION

Due to the development of technology, e-commerce had been advanced rapidly so million or trillions of users are acquiring and offering goods by online over internet. In which the significant part of e-commerce is sentiment analysis. Social media including Twitter, Facebook, LinkedIn are the popular free public accessible network for convey opinion on a specific item. One of the most visited social networking sites by millions of users is twitter.In that site they imparted their opinion about politics, brands and products.

With millions of tweets (feeds) day by day, there is aabundance of data exists. Due to the advanced technology twitter had attained a lot of attention with a user limit of 140 characters and become a many source for sentiment analysis and belief mining. The various incidents, sentiments to anticipating stock markets, events and trends are analyzed which offers a considerable measure of data for mining and contextual analytics.

Sentiment analysis refers to the task of identifying opinion from reviews. Sentiment analysis comprises of three stages includes document level, sentence level and entity-aspect level. In document level, the whole opinion is analyzed and the sentence in the opinion is recognized by sentence level analysis. In entity aspect level the main concentration is given on the opinion.

Basically, it is performed in two levels which differ from coarse level to fine level. In which coarse level deals with identifying the sentiment of an entire document. Fine level deals with attribute level sentiment analysis. Sentence level appears between these two levels.

Many research works are conducted by sentiment analysis which are primarily utilize in modelling and tracking public sentiments. It has a big effect on the business intelligence field. In case when the merchant needs to know why people are not purchasing his product, it turns out to be exceptionally hard to review clients who don't get it. Subsequently merchant utilizes sentiment analysis to search the web for opinions and reviews of this and competing products by using Blogs, Amazon and twitter like microblogging sites.

Previous work are establishes a efficient work but need to improve the tracking methods. Due to the advancement of the technology, this process used for mining and summarizing products reviews to solve the polarity issues for sentimental analysis.

In previous work manual tracking and extracting the data from thehuge measure of data is relatively incomprehensible. Sentiment analysis of user posts is required to help taking business decisions. It is a procedure which removes sentiments or opinions from reviews which are given by users over a particular subject, area or product in online. The sentiment can be categorize into two types are positive and negative that determine the general attitude of the people to a particular topic. The main aim is to correctly detect sentiment of tweets as more as possible.

II. LITERATURE SURVEY

Wang and Castanon (2015) analysis about the sentiment expression based on the machine learning algorithms. Emotions are generally utilized in social media to express the sentiments. The datasets are gathered from the twitter link. The sentiment polarity based on emotions is characterized and words clusteringand emotions are differentiated. Thesentiment polarity arealtered and expelled from the text based on machine learning algorithms. The performance results demonstrates the separation of emotions based sentimental analysis.

Go and L.Huang (2009) proposed a technique for sentiment analysis for twitter data by utilizing distant supervision, in which their training data consisted of tweets with emojis which filled as noisy labels. Naive Bayes, MaxEnt and Support Vector Machines (SVM) are mainly utilized for classification and the feature space includes unigrams, bigrams and POS. The performance results demonstrate that SVM achieves the efficient results and that unigram models obtain effective result for feature extraction.

Po-Wei Liang et.al (2014) developed a Twitter API to collect twitter data and classification of positive and negative tweets. The training data are collected from the public available database such as camera, movie and mobile. The algorithm is applied to filter the opinions based on tweets comments. Unigram Naive Bayes is employed to simplifying the datas by eliminating certain features based on Mutual Information. Feature extraction by Chi square and the algorithm is applied to detect the positive and negative tweet.

Xia et al proposed an efficient framework for classification of sentiment by integrating certain features. The Part-of-speech information and Word-relations are used as feature sets for classifying the positive and negative comment based on certain classifiers Naive Bayes, Maximum Entropy and Support Vector Machines. The ensemble approaches such as fixed combination, weighted combination and Meta-classifier combination for sentiment classification. The performance results show that the proposed method achieves efficient accuracy.

Davidov et al proposed asupervised sentiment classification system to separate the hashtags and smileys. The 50 Twitter tags and 15 smileys are utilized as sentiment labels. The different kinds of sentiment classification are utilized to recognize untagged sentences. The performance results of the quality of the sentiment identification are assessed by human judges. The expected results demonstrate that the proficient classification of the smileys based on cross validation.

