MULTIMEDIA SENTIMENT ANALYSIS ONPUBLIC SOCIAL NETWORK USING NLP

[1]Swamy TN, [2]Lakshmi Janaki K, [3]Uma Maheswari P, [4]Reshma D, [5]Jayanth D

[1]Assistant Professor,[2,3,4,5]Student
[1,2,3,4,5]Department of Electronics and Communication Engineering
[1,2,3,4,5]Dr. Ambedkar Institute of Technology, India

Abstract – Social Networks are the platforms that engage large number of users where the necessity of using sentiment analysis is highly critical. Sentiment is an approach used in natural language processing (NLP), sometimes known as opinion mining. It is the technique of identifying if multimedia, such as text, images, and coded communications, has positive or negative emotion. Usage of hate speech which includes sexist and racist comments that are likely to provoke issues between people or states or nations is high inpublic social networks. The main goal is to detect hate speech in public social networks like Facebook, Twitter, YouTube etc., that can be filtered and reported. The project proposes multimedia sentiment analysis on public social networks using natural language processing (NLP). The analysis is based on the Kaggle dataset which is a labelled dataset with the correct combinations of texts or messages and images and dataset collected from public accounts. The dataset is for each media is unique and provided in the form of a suitable file formats required for multimedia processing of data. This project filters out the messages using natural language processing technique not only in text but also in images. The messages will not be removed radically. The simulation is tested with python platform and various collaborative tools. The data is visualized and committed to provide the maximum accuracy and degree of usability to take decisions in real time without violating the user's freedom of thought.

Index Terms—*Multimedia, Sentiment Analysis, Natural Language Processing, Keras, ReLU, TF-IDF Vectorization, SMOTE, ClassImbalance*

I. INTRODUCTION

An NLP tool or algorithm called sentiment analysis analyses and categorizes the emotions expressed in the text. To categorize individuals into the lovely or the nasty, is a simple way to assign feelings. The process of gathering information about customer sentiment toward a certain product, service, or brand is known as sentiment analysis. Consider the feedback left by your clients as an illustration. When customers textually describe the characteristics of a service or product, sentimentanalysis, a sort of text analytics, determines how they feel about those aspects. This usually involves looking at a textual element—a line, a remark, or a whole document—and giving the text a "score" that indicates how positive or unfavorable it is. Additionally, sentiment analysis has been referred to as opinion mining and, less commonly, emotion AI. When we are dealing with data overload in the contemporary climate, businesses may have gathered mountains of client feedback (althoughthis does not equal better or deeper insights). However, even for humans, it would be impossible to carefully review it withoutbias or error of any type. Even the best-intentioned businesses usually struggle with a lack of insights. You understand that information is important when making decisions. And you realize that you don't have them. But you're not sure how to go about getting them. Sentiment analysis sheds light on the most pressing issues.

Rule-based sentiment analysis and Machine Learning (ML)-based sentiment analysis are the two different methods of sentiment analysis. Using a training set that has been labelled with the intended sentiment, we train the machine learning (ML)model to infer the sentiment based on the words and their order in this example. The quality of the training data and the selected algorithm have a significant impact on this method.

II. BACKGROUND

a. Rationality behind choosing the project

The computer method of locating and extracting opinions from the text is known as sentiment analysis. You may monitor what people are saying about your business on social media, in news stories, and in online reviews to learn how to enhance your brand, service, or product and boost sales. Social media generates a lot of data every day. With the help of sentiment analysis, you can make sense of this massive amount of data and draw out crucial details about how people feel about your business in order to make wise decisions. Businesses may learn more about the opinions and preferences of their consumers if they have access to enough pertinent data. Additionally, it aids in removing racist and sexist remarks from social media platforms. This inspired the creation of an automated algorithm that could analyze text and visual material to extract sentiments.

