RT&A, Special Issue № 1 (60)
Volume 16, Janyary 2021

SENTIMENT ANALYSIS OF PRODUCT REVIEWS USING SUPERVISED LEARNING

Arkesha Shah
•
GCET, Vidyanagar
arkeshashah1000@gmail.com

Abstract

Today, Online Reviews are global communications among consumers and E-commerce businesses. When Somebody wants to make a purchase online, they read the reviews and comments that many people have written about the product. Only after customers decide whether to buy the product or not. Based on that, the Success of any Products directly depends on its Customer. Customer Likes Products It's Success. if not, then Company needs to improve it by making some changes in it. For that, the need is to analyze the customers' written reviews and find the sentiment from that. the task of Classifying the comments and the reviews in positive or negative is known as sentiment analysis.in this paper, A Standard dataset reviews have been classified into positive and negative sentiments using Sentiment Analysis. For that different Machine Learning and Deep Learning Technique is used and also Compared the performance of word2vec-CNN Model with FastText-CNN Model on amazon unlocked mobile phone Dataset.

Keywords: Sentiment Analysis, SVM, Naïve Byes, FastText word-Embedding, CNN Model

I. Introduction

Sentiment analysis is one of the fastest-growing research areas, which helps customers to make better-informed purchase decisions from the web and social media. It also provides organizations the ability to measure the impact of their social marketing strategies by identifying the public sentiments towards the product [2]. Sentiment Analysis is a computational study to extract subjective information from the text[5].

Nowadays shopping for Mobile phones from online sites like Amazon and Flipkart Increases. With an ever-increasing demand for smartphones, the mobile phone market is expanding. With such a boom in the smart-phone industry, there is a need to extract meaningful information from the review of the brand and the model of phone [3]. There are numerous brands present in the market, out of which the brand phone Model select is very confusing for the customer. They read Various Reviews available on E-Commerce sites which act as a guiding tool for the customers and help to make a decision which brand phone model should buy. From the manufacturer's point of view, helpful online reviews mining customer requirements improving a product or designing a new product [5]. Sentiment analysis is a classification problem. This Problem solved using the following approach.

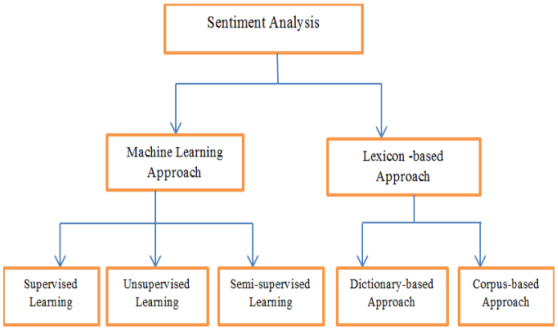


Fig.1: *Sentiment Analysis*

Machine Learning is a scientific discipline that explores the construction and study of algorithms that can learn from data [1]. In Machine learning, Model Learns by giving input as a history of past data based on that Model is capable of predictions or make decisions. The machine learning model is a Mathematical Model. Machine Learning Model classified into supervised Learning, unsupervised learning, Reinforcement Learning.in supervised Learning Model Classifies the data based on Known Labels that are given.in unsupervised Learning Model Classifies the data based on Similar Characteristics in Data. In Reinforcement Learning Model learns by trial and error using feedback from its actions and experiences.

II. LITERATURE REVIEW

There is a rapid growth of E-commerce Shopping, from that online Customer reviews playing an Important role in sales of a product. Research[5], focus on data analysis of mobile reviews data set. They find the relationship between different features. and Perform sentiment classification in positive and Negative which is helpful to consumers and the manufacturers. they take the unstructured Data to perform Pre-processing for sentiment analysis of mobile phone reviews. They used the Support Vector Machine (SVM) Machine Learning Model and got 84.87% Accuracy after cross-validation. In Research[1], They Used the Support Vector Machine (SVM) Classification Technique to classify text in positive and Negative from the smartphone product Reviews. the model Performance Measured using Precision, Recall, F-measure. the predicted Model obtained High Accuracy.

In Research[3] They performed Classification of Mobile phone Reviews in positive and Negative Sentiment. which are taken from amazon.com. Reviews are classified using different Machine Learning models such as Naïve Bayes, Support Vector Machine (SVM), and Decision Tree. These three Classifiers had cross-validated to find the best classifier from them. They found Support Vector Machine (SVM) as the best Classifier as they obtained Higher accuracy 81.87%.in Research[4] They performed sentiment analysis of the smartphone Reviews. they used the Machine learning approach to classify the reviews in positive and Negative. Machine learning Techniques Like Naïve Byes and Support Vector Machine (SVM) used. And found the SVM Technique reliable for mining of data.