Gamello and Garcia proposed a Naivesbayes for sentiment analysis for the recognition of the English tweets. This classifier is utilized to differentiate the positive and negative tweets. The framework performs in basic rule for polarity words for analyzing tweets. The datasets is preprocessed to remove the noise and background illumination and then fed to the naivesbayes classifier for differentiating the tweets. The experimental results show that the classifier accomplishes the effective consequence of 63% f-score for tweets polarity classification.

Jadav and Vaghela developed a machine learning method based on computer assisted techniques for efficient analysis of sentiment. The preprocessing strategy is executed to evacuate the punctuation, tags and transformation into structured form. The lexicon techniques are employed to accomplish numerical score and the feature selection utilized for sentiment analysis. The support vector machine and RBF kernel is applied for the classification which achieves the efficient classification accuracy.

Utami and Luthfi proposed a novel framework for tax comments analysis based on text mining. The input dataset is collected from the facebook and twitter for processing tax comments. Initially preprocessing is applied to the conversion of structured format and the mining of text involves the phases of text (tax) comments and the feature selection is implemented to identify the relevant features and classification by the Support Vector Machine (SVM) for analysing the tax comments.

III. PROPOSED METHODOLOGY

Sentiment Analysis in twitter is complicated because of minimum length. Generally, people shared their opinions in the form of unstructured representation includes blogs, forums etc. In this paper the analysis is made to differentiate the tweeter comments the first one is to remove the unstructured blogs and tags and in the second classify the tweets by learning algorithms. The computer assisted processing techniques such as preprocessing; feature extraction and classification algorithms are implemented to classify the tweets as positive, negative and neutral.

Image Acquisition

The input image twitters (datasets) are collected from the publicly available datasets of twitter link. Data in the form of raw tweets is acquired by using the function“tweets stream” which provides a package for simple twitter streaming.
emoticons, slang words and misspellings in tweets forced to have a preprocessing step before feature extraction.

Preprocessing

The preprocessing is mainly utilized for smoothing, background homogenization and reduce the noise in the input image to enhance the image contrast. During acquisition and transmission, the image gets corrupted by noise, so it is mandatory toreduce noises by retaining the significant image features. The main process in the preprocessing step is the removal of emoji’s and URL. A tweet contains a lot of opinions about the data which are expressed in different ways by different users. The raw data having polarity is highly susceptible to inconsistency and redundancy. The main process of preprocessing steps is

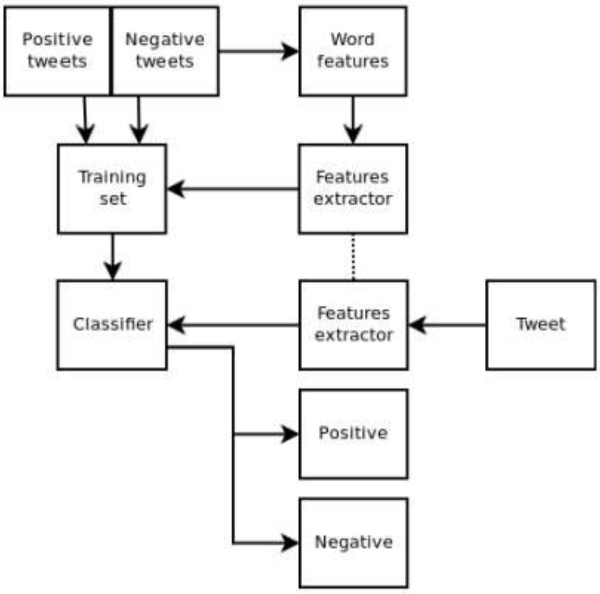


Figure 1 Process of Sentiment Analysis

- Each URLs (@username) and tags are removed
- Correct the spellings and repeated characters
- Punctuations,symbols, numbers are removed

Punctuation are removed by utilizing the constant function which eliminates characters such as Unicode punctuation or symbol classes such as connector, dash, close, quotes, open, currency, modifier and math

Feature Extraction

This process is to extract the known feature which contains some information about the image. This process is to extract the known feature which contains some details about the given input image.The various feature extraction techniques are available to gather the relevant features from text which can be extracted by two steps. In the first step twitter specific features are extracted to create normal text. Then the feature extraction is thepreprocessed dataset consists of unique properties are forwarded to the feature extraction method to extract the aspects of the processed data to differentiate the positive, negative and neutral.
Raw datasets are collected from various resources are preprocessed to extract the features. The feature extraction steps are tokenization, stop word removal, stemming and N-gram

