Sentiment Analysis on violence against children and women on social media and news comments in Bangladesh

BY

Azmain Iqutedar Anik
191-15-12988
AND

Md. Rayhan Mia
191-15-12991
AND

Md. Zia-Ul-Kabir Khan
191-15-12989

This Report Presented in Partial Fulfillment of the Requirements for the Degree of Bachelor of Science in Computer Science and Engineering.

Supervised By

Nishat Sultana
Lecturer
Department of CSE
Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY
DHAKA, BANGLADESH
JANUARY 2023

APPROVAL

This Project titled “Sentiment Analysis on violence against children and women on social media and news comments in Bangladesh”, was submitted by Azmain Iqutedar Anik, Md. Rayhan Mia and Md. Zia-Ul-Kabir Khan to the Department of Computer Science and Engineering, Daffodil International University, has been accepted as satisfactory for the partial fulfillment of the requirements for the degree of B.Sc. in Computer Science and Engineering and approved as to its style and contents. The presentation has been held on 26 January 2023.

BOARD OF EXAMINERS

Chairm

Dr. Touhid Bhuiyan
Professor and Head
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Exami

Sazzadur Ahmed
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Internal Exami

Ms. Sharmin Akter
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

External Exami

26.1.2023

Dr. Ahmed Wasif Reza
Associate Professor
Department of Computer Science and Engineering
East West University

DECLARATION

We hereby declare that this project has been done by us under the supervision of **Nishat Sultana, Lecturer, Department of CSE** Daffodil International University. We also declare that neither this thesis nor any part of this thesis has been submitted elsewhere for the award of our degree.

Supervised by:

(Nishat Sultana)
Lecturer
Department of CSE
Daffodil International University

Submitted by:

Anik

(Azmain Iqutedar Anik)
ID: -191-15-12988
Department of CSE
Daffodil International University

Rayhan

(Md. Rayhan Mia)
ID: -19-15-12991
Department of CSE
Daffodil International University

Zia

(Md. Zia-Ul-Kabir Khan)
ID: -19-15-12989
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First we express our heartiest thanks and gratefulness to almighty God for His divine blessing made us possible to complete the final year thesis successfully.

We are grateful and wish our profound indebtedness to **Nishat Sultana**, **Lecturer**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "Machine Learning (ML), Android Applications, and Artificial Intelligence (AI)" to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts,s and correcting them at all stages have made it possible to complete this project.

We would like to express our heartiest gratitude to Professor Dr. Touhid Bhuiyan the Head of**,** the Department of CSE, for his kind help in finishing our project and also to other faculty members and the staff of the CSE department of Daffodil International University.

We would like to thank our entire coursemates at Daffodil International University, who took part in this discussion while completing the coursework.

Finally, we must acknowledge with respect the constant support and patients of our parents.

ABSTRACT

Sentiment analysis is a natural language processing task that determines the sentiment or emotion expressed in a text. In the context of violence against children and women, sentiment analysis could be used to identify comments in news articles that express negative emotions or sentiments related to positive ones. This could potentially be useful for understanding public opinion or identifying language patterns commonly used to discuss violence against children and women. Crime is now gradually expanding in Bangladesh. Various newspapers, Facebook groups, Instagram pages, and youtube videos on social media posting abuse news in post people leave comments on their expressions. Comments are expressed by their writing style and news types. This report is a research-based project on Violence against children and women sentiment analysis on social media news comments. This report will analyze people's sentiments on abusive news comments. The research will detect the sentiment of positive or negative violent news comments. Some algorithms will apply to our collected data but this paper focuses on the LSTM algorithm. In this report, the dataset will collect our own collected dataset for an expected good result. Besides the LSTM this paper contains some algorithms beside our focus. Outside LSTM there was SVM, Logistic Regression, Naive Bayes, and Naive Bayes, Random Forest was used in this report.

TABLE OF CONTENTS

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Introduction

Crime is now gradually expanding in Bangladesh. Before the pandemic, situation crimes were controlled by the government but in the situation of covid-19, the record is passed through the previous record of the previous 10 years, from 2011-1013 violence decreased by 87% to 77 % but after the pandemic [1]. survey result of 27 out of 64 districts 2019-2020 just for the April of 2022 4,249 women, and 456 children are the subject who face domestic violence [2]. People of Bangladesh are now frustrated because of the market product price; some lost their job and others were in a financial crisis. The behavior of males is now aggressive toward children and women also cos of violating a woman or a child. on the other side, women now lack money to maintain their families on the other side husbands are forced sometimes to bring some money from their parents, and when they can’t they are violated by their husbands, as a result, children also sometimes fill in the violated situation of their family. Families are barking in this situation according to the photom alo news most of the victims of families barracked for the financial crisis. when a family breaks their financial, and mental strength, violations started to break their family. Various newspapers, Facebook groups, Instagram pages, and YouTube videos on social media posting abuse and domestic violence news posts people leave comments on their expressions. Comments are expressed by their writing style and news types. This report is a research-based project on Violence against children and women sentiment analysis on social media news comments. This paper will analyze people's sentiments on abusive news comments. The research will detect the sentiment of positive or negative violent news comments. Now (Long short-term memory) LSTM was most popular for its classification which was the updated version of RNN (Recurrent Neural Networks) [3]. Some algorithms will apply to our collected data but this paper focuses on the LSTM algorithm. In this report, the dataset will collect our own collected dataset for an expected good result.

1.2 Motivation

Understanding public opinion: By analyzing the sentiment of comments related to violence against children and women, it may be possible to get a sense of how the general public feels about this issue. Understanding the prevalence of violence: By analyzing the sentiment of comments related to violence against children and women in Bangladesh, it may be possible to get a sense of the extent to which this issue is present in the country. This could be useful for policymakers and advocacy groups seeking to understand the scale of the problem and develop strategies to address it. This could be useful for policymakers, advocacy groups, and others who are working to address this issue. Since from 2016 to 2020 by observing violence against women a decent amount for that case day by day crimes against women and children are being controlled by the government arm police battalion but at covid-19 pandemic which was expanding gradually on the other side if you look at the financial crisis Saudi foreign women of Bangladesh are suffering from violence government try to back them at country but the violation now spread around the whole country while any author and publisher publish their violence contain on social media or the news like violence publish people of Bangladesh express their valuable filling on that for reason various types of comments arrives like negative types, positive types or the neutral types means positive, negative or neutral. Neutral means the textual comments were not on the topic of the news. This type of data was needed to analyze the sentiment of Bangladesh people.

1.3 The rationale of the Study

In This study needs to express that thinking about violence against women and children of the Bangladeshi people. Understanding the prevalence and nature of violence against children and women in Bangladesh: By analyzing comments related to this issue, it may be possible to get a sense of the extent to which violence against children and women is a problem in Bangladesh, as well as the forms that this violence takes. Identifying contributing factors: By analyzing the language used in comments related to violence against children and women in Bangladesh, it may be possible to identify factors that contribute to this issue in the country. This could include cultural, social, or economic factors, among others. The study needs to on emotion and expression of the comments on social media everyone expresses their feelings. For social media where anyone expresses their opinion about what to do and why for that based on their comments this study needs to sentiment analysis on the social media textual comments.

1.4 Research Questions

Some possible research questions could be addressed through sentiment analysis of comments related to violence against children and women in Bangladesh.

1. How prevalent is violence against children and women in Bangladesh, as reflected in news comments?

2. What are the most common themes or patterns in the language used to discuss violence against children and women in Bangladesh?

3. What factors seem to be most strongly associated with violence against children and women in Bangladesh, as reflected in the sentiment of comments?

4. Are there any specific geographic areas in Bangladesh where violence against children and women seems to be more prevalent, based on the sentiment of comments?

5. How do public attitudes towards violence against children and women in Bangladesh change over time, as reflected in the sentiment of comments about?