III. LITERATURE SURVEY

a. Evaluating Annotated Dataset of Customer Reviews for Aspect Based Sentiment Analysis, Dimple Chehal Parul Gupta Payal Gulati

Aspect-based sentiment analysis (ABSA) helps in identification of key characteristics and their polarities to better understand their relevance to the product of interest. This paper discusses the method which can replace the traditional rating- based recommendation process. The authors have proposed supervised classification on labelled dataset. Consequently, ML methods such Naive Bayes, Support Vector Machine, Logistic Regression, Random Forest, K-Nearest Neighbor, and Multi- Layer Perceptron were used, as well as a sequential model created with the Keras API. Finally, the accuracy resulted to be in the range of 67.45 - 79.46 % which is not acceptable for the text-based analysis. [11]

b. Comparative Study of Sentiment Analysis on Mask-Wearing Practices during the COVID-19 Pandemic, Ghulam Mujtaba, Zahid Hussain Khand, Javed Ahmad Zafar Ali

COVID -19 pandemic caused harm to human resource in many countries. The outbreak can be controlled by taking necessary actions such as wearing mask and practicing personal hygiene. The study aims to analyze the tweets related to the best mask wearing behaviors. The tweets were labelled and text preprocessing techniques used in this study were stemming, tokenization,normalization and stop words removal. The study highlights to have used basic components of NLP and feature extraction. Thepaper proposed the use of 5 classification models on 228 feature vectors. Naive-Bayes performed well in terms of macro accuracy, precision, macro F1 measure [10].

c. Deep Bidirectional LSTM Network Learning-Based Sentiment Analysis for ArabicText, Hanane Elfaik, ElHabib Nfaoui

Sentiment analysis on Arabic texts is difficult due to complexity, ambiguity, lack of resources and explicit emotion terms usage. The paper discusses the possibilities of analyzing the Arabic Texts with Deep learning models. According to the authors Bi-directional LSTM Network can understand the ambiguity better and predict the sentiment. But the study has proved that themodels using LSTM need a lot of time to train on dataset and has a problem of overfitting. Due to its complexity, dropout cannot reduce the layers as in sequential model. [9].

d. Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism, Ramanathan and Meyyappan

In this study, they provide a new sentiment analysis technique based on information from Oman tourism ontology. The information consists of conversation and comments made by customers on Oman tourism. The purpose of this study is to improve the weak areas that have been overlooked during the specified period as considered by the ontology. Lexicon based approach and techniques of NLP was used. Machine learning techniques thus considered in latter stage was not effective in terms of the accuracy recorded by the classification models. The conclusion provided had the biased opinions relative to analysis performed in the study. [8].

e. Social media metrics and sentiment analysis to evaluate the effectiveness on social media posts, Poecze, Ebster, Strauss & Christine

The ideal group for study is the role of YouTube gamers in self-marketing their content. This paper proposed the importanceof social media marketing with photos received more likes and comments than the videos in YouTube uploaded by the gamers.The posts were analyzed on the number of likes and hate comments. This paper proposed the use of ANOVA and sentiment analysis along with supervised learning technique called K-NN classification algorithm with accuracy of 82.3 % outperformed other classification models. The authors concluded that reposted videos and photos gained less attention [7].

f. Sentiment Mining on Community Development Program Evaluation Based on social media, Yuliyanti, Djatna & Sukoco.

The authors above created a model named "Design and Implementation of Web-based GPS-GPRS Vehicle Tracking System". Here, they tried to use an internet-based GPS-GPRS for their vehicle tracking system and implemented it. Keeping an enterprise in mind, this concept was created. Hence, this model enabled the enterprise owners to locate the target vehicle onGoogle maps. Here, both the present and past locations were stored in the device hence it helped the owners in tracking the vehicle with the detailed journey of it. The HTTP protocol was used to send the GPS data, and the server kept the data in a separate database that could be accessed whenever necessary.

IV. OBJECTIVES

The objective of this task is to classify the sentiments in multimedia (text and image). For ease of use, we define a message (text) as containing hate speech if it expresses racist or sexist attitude. So, the task is to develop a classification model using machine learning for racist or sexist messages from the available public social network's dataset. Formally, given a training sample of messages and labels, where label '1' denotes the message is negative and label '0' denotes the message is not negative. Next, forimage the task is to develop a CNN classification model to predict the sentiments. The goal is to attain maximum accuracy and score based on the sentiments.