- Tokenization is the process which is utilized to rupture a sentence into words, phrases, symbols or other meaningful tokens by eliminating the punctuation marks.
- Stop words are the common used words which appear as small for helping to select documents matching a user word examples: ABOUT, ABOVE, etc.
- Stemming is the process of transforming a word into normalized form by removing the suffixesthe word ended with ed, ion, ions, ing, le format Example: automate, automatic, automation all reduced to automat.
- Ngram

Porter Stemming

Porter Stemming is the process of removing suffixes from English words based on defined functions. This process is mainly used for information retrieval. It is represented by vector of words or certain terms. This algorithm is generated by certain assumptions without the stem dictionary to improve the English sentence.

The process to remove the suffixes from two words R1 and R2 to generate the single stem S. In some cases if there is no difference between two statement, then it is represented as " a document about R1" and a document about R2". For an instance R1= CONNECTION and R2 = CONNECTIONS are defined into single stem. But if R1 = RELATE and R2 = RELATIVITY are defined as unreasonable because R2 is related with physics. Both the cases are seems different

The rules for removing a suffix are given as

$$(Condition)\ R_1 \rightarrow R_2$$

It defines that the word ends with suffix R_1 and the stem beforeR_1 satisfies the condition andR_1 replacedby R_2

For example the word ends with EMENT

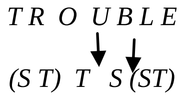
$$(m > 1)\ EMENT$$

R_1 is EMENTandR_1 is null. This condition map the REPLACEMENT to REPLAC

The Porter stemmer makes a use of a measure, *m*, of the length of a word or word part. If *S* is a sequence of one or more consonants, and *T* a sequence of one or more vowels, any word part has the form

$$[S](ST)^{P}[T]$$

which is to be read as an optional *S*, followed by *m* repetitions of *ST*, followed by an optional *T*. This defines *P*



Most of the rules for suffix removal involve leaving behind a stem whose measure exceeds some value. For example

$$(m>0)\ eed \rightarrow ee$$

The above equation defines that replace '*eed'* with '*ee'* if the stem before '*eed'* has measure *m* > 0. This porter stemmer obtains routine which is computed m for each time with removal for certain cases.

In certain case the porter stemmer are represented as *m* > 0, *m* > 1 with respect to m=1. It denotes that the significant positions in the representation of the word which moves from left to right m>0 indicates true after the first consonant followed by vowels and the point at m>1 indicates true when the first consonant following a vowel, consonant and vowels.

This algorithm is implemented not only to eliminate the suffix even when the stem is too short and the length of the stem is measured as m. This method is efficient for removingthe suffix.

Ngram

Ngram defines the combinations of neighboring words or length of letters *n* in given text. An n-gram represents the group of n words or characters (indicated as grams denotes grammar) which follows one another. N-grams can be used to predict the next word given the previous N -1 words.N-grams are widely used in data mining and word processing tasks.

For an instance If " I work in India " denotes n=4 (number of words from the sentence. Which is used to create an index of how often words follows one another.

The n-gram can be represented as

$$Ngrams_k = X-(N-1)$$

The *X* represents number of words in given sentence and *K* denotes number of grams for K sentence. N-grams are widely utilized for various word processing tasks. The language model is generated by using n-grams to develop unigrams, bigram and trigram models.

The basic operation of the n-grams which captured the language representationfrom the arithmetic view indicates the letter or word given in the input structure. If the n-gram is larger (n is higher) denotes more number of text. The length of the word can be represented based application. If n-grams are minimum the it is fail to detain the differences. Suppose n-grams are long, then it is fail to detain the particular cases. The figure shows the example for N-gram