In Research[6], Proposed Deep Learning approach for sentiment analysis of smartphone reviews using Twitter Corpus. For the Deep Learning approach, They Used Convolutional Neural Networks. (CNN). They also used Machine Learning Models like Naïve Byes and Support Vector Machine (SVM). They found that the deep learning technique is efficient than Machine Learning techniques.In Research [7], they Proposed a sentiment analysis approach through Deep Learning Technique. They Classified the Movie Reviews using Different Supervised Machine Learning Models and after the Compared the results with Used Convolutional Neural Networks (CNN) a

Deep Learning Technique. They found that deep learning model CNN gives reliable performance.

In Research[8] Proposed Deep Learning Approach for Sentiment Analysis that Applied on Hotel's Reviews. They applied Convolutional Neural Networks(CNN) on Hotel Reviews and Classified in Positive, Negative, and Neutral Sentiments. They Compared the Results of the CNN model with Other Models and found that Convolutional Neural Networks gives Better Performance.In Research[9] They Performed Sentiment Analysis of Restaurant Reviews given by the customers. They used Different Supervised Machine Learning Models to classify the Reviews. After that, They compared their Performances. They found that SVM Model resulted in the Highest accuracy of 94.56%.

In Research[11], they used Machine learning Classifiers such as Decision tree, Naïve Byes, KNN for prediction of sentiment. They compared the results of classifiers and find the best classifier from that they found naïve byes is the best classifier. In Research[12], They described the process of sentiment analysis on twitter data using machine learning for that they used Decision tree and Naïve Byes they found that the Decision tree performed well. In Research[13], They used Naïve byes and SVM Classifiers for sentiment analysis, compared the result, and found that Naïve byes performed well than SVM.In Research[14], They used unigram, bi-gram approach with Machine learning Classifiers such as Naive Bayes, maximum entropy classification, SVM. they measured the performance of classifiers using precision, recall, and F-measure. Found that bi-grams gave a good result. In Research[15], They used different supervised learning methods on a large scale amazon dataset for classifying their sentiment and get good accuracy.

In Research[2], Proposed Deep Learning approach for sentiment analysis of Product Reviews of Mobile Phones extracted from the Amazon. They used word2vec word-embedding Technique with CNN Model to classify the Mobile Phone Reviews in Positive and Negative sentiment. They used Google's word2vec pre-trained Model to Convert Text to word vectors of 300 Dimensional. if words that are not available in the pre-trained model they created the word vectors by randomly taking the Values between -0.25 to +0.25.CNN Model is Trained and Predict the Sentiment of new Mobile phone Reviews. They also used different Vectorization Techniques with Machine Learning models like Naïve Byes. A comparison is done and found that the Proposed approach has better accuracy of 0.9123 than the Machine learning model. they implemented CNN Model using deep learning Library Keras and Tensorflow.

III. Proposed System for Sentiment Analysis

The Steps Followed for Sentiment analysis of Mobile Phone Reviews are Shown in the figure as follows:

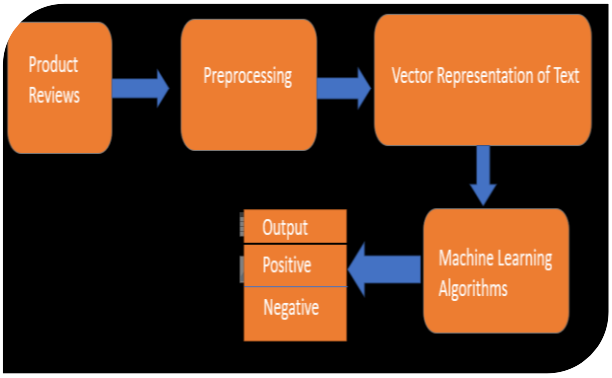


Fig.2: *Proposed Model*

A. Data Collection

Dataset used in this work is Smartphone Product Reviews released by amazon website [https://www.kaggle.com/ Prompt Cloud HQ/ amazon reviews-unlocked-mobile-phones/data] which is static Dataset [2]. In this Dataset, there are a total of 4,00,000 reviews and 6 Columns. Customers gave 1 to 5 Rating for their Reviews so based on that in this project Created Separate Column name as Sentiment in which Reviews above 3 Rating have assigned Value 1 that represents positive Sentiment and Value 0 represent negative Sentiment. The idea is to evaluate the performance of Machine Learning algorithms like SVM and Naïve Byes as well as the FastText-CNN Deep learning Model on Amazon product reviews (mobile phone reviews).