V. METHODOLOGY

a. General block diagram

The block diagram of the general steps employed for the multimedia sentiment analysis is shown in the below Fig 1. The block diagram consists of the following steps:

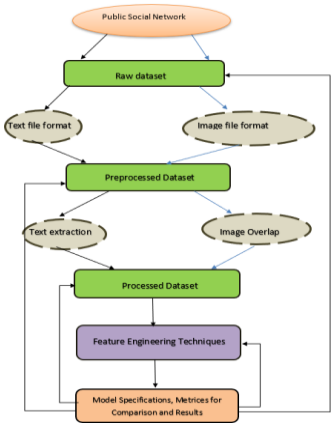


Fig 1. General Block Diagram

i. Raw Dataset:

This is the step where the collection of legit data or messages in multimedia like text, image from the public social networksis done. The dataset is in its original format without any modifications.

ii. Pre-processed Dataset:

The raw dataset thus received is organized based on the suitable file formats for preparing the dataset to be used in the platform thus considered based on the efficiency and usability.

iii. Processed Dataset:

In this the dataset obtained is a result of the combination of data vectorization and data normalization. This dataset obtained can be used to apply the required techniques.

iv. Feature Engineering Techniques:

This process makes use of ML to generate new values that are not present in the training dataset. It may include properties of both supervised and unsupervised learning in an effort to improve model accuracy while also streamlining and acceleratingdata conversions.

v. Model Specifications, Metrices for Comparison and Results:

In this step we select a suitable model and calculate the model's performance in terms of the metrics like confusion matrix and score. Later, we compare the results to analyze the approach avoiding the scenario of over fitting in the results thus obtained or repeat the process from data collection.

b. Text Analysis:

Initial step of collection of raw textual dataset is done. The collected dataset is processed in 2 steps a) Data Cleaning and b) Data Analysis. The Data Cleaning includes the process of Data Vectorization and Data Normalization. Data Analysis is done using the process called TF – IDF Vectorization.

i. TF-IDF Vectorization:

TF-IDF stands for Term frequency and inverse document frequency and is among the most widely used and successful methods of natural language processing. Using this method, you may gauge the term's value in relation to all other terms in a text collection. A phrase has greater significance for a text if it occurs frequently in that text and infrequently in all other texts. TF stands for Term Frequency. It is comparable to a normalized frequency score. The formula used to compute it is as follows:

$$TF = \frac{Frequency\ of\ word\ in\ a\ document}{Total\ number\ of\ words\ in\ that\ document}$$

We now determine how frequently a word occurs in relation to all of the other terms in a document based on the assumption that this number will always be equal to or less than 1. The opposite of document frequency is IDF, and thefollowing formula generates the final IDF score:

$$IDF = \log(\frac{Total\ number\ of\ documents}{Documents\ containing\ word\ W})$$

ii. Class Imbalance:

This situation occurs when the observations in one class is much smaller than the number of observations in the other classes. The prediction model created using traditional machine learning methods in this case may be unreliable and biased. This occurs as a result of the fact that ML algorithms are often created to increase accuracy by decreasing the error. As a result, they disregard the distribution, proportion, or balance of classes. Class balance's primary objective is to either increase the frequency of the minority class or decrease the frequency of the dominant class. This is done to ensure that each type has about the same numberof occurrences. To address this class difference, we use the Synthetic Minority Over-sampling Technique (SMOTE).

When precise replicas of minorities appear in the larger dataset, overfitting occurs, and SMOTE is employed to prevent this. The minority class's subset of data is used as an example before additional artificial instances that are comparable to it are produced. The initial dataset is then updated with these created instances. The classification models are trained using a sample from the fresh dataset.

