iJRASET
International Journal For Research in Applied Science and Engineering Technology

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY

Volume: 11 **Issue:** V **Month of publication:** May 2023

DOI: https://doi.org/10.22214/ijraset.2023.51748

www.ijraset.com

Call: 08813907089 | E-mail ID: ijraset@gmail.com

International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

Identification of Positive and Negative Tweets

Kunti Dongare[1], Neha Mahajan[2], Gayatri Takawale[3], Akshada Thange[4], Prof. Runal.P. Pawar[5]
[1, 2, 3, 4,5]*Department Of Computer Science Engineering Sinhgad College Of Engineering, Vadgaon(BK), Pune, India*

Abstract: ***With the development of the Internet, people can obtain and share information almost instantly from a wide array of sources, for example, online news outlets and fast-growing social networks. Over the past few decades, the explosive growth of textual data far outpaces human beings' speed of understanding its content. Indeed, we have seen the emergence of new types of textual data that reflect social interactions in online settings. The production of this socially-generated content is accelerated by the wide adoption of social media sites, such as Facebook, Twitter, Yahoo! Answers, and Reddit. Text summarization helps in reducing the size of a text while preserving its information content. Text Summarization can be derived as shortening the source text into a version that it's information content and overall meaning is preserved. It is very difficult for human beings to manually summarize large documents of text. Sentiment analysis on the other hand is the process of computationally identifying and categorizing opinions expressed in text to clarify someone's' attitude towards a topic is positive or else negative or even neutral. In this paper we will discuss the use Extractive summarization for the text summarization followed by SVM-algorithm for sentiment analysis.***

Keywords: Text-Summarization, Sentiment Analysis, Extractive summarization, SVM-algorithm, NLP, Machine Learning, Social-Media.

I. INTRODUCTION

Inshorts is one of numerous text-summarization tools available on the internet. It offers its users a summary of the news articles' substance. Being confined, it might not be helpful for those looking for a text summarizer when it comes to news stories. Additionally, the current technology does not offer sentiment analysis for news stories. Due to these two short comings of the current system, we came up with the notion of offering consumers text summary combined with sentiment analysis. Text summarization is one of the methods of identifying the important meaningful information in a document or set related document and compressing them into a shorter version preserving its overall meanings. It reduces the time required for reading whole document and also it solves problem that is needed for storing large amount of data.

In Automatic Text summarization has two approaches 1) Abstractive text summarization and 2) Extractive text summarization. An extractive text summarization means an important information or sentence are extracted from the given text file or original document. An extractive text summarization approach uses linguistic or statistical features for selecting useful informative sentence. An Abstractive text summarization will try to understand the input file or original file and re-generate the output in few words by identifying the main concept of the input file. In many research papers they have mentioned that extractive text summarization is sentence ranking. Extractive text summarization is divided in two phases: 1) Pre-processing 2) Processing. In this paper we are explaining extractive text summarization on single document.

Sentiment analysis is an original work in Natural Language Processing (NLP). The main goal of sentiment analysis or opining mining is to extract the attributes or views regarding political opinion on social media, draft trading strategies from market sentiment, or collect product insights from customer review. To enhance decision-making, business and academic researchers have suggested a variety of ways to address the issue. Now days, there is a huge increase in number of peoples who have been accessing many social networking sites and microblogging websites which open a new door to the impression of today's generation. Various user reviews for a specific product, company, brand, individual, forums and movies etc. have been much helpful in judging the perception of people. This automated classification mechanism is referred as Sentiment Analysis. The primary aim of this paper is to apply Support Vector Machine (SVM) machine learning algorithm to classify the sentiments and texts analyses different datasets used for classification of sentiments and texts. Furthermore, various data sets have been utilized for training as well as testing and implementing the Support Vector Machine learning algorithm to find the polarity of the ambiguous sentiments.

A. Problem Statement

Twitter is one of the most popular social media platforms to send and read posts to communicate with others. It allows people to write their own opinions about products or share their moments, even influence politics and companies.

Using machine learning, the online trail of data that a person leaves behind can be used to gain insights on the behaviour and psychological status. Given a tweet or a comment of that tweet, classify whether that is of positive, negative, or neutral sentiment.

B. Objective

1) Classifying whether tweet or comment is of positive, negative, or neutral sentiment is like opinion mining, i.e., analysing conversations, opinions, and sharing of views (all in the form of tweets) for deciding business strategy, political analysis, and also for assessing public actions.
2) To help people/ businesses get honest public review about their product or services and improve it.