1.5 Expected Output

This study expects to find out the sentiment of the texts in that people express their filling against violence were applying deep learning and machine learning algorithms to teach the system to classify that sentiment as negative or positive. The system needs to classify those sentiments via the classifying algorithm, after that, the system needs to learn according to the sentiment text collected in this study. Enhanced social media monitoring: As in other countries, analyzing the sentiment of comments related to violence against children and women on social media in Bangladesh could be useful for organizations working to prevent violence and abuse, particularly if they are able to respond to incidents more quickly as a result. By identifying areas where the risk of violence against children and women is particularly high, based on the sentiment of comments related to this issue, it may be possible to prioritize resources and efforts to prevent violence in these areas. It was believed that this study would be able to identify the line by those models from the text.

CHAPTER 2

Background

2.1 Background studies

Violations against women and children are quite rare in Bangladesh. LSTM was often frequently used to characterize feelings in texts. For a good understanding of the LSTM, one must first comprehend the RNN. Recurrent neural networks (RNNs) are primarily based on the idea of memory, which may be used to analyze sequential input and develop long-term relationships. its output is produced based on the results of earlier calculations.[4]. A deep learning-based method called LSTM proved successful in predicting the emotion of the text [5]. With long-term dependence, it is possible. it has a neural network that can process text data. [5]. In order to collect routine blood analysis parameters, Mücahid Mustafa Saritas and Ali Yasar used the Naive Bayes and ANN techniques. Naive Bayes provides a very detailed explanation of how to calculate the Bayes thorium by loading data from the calculation of the confusion matrix to generate the probability of a data set. [6]. Hare Sarita's studies Nave Bays for the example of their study generate a great understanding of the properties of their algorithm working alignments based on the bay's theorem manage the P(A|B) where probability calculation determines the outset of total class [6]. By dividing the probability by P(B) calculate the P(A) and multiply the P(B|A) to generate the predictive outset for the prior property of classification with the marginal property of class probability how do they generate the Posterior probability to the calculation?

On the other side, its Logistic Registration technique is used by Menard [11], and draisaitl [7] where they discuss the sigmoid function working mechanism to explain how to produce output through the mathematical term of the bay's functionality mapping value and the range functionalities. The curve size and the shape would be the s curved shape function determined by them in this report. Threshold calculation by devising those such as the 1 to divide the maximum classes that we can find the probability of the sophisticated data set the probability to minimize further ai by the deep learning technique's trends to calculate the probability of measurement through 0 as the minimum value or its maximum prediction goes up to a total one there for higher value gets the high priority for its detection of the separated class [12] that it needs to be targeted to determine the value as model needed by work Jakkula [9] a great studied technique is shown.Menard [11] delivers a pitching technique of study by defining the functionality of a Random Forest where trains data multiple line inputs by through the tree of the decision as they calculate the value of texts can provide a tree as they are building for to make then pass the decision trees as their outgoing parameters as their shown thor studies [11]. Gathering all value from the tree calculation start for the best of the results as they needed to select by the average of the voting section where tree numerical value is calculated by the maximum value for predicting the class and other hand removing overfitting the textual classification model by expressing useful study this report inspired [13].

Mechanisms like the logistic regression of Zhu describe the multiple logistic regression can work that understand the need to learn working flow by knowing single work at a time predicting the founded value from data district numbers was necessary for regression method that works on machine by learning them by solving classification hare according to them logistic regression calculate some values based on a total number that can divide as a class where classes value depend on the number of s shape sigmoid function by predicting classes [14].

A deep study on tokenizers understands the basic ai most components described by Webster [16] In his study tokenizers convert the whole textual text into a large number of arrays that was defined as a token. Tokenizer converts a natural language into a floating number of the token where the token passes through the embedding process that supports the vector formatting for analyzing text [5].

By LSTM text classification will be tokenized by the tokenizer as the word separate from the sentence. Tokenizer will tokenize and convert the word to an integer value after that embedding the integer value to convert real-valued vectors then LSTM process sequence of arbitrary length as a final output SoftMax output layer which calls dense output layer predicts the output of negative or positive sentiment of a text.

2.2 Comparative Analysis and Summary

A better understanding of the machine learns learning system was to practice machine learning to go on deep learning method. Machine learning methods like SVM, Random Forest, and naive bays are the meshing learning classification whereas LSTM was a combination of machine and deep learning classification for the learning systems. Analyzing the data was better for statistical analysis on machine learning and whatever machine learning algorithm runs faster than the deep learning algorithm but, on another side, the deep learning algorithm was better to learn a system because of better accuracy and better training for a model. Machine learning algorithms easier to understand the nodes and the flow of data but the complexity of deep learning hidden lays and the calculation methodology.

2.3 Scope of the Problem

Hare's scope of the problem can be they will be for the shortest of data neural network was there to pass through them but the data will not be enough to learn the system properly for the season can predict wrong decision. That makes the study invaluable in that case deep learning algorithm can help to generate a better output case between low data. Approximately deep learning needs a vast amount of data that will process and some data will fall in loss but the loss amount was a problem to maintain accuracy in that case to decrease that loss amount neural network and model need to be accurate to better performance for the textual language processing.

2.4 Challenges

Have challenges when anyone tries to do some work that wasn't before done by anyone as this work. This work was done with our own data set. We fill challenges where we tried to find the sentiment data of this work hard at Bangladesh prepaid or existing data was not available for the test, That's why we need to spend time day by day to collect that news and comment by selecting them with their proper meaning of sentiments. For data collection in Bangladesh abuse, and news comments are not precise some of the data are types of positive words but according to the news that was a negative word this was a challenge for our algorithm to learn whether the sentence was negative or positive. where the challenges of the data collection also training the right expression for the algorithm most challenging work for the traditional algorithm like LSTM, SVM, Logistic regression, etc. If a challenge does not come to life there was no solution for that one more challenge was to collect data for uniqueness. In Bangladesh, the comet type of the people was as same types such that for news according to the comments there says about six or seven times for Allah to help them or punish them, that’s why collecting unique data was very hard for us in this case. On another side, the news was limited have to wait for the next news because when news publishes there need someday to comment that's why collecting data was very slow to complete the overall work.

CHAPTER 3

Research Methodology

3.1 Data Collection Procedure/Dataset Utilized

In this paper we collect data from social media news pages, groups newspaper line prothom alo, shomoy tv, daily star, Jamuna tv, etc. on other side collected data are also collected from Facebook, Instagram, YouTube, and Twitter posts. As we know the national language of Bengali was Bangla, in social media Bengali use their native language for posts or comments. Bangali use the Bangla language to express their filling or the English language mixed with Bangla it's called "banglish" language for social media so we convert those lines from Bangla to English by ower own typing and others where the comments are in English with the help of google translate, we convert those line Bangla to English. Our LSTM model work in the English Language so conversion was necessary to identify the sentiments of the line that want to express through the comments. Data are stored in Excel format contained in data news publisher/news channel/newspaper name, Date of that news publisher, and Headline of the posted news, news types are in 4 categories Eve teasing, Woman Abuse, Sexual assault, and Domestic violence. Columns of sentiment text and sentiment types are side by side better to understand the sentiment type's positive, negative and neutral expressions. in this report data are collected on our own by searching the news post and videos on social media to collect the unique comments, the collected comments are taken from every news at list 5 comments. Data are between 2015 to 2022 most of the data are recently collected between 2019 to 2022 because this time was the pandemic situation of covid-19 most of the crime occurs during this period. The data collection process is searching news related to the abuse of women and children, taken from the news title was defined by the news abuse type that Eve teasing, women abuse, sexual assault, or child abuse. by defining the type based on comments understand whether the comments type is positive or negative. data are arranged by the row and column of a person's comment. To apply any algorithm data, need to arrange row and column for that row and column maintained in the dataset to better understand the comment type according to the sentiment of the news. The comments do not use a fixed length according to the news comments arranged on the data set. The length of the comment is between one word to 30 words with their actual meaning translation work done very precisely and accurately to identify by the algorithm easily. Most of the comments are similar words used and in some of the news portals, all comments are the same types as that type was collected previously so that comments are avoided. Here, we also used sentiment data collected from the kaggle to justify our work because of the shortest of our data.

data preprocessing was a good strategy for learning a system there are many unnecessary characters that stay in data that need to be removed. If there are any null values or empty data found then need to drop that row to clean the data unnecessary characters like “# ?!) +=-” etc. will decrease the learning capabilities of the model that will train.