B. Data Pre-Processing

It is a necessary process before Data feed to Machine Learning Model. It cleans the unnecessary Data so that the Machine will Learn batter and also performed well. It is the process that converts raw data into to clean dataset. In this work, Cleaning data by removing rows having 'null' values. The final outcome of this project is positive and negative sentiment. there is no requirement of neutral sentiment. So, it is removed from the dataset. The steps that are carried out in pre-processing of data are as follows-

- Remove Punctuation: -It removes special Characters that do not contain any meaningful information.so it is removed from data.
- Remove Stop words: Stop words like a, an, the, he, she, was, etc. has importance in English grammar. but in Machine learning data there is no need for that because these stop words not provide any meaningful information Regarding Sentiments.
- Case Conversion: All texts converted into either Lower Case or Upper Case. So that any case sensitive issue not occur.
- Perform Tokenization: This divides the data into small units.in this case, divides the sentences into individual words, it gives structure to unstructured data. which means it assigns a list of integers to the sequence of unique words.
- Lemmatizing: It converts the word into its root word. Root words are words that have no prefix and suffix. For that, it used a dictionary-based approach.eg. pleased => please. In this work, I have used the wordnet library. For Text pre-processing, I have used the NLTK Toolkit of Python Language.

C. Vectorization

In this process, Text is converted into Numeric Form so that the Machine learning model can understand, the machine learning model is a mathematical model. It only understands the numerical information. For that different Techniques are used. In this work, I have used Techniques that are listed below:

- Term Frequency – Inverse Document Frequency (TF-IDF):
  It measures how important a word in a given document with a collection of documents. TF measures the occurrences of the word in a particular document. And IDF measures informativeness of word in a collection of documents. In this work, we have used TF-IDF with Bi-grams which means a sequence of 2 words instead of a single word.it creates the feature vector of two words.

TF calculated by:

TF = (Number of time the word occurs in the Text) / (Total Number of Words in Text)

IDF Calculated By:

IDF = (Total Number of documents / Number of documents with word t in it)

TF-IDF Calculated By:

TF-IDF = TF * IDF

- Bag of Words model using Count- vectorizer:

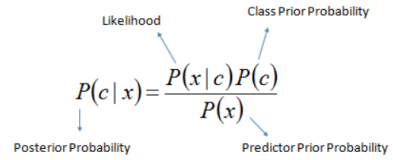
It is a vectorization technique. It creates a dictionary of unique words from the documents. Compare the dictionary with each document. if word present in the dictionary then it gives 1 otherwise 0 in this way, it makes Text into fixed-length Vectors by counting Occurrences of each word appears. In this work, to implement the Bag of words model We have used the Countvectorizer Built-in Function which is provided in the Scikit Python Library.

- FastText Word - Embedding:

Word-Embedding is a vector Representation of word. It represents the word in vector format.it is used to find the similarity between the words. FastText Model is an extension of the word2vec model. Instead of learning vectors for Individual words directly, FastText utilizing internal structures of the words and represents each word as an n-gram of characters. for example, take the word, "artificial" with n=3, the FastText representation of this word is <ar, art, rti, tif, ifi, fic, ici, ial, al> This helps capture the meaning of shorter words. So even if a word wasn't seen during training, it can be broken down into n-grams to get its embeddings. In this work, Created The word Vectors using FastText Pretrained Model wiki.simple.bin File. Which has 300 dimensions for each word.

D. Models that are used for Sentiment Classification are as follows:

- Naïve Byes: It is a classification technique based on Bayes' Theorem with an assumption of independence among predictors. This model is easy to build and useful for very large data sets. It works on Bayes theorem of probability to predict the class of unknown data sets.



$$P(c|X) = P(x_1|c) \times P(x_2|c) \times \cdots \times P(x_n|c) \times P(c)$$

Here,

- ***P*(*c*|*x*) is the posterior probability of *class* (*target*) given *predictor* (*attribute*).**
- ***P*(*c*) is the prior probability of *class*.**
- ***P*(*x*|*c*) is the likelihood which is the probability of *predictor* given *class*.**
- ***P*(*x*) is the prior probability of *predictor*.**

But, in this work, if there are n independent features which represented as vector y=(y1,y2,…,yn) then Naïve byes use the following Equation and Classify the Results:

$$p(Lk \mid y1,y2,\ldots,yn) = p(Lk) \prod n\ i=1\ p(yi \mid Lk),$$

Lk represents kth number class.

