International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

Text Based Sentiment Analysis Using Simple to Advanced ML Processes

Varna C V
Research Scholar, KSOU, Mysore, Karnataka, India

***Abstract*-- Today is an era of Artificial Intelligence and Machine Learning. The technique of Emotion Recognition has been improving businesses and organizations through the study of Reviews or the feedback received in the form of text which contains positive or negative emotions in it. Customers expresses their feedback by posting reviews or through comments. Sometimes, emotions towards a product, person or an event can be expressed through images, Audio and videos too in online mode. This paper is concentrated on the study of emotions from simple polarity method to applying tokenizers and transformers to recognize the better model. These different models are applied on the preprocessed and cleaned text data and Amazon preprocessed cleaned datasets to compare their performance accuracy vectorized SVM and polarity brings highest accuracy. SVM, RNN, CNN, Transformer methods are applied on the Kaggle datasets to illustrate the journey of emotion Recognition.**

***Keywords*---Stemming, Lemmatizing, polarity, Tokenizer, TF-IDF vectorizer**

I. INTRODUCTION

Sentiment polarity is a numeric score that is assigned to both the positive and negative sentiments hidden in the text. The sentiments thus hidden would be retrieved by performing specific pre-processing operations on the words, phrases, special characters and unnecessary characters which obstructs the process of taking out hidden sentiments. A simple approach to start with is using the text blob library to get the polarity of the text, which is a value between -1 (negative) and 1 (positive) using which 62% accuracy has been achieved. For Example, the text: I love my city has a positive sentiment. This yields the output Sentiment as Positive. Polarity is 0.5 and the Subjectivity as 0.6 for this text. The text: I hate lazy people provides the negative sentiment and the Polarity would be -0.525 Subjectivity: 0.95. Similarly, the text: I am neutral towards watching political news. Yields the Sentiment as Neutral and the would-be Polarity: 0.0and Subjectivity: 0.1. The text: I'm disappointed by today's happenings yields the Sentiment: Negative Polarity: -0.75 and Subjectivity: 0.75. Advanced Emotion Detection with Machine Learning is done for joy, fear, anger, sadness, surprise, one typically needs a labeled dataset to train a machine learning model.

This process usually involves steps such as Collecting a dataset of text samples. Text Preprocessing and Tokenization is done by Cleaning the text and breaking it down into individual words or tokens. Feature Extraction Converts text into a numerical format the model can understand using TF-IDF vectorizer or word embeddings. Model Training is done Using deep learning models to train on the data.

II. LITERATURE SURVEY

In this paper, authors have discussed the extraction of sentiment analysis on tweets. They have used Hadoop Framework for processing movie data set that is available on the twitter website in the form of reviews, feedback, and comments. Results of sentiment analysis on twitter data has displayed as different sections presenting positive, negative and neutral sentiments [1]. Authors in this paper have performed sentiment analysis to explore twitter data referring to tweets relating to donations, fundraising or charities. This paper covers techniques and approaches to capture polarity of sentiments of people towards donating for any cause under exploratory data analysis. By using Natural Language Processing Toolkit (NLTK) they have determined whether a tweet is of neutral, positive or negative polarity [2]. This study presents a comparison of different deep learning methods used for sentiment analysis in Twitter data. convolutional neural networks (CNN), in the area of image processing and recurrent neural networks (RNN) which are applied with success in natural language processing (NLP) tasks. In this work authors have compared ensemble techniques and combinations of CNN and a category of RNN the long short-term memory (LSTM) networks. This study contributes to the field of sentiment analysis by analyzing the performances, advantages. [3]. This paper compares between two techniques for Arabic text classification using WEKA application. These techniques are Support Vector Machine (SVM) and Naïve Bayesian (NB), Authors have investigated the use of TF-IDF to obtain document vector. The main objective of this paper is to measure the accuracy and time to get the result for each classifier and to determine which classifier is more accurate for Arabic text classification.