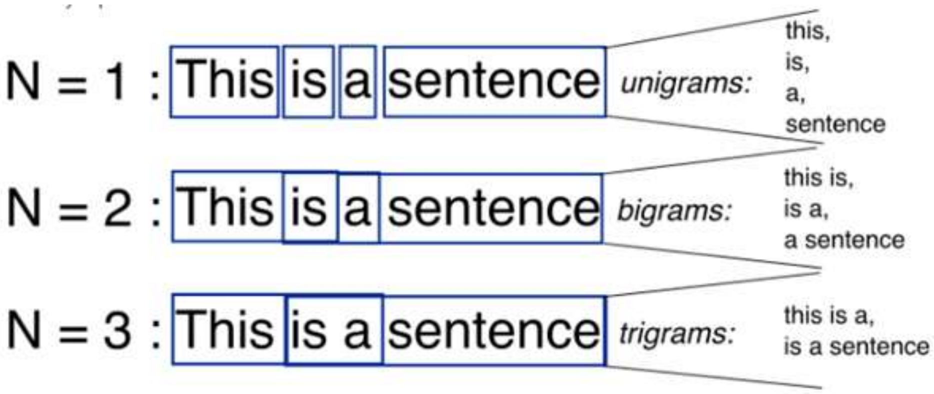


Figure 2 Example for Ngram

Classification

The extracted feature values from the Ngram are forwarded to the classifier to distinguish the positive, negative and neutral. The learning algorithms are mainly utilized to solve the classification issues. The raw datasets undergoes training by classifier to distinguish the positive, negative and neutral comments. Convolutional Neural Network is used as classifier to classify the positive, negative and neutral comments. The CNN contains neurons in 3dimension which are spatial dimensionality of the input layer and depth.

CNN consists of three layers such as convolutional layer, pooling layer and fully connected layer. CNN structure is formed by stacking these layers.

Pooling layer –It minimizes the dimensionality of the representation and also minimize the number of parameters and the computationalcomplexity of the mode

Fully-connected layer – This layer includes certain neurons are directly connected tothe neurons in the neigh boringlayers, without connecting to any layers.

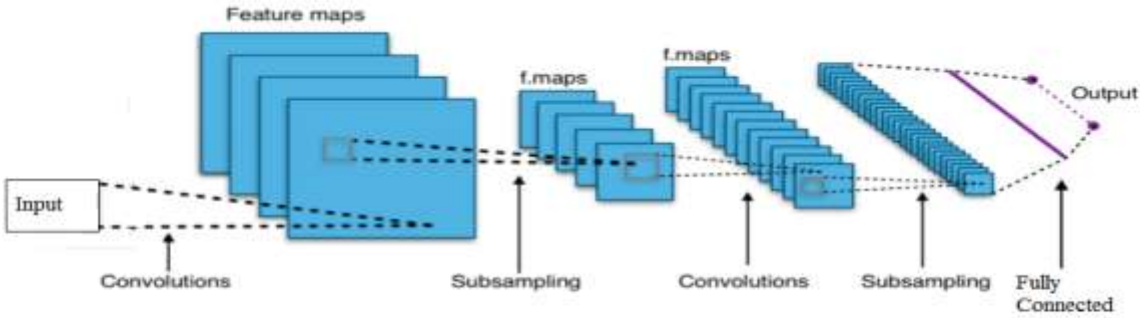


Figure 3 Structure of CNN

The main function of CNN is

- The input layer maintain the input image pixel value
- The convolutional layer identify the neurons output which are linked with the input by weights. The activation functions which are sigmoid to the output are generated by the previous layer.
- The pooling layer operates down sampling with the spatial dimensionality of the input layer by reducing number of parameters
- The main function of the fully connected layer to generate the activation function used for classifying the registered users and non-registered users.

When the input data forwards to the convolutional layer, then the each layer convolves the filter across the across the spatial dimensionality of the input to generate the 2D activation map.CNN comprises of neurons that self-optimise by learning. Every neuron will receive input and operated based on scalar product of non-linear function. For CNN operation, the input and the output layer are expressing a single activation function (weight). The final layer performs loss function linked with the activation weight.

IV. EXPERIMENTAL ANALYSIS

The performance of the proposed CNN method is simulated in MATLAB Tool under windows environment. The evaluation of the proposed method is measured by certain performance metrics such as accuracy, sensitivity, specificity and F-measure which is obtained by confusion matrix.