LIST OF DATA

Data	Year	Quantity	Comments type
Sentiment	2017	13871	Twitter Sentiment
Synthetic Data	2022	3577	social media abuse news sentiment

3.2 Statistical Analysis

statistically significant analysis was on the data set of our own that the data contain three types of sentiment but for classifying those hares used two types of data one ware positive and the other negative from our analysis we get that most of the people's comments like textual data of our datasets are accurately precise. For the data of sentiment which contained Twitter data are total of 13871 data in this 2236 data were positive and 8493 data were negative. Hare negative data contain more than the positive data where neutral was about 3142 so data was not well-balanced but for the system, it’s enough to contain data.

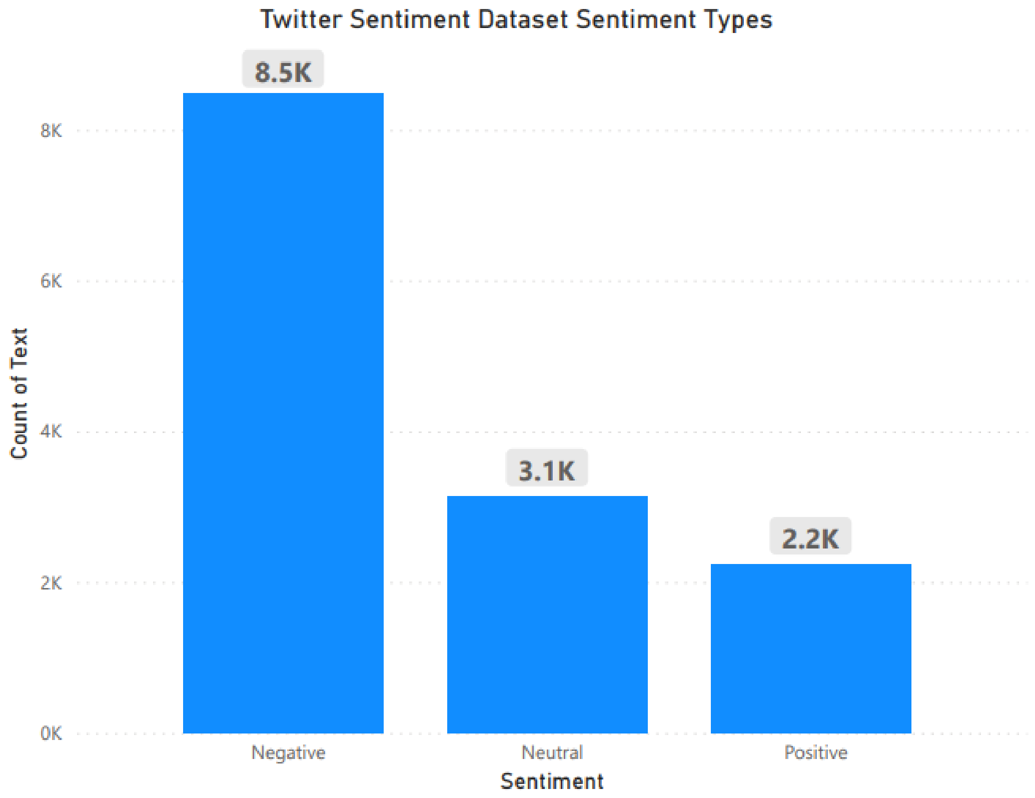


Figure 3.1: Twitter data by sentiment.

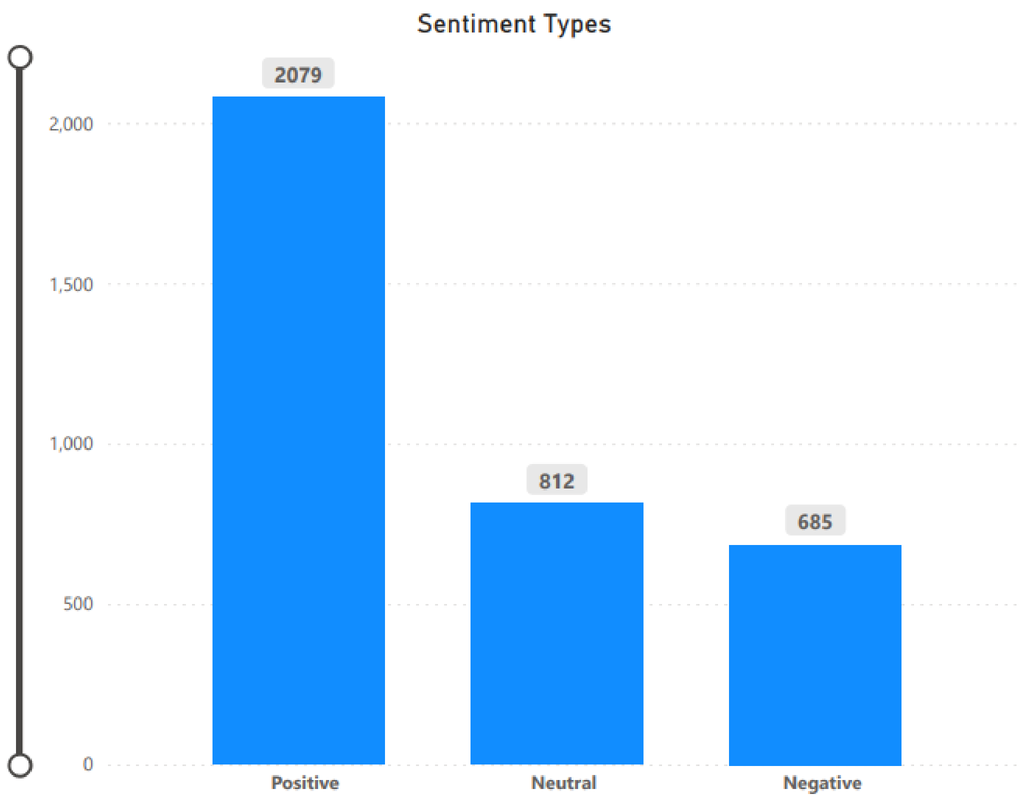


Figure 3.2: Synthetic dataset collected by own contain sentiment.

3.3 Implementation Requirements

Where implementation requirements are the python knowledge about the machine learning functional knowledge are the basic requirements for high-level requirements need to learn the working technique of the nodes hidden layers tokenizer, embedding, and SoftMax function for understanding the neural network build like RNN LSTM work similarly to the but LSTM work better the RNN. To understand those who need knowledge about the working platform colab. Hare easily can run those for making a project.Colab library was now on python version three that's why the functions can be easy to understand. By side need the requirements of code manipulation technique to manipulate them as they need. For this whole work need the data to train them in our workstation data needs to prepare for analysis.

3.4 Proposed Methodology

1. Naive Bayes

Classifier like Naive Bayes is capable of classifying by Bayes Theorem its not a single algorithm it has its own different algorithm with a total its call the Naive Bayes. It's a probability classifier that can calculate the probability of the output by counting the frequency and combination values that exist in the data set. Hare all variables are independent as their own considering the value of the class variable. Algorithms learn quickly from the data by probability equation on Bayes Theorem [6].

$$P(A|B) = \frac{P(A)\,P(B|A)}{P(B)}$$

By applying the Naive Bayes, the confusion matrix was the way to predict the output of a particular data set.

2. SVM (Support vector machines)

When classification and nonlinear characteristics are spatially mapped to the kernel function, support vector machines are supervised learning methods [8]. For solving the optimization problem linear classifiers on SVM use the machine learning technique to predict and classify the classes Also pixel mapping and picture-based classification are vastly used for it classification pixel mapping technique. Handwriting recognition SVM many applications were useful for object classification on pictures, face analysis, and pattern classification are the features for separating those objects from all [7]. Traditional machine learning algorithms like the SVM solve the classification problem as well as the regression problems. multi problem-solving capabilities makes it separate from others.

It's a high dimensional feature space for Vector machines defining as well as linear classifiers making it a good usable algorithm for the world. Choosing SVM can work with unsupervised and supervised also for machine language processing properties including multiple input and output make special and favorite algorithms for those who use to predict multiple classes [9]. SVM divided the data with a line or a space on that classes are separate from each other to better identify class values and noise data are needed to remove in that case for smooth classification. In The kernel, Data are in High-Dimensional

Space If Any Data is Interstate Maximize values are needed to optimize properly to classify classes by separating them.