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

Comparison reported in this paper shows that the Naïve Bayesian method is the highest accuracy and the lowest error rate.[4]. Authors have proposed a multimodal sentiment analysis model to determine the sentiment polarity and score for any incoming tweet, both text and images. Image sentiment scoring is done with convolution neural network (R-CNN). Text sentiment scoring is done using a novel context-aware hybrid (lexicon and machine learning) technique. Multimodal sentiment scoring is done by separating text from image using an optical character recognizer and then aggregating the independently processed image and text sentiment scores. High performance accuracy of 91.32% is observed for the random multimodal tweet dataset used to evaluate the proposed model.[5]The context plays an important role in emotion perception, and when the context is incorporated, one can infer more emotional states. In this paper authors have presented the Emotions in Context Database (EMCO), a dataset of images containing people in context in non-controlled environments. In these images, people are annotated with 26 emotional categories. With the EMCO dataset, trained a Convolutional Neural Network model that jointly analyses the person and the whole scene to recognize rich information about emotional states.[6]This paper analyzes the efficacy of BERT, RoBERTa, DistilBERT, and XLNet pre-trained transformer models in recognizing emotions from texts. The paper undertakes this by analyzing each candidate model's output compared with the remaining candidate models. The implemented models are fine-tuned on the ISEAR data to distinguish emotions into anger, disgust, sadness, fear, joy, shame, and guilt..[7]

In this paper, by combining the emotionally fine-tuned embedding with contextual information-rich embedding from pre-trained BERT model, the emotional features underlying the texts has been more effectively captured in the subsequent feature learning module, which in turn leads to improved emotion recognition performance. The knowledge-based word embedding fine-tuning model is tested on five datasets of emotion recognition, and the results and analysis demonstrate the effectiveness of the proposed method.[8]

III. Methodology

All the necessary libraries are imported for each model, then the task of preprocessing is done on the imported Amazon dataset and lexicon-based approach is applied then the accuracy obtained is 64% in this method [fig-1]. While the tokenization approach with simple sequential model and simple RNN yields 62% on the loaded preprocessed data [fig-2].

On the same dataset SVM is applied with simple sequential model both of which have achieved the accuracy rate of 98%. This test declares Vectorized SVM and Lexicon based approach are the better methods for the sentiment Analysis. Data is split in to training and testing sets, model built with various CNN layers and training and evaluation is done to analyze the sentiment or emotion hidden in the Reviews.

Preprocessing

We need to first load data from different sources such as text files, pdfs and csv files. Text mining is the process of evaluating large amount of textual data to produce or extract opinions and sentiments. Text mining involves the steps such as gathering textual data, then perform preprocessing on that data, perform shallow parsing, identify stop words, stemming and lemmatizing is done. The text file called nlpl.txt is prepared first. Text document. translate(punctuation). lower(). split() is used to split the words based on space between the words and punctuation marks. [fig-1]

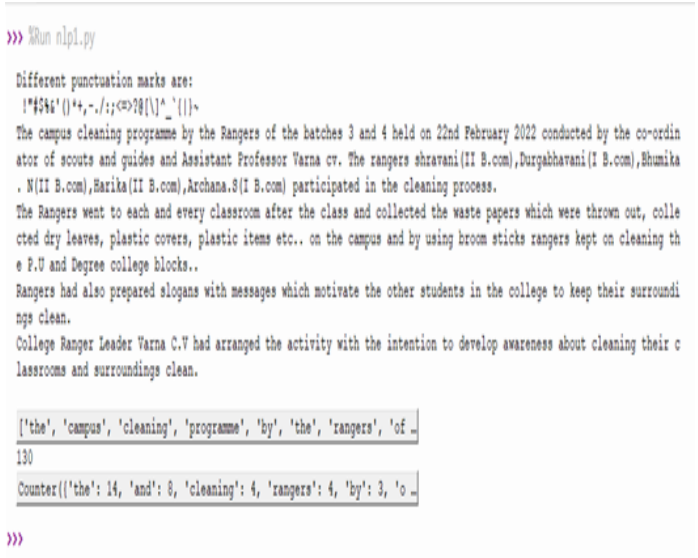


Fig-1

In the preprocessing stage, unwanted characters from the document are removed using re library function and converted to lower case and leading and trailing spaces are removed to get a preprocessed document. [fig2].