After the step of feature engineering technique, classification models are used to perform the analysis on text dataset that is available. The classification models here include the Logistic Regression, Decision Tree, Random Forest and KNN Neighbors. The model's performance is evaluated using the Random Search CV method. It is a method where the optimal answer for the created model is found by using random combinations of the hyperparameters. It has consistently produced superior outcomes.

c. Image Analysis:

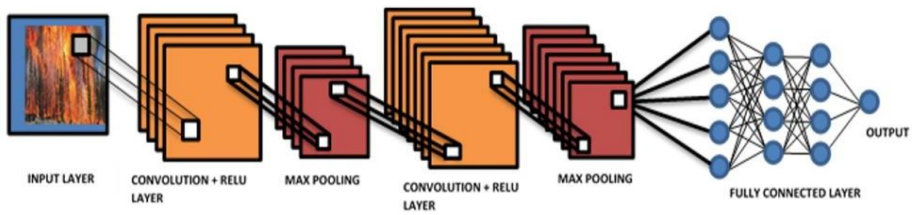
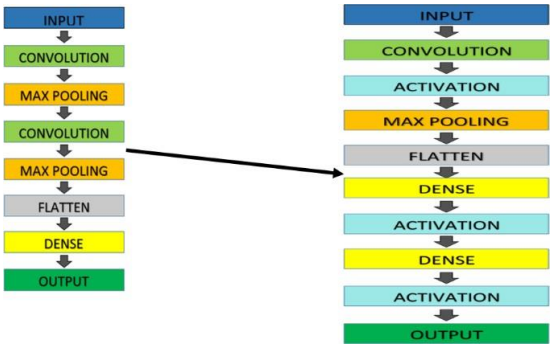


Fig 2. Block Diagram Of CNN

Raw picture datasets (both positive and negative) are first collected from social media. The Convolution Neural Network (CNN) is used to process the collected dataset. The CNN has 3 layers a) Input layer b) Hidden layer c) Output layer. The Hidden Layer has 3 sub-layers of process under it. Tools named Keras is utilized to perform CNN and ReLU to perform the activation function.

An open-source neural network library made in Python termed Keras can be utilized on top of TensorFlow or Theano. A full framework is provided by Keras to build any kind of neural network.



.

Fig 3. Refined model Architecture of Keras Tool

After the CNN is completed, Data Analysis is done using the process called Back Propagation Algorithm. By using a procedure known as the delta rule or gradient descent, this method identifies the weight space with the lowest value of the errorfunction. The learning issue is thus thought to have an answer in the weights that minimize the error function.

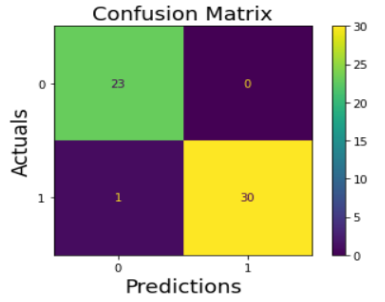
Finally, classification models are used to perform the analysis on text dataset that is available. The classification models here include the K-Nearest Neighbors, Random Forest, Decision Tree, and Logistic Regression. The model's performance is evaluated using the Random Search CV method. It is a method where the optimal answer for the created model is found by using random combinations of the hyperparameters. It has consistently produced superior outcomes.

Here, we have built classification models for image and text dataset which can be used effectively in performing the multimedia sentiment analysis.

VI. OUTCOME OF PROPOSED RESEARCH

i. Text Sentiment:

The confusion matrix for the classification models were generated. Random Forest Classifier was observed to givethe best accuracy compared to other classification models. The below figure shows the confusion matrix for RandomForest



	precision	recall	f1-score	support
0	0.96	1.00	0.98	23
1	1.00	0.97	0.98	31
accuracy			0.98	54
macro avg	0.98	0.98	0.98	54
weighted avg	0.98	0.98	0.98	54

Fig 1. Confusion Matrix for random Forest

ii. Image Sentiment:

The results can be further explained and stated in the form of model accuracy and loss for each number of epochsthus shown in Figure 2 for both Training and Testing accuracy and loss. In the Figure the Accuracy showed the validresults as there is increase in accuracy with increase in the number of epochs. But, in the case of Model Loss we can see a clear decrease in the loss which is the correct indication that the model i]s working well for the image's dataset collected in the data collection step and pre-processed.