3. Random Forest

Machine learning techniques like the random forest are used for regression and classification by its concept of ensemble learning technique on working with multiple problems and used to improve performance by solving complex problems. In a random forest, classification using multiple numbers of decision trees was used to improve the accuracy of the data. Random forests take different types of predictions from the different trees for that by calculating the majority votes of predictions to take the best one to predict the class. Random forest is supervised learning that's like a supervisor as work majority calculation is done by itself for avoiding overfitting issues. over the calculation of the decision on the hand of the random forest by calculating the major part itself to overfit avoiding that makes it the most in-need algorithm for all. Providing a unique combination of prediction based on itself can predict all features of prediction for better understanding the building model to get maximum output of accuracy by training the model. For feature selection techniques to avoid overfitting increase, the model accuracy goes deep into the model. Random Forest working sequence was to take input multiple training data to analyze the data to produce a decision tree for each input calculation of voting averaging the tree to make a decision.

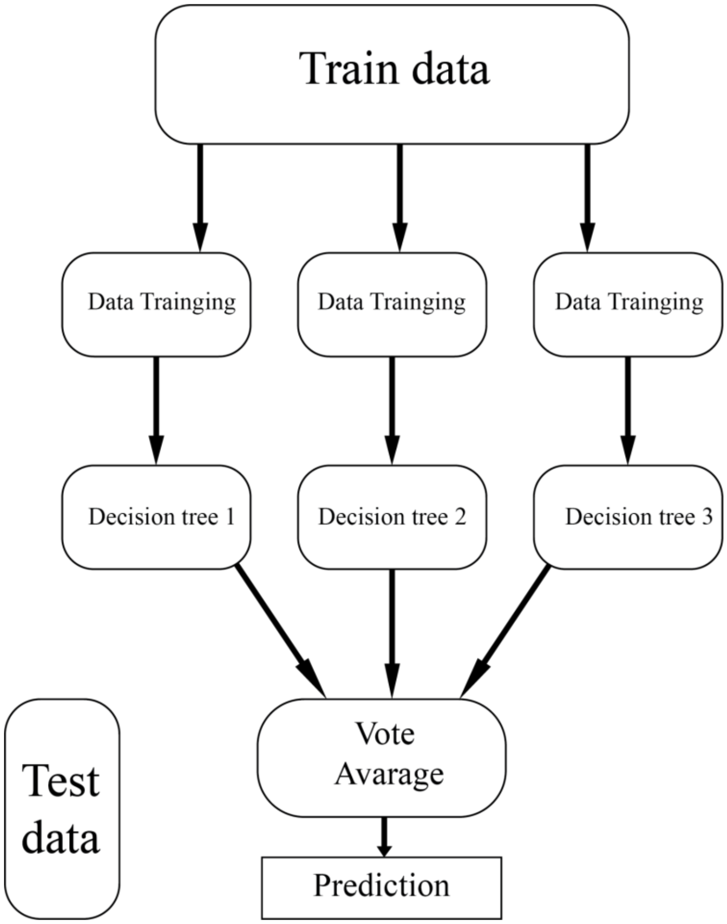


Figure 3.3: Random Forest training data to output prediction.

4. Logistic Regression

Predicting was a categorical dependent variable hare Logistic regression was one of the best algorithms to detect independent variables. Its categories the dependent variable for that output was the absolute value of districts are either two-sided, yes or no categories It's similar to the linear regression for the working method regression method solving was the main purpose of this algorithm. Predicting the maximum value of total one can be the to calculate function indicated likelihood calculation makes Logistic Registration a capable algorithm to classify those classes and determine the most effective way to use the classification. the sigmoid function of the metaethical function term that is used to map the predicting value of the probability [11]. Logistic Regression stays at its limit which does not go up to its limit curve s formatted structure produced by the sigmoid

function. where y=Predefine and x can increase by its shape where X and Y interstate the curve behaves like linear that's why the dependent variable is in natural values. Multiple regression can apply to classify those classes which one was targeted to classify [11].
Like the Supervised Learning technique as well as predicted Logistic Regression popularity of the algorithm was around the data science, categorical dependent variables are capable of finding the output of categorical discrete value where the value can be 0 or 1 based on the data classification of the sets. It's similar to Linear Regression but performs according to solving regression problems. To solve a classification problem Logistic Regression Work pretty well-known algorithm. It uses a parametric method where some time is called the semi-parametric or non-parametric artificial neural nets. The distinction is important because the parameters in logistic regression contribution can be interpreted but not always neural network interpreted by parameters [7]. Hare Multinomial classification can find multiple classes where we can find the three basic options to classify the three sentiments.

5. LSTM (Long short-term memory)

Long short-term memory (LSTM) was the updated version of RNN-based architecture that works on deep learning techniques. The networks are suited for classifying the text process and predicting static data.LSTM was designed as the traditional RNA network benefits of LSTM using the hidden Markov model and other numerous applications. There are many different architectures of the LSTM.LSTM architecture consists of a cell unit (memory part of the LSTM) [3]. LSTM unit has three regulators called units input gate, output gate, and forget LSTM doesn't have more than one gate. Processing the sequence data makes it easier to work with LSTM working methods. Based on requisition neural network can provide better accuracy to this textual work to identify that two sentiments were negative and positive expressions contained. For solving the complex problem sequential data LSTM used a vast amount there where peach to recognize and text to classify. Individual time steps can be the problem-solving solution for machine learning and deep learning. Words count in sentiment was the efficient solution for those where recursion neural network.

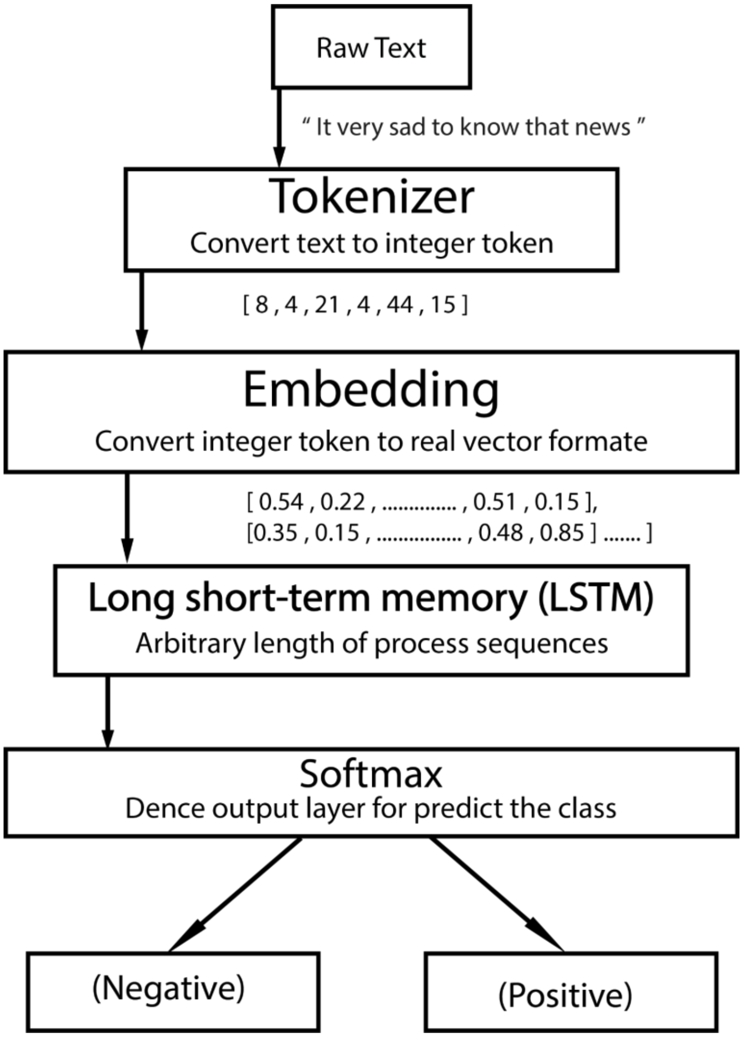

Raw Text
" It very sad to know that news "
Tokenizer
Convert text to integer token
[8 , 4 , 21 , 4 , 44 , 15]
Embedding
Convert integer token to real vector formate
[0.54 , 0.22 , ... , 0.51 , 0.15],
[0.35 , 0.15 , ... , 0.48 , 0.85] ....]
Long short-term memory (LSTM)
Arbitrary length of process sequences
Softmax
Dence output layer for predict the class
(Negative)
(Positive)

Figure 3.4: LSTM workflow figure indicates the working step.