II. LITERATURE SURVEY

A. Automatic Sarcasm Detection Using Feature Selection

Automatic sarcasm detection refers to the detection of sarcasm in the text written in natural language. This paper focuses on various sarcasm analyzing techniques employed for filtering of sarcastic statements from a text and the use of Automatic sarcasm detection in the categorization of tweets and product review texts. The architecture for the automatic sarcasm detection followed creating training and testing data sets, processing the dataset, Feature engineering, Feature selection, classification and model evaluation.

B. Twitter Sentiment Analysis Using Deep Learning Models

The purpose of this paper is to compare the tweet sentiment classification using Google BERT, attention based Bidirectional LSTM and Convolutional Neural Networks (CNNs). We evaluated the efficiency of the classifiers on SemEval-2016 test set containing 20,632 tweets. The experiment shows that the Google BERT outperforms both Bi-Attentive LSTM and Convolutional Neural Networks.

C. Sentiment Analysis on COVID-19 Twitter Data

This paper analyzed the tweets regarding COVID-19 from November, 2019 to May, 2020 in India and its affect. All tweets are categorized into 3 categories (Positive, Negative and Neutral). TextBlob is used to find the polarity of scraped tweets and Natural Language Toolkit (NLTK) for word frequency. The analysis phase of the process gave us great insights into the emotional state of people in India, and also, how sentiments varied from state to state on daily basis.

D. Opinion Mining Using Twitter Data Set

The main aim of this project is to develop a functional classifier for accurate and automatic sentiments classification of an unknown tweet stream. We conclude that using different NLTK classifier like Bernoulli NB and Random Forest it is easier to classify the tweets and moreover we improve the training data set to give accurate results.

E. Identification Of Informative Tweet During Disasters

This research paper proposes a model that specifies Bi-LSTM and CNN combined approach for text-based classification while VGG-16 architecture for image classification. The objective is to construct a model by combining Bi-directional Long Short-Term Memory (Bi-LSTM) and Convolutional Neural Network (CNN)to categorize the textual content for the tweets. The output of the text-based model and the image-based model will be consolidated using the late fusion technique to predict the tweet label.

F. Extractive Text Summarization Using Sentence Ranking

The paper is divided into the following parts: background describing the theoretical terms related to the working of a text extractor and summarizer, highlights of the methods under review, comparison and evaluation of methods based on common efficiency parameters, analysis of working of best compared methods, results to identify the best components of each method to form an efficient system and conclusion outlining the requirements and features of a good automatic summarizer.

III. PROPOSED SYSTEM

After doing the literature and understanding the requirements of our project we decided to use Extractive text summarization and SVM algorithm for sentiment analysis.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

The application would be providing the user with two options i.e., to do separate text summarization and then analyse the sentiments of the text in real time and another option would be the user would be able to upload a dataset to be analysed and our application would produce a detailed report on the sentiments observed in the dataset provided by the user.
We will be taking the tweet input and then analyse if the entered input is positive or negative. For the analysis we have first trained our model using SVM algorithm, this model is trained on 21k comments from social media, it was cleaned, pre-processed, 80% for the dataset was used for training the model and 20% was used for testing purpose. Other things like feature extraction, classification and finally identifying if the tweet is positive or negative is obtained as visible in Fig 1.