- Accuracy measures the exact classification of tweets comments

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FN} + \text{FP}}$$

- Specificity measures the negative classification of the tweet comments when the condition is actually not present. It is recognize as false-positive rate

$$\text{Specificity} = \frac{\text{TN}}{\text{TN} + \text{FP}}$$

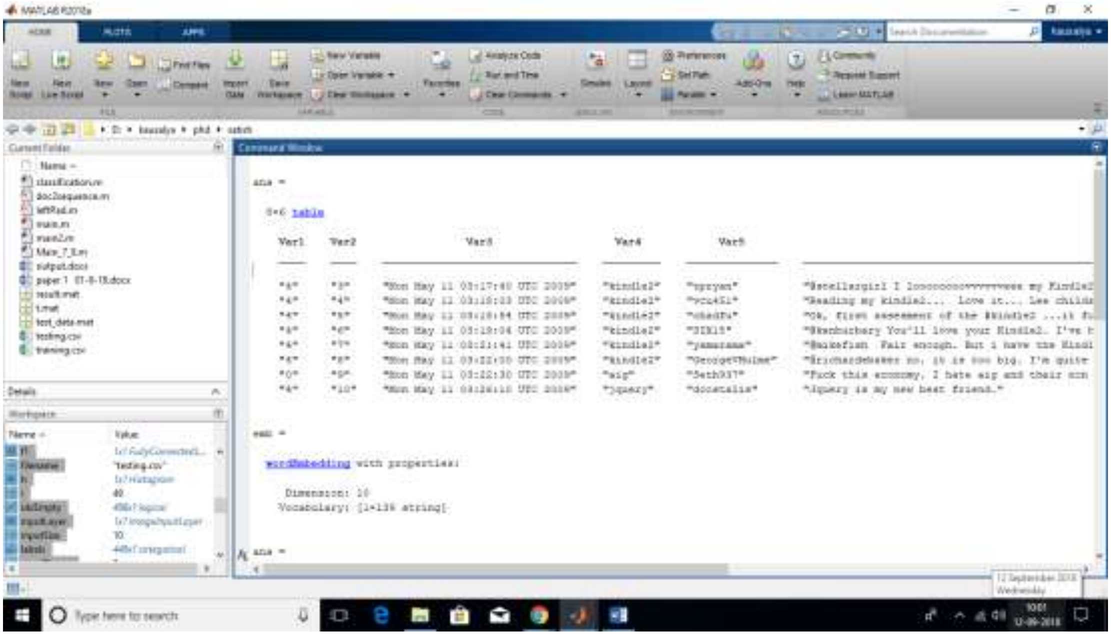
- Sensitivity measures the positive classification of tweet comments when the condition is actually present. It is represented as false-negative rate,

$$\text{Sensitivity} = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

Where True Positive (TP) denotes positive result of tweet classification
True Negative (TN) denotes negative result of tweet classification
False Positive (FP) shows the positive result for negative tweets classification
False Negative (FN) shows the negative result for positive tweetsclassification

The data sets are collected from the publicly available twitter datasets from Stanford University. The labelled datasets are analyzed by preprocessing and feature extraction techniques. Preprocessing is employed to raw datasets and deep learning networks are applied to train the dataset with the feature vectors from the large dataset to classify the twitter datasets as positive, negative and neutral. The performance of the proposed method is compared with the Naive Bayes Algorithm with respect to certain performance metric to evaluate the classification accuracy.

The raw data of the tweets of sentiment analysis is shown in the below figure 4.



```
ans =

  8×6 table

    Var1    Var2              Var3                   Var4           Var5
    ____    ____    ______________________________    __________    ______________    __________________________________________

    "4"     "3"     "Mon May 11 03:17:40 UTC 2009"    "kindle2"     "tpryan"          "@stellargirl I loooooooovvvvvveee my Kindle2
    "4"     "4"     "Mon May 11 03:18:03 UTC 2009"    "kindle2"     "vcu451"          "Reading my kindle2...  Love it... Lee childs
    "4"     "5"     "Mon May 11 03:18:54 UTC 2009"    "kindle2"     "chadfu"          "Ok, first assesment of the #kindle2 ...it fu
    "4"     "6"     "Mon May 11 03:19:04 UTC 2009"    "kindle2"     "SIX15"           "@kenburbary You'll love your Kindle2. I've h
    "4"     "7"     "Mon May 11 03:21:41 UTC 2009"    "kindle2"     "yamarama"        "@mikefish  Fair enough. But i have the Kindl
    "4"     "8"     "Mon May 11 03:22:00 UTC 2009"    "kindle2"     "GeorgeVHulme"    "@richardebaker no. it is too big. I'm quite
    "0"     "9"     "Mon May 11 03:22:30 UTC 2009"    "aig"         "Seth937"         "Fuck this economy. I hate aig and their non
    "4"     "10"    "Mon May 11 03:26:10 UTC 2009"    "jquery"      "dcostalis"       "Jquery is my new best friend."


emb =

  wordEmbedding with properties:

     Dimension: 10
    Vocabulary: [1×138 string]


fx ans =
```