A. Raw Text

Using our own dataset, the total dataset contains positive, negative, and neutral sentiment texts which by training and validation of our model to learn the sentiment of violence. To learn the model data, one needs to pre-process the case text needs to lower case on the other side needs to remove punctuation, and have to remove those characters that were not used in the language or represent an expression like any symbol that replaces the empty space.

B. Tokenizer

Tokenizer was the step to token the string and the string will be the individual textual token. the token can be taken from a sentence and the sentence can be a token from a paragraph. Tokenizer was the most useful option for NLP but taking a token as a sentence or paragraph was most useful for analyzing the sentiment of any sentence [17]. The tokenizer splits a string into a block of words. sentiment sentences need to separate from a sentence to classify to represent a single cluster. The Collections library assigned lower indexed words for indexing to Create Vocab to Int mapping dictionary.

Next was to encode the sentiment textual word the encoder encodes the word to a block of integer or floating numbers by creating a list of numbers. Each comment's word text will be just a list of numbers and every word makes the list a huge numerical list.

Encoding a table will generate the two outputs because we are working on 2 classes negative or positive that represent negative and 0 and positive representing 1. Classes were allowed to vectorize by turning each text into a sequence. The vector represents the binary-based word.

C. Embedding

Word embedding was used as the neural text processing. In AI (Artificial Intelligence), ML (Machine Learning), and DL (Deep Learning) embedding was the text-mining technique for NLP. Embeddings were dense vector representations that capture relationships in language. The syntactic and semantic words come from the context in which Embedding is used. The suggested word is semantically similar. counter based embedding and prediction-based embeddings are used in the embedding to embed a word. The embedding layer converts any integer values to dense vectors of length 128 [2]. For the input length input sequence was the max length. vocabulary was the most frequent word used in embedding. For output dimension, dense embedding is used as a vector space where words are embedded. In that way, the related words stay together or another word can say that close enough to represent a relationship between them.

A framework like TensorFlow and Keras represents the handling of the indexes of the dense vector.

D. Embedding Layer

Fore language processing word embedding and neural network models are learned together for a particular language processing or document classification. For embedding, layer text needs to be clean as it can be for encoding. an embedding layer has been used at the end of the neural network. A lot of embedding layers are the cost of slowing down the work for the work of text classification or NLP text identifiers.

E. Using Word Embedding

- Hare word embedding was the way to use neural network processing of the model to train it properly. How to learn an embedding was the way to understand the data being processed that can say how much data will processed word can be in million or could in billion so the option needs to be maintained when building or training.
- Embedding will determine whether the model was trained or not; it saves the information and passes through the information for multiple model embedding.
- To Understand the embedding, you have to learn the embedding working mechanism and modeling jointly for giving a tusk for a specific model.

F. Embedding Reusing

was an option to use a pre-train embedding that also has permission to use them on any project like world VCC times. It's better than learning or building your own embedding to reuse those on the project for better output and time-saving. Overall, pertaining embedding can be used as the embedding reusing on the project for time improvement and unnecessary interruption in building the embedding.It was the honest option to see the output of the embedding during the updating pre-trained embedding of the model which will be an acceptable approach to determining whether embedding was good and suited for your problem-solving.

G. SoftMax

The SoftMax function was a function that calculates the chances distribution of classes and its works for all possible target classes, it works with the probability of 0 and 1 for that reason it's not classified as just o and 1 or yes or no that's not the way of SoftMax we call distributed the SoftMax to identify those classes. An example you have three classes but you can only predict 2 classes 0 or 1 which is not right soft max hare one min total values are one if we can divide them into several portions just like output = [0.50,0.25,0.25] it can be divided into three classes for classifying that class. SoftMax can work with multiple classes of the trained model. It deals with the higher value of the target classes class could be a multiple of target classes the higher value will go the better probability of the other values of the classes. For identifying an object such as a flower or an animal by a picture it might work individually to classify those objects by every time given inputs as the picture but for the SoftMax, it can be easier to identify those objects with their multiple classes using the neural network works will be in dynamically not in a single layer.

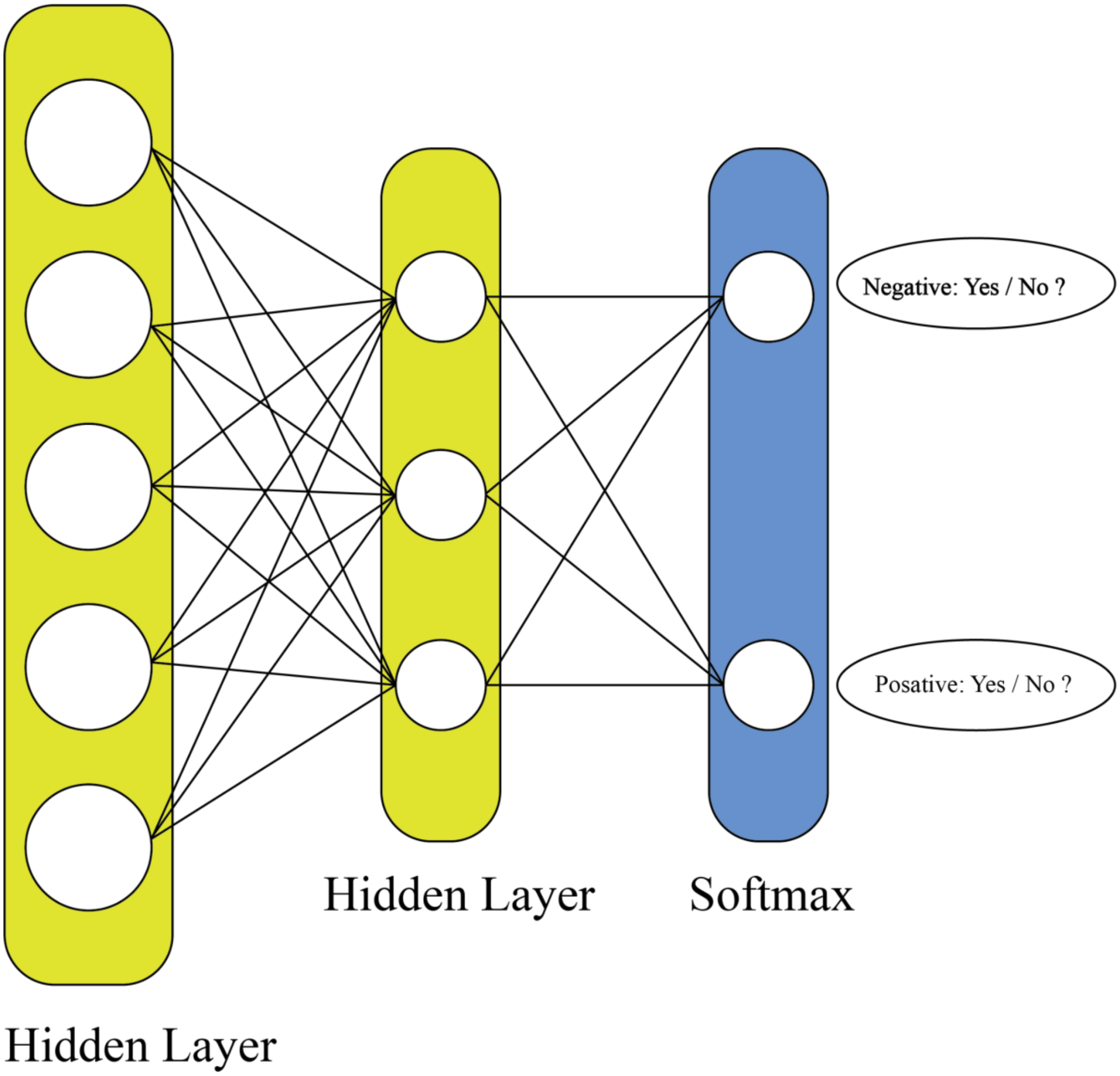


Figure 3.5: SoftMax working with hidden layers.