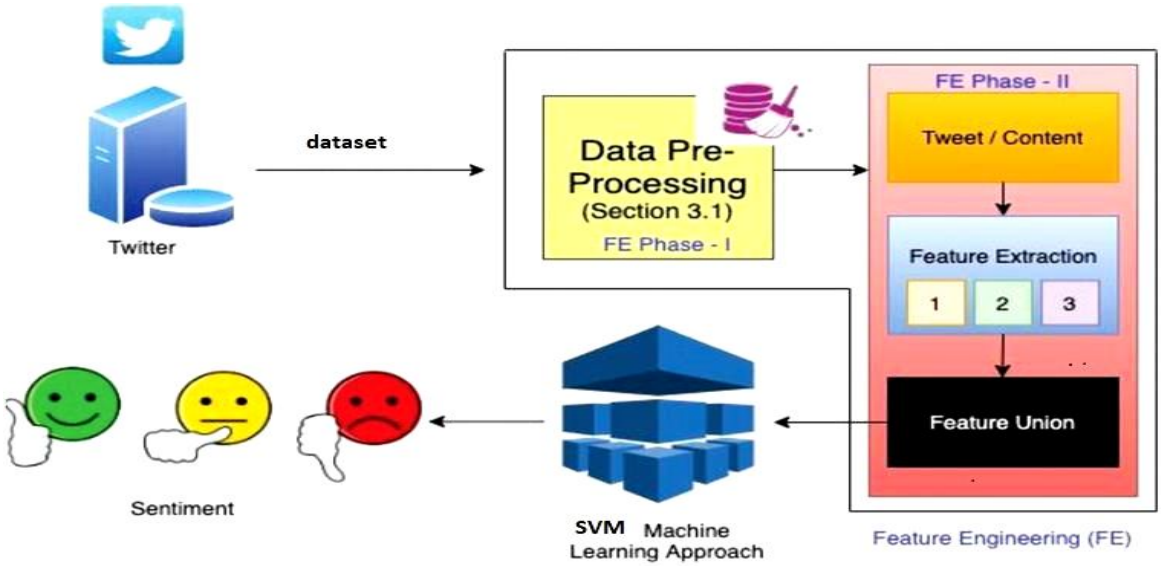


Fig: 1 Data flow diagram

This flowchart explains deliberately about the process flow of proposed methodology for sentiment classification. The processing is given in the architecture which is shown in figure 1.

1) *Training Data*: - Collecting dataset (tweets) that are done by users. Here we have collection of 40,000 datasets for the training purpose gathered from GitHub, Kaggle.
2) *Pre-processing:* - With respective of improving the performance of analysis, the compulsion is to do pre-processing of data before exploring it. At first the process of tokenize (splitting of sequence of a string) is initiated, the input stream converted in to separate words. Typically, Input texts contain special stuffs like "URL's username, hashtag, punctuation and additional white space", which does not bear any sentiment. Simply It can be state that stop words from the datasets can be avoided that does not convey any emotion using NLTK stop word corpus. Converting all the uppercase into lowercase. After that filtration process is done and then POS tagging is done for classifying words based on their parts of speech.
3) *Features*- For making the sentiment analysis model, we have to extract every single feature from the text data which are widely categorized into morphological features, word N-gram features. Morphological Features: - Here we use morphological features which determines existence of elongated words (such as dull, fooool, byeeeee etc.), time and date expression, punctuation marks. It also counts every single feature like elongated words, punctuation marks, fully and partially capitalized tokens.
4) *Bag of Words Features*: This means extracting features from text data for modelling purpose with the help of machine learning algorithms. "It is a representation of text that describes the occurrence or frequency of words within a document". Two things can be used and given as

a) A terminology of recognized words.
b) A measure of the presence of known words.

5) *Classification Model:* - In our Sentiment Analysis experiment model, support vector machine is used as a classifier and trained this classifier over the training sample.

Afterwards completion of the first process, the trained classifier has been applied on the test sample of datasets, tend the new tweets can be characterized into fine grained emotions as worry, hate, love, sadness, happy, neutral, anger, boredom, surprise, relief, enthusiast, fun, empty.

IV. EXTRACTIVE TEXT SUMMARIZATION

The extractive approach involves picking up the most important phrases and lines from the documents. It then combines all the important lines to create the summary. So, in this case, every line and word of the summary actually belongs to the original document which is summarized. Extractive summarization techniques vary, yet they all share the same basic tasks:

1) Construct an intermediate representation of the input text (text to be summarized)
2) Score the sentences based on the constructed intermediate representation
3) Select a summary consisting of the top *k* most important sentences

Tasks 2 and 3 are straightforward enough; in sentence scoring, we want to determine how well each sentence relays important aspects of the text being summarized, while sentence selection is performed using some specific optimization approach. Algorithms for each of these 2 steps can vary, but they are conceptually quite simple: assign a score to each sentence using some metric, and then select from the best-scored sentences via some well-defined sentence selection method.