Figure 4 Raw data of tweets

The distribution of tweets as positive (4), negative (0) and neutral (2) is shown in below figure 5.

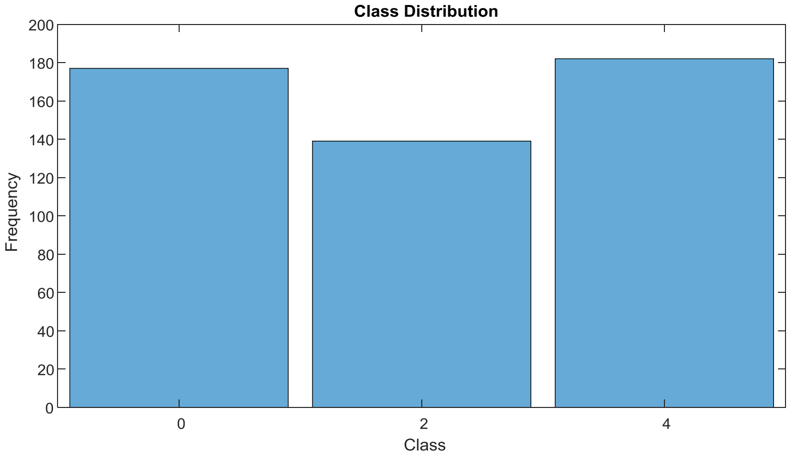


Figure 5 Distribution of Tweets

To extract the feature from the tweets the stemmed words are converted into tokenized documents. The distribution of tokenized documents are shown in the figure 6.