For all methods, there is SoftMax that doesn't need to but limits the scope of its calculations for a single class. For example, determining an animal to a particular set of classes probably doesn't calculate for each different class. SoftMax remembers the multiple classes, and SoftMax assumes one object for the classes where an object belongs to multiple classes SoftMax doesn't work on that particular work then it uses logistic regressions multiple times.

There are some properties of the SoftMax function that are used to calculate the probability of the output classes.

- The probability calculation of the SoftMax function between the 1 to 0 range.
- Sum of the total probability will be 1 of all target classes.
- Use logistic regression for multiple classes in the model.
- For building the neural network for SoftMax it takes to use different layer levels.

H. Algorithm

Recurrent neural networks can be implemented by LSTM architecture step by step.

First of all

- Load in and visualize the data shape, quantity, and classes.
- Data preprocessing by removing Punctuation from every text comment.
- Encode the text word block and label those.
- Split data into three sections: test, valid, and training data.
- By defining the LSTM architecture, build the model.
- Training the network using the dataset.
- testing (Test the data to generate accuracy).

I. Working of LSTM Network

1. Take input from the previous hidden state and previous internal cell to prepare the calculation.
2. Every node will calculate the input and the previous input of the parameter vector as a hidden layer; all the nodes are connected to watch others for respect weights for each gate. The respective activation function for the gate all get activated with the activated function.
3. First calculate the internal cell multiplication vector will be calculated element-wise as a sequential input gate and the input modulation gate last calculate the forget gate and the previous internal cell will add as a marge vector.

4. Hidden state calculation was element-wise and for the current internal cell will calculate by vector and then perform element-wise with the output gate by multiplication them.

Like RNN, LSTM produced output continuously, and the output was utilized to compute each stape. Each stape was then used to train the neural network. In that particular instance, RNN was the back-propagation technique used in LSTM term memory networks, which is based on a mathematically fictitious turn.

CHAPTER 4

Experimental Results and Discussion

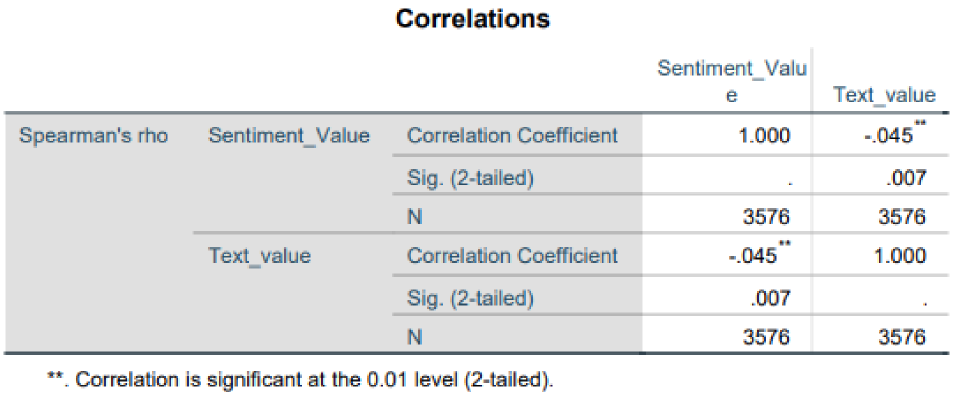
4.1 Experimental Setup

Need to be set up for this study was to identify textual comments to identify sentiments that need the environment to set up. This project based on python for that environment needs to be set up for most work done on the Google colab so that a Virtual environment like an online colab provides the full setup for itself hare python 3 is used on colab where sklearn, TensorFlow like function libraries available to work with them. All the algorithms need to be called by for that reason it helps to call those at a single place with those who manipulate our textual data was easier to work on that. Necessarily required libraries are able to this environment and easily playable with those. That case from all-direction colab using was the easiest environment for the experiment. Work with StackOverflow that's why any error Oscars are easily solvable for any problem while coding the experiment. Providing all versions of the libraries by that environment by calling them we can easily manipulate any data by it. Machine learning, deep learning, and artificial ai types of work are easily done by the environment where its development the neural network visualization but they are also easier for matplotlib graph by the environmental setup on the colab.

4.2 Experimental Results & Analysis

Algorithm	Data	Quantity of Data	Positive Data	Negative Data	Neutral data	Accuracy
LSTM	Sentiment	13871	2236	8493	3142	85 %
SVM						39 %
Logistic Regression						41 %
Naive Bayes						34 %
Random Forest						24 %
LSTM	Abuse Sentiment (own)	3577	2072	701	803	79 %
SVM						25 %
Logistic Regression						29 %
Naive Bayes						31 %
Random Forest						20 %

Correlation Report

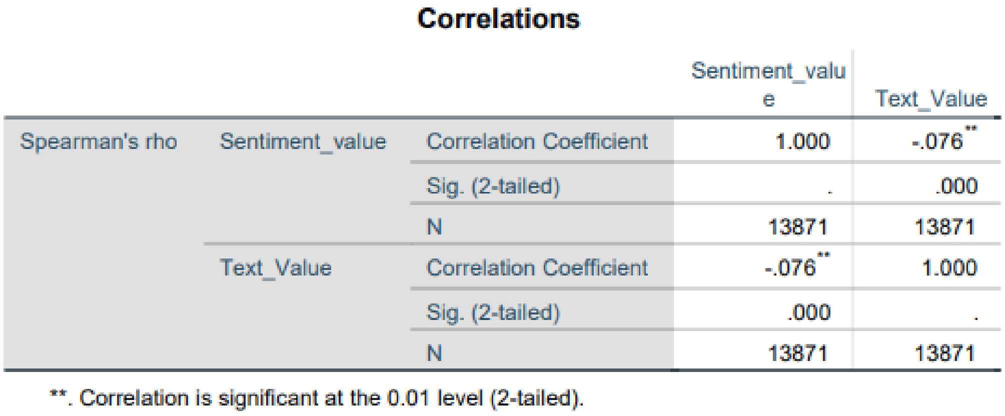


Correlations

			Sentiment_Value	Text_value
Spearman's rho	Sentiment_Value	Correlation Coefficient	1.000	-.045**
		Sig. (2-tailed)	.	.007
		N	3576	3576
	Text_value	Correlation Coefficient	-.045**	1.000
		Sig. (2-tailed)	.007	.
		N	3576	3576

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 4.1: Correlation Report by SPSS spearman model on Twitter sentiment data.



Correlations

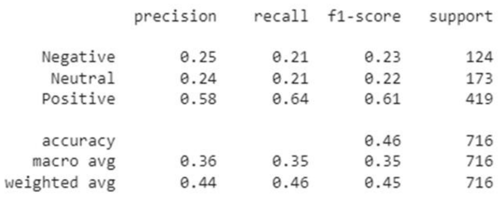
			Sentiment_value	Text_Value
Spearman's rho	Sentiment_value	Correlation Coefficient	1.000	-.076**
		Sig. (2-tailed)	.	.000
		N	13871	13871
	Text_Value	Correlation Coefficient	-.076**	1.000
		Sig. (2-tailed)	.000	.
		N	13871	13871

**. Correlation is significant at the 0.01 level (2-tailed).

Figure 4.2: Correlation Report by SPSS spearman model on synthetic sentiment data of own.