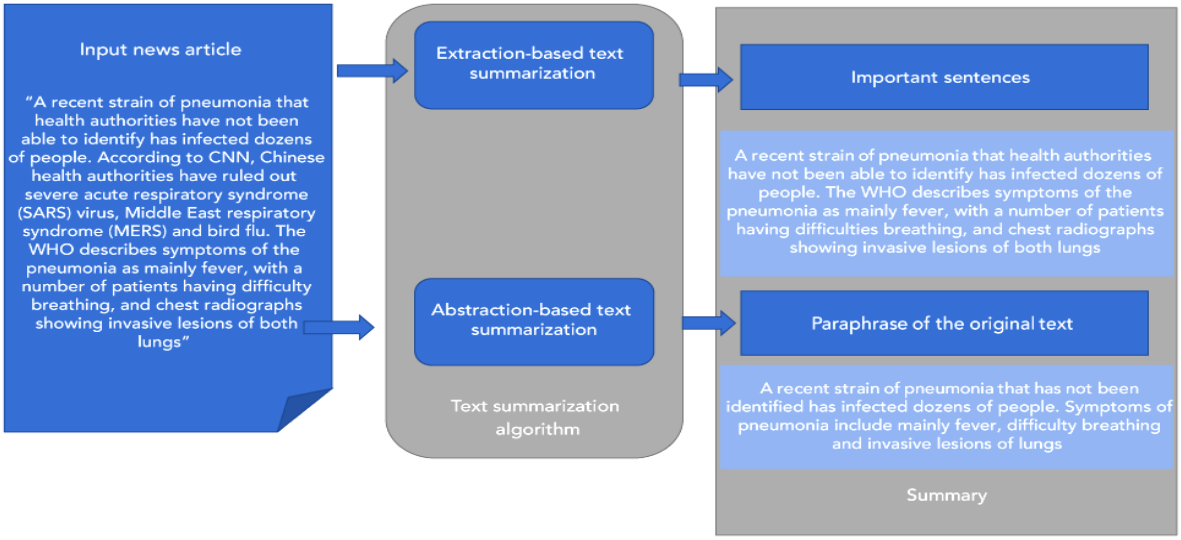


Fig 2. Extractive summarization sample

In this proposed approach, we are using extractive method to get summary of given input. We are taking input as text file .txt.

1) Firstly, the file which is given as input is tokenized in order to get tokens of the terms
2) The stop words are removed from the text after tokenization. The words which are remained are considered as a key word
3) The key words are taken as an input for that we are attaching a part of tag to each key word
4) After completing this pre-processing step, we are calculating frequency of each keyword like how frequently that key word has occurred from this maximum frequency of the keyword is taken.
5) Now weighted frequency of the word is calculated by dividing frequency of the keywords by maximum frequency of the key words
6) In this step we are calculating the sum of weighted frequencies. Finally, summarizer will extract the high weighted frequency sentences and the extracted sentences are converted into audio form

V. SUPPORT VECTOR MACHINE (SVM) CLASSIFICATION ALGORITHM

Support Vector Machine or SVM is one of the most popular Supervised Learning algorithms, which is used for classification as well as Regression problems. However, primarily, it is used for Classification problems in Machine Learning. The goal of the SVM algorithm is to create the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category in the future. This best decision boundary is called a hyperplane.

ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 7.538
Volume 11 Issue V May 2023- Available at www.ijraset.com

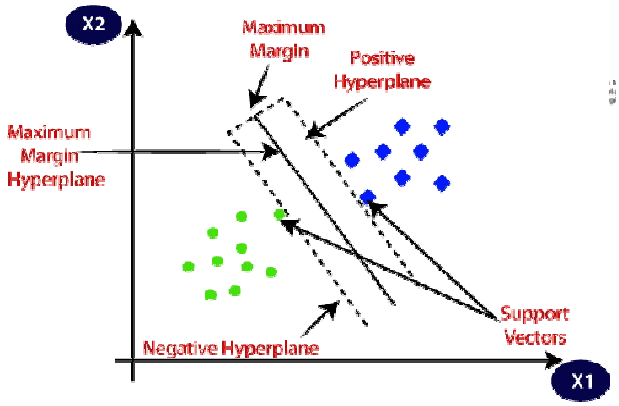


Fig 3: SVM Algorithm graphical representation

SVM chooses the extreme points/vectors that help in creating the hyperplane. These extreme cases are called as support vectors, and hence algorithm is termed as Support Vector Machine. Consider the Fig 3 in which there are two different categories that are classified using a decision boundary or hyperplane.

Hyperplane: There can be multiple lines/decision boundaries to segregate the classes in n-dimensional space, but we need to find out the best decision boundary that helps to classify the data points. This best boundary is known as the hyperplane of SVM. The dimensions of the hyperplane depend on the features present in the dataset, which means if there are 2 features (as shown in image), then hyperplane will be a straight line. And if there are 3 features, then hyperplane will be a 2-dimension plane. We always create a hyperplane that has a maximum margin, which means the maximum distance between the data points.

Support Vectors: The data points or vectors that are the closest to the hyperplane and which affect the position of the hyperplane are termed as Support Vector. Since these vectors support the hyperplane, hence called a Support vector.

Hence, the SVM algorithm helps to find the best line or decision boundary; this best boundary or region is called as a hyperplane. SVM algorithm finds the closest point of the lines from both the classes. These points are called support vectors. The distance between the vectors and the hyperplane is called as margin. And the goal of SVM is to maximize this margin. The hyperplane with maximum margin is called the optimal hyperplane.