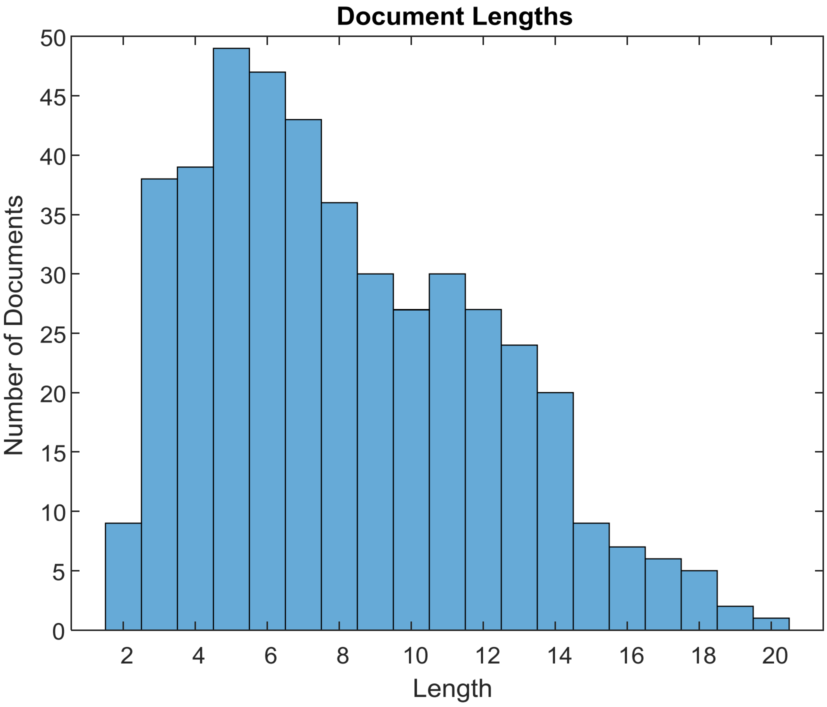
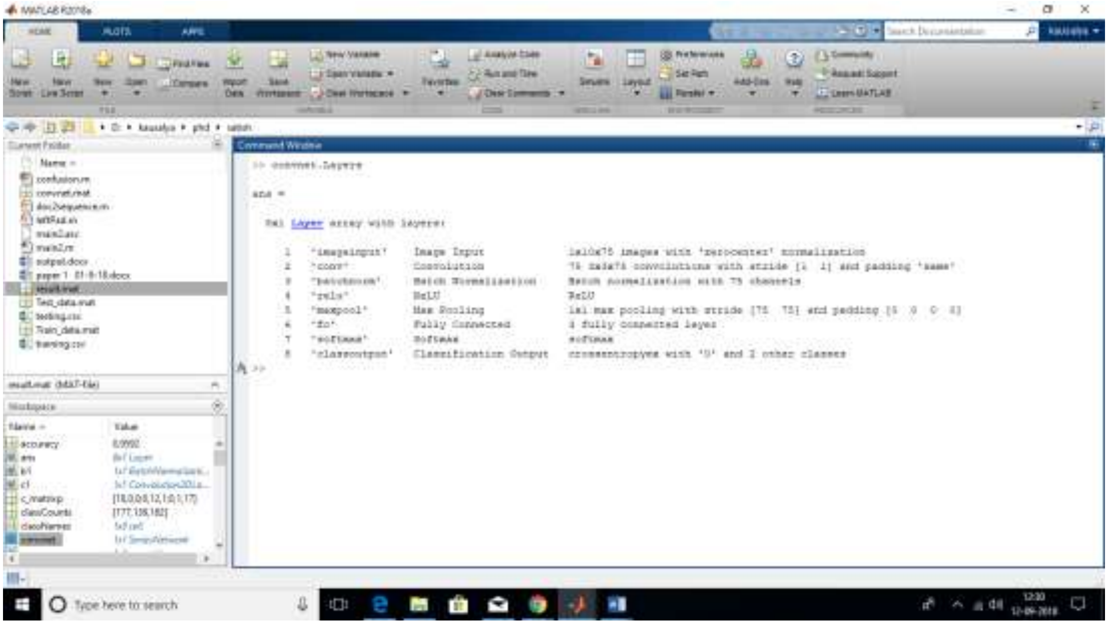


Figure 6 Distribution of tokenized documents in the tweets

This tokenized documentis converted into sequence for the training of deep learning convolutional neural network. The convolution neural network parameters are shown in figure 7.



```
>> convnet.Layers

ans =

  8x1 Layer array with layers:

     1   'imageinput'    Image Input             1x10x75 images with 'zerocenter' normalization
     2   'conv'          Convolution             75 1x1x75 convolutions with stride [1  1] and padding 'same'
     3   'batchnorm'     Batch Normalization     Batch normalization with 75 channels
     4   'relu'          ReLU                    ReLU
     5   'maxpool'       Max Pooling             1x1 max pooling with stride [75  75] and padding [0  0  0  0]
     6   'fc'            Fully Connected         3 fully connected layer
     7   'softmax'       Softmax                 softmax
     8   'classoutput'   Classification Output   crossentropyex with '0' and 2 other classes
>>
```

Figure 7 CNN properties

The dataset is classified as training and testing data. The training data is trained through CNN with the above properties and the network is saved for testing purpose. The evaluation of the proposed method is based on the mentioned metrics and the accuracy of the proposed method is compared with the existing method.

Method/Metric	Accuracy
Naïve Bayes	74.56
Support Vector Machine	76.68
Maximum Entropy	74.93
Proposed CNN	95.92

Table 1 Comparative analysis of proposed Method

Method/Metric	Proposed CNN
Accuracy	95.92
Sensitivity	95.58
Specificity	98.00

Table 2 analysis of the proposed Method

From the above tables it is observed that the proposed method outperformed the existing method as well as it produces high accuracy rate as compared to the other methods.

V. CONCLUSION

In this paper a novel framework is developed for sentimental analysis for the classification of the twitter comments. Initially twitter datasets are preprocessed to remove the hash tags, punctuations and to enhance the tweet observations. Feature extraction is employed to process the stemming and n-gram. Convolutional neural network is employed to classify the tweets such as positive, negative and neutral. The performance results show that the proposed method achieves the efficient result with classification accuracy.