Synthetic data

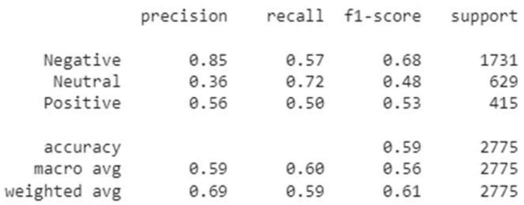
Twitter Data



```
              precision    recall  f1-score   support

    Negative       0.25      0.21      0.23       124
     Neutral       0.24      0.21      0.22       173
    Positive       0.58      0.64      0.61       419

    accuracy                           0.46       716
   macro avg       0.36      0.35      0.35       716
weighted avg       0.44      0.46      0.45       716
```

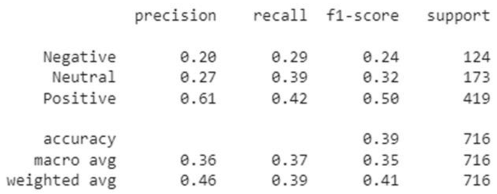


```
              precision    recall  f1-score   support

    Negative       0.85      0.57      0.68      1731
     Neutral       0.36      0.72      0.48       629
    Positive       0.56      0.50      0.53       415

    accuracy                           0.59      2775
   macro avg       0.59      0.60      0.56      2775
weighted avg       0.69      0.59      0.61      2775
```

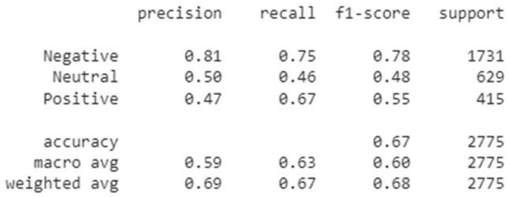
Naive Bayes



```
              precision    recall  f1-score   support

    Negative       0.20      0.29      0.24       124
     Neutral       0.27      0.39      0.32       173
    Positive       0.61      0.42      0.50       419

    accuracy                           0.39       716
   macro avg       0.36      0.37      0.35       716
weighted avg       0.46      0.39      0.41       716
```

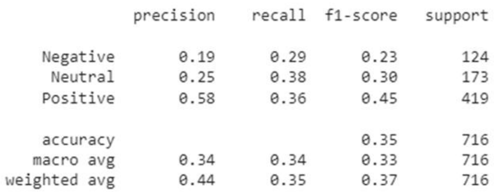


```
              precision    recall  f1-score   support

    Negative       0.81      0.75      0.78      1731
     Neutral       0.50      0.46      0.48       629
    Positive       0.47      0.67      0.55       415

    accuracy                           0.67      2775
   macro avg       0.59      0.63      0.60      2775
weighted avg       0.69      0.67      0.68      2775
```

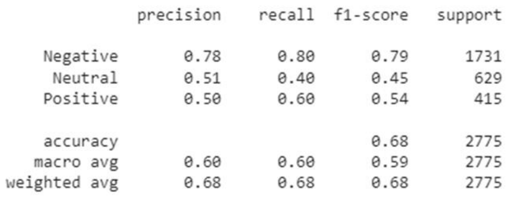
Logistic Regression



```
              precision    recall  f1-score   support

    Negative       0.19      0.29      0.23       124
     Neutral       0.25      0.38      0.30       173
    Positive       0.58      0.36      0.45       419

    accuracy                           0.35       716
   macro avg       0.34      0.34      0.33       716
weighted avg       0.44      0.35      0.37       716
```

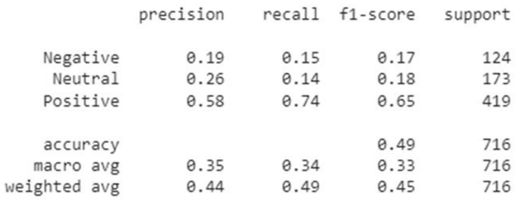


```
              precision    recall  f1-score   support

    Negative       0.78      0.80      0.79      1731
     Neutral       0.51      0.40      0.45       629
    Positive       0.50      0.60      0.54       415

    accuracy                           0.68      2775
   macro avg       0.60      0.60      0.59      2775
weighted avg       0.68      0.68      0.68      2775
```

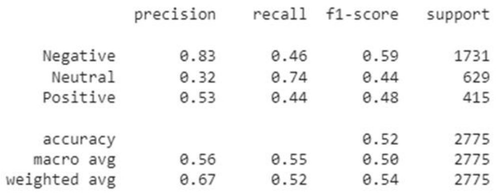
SVM



```
              precision    recall  f1-score   support

    Negative       0.19      0.15      0.17       124
     Neutral       0.26      0.14      0.18       173
    Positive       0.58      0.74      0.65       419

    accuracy                           0.49       716
   macro avg       0.35      0.34      0.33       716
weighted avg       0.44      0.49      0.45       716
```



```
              precision    recall  f1-score   support

    Negative       0.83      0.46      0.59      1731
     Neutral       0.32      0.74      0.44       629
    Positive       0.53      0.44      0.48       415

    accuracy                           0.52      2775
   macro avg       0.56      0.55      0.50      2775
weighted avg       0.67      0.52      0.54      2775
```

Random Forest

Figure 4.3: precision, recall support score in the visualization of Random Forest, SVM, Logistic Regression, Nive Bayas.

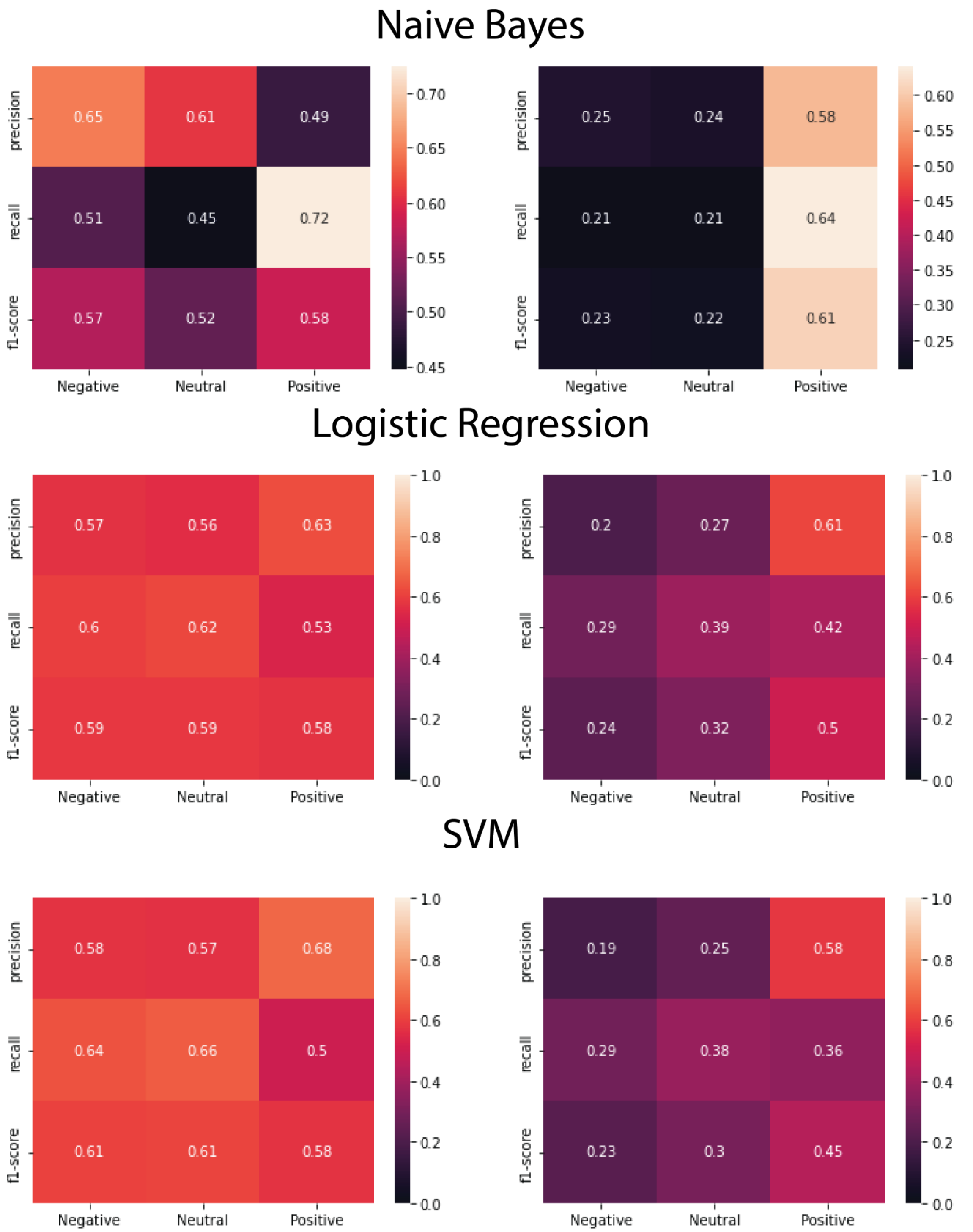


Figure 4.4: Confusion matrix of SVM, Logistic Regression, SVM of the synthetic dataset.