- Support Vector Machine (SVM) : SVM (Support Vector Machine) is a supervised machine learning algorithm that is mainly used to classify data into different classes. SVM makes use of a hyperplane to separate various classes. SVM can be used for both Linear Data and Non-linear Data. In this work, We have used linear SVM to classify the Data because Data is linearly Separable, like whether Positive or negative.

- Convolutional Neural network (CNN): Convolutional neural network (ConvNets or CNNs) is Used for image classifications, object detections, recognition, text classification. There are some Components in CNN which are the following:

  i. Input Layer: Reviews are converted in to feature vectors from the available word vectors of the fasttext pre-trained model. and target column which is sentiment 1 and 0, as a vector. These two are given to input to CNN Model.
  ii. Convolution Layer: Convolution is the first layer to extract features from given input.it is a sliding window that slides over the list of word embedding in sequence and captures the meaningful information regarding sentiments. applied multiple filters on the given set of input so that learn more about the features.in this work, We have used a 1D convolution layer because it slides over only in one dimension.
  iii. Embedded layer: It is defined as the primary hidden layer of a network system, it is instated with arbitrary weights and will learn an embedding for all of the words in the training dataset[2].the main work of the embedding layer is to convert a list of word indexes into a dense vector of fixed length.
  iv. Pooling Layer: Reduce the computational complexity, CNNs use pooling to reduce the size of the output from one layer to the next in the network.
  v. Fully Connected Layer: Fully Connected Layer is a fully connected neural network. These vectors are fed it into a fully connected layer like a neural network for Training.

here, each layer connected to other layer. final output is the probability of each label class.

vi. Dropout: to avoid overfitting of the model dropout layer is used.

vii. Output Layer: In this work, the sigmoid activation function is used that transform value between 0 and 1. which is good for binary classification problem. This layer returns the classification result.

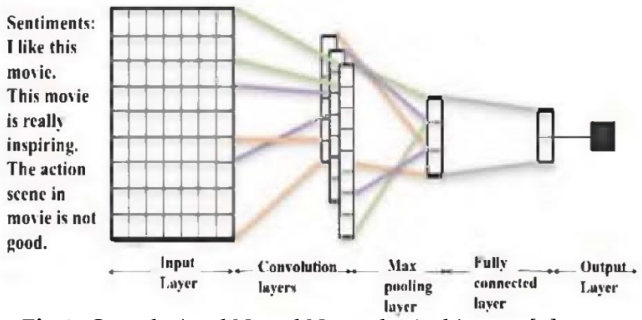


Fig.3: *Convolutional Neural Networks Architecture[7].*

IV. EXPERIMENTS AND EVALUATION

A. Experiment:

In this work, the input is Customer Reviews and output is the Sentiment column which derived from the Rating Column. Data Pre-Processing applied to Reviews. After that Feature Vectors are created using vectorization techniques, in this work feature vectors first created using TF-IDF Bi-grams, Countvectorizer. created feature vectors first applied as input to Machine Learning Models like SVM and Naïve Byes. From the Dataset 80% of data is used for Training Purpose and the remaining 20% data is used for Testing the model. after training the machine learning model they predict the sentiment of reviews. For FastText word-embedding, FastText pre-trained model wiki.simple.bin is used to create the word vector of Text Reviews. which generates a 300-dimensional word vector.in the dataset different lengths of reviews are present.so to make the equal length of reviews zero-padding used, the maximum sequence length used is 800.so review matrix size becomes 800* 300. The Review matrix and target label are given input as CNN deep learning model. CNN Model is implemented using deep learning Framework Keras and Tensorflow. after a trained CNN Model it is able to predict unseen data.

B. Result:

To find the best classifier different evaluation metrics used which are confusion Matrix, Precision, Recall, F1-score, etc.

- Confusion Matrix: it is the visualization of Machine learning algorithms on Test data. It defines how accurately Machine Learning Model Classifies the unseen Data.

- TP (True positives): No. of positive reviews that are correctly Predicted by the Model. Which are the same in actual data in the dataset.

- TN(True negatives): No. of negative reviews that are correctly predicted by the Model same as actual Data. FP(False positives): No. of negative reviews that are incorrectly predicted as positive by the Model.