REFERENCES

[1]. Liang, P.W. and Dai, B.R., 2013, June. Opinion mining on social media data. In Mobile Data Management (MDM), 2013 IEEE 14th International Conference on (Vol. 2, pp. 91-96). IEEE.

[2]. Go, A., Bhayani, R. and Huang, L., 2009. Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford, 1(12).

[3]. Xia, R., Zong, C. and Li, S., 2011. Ensemble of feature sets and classification algorithms for sentiment classification. Information Sciences, 181(6), pp.1138-1152.

[4]. Davidov, D., Tsur, O. and Rappoport, A., 2010, August. Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of the 23rd international conference on computational linguistics: posters (pp. 241-249). Association for Computational Linguistics.

[5]. Wang, H. and Castanon, J.A., 2015. Sentiment expression via emoticons on social media. arXiv preprint arXiv:1511.02556.

[6]. Gamallo, P. and Garcia, M., 2014. Citius: A naive-bayes strategy for sentiment analysis on english tweets. In Proceedings of the 8th international Workshop on Semantic Evaluation (SemEval 2014) (pp. 171-175).

[7]. Jadav, B.M. and Vaghela, V.B., 2016. Sentiment analysis using support vector machine based on feature selection and semantic analysis. International Journal of Computer Applications, 146(13).

[8]. Utami, E. and Luthfi, E.T., Text Mining Based on Tax Comments as Big Data Analysis Using SVM and Feature Selection.

[9]. Kharde, V. and Sonawane, P., 2016. Sentiment analysis of twitter data: a survey of techniques. arXiv preprint arXiv:1601.06971.

[10]. Pak, A. and Paroubek, P., 2010, May. Twitter as a corpus for sentiment analysis and opinion mining. In LREc (Vol. 10, No. 2010, pp. 1320-1326).

[11]. Bifet, A. and Frank, E., 2010, October. Sentiment knowledge discovery in twitter streaming data. In International conference on discovery science (pp. 1-15). Springer, Berlin, Heidelberg.

[12]. O'Shea, K. and Nash, R., 2015. An introduction to convolutional neural networks. arXiv preprint arXiv:1511.08458.

[13]. Fernández Anta, A., NúñezChiroque, L., Morere, P. and Santos Méndez, A., 2013. Sentiment analysis and topic detection of Spanish tweets: A comparative study of NLP techniques.

[14]. Baccianella, S., Esuli, A. and Sebastiani, F., 2010, May. Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Lrec (Vol. 10, No. 2010, pp. 2200-2204).

[15]. Vinodhini, G. and Chandrasekaran, R.M., 2012. Sentiment analysis and opinion mining: a survey. International Journal, 2(6), pp.282-292.

[16]. Davidov, D., Tsur, O. and Rappoport, A., 2010, August. Enhanced sentiment learning using twitter hashtags and smileys. In Proceedings of the 23rd international conference on computational linguistics: posters (pp. 241-249). Association for Computational Linguistics.

[17]. Medhat, W., Hassan, A. and Korashy, H., 2014. Sentiment analysis algorithms and applications: A survey. Ain Shams Engineering Journal, 5(4), pp.1093-1113.

[18]. Cambria, E., Havasi, C. and Hussain, A., 2012, May. SenticNet 2: A Semantic and Affective Resource for Opinion Mining and Sentiment Analysis. In FLAIRS conference (pp. 202-207).

[19]. Dedhia, C. and Ramteke, J., 2017, January. Ensemble model for Twitter sentiment analysis. In Inventive Systems and Control (ICISC), 2017 International Conference on (pp. 1-5). IEEE.

[20]. Wang, S., Li, D., Song, X., Wei, Y. and Li, H., 2011. A feature selection method based on improved fisher's discriminant ratio for text sentiment classification. Expert Systems with Applications, 38(7), pp.8696-8702.

[21]. Ahmad, M. and Aftab, S., 2017. Analyzing the Performance of SVM for Polarity Detection with Different Datasets. International Journal of Modern Education and Computer Science, 9(10), p.29.

[22]. Fatima, S. and Srinivasu, B., 2017. Text Document categorization using support vector machine. International Research Journal ofEngineering and Technology (IRJET), 4(2), pp.141-147.