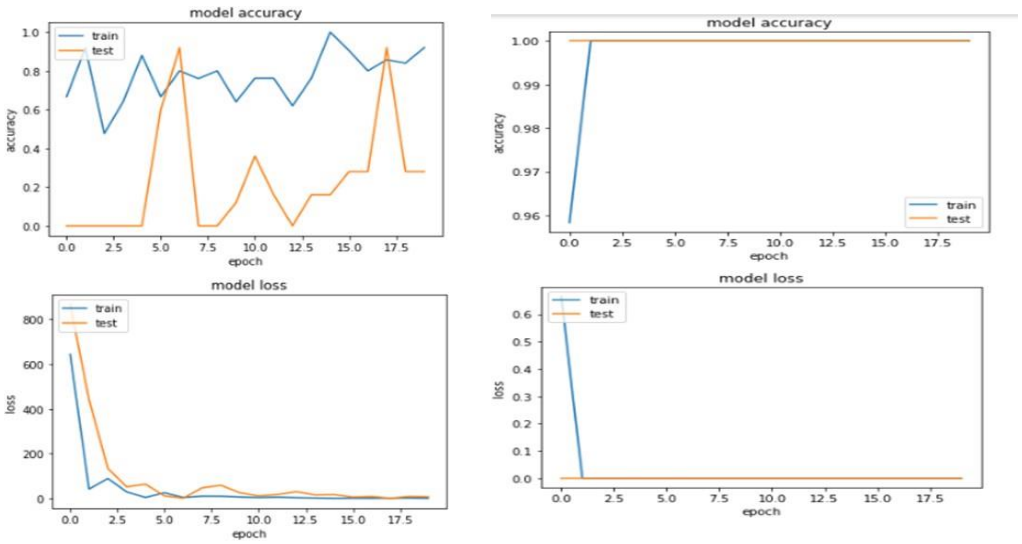


Fig 2. Training and Testing Accuracy and Loss

The confusion matrix for the classification model of image is show in the below figure:

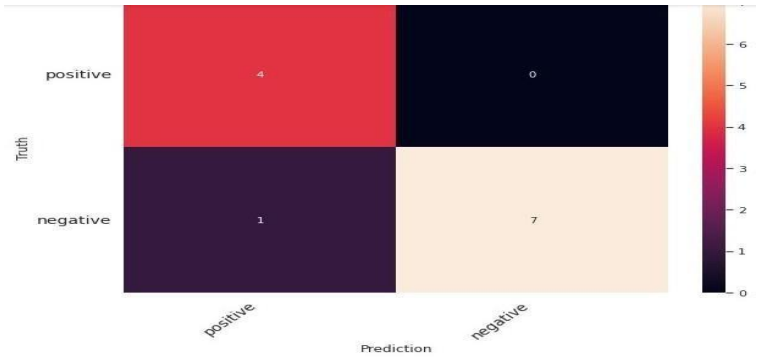


Fig 3. Confusion Matrix for image

VII. DRAWBACKS AND FUTURE SCOPE

Sentiment analysis cannot completely replace reading survey responses, however how crucial it is. Frequently, the comments themselves provide beneficial nuances. The time needed to finish the training is the deep neural network's final difficulty. With the number of hidden layers, the number of weights grows exponentially. More layers require large training data and increase training time. From a practical perspective, it is a serious issue. The deep neural network takes 5 days to train and we can modify it 6 times a month. To complete the training process faster, we can use high-performance hardware such as GPU and instead of CPU.

SOC reconfiguration to ensure the embedded assistance to control the hate speech from source as in smartphones cuts down the effort in monitoring and storage.

VIII. SUMMARY

Brands trying to understand more about the thoughts and feelings of their consumers have found sentiment analysis to be a valuable tool. It is a very straightforward type of analytics that assists companies in identifying their most important areas of weakness (negative attitudes) and strengths (positive sentiments). Sentiment research is currently gaining traction in various businesses. The best model was chosen because it performed well in terms of accuracy and precision and could, up to a point, avoid the disadvantages mentioned. The outcomes were appropriate for the setting indicated as part of the purpose, which is thekey benefit. While testing the correctness of the Images dataset, we found that it ranged from 92 to 95.83 percent. We also observed the accuracy of Approx. 98 % accuracy for training and testing text from twitter based on sentiments prediction.