- FN(False negatives): No. of positive reviews that are incorrectly predicted as negative by the Model.

The figure shows the confusion matrix of FastText Word-Embedding with the CNN Model. When the model trained with 4608 reviews out of 5761 reviews and tested with 1153 reviews, the model correctly classified 1091 reviews and the accuracy achieved to 94.62%

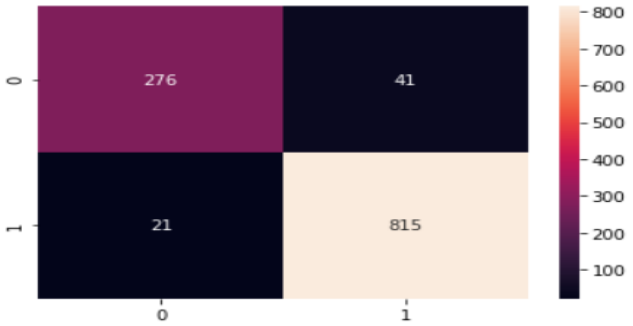


Fig.4: Confusion Matrix of CNN Classifier

Accuracy Calculated by:

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

[2]

ROC Curve:
It is a probability curve for different classes. It tells how well the model Separates given target classes. The following is the ROC Curve for CNN Model.

True-positive rate(TPR) is also called as sensitivity, recall(ratio of no. of correctly identified positives, and total no. of positives)[2].

$$Recall = TPR = \frac{TP}{(TP + FN)}$$

[2]

False-positive rate(FPR) is also called as the fall-out,1-specificity (ratio of no. of incorrectly identified negatives and total no. of negatives)[2].

$$FPR = \frac{FP}{(FP + TN)}$$

[2]

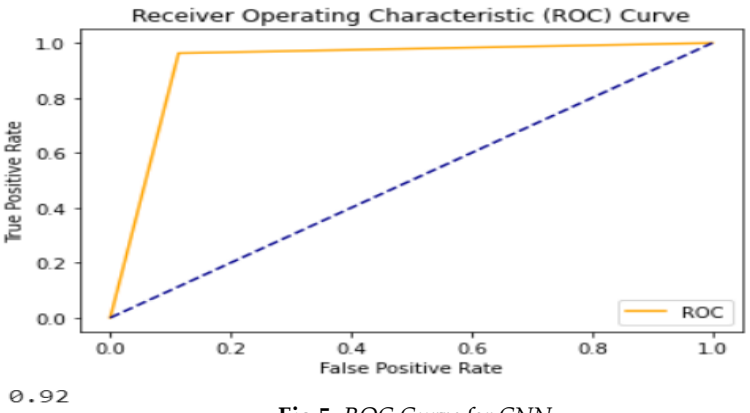


Fig.5: *ROC Curve for CNN*

In this work Performance of Machine learning Models like Naïve Byes and SVM compared with CNN Deep learning Model. and also compare the result of the FastText-CNN Model with the word2vec-CNN Model[2]. The accuracy of these models is listed below.

NO.	Model	Accuracy	Precision	Recall	F1-Score	Correctly classified
1.	TF-IDF Bi-grams + SVM	0.92107	0.91	0.98	0.95	1062
2.	TF-IDF Bi-grams + Naïve Byes	0.9071	0.89	1.00	0.94	1046
3.	Countvectorizer + SVM	0.9358	0.96	0.95	0.96	1079
4.	Countvectorizer + Naïve Byes	0.9176	0.95	0.94	0.94	1058
5.	FastText + CNN	0.9462	0.95	0.97	0.96	1091

Fig.6: *Machine Learning Models Results*

From the table, we can see that, FastText with CNN Model achieved the highest accuracy than the other Models. We have also Performed a comparison with Research Paper[2].

Model	Accuracy	No of Reviews
Word2-vec + CNN[2]	0.91323	5761
FastText + CNN	0.9462	5761

Fig.7: *Comparison Result*

We have Found that FastText Word Embedding with CNN Model Performance is Better than the Word2-Vec with CNN. Because In Paper[2], Authors created word vectors using random values for the words that are not present in the pre-trained Word2Vec model. which affected the performance of the CNN model and also they used Softmax activation function at the output layer which is generally suited in the Multiclass classification problem. As well as they used Adadelta optimizer which has slow convergence in comparison with Adam Optimizer. all these effects on word2vec-CNN Model Performance. But in this work, all the words available in the FastText pre-trained model, No Random values of word vectors are generated, As FastText functionality, it creates a vector of unseen word by adding n-grams of characters. which increases the performance of the CNN model. Also, We have used sigmoid function at the output layer, which is most generally used for binary classification problems in this work model classified into Positive and Negative Reviews. the optimizer used is Adam. all these improve the FastText-CNN Model Performance in comparison with word2vec- CNN Model.