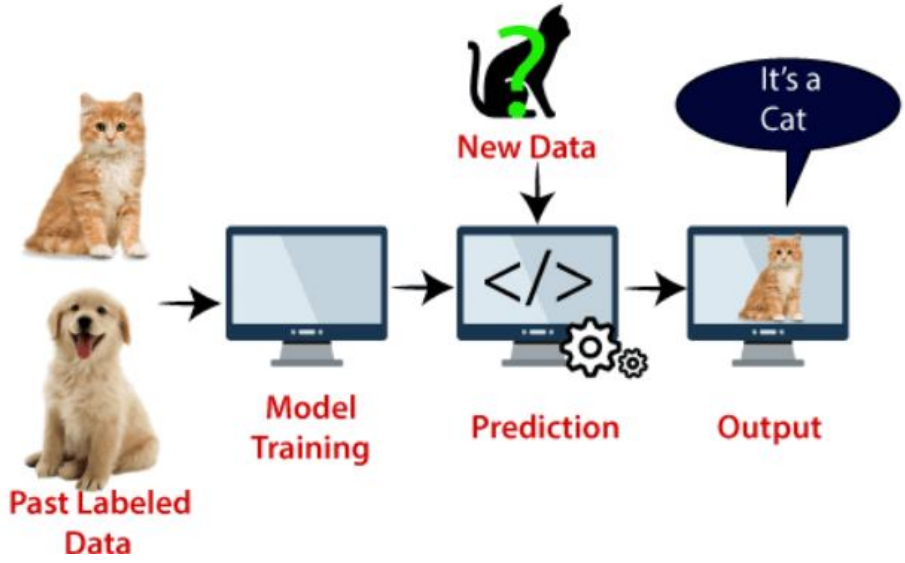


Fig 4: Working of SVM algorithm

Looking at Fig 4SVM can be understood with the example that we have used in the KNN classifier. Suppose we see a strange cat that also has some features of dogs, so if we want a model that can accurately identify whether it is a cat or dog, so such a model can be created by using the SVM algorithm. We will first train our model with lots of images of cats and dogs so that it can learn about different features of cats and dogs, and then we test it with this strange creature.

So as support vector creates a decision boundary between these two data (cat and dog) and choose extreme cases (support vectors), it will see the extreme case of cat and dog. On the basis of the support vectors, it will classify it as a cat. Consider the below diagram:

VI. RESULT

We have even tried to do the text summarization possible in marathi and hindi language as depicted in Fig 5.

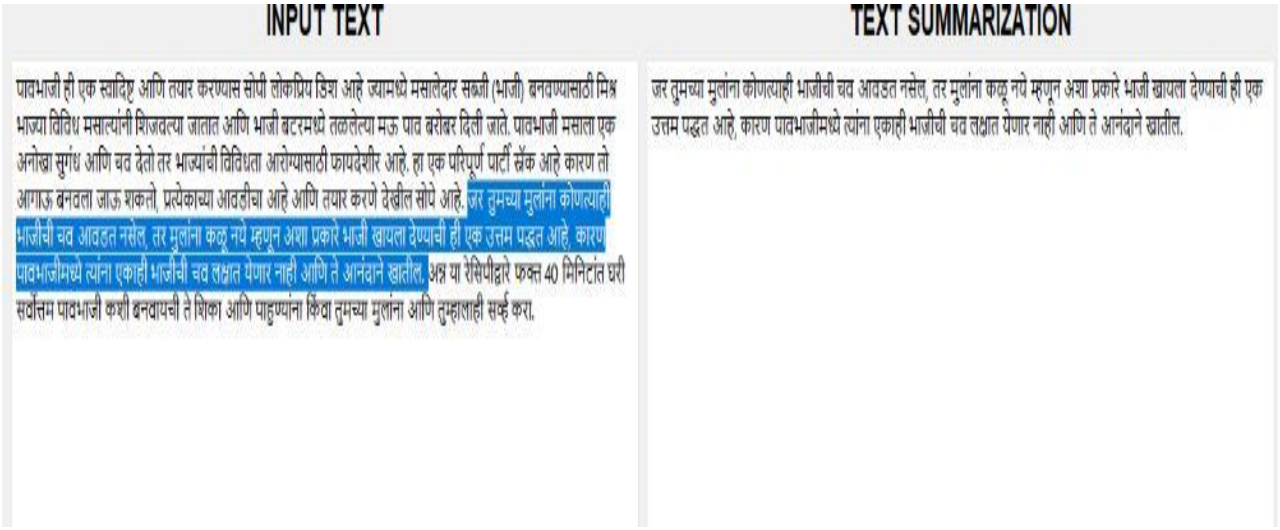


Fig 5: Text summarization in marathi

Our paper focuses on extractive text summarization and sentiment analysis. So, Fig 6 and Fig 7 shows the output of our application after the user enters his input which is to be analysed. We are using python GUI for creating our application and the precision of our application is 91.3% for sentiment analysis.