International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

Fig-2

After removing unnecessary characters length of the text document reduced from 887 to 852. [fig4].Shallow parsing is done also called as chunking which is an operation to group words into noun phrases, verb phrases and prepositional phrases. It is used to analyze the structure of a sentence to break it down into its smallest constituents called tokens and grouping them into higher level phrases. This includes POS tags as well as phrases in a sentence. The process of POS involves classifying words into parts of speech and labeling them accordingly is known as part-of-speech tagging or POS-tagging.

[('the', 'DT'), ('campus', 'NN'), ('cleaning', 'VBG'), ('programme', 'NN'), ('by', 'IN'), ('the', 'DT'), ('rangers', 'NNS'), ('of', 'IN'), ('the', 'DT'), ('batches', 'NNS'), ('3', 'CD'), ('and', 'CC'), ('4', 'CD'), ('held', 'VBD'), ('on', 'IN'), ('22nd', 'CD'), ('february', 'JJ'), ('2022', 'CD'), ('conducted', 'VBN'), ('by', 'IN'), ('the', 'DT'), ('coordinator', 'NN'), ('of', 'IN'), ('scouts', 'NNS'), ('and', 'CC'), ('guides', 'NNS'), ('and', 'CC'), ('assistant', 'NN'), ('professor', 'NN'), ('varna', 'NN'), ('cv', 'VBD'), ('the', 'DT'), ('rangers', 'NNS'), ('shravaniii', 'VBP'), ('bcomdurgabhavanii', 'JJ'), ('bcombhumika', 'FW'), ('nii', 'JJ'), ('bcomharikaii', 'NN'), ('bcomarchanasi', 'NN'), ('bcom', 'NN'), ('participated', 'VBD'), ('in', 'IN'), ('the', 'DT'), ('cleaning', 'NN'), ('process', 'NN'), ('the', 'DT'), ('rangers', 'NNS'), ('went', 'VBD'), ('to', 'TO'), ('each', 'DT'), ('and', 'CC'), ('every', 'DT'), ('classroom', 'NN'), ('after', 'IN'), ('the', 'DT'), ('class', 'NN'), ('and', 'CC').

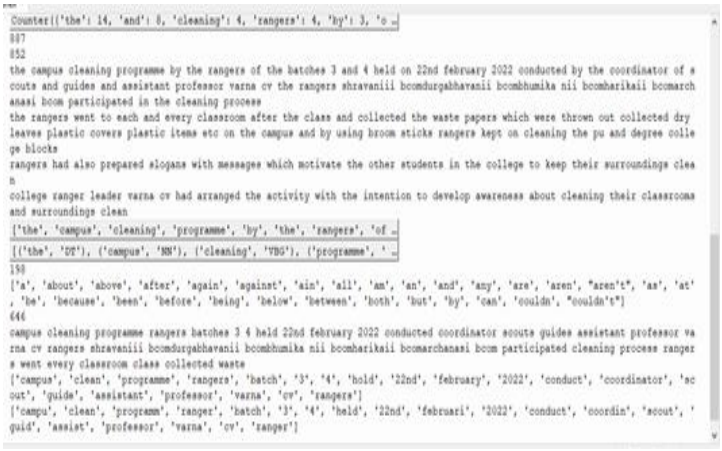
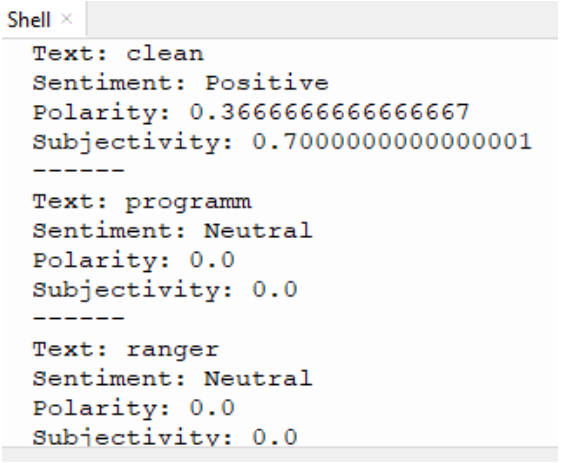