REFERENCES

1. The complete guide to Sentiment Analysis, Rob Dumbelton, https://getthematic.com/insights/sentiment-analysis/
2. Google colab official website, https://colab.research.google.com/notebooks/intro.ipynb#scrollTo=5fCEDCU_qrC0>
3. IogrBobriakov TOP 10 NLP Algorithms and Concepts, https://www.datasciencecentral.com/top-nlpalgorithmsamp-concepts/
4. Tanmay Thaker, Kaggle Discussions, https://www.kaggle.com/general/262641
5. XENONSTACK,Stack Innovator https://www.xenonstack.com/blog/difference-between-nlp-nlu-nlg
6. Yuliyanti, Djatna & Sukoco. (2017), Sentiment Mining of Community Development Program Evaluation Based on social media , TELKOMNIKA,
7. Vol.15, No.4, December 2017, pp. 1858~1864ISSN: 1693-6930, accredited A by DIKTI, Decree No: 58/DIKTI/Kep/2013

8. Poecze, Ebster, Strauss & Christine (2018), Social media metrics and sentiment analysis to evaluate the effectiveness of social media posts ,https://doi.org/10.1016/j.procs.2018.04.117
9. Ramanathan & Meyyappan (2019), Twitter Text Mining for Sentiment Analysis on People's Feedback about Oman Tourism https://ieeexplore.ieee.org/document/8645596
10. Hanane Elfaik, El Habib Nfaoui (2020), Deep Bidirectional LSTM Network Learning-Based Sentiment Analysis for ArabicTexthttps://www.degruyter.com/document/doi/10.1515/jisys-2020- 0021/html?lang=en
11. Ghulam Mujtaba, Zahid Hussain Khand, Javed Ahmad Zafar Ali (2020), A Comparative Study of Sentiment Analysis on Mask-Wearing Practices during the COVID-19 Pandemic https://www.researchgate.net/publication/350961917_A_Comparative_Study_of_Sentiment_Analysis_on _Mask-
12. Wearing_Practices_during_the_COVID-19_Pandemic
13. Dimple Chehal Parul Gupta Payal Gulati (2021), Evaluating Annotated Dataset of Customer Reviews for Aspect Based Sentiment Analysis DOI:https://doi.org/10.13052/jwe1540-9589.2122
14. NicoVerwer,Plain textprocessinginstructureddocuments,http://www.cs.cmu.edu/~rcm/papers/proposal/proposal.html, (2020) MULTIMEDIA SENTIMENT ANALYSIS ON PUBLIC SOCIAL NETWORK USING NLP 2021-2022 Department of ECE, Dr. AIT, Bengaluru-56 Page 58
15. Justin E. Tang, Varun Arvind, Christopher A. White, Calista Dominy, Jun S. Kim, Samuel K. Cho, Amanda Walsh,- Using Sentiment Analysis to Understand What Patients Are Saying About Hand Surgeons Online, Using Sentiment Analysis to Understand What Patients Are Saying About HandSurgeons Online | ScienceGate , (2021)
16. G. Shyam Chandra Prasad, K. Adi Narayana Reddy, Sentiment Analysis Using Multi-Channel CNNLSTM Mode, https://www.jardcs.org/abstract.php?id=3847 , (2019)
17. Ho-Seung Kim, Jee-Hyong Lee, Sentiment Analysis Using Mixed Feature Vector combined with the Sentiment Dictionary Information, Sentiment Analysis Using Mixed Feature Vector combined with the Sentiment Dictionary Information | ScienceGate, (2020)
18. Jurgita Kapočiūtė-Dzikienė, Robertas Damaševičius, Marcin Woźniak, Sentiment Analysis of Lithuanian Texts Using Traditional and Deep Learning Approaches, https://www.sciencegate.app/document/10.3390/computers 8010004 , (2019)
19. Suman, Gupta & Sharma, Analysis of Stock Price Flow Based on Social Media Sentiments, https://ieeexplore.ieee.org/document/8520311 , (2017)