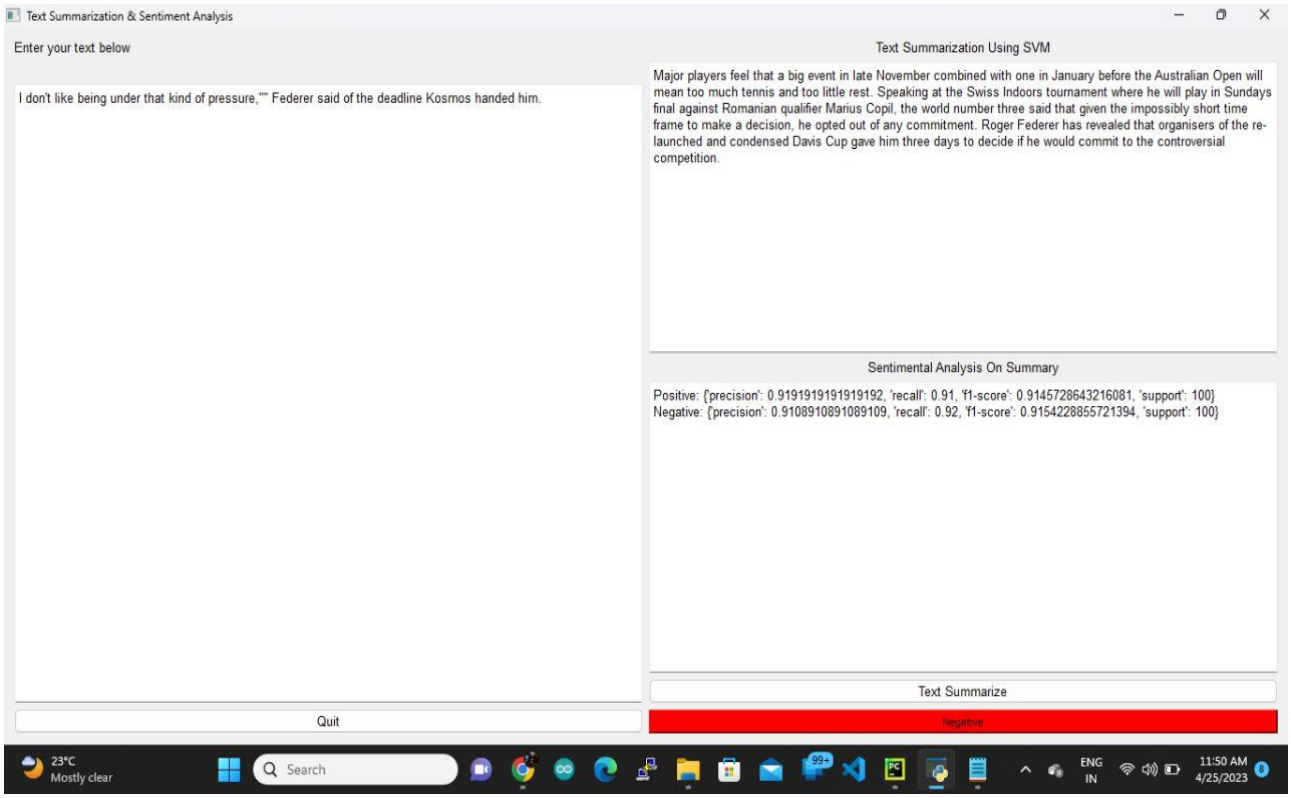


Fig 5: Output of our application identifying a negative tweet after text summarization.