Fig-3

Text may contain stop words such as a, about, and, any, by, the. These stop words are considered as noise in the text and should be removed. Before analyzing the text data, we should filter out the list of tokens from these stop words. [fig3]. Stemming and lemmatization consider another type of noise in the text which reduces the list of words with the same base to the common root word. Stemming is the process of gathering words of similar origin into one word. This above cleaned preprocessed dataset is provided as input to the Lexicon based polarity approach of text review processing and tokenizer approach of text processing with sequential model.

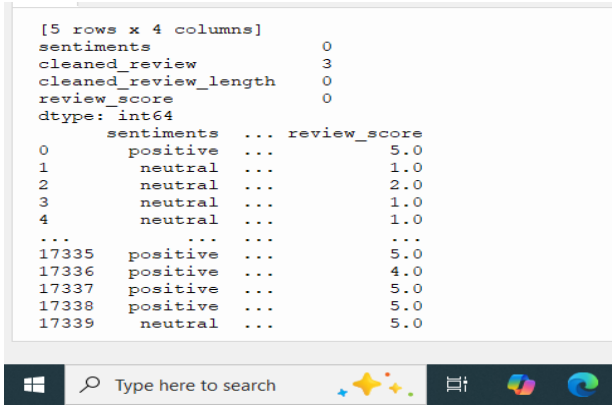
Lexicon based approach on Text dataset



International Journal of Recent Development in Engineering and Technology
Website: www.ijrdet.com (ISSN 2347-6435(Online) Volume 14, Issue 12, December 2025)

Tokenizer Method

For the tokenizer method of processing the text, Tokenizer, pad_sequences and Embedding needs to be imported from the tensorflow.keras. Regex module is imported for the text filtering and manipulation purpose. Selected columns will undergo case conversion process from uppercase to lowercase. Regex class will perform its operation. Then the tokenizer splits the words in to padded sequences. Train and test datasets would be generated. Simple sequential model with embedding with RNN is done. Adam optimizer is applied to optimize the accuracy rate. [fig-4].



The sentiments obtained in tokenization approach with CNN on amazon pre-cleaned dataset named as cleaned_reviews.[fig5].

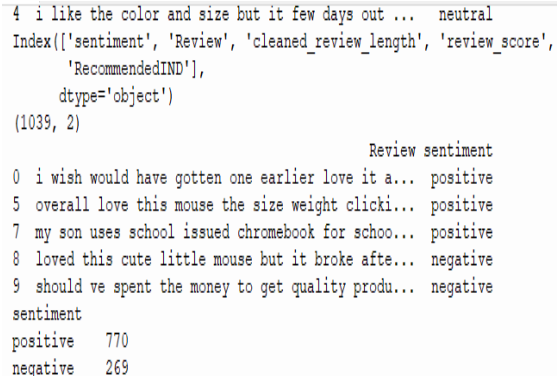


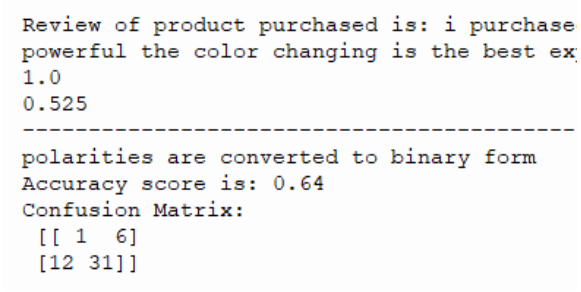
Fig-5

Vectorizer SVM Method

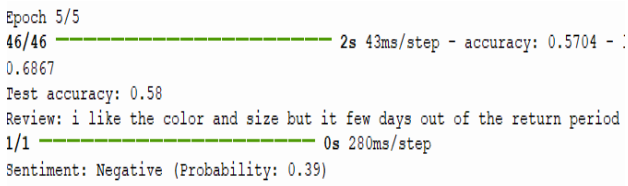
In this method, Term frequency-inverse documents frequency (TFIDF) is a statistical measure used to evaluate how many number of times a particular word appears in the document is used.

vectorizer = TfidfVectorizer (max_features=150, ngram_range=(1,3))
train_ = vectorizer. fit_ transform(train)
test = vectorizer. transform (test) is the code applied Model is SVM.