V. CONCLUSION

In this paper, sentiment analysis of mobile phone Reviews are done using different Machine learning Models as well as also using FastText word embedding with CNN Depp learning Model. Comparisons of all these algorithms are done and also compared the Performance of FastText-CNN Model with the word2vec-CNN Model of Research Paper[2]. We have used various vectorization Techniques with machine learning models like naïve byes , SVM, and compare there performance with FastText deep learning model. We found that FastText-CNN Model achieved higher accuracy than machine learning models. Which means deep learning model performs better than the machine learning model. use of Different Vectorization Techniques affects the performance of the Models. With an increase, the size of data accuracy also increases. SVM based Classifier gives better accuracy than the Naïve byes. Classification of Positive & Negative Reviews influence consumer buying patterns.Use of different learning rate, no of epochs, different window sizes, different embedding sizes, number of filters and other optimization algorithms, the maximum Sequence length of Reviews, different activation functions, different batch size, etc. affects the Performance of the FastText-CNN Model. Implemented Machine Learning Techniques on Local CPU Takes a longer time than GPU.

References

[1] Upma Kumari, Dr. Arvind K sharma,Dinesh Soni, “Sentiment analysis of Smart Phone Product Review Using SVM Classification Technique” 2017 IEEE

[2] Jagadeesh Panthati, Jasmine Bhaskar , Tarun Kumar Ranga , Manish Reddy Challa, “Sentiment Analysis of Product Reviews using Deep Learning” 2018 IEEE

[3] Zeenia Singla, Sukhchandan Randhawa, Sushma Jain,“Sentiment Analysis of Customer Product Reviews Using Machine Learning ” 2017 InternationalConference on Intelligent Computing and Control (I2C2)

[4] Shilpi Chawla, Gaurav Dubey, Ajay Rana , “Product Opinion Mining Using Sentiment Analysis on Smartphone Reviews” IEEE-2017

[5] Zeenia Singla, Sukhchandan Randhawa, and Sushma Jain, “STATISTICAL AND SENTIMENT ANALYSIS OF CONSUMER PRODUCT REVIEWS” IEEE-2017

[6] R. Pavithra and A. R. Mohamed Shanavas ,“Sentiment Analysis about Smart Phones Using Twitter Corpus by Deep Learning Approach “ Asian Journal of Computer Science and Technology

ISSN: 2249-0701 Vol.8 No.S2, 2019

[7] Tanushree Dholpuria, Y.K Rana, Chetan Agrawal, "A Sentiment analysis approach through deep learning for a movie review"IEEE-2018

[8] Joana Gabriela Ribeiro de Souza, Alcione de Paiva Oliveira,"Deep Learning Approach for Sentiment Analysis Applied to Hotel's Reviews "Springer International Publishing 2018

[9] Akshay Krishna, V. Akhilesh, Animikh Aich and Chetana Hegde," Sentiment Analysis of Restaurant Reviews Using Machine Learning Techniques" Springer 2019

[10] https://fasttext.cc/

[11] Piotr Bojanowski and Edouard Grave and Armand Joulin and Tomas Mikolov " Enriching Word Vectors with Subword Information" Facebook AI Research

[12] Achmad Bayhaqy, Sfenrianto Sfenrianto, Kaman Nainggolan, Emil R. Kaburuan," Sentiment Analysis about E-Commerce from Tweets Using Decision Tree, K-Nearest Neighbor, and Naïve Bayes" IEEE

[13] Anuja P Jain,Asst. Prof Padma Dandannavar, "Application of Machine Learning Techniques to Sentiment Analysis"IEEE 2016

[14] Satuluri Vanaja,Meena Belwal,"Aspect-Level Sentiment Analysis on E-Commerce Data "IEEE 2018

[15] Dilip Singh Sisodia, N. Ritvika Reddy," Sentiment Analysis of Prospective Buyers of Mega Online Sale using Tweets ",IEEE 2017

[16] Tanjim Ul Haque, Nudrat Nawal Saber,Faisal Muhammad Shah,"Sentiment Analysis on Large Scale Amazon Product Reviews" IEEE 2018