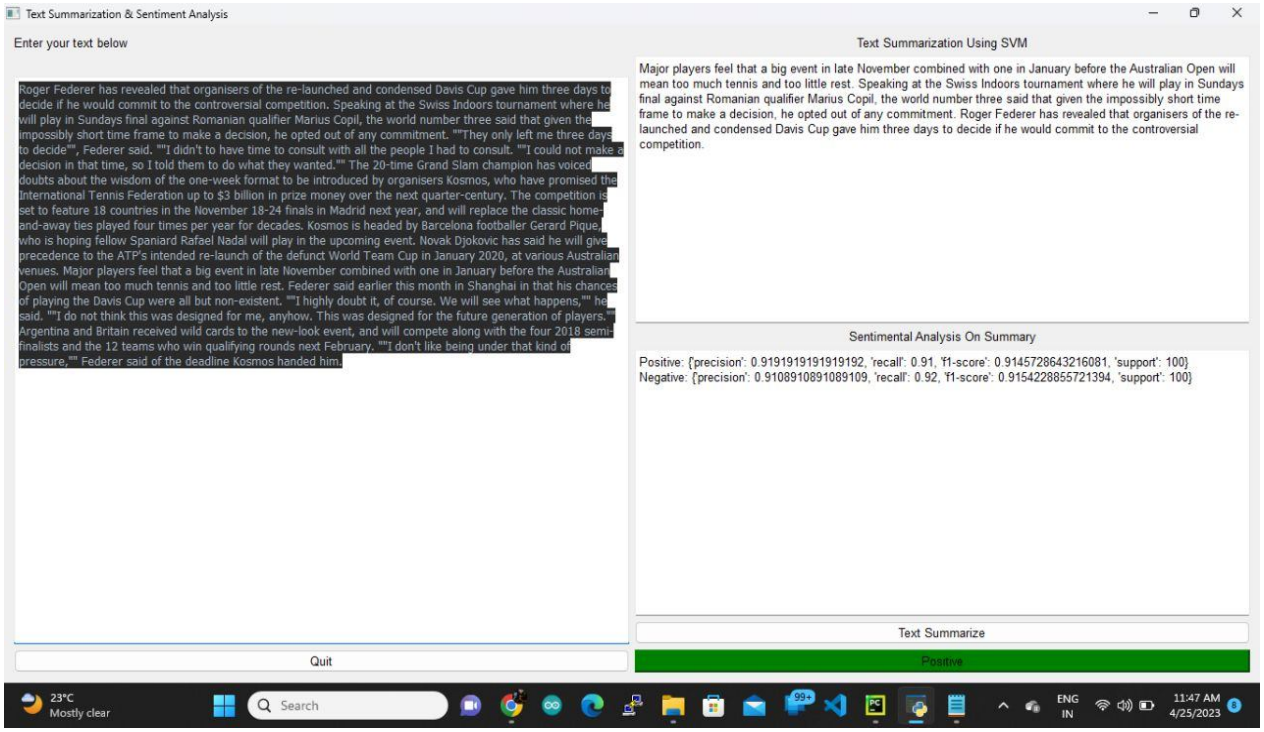


Fig 6: Output of our application identifying a Positive tweet after text summarization.

VII. CONCLUSIONS

Automatic text summarization is a complex task which contains many sub-tasks in it. Every subtask has an ability to get good quality summaries. The important part in extractive text summarization is identifying necessary paragraphs from the given document. In this work we proposed extractive based text summarization by using statistical novel approach based on the sentences ranking the sentences are selected by the summarizer. The sentences which are extracted are produced as a summarized text and it is converted into audio form. The proposed model improves the accuracy when compared traditional approach.

In this project we tried to show the basic way of classifying tweets into positive or negative category using SVM as baseline and how language models are related to the SVM and can produce better results. We could further improve our classifier by trying to extract more features from the tweets, trying different kinds of features, tuning the parameters of the SVM classifier, or trying another classifier all together.

VIII. ACKNOWLEDGMENT

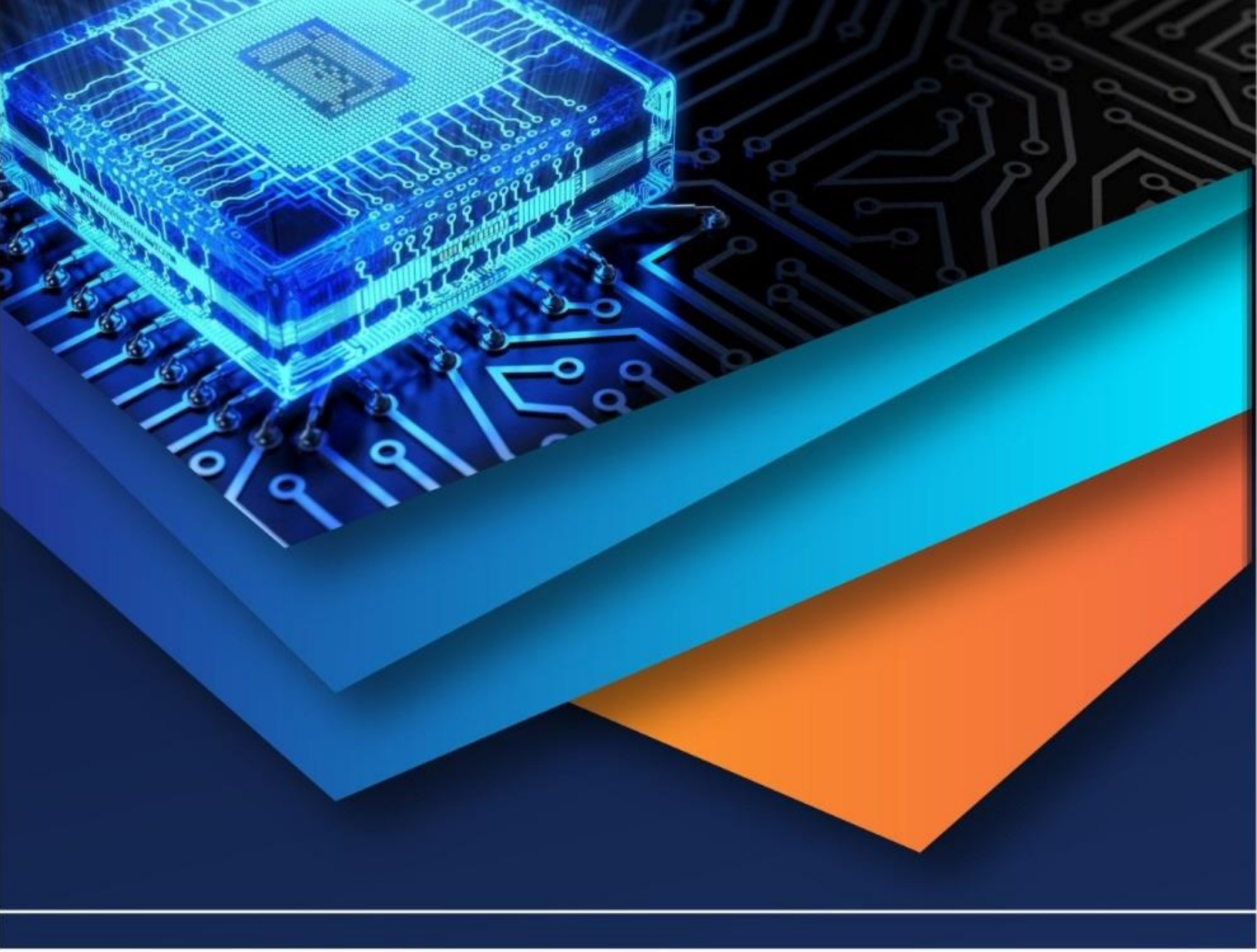
I thankful to all my teachers who helped me in framing this research paper. I would like to thank my friends for their encouragement, which helped me to keep my spirit alive and to complete this paper work successfully. Above all, I owe my gratitude to the Almighty for showering abundant blessings upon me; my acknowledgement would not be complete without acknowledging my gratitude to my beloved parents who have been pillars of support and constant encouragement throughout this work.