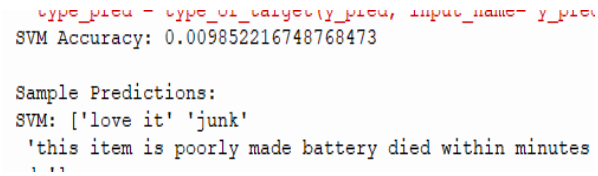
IV. Results And Discussion



Lexicon Polarity Method Result for amazon pre-cleaned dataset named as cleaned_reviews



CNN Tokenizer prediction on amazon pre-cleaned review dataset



SVM prediction using TFIDF vectorizer on pre-cleaned amazon dataset

V. Conclusion

Initially, preprocessing stage is done The preprocessed data is kept in a separate file and provided as an input. An already preprocessed data is provided as an input which was processed separately on the amazon cleaned kaggle dataset in the tokenization approach. TFIDF Vectorizer turns text into numbers. This code splits the clean text and polarity columns into training and testing sets using an 80/20 split. ensures reproducibility. This code creates a TFIDF vectorizer that converts text into numerical features. It fits and transforms the training data and transforms the test data and then prints the shapes of the resulting TFIDF vectors.

Here we train a SVM classifier model on the TFIDF features from the training data. It predicts sentiments for the test data and then prints the accuracy. Finally, polarity approach and vectorizer with SVM model applied predicts highest accuracy rates.

REFERENCES

[1] H. Parveen and S. Pandey, "Sentiment analysis on Twitter Data-set using Naive Bayes algorithm," 2016 2nd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT), Bangalore, India, 2016, pp. 416-419, doi: 10.1109/ICATCCT.2016.7912034.

[2] A. Shelar and C. -Y. Huang, "Sentiment Analysis of Twitter Data," 2018 International Conference on Computational Science and Computational Intelligence (CSCI), Las Vegas, NV, USA, 2018, pp. 1301-1302, doi: 10.1109/CSCI46756.2018.00252.

[3] D. Goularas and S. Kamis, "Evaluation of Deep Learning Techniques in Sentiment Analysis from Twitter Data," 2019 International Conference on Deep Learning and Machine Learning in Emerging Applications (Deep-ML), Istanbul, Turkey, 2019, pp. 12-17, doi: 10.1109/Deep-ML.2019.00011.

[4] RESEARCH-ARTICLE Political Sentiment Analysis Using Twitter Data Authors: Tarek Elghazaly, Amal Mahmoud, Hesham A. HefnyAuthors Info & Claims ICC '16: Proceedings of the International Conference on Internet of things and Cloud Computing https://doi.org/10.1145/2896387.2896396 Published: 22 March 2016

[5] Kumar, A., Garg, G. Sentiment analysis of multimodal twitter data. *Multimed Tools Appl* **78**, 24103–24119 (2019). https://doi.org/10.1007/s11042-019-7390-1

[6] Emotion Recognition in Context Ronak Kosti, Jose M. Alvarez, Adria Recasens, Agata Lapedriza; Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 1667-1675

[7] A. F. Adoma, N. -M. Henry and W. Chen, "Comparative Analyses of Bert, Roberta, Distilbert, and Xlnet for Text-Based Emotion Recognition," 2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP), Chengdu, China, 2020, pp. 117-121, doi: 10.1109/ICCWAMTIP51612.2020.9317379.

[8] Zixiao Zhu, Kezhi Mao,Knowledge-based BERT word embedding fine-tuning for emotion recognition,Neurocomputing,Volume 552,2023,

[9] Data Analysis using python Text book by the Author-Bharti Motwani