We have great pleasure in presenting report on Identification of Positive and Negative Tweets. Completing a task is never a one-man effort. It is often a result of invaluable contribution of a number of individuals in direct or indirect manner. I would like to express deepest appreciation towards S.D. Lokhande, Principal Sinhgad College of Engineering. And Dr.M.P. Wankhade, HOD Computer Department, whose invaluable guidance supported me in completing this report. I am profoundly grateful to Prof. Runal.P.Pawar for his expert guidance and continuous encouragement throughout to see that this project report rights its target since its commencement to its completion. At last, I want to express my sincere heartfelt gratitude to all the staff members of Computer Engineering Department who helped me directly or indirectly during this course of work

REFERENCES

[1] Paras Dharwal,Tanupriya, Choudhury,Rajat, Mittal,Praveen Kumar ,"Automatic Sarcasm Detection using feature selection", International Conference on Applied and Theoretical Computing and Communication Technology, 978-1-5386-1144-9$31.©c 2020 IEEE .

[2] Arnab Roy, Muneendra Ojha, “Twitter sentiment analysis using deep learning models” University of Technology Sydney, 2020 IEEE 17th India Council International Conference (INDICON).
[3] Tanmay Vijay, Ayan Chawla, Balan Dhanka , Purendu Karmakar “Sentiment Analysis on COVID-19 Twitter Data”, IEEE International Conference on Recent Advances and Innovations in Engineering- ICRAIE 2020 (IEEE Record#51050)
[4] Gururaj S, Ayesha Sameen, Angana Prasad, Nida Tahreem, “Opinion Mining Using Twitter Data Set”, International Research Journal of Engineering and Technology (IRJET), july 2020
[5] Prof. R.B. Murumkar, Vinay M. Harwani, Namita Bhalerao, Nilambari K. Rathi, Rutuja D. Mahajan,”Identification of Informative Tweet during Disasters” ,IEEE, 2020
[6] J. N. Madhuri and R. Ganesh Kumar, “Extractive Text Summarization Using Sentence Ranking,” 2019 Int. Conf. Data Sci. Commun. IconDSC 2019, pp. 1–3, 2019, doi: 10.1109/IconDSC.2019.8817040.
[7] Shivangi Singh, Aayush Singh, Sudip Majumder,Sanjay Deshmukh,”Extractive test summarization techinques of News Artuckes: Issues, Challenges and Approaches” IEEE, 2019



10.22214/IJRASET

45.98

IMPACT FACTOR: 7.129

IMPACT FACTOR: 7.429

INTERNATIONAL JOURNAL FOR RESEARCH

IN APPLIED SCIENCE & ENGINEERING TECHNOLOGY