4.3 Discussion

The majority of models are machine learning-based, while the combined LSTM model used both machine learning and deep learning to train and evaluate models using textual data. SVM, Logistic Regression, Naive Bayes, and Random Forest have average accuracy because there is little correlation between them, which is why machine learning algorithms don't perform well on these models. Because of its block-wise learning methodology, LSTM has more accuracy than other models.

CHAPTER 5

Impact on Society, Environment, and Sustainability

5.1 Impact on Society

Studies on abuse textual data (comments on abuse textual data) were very useful for building a good society. It predicts society's future and what the next generation will face easily identified by sentiment analysis. Society will improve by this study this technique can easily identify those negative and positive sentiments that data analysis can use that for their work for that reason from that will give us a better output for seeing where society is failing to build good points society can improve by using the sentiment analysis it detects the sentiment of the text while the artificial ai detecting every movement of our every lifestyle there it can be easier to develop society by suggesting them who said the negative word for the comment by providing development and improvement of the punishment news through the social media its cos the mind change can possible for society.

5.2 Impact on Environment

However, it is possible that the results of such a sentiment analysis could be used to inform policies or initiatives that could impact the environment. For example, if the analysis identified patterns or trends in the language used to discuss violence against children and women that were related to certain environmental factors, this information could be used to develop strategies to address these issues and potentially reduce the risk of violence in certain areas. It can represent a good environment to society in that case its detection technique can help to manipulate the environment by making a custom environment that can change or avoid a bad environment. People are now commenting on abusive news of frustration without knowing the situation commenting was not a solution for the environment. In that type of case, the environment spread badness for a reason environment needs to be custom-built by applying this algorithm to protect the environment, this study can find out the text that abusive comments where can be spreading goodness to change the environment and it can great solution for environment development.

5.3 Ethical Aspects

Ethically sentiment of abuse text analysis was necessary to understand the ethics of the text. If thinking ethically the comments are on abuse so that machines can understand the properties of the comments and machine can understand whether the news comment was ethical or not for avoiding violence it can be a good subject to study. Making scientific knowledge about machine learning and deep learning was ethically beneficial. How a machine understands the abuse comments on what is necessary to understand the machine to extract the features for future application. Machine learning can classify those aspects but via deep learning, this paper show that works on the sentiment data and abuse data can be learned by a machine that’s why this was an ethical aspect of this study that it can help to understand the machine. For now, data transformation is in English but English works well for all aspects can see the benefits of this study will help to develop more in future work also.

5.4 Sustainability Plan

Machine learning and deep learning on abused text data were not so popular. Still, this work already shows those aspects where in the future machine will be precise by this work for social media. Now the tradition was to post on social media all this sustain post analysis the sentiment was popular now for this aspect according to the popularity of sentiment analysis can make the work popular for future of this work done by this report. In the future, this work will extend to the Bengali language this work converts the data to the English language for that analysis in English, but the future plan was to apply sentiment analysis for the Bangla language work will be sustainable for the future, not this work.

CHAPTER 6

Summary, Conclusion, Recommendation, and Implication for Future Research

6.1 Summary of the Study

This study was for the textual comments line for training the meshing with the data set of our own and with the sentiment Twitter data to identify the sentiment of abuse news. News portals online post various news on social media day by day and also on Facebook, YouTube, Twitter, and other places where comment expressions will analyze to predict the sentiment positive or negative. Nave Bayes, SVM, Random Forest, and Logistic Regression were used here for machine learning approaches on the other side machine and deep learning LSTM was used to train the system with proper description in this study. Data of this study was not larceny collected in this work so for the test algorithm on sentiment data we used sentiment analysis on Twitter data to learn the machine so that we can prove our algorithm works well as well as the machine and deep learning algorithm. This study describes the LSTM very precisely step by step from input, tokenizer, embedding, and SoftMax to predict two classes that can understand for easter for all views.

6.2 Conclusion

The algorithm works properly on this project which is shown in this particular work. This work proved the sentiment analysis on social media abuse on women and children textual comment data could learn a system for predicting the sentiment as well as can detect those by the LSTM model. Because of the shortest data collection, the accuracy was low in this project if the data quantity will largely amount then it's possible to get better output results from this sentiment analysis. Three feelings are present in the data for this article, however, a typical algorithm like LSTM does not adequately address these emotions.

6.3 Implication for Further Study

Implication feature study will be on the Bangla language in this paper containing data was at English by translating in google translate for this reason Bangla text in the paperwork was done in this paper. In the future, this work extends to the Bangla data set and various algorithms will apply for better accuracy than this work. This paper's data are in limited amounts that's why accuracy doesn't match satisfactory accuracy depending on the data used in this paper. Data need around 20000 to 50000 for math the expected outcome, in that case, data will increase for this work as the data increase the accuracy, prediction, and classification will increase and work accurately. With those two datasets, this study shows how an abusive sentiment in translated Bangla data set can use to learn a system easily to predict the sentiment of the textual data sets. Hare for data on sentiment by Twitter where the data set was also not stable hare negative data was a large amount than the positive data. Traditional algorithms like the LSTM can understand positive and negative data better than neutral ones. So future work with neutral data will make this study more efficient for future studies on sentiment analysis. Bangla is a complex language with many contextual and syntactic nuances that can affect the sentiment of a given piece of text. Future research may concentrate on creating sentiment analysis algorithms that are better equipped to take these contextual elements into account. algorithms for machine learning, such as SVM, Nave Bayas, Random Forest, and Logistic Regression.

Reference

[1] Rahman, K.F., 2019. Focus on domestic violence in Bangladesh: a study from criminological perspectives. *Journal of international women's studies*, *20*(3), pp.98-115.

[2] Sifat, R.I., 2020. Impact of the COVID-19 pandemic on domestic violence in Bangladesh. *Asian journal of psychiatry*, *53*, p.102393.

[3] Murthy, G.S.N., Allu, S.R., Andhavarapu, B., Bagadi, M. and Belusonti, M., 2020. Text-based sentiment analysis using LSTM. *Int. J. Eng. Res. Tech. Res*, *9*(05).

[4] Wang, X., Jiang, W. and Luo, Z., 2016, December. Combination of convolutional and recurrent neural networks for sentiment analysis of short texts. In *Proceedings of COLING 2016, the 26th international conference on computational linguistics: Technical papers* (pp. 2428-2437).

[5] Gandhi, U.D., Malarvizhi Kumar, P., Chandra Babu, G. and Karthick, G., 2021. Sentiment analysis on Twitter data by using convolutional neural network (CNN) and long short-term memory (LSTM). *Wireless Personal Communications*, pp.1-10.

[6] Saritas, M.M. and Yasar, A., 2019. Performance analysis of ANN and Naive Bayes classification algorithm for data classification. *International Journal of Intelligent Systems and Applications in Engineering*, *7*(2), pp.88-91.

[7] Dreiseitl, S. and Ohno-Machado, L., 2002. Logistic regression and artificial neural network classification models: a methodology review. *Journal of biomedical informatics*, *35*(5-6), pp.352-359.

[8] Wang, H. and Hu, D., 2005, October. Comparison of SVM and LS-SVM for regression. In *2005 International conference on neural networks and brain* (Vol. 1, pp. 279-283). IEEE.

[9] Jakkula, V., 2006. Tutorial on support vector machine (SVM). *School of EECS, Washington State University*, *37*(2.5), p.3.

[10] Qi, Y., 2012. Random forest for bioinformatics. In *Ensemble machine learning* (pp. 307-323). Springer, Boston, MA.

[11] Menard, Scott. *Applied logistic regression analysis*. No. 106. Sage, 200

[12] Chauhan, Vinod Kumar, Kalpana Dahiya, and Anuj Sharma. "Problem formulations and solvers in linear SVM: a review." *Artificial Intelligence Review* 52, no. 2 (2019): 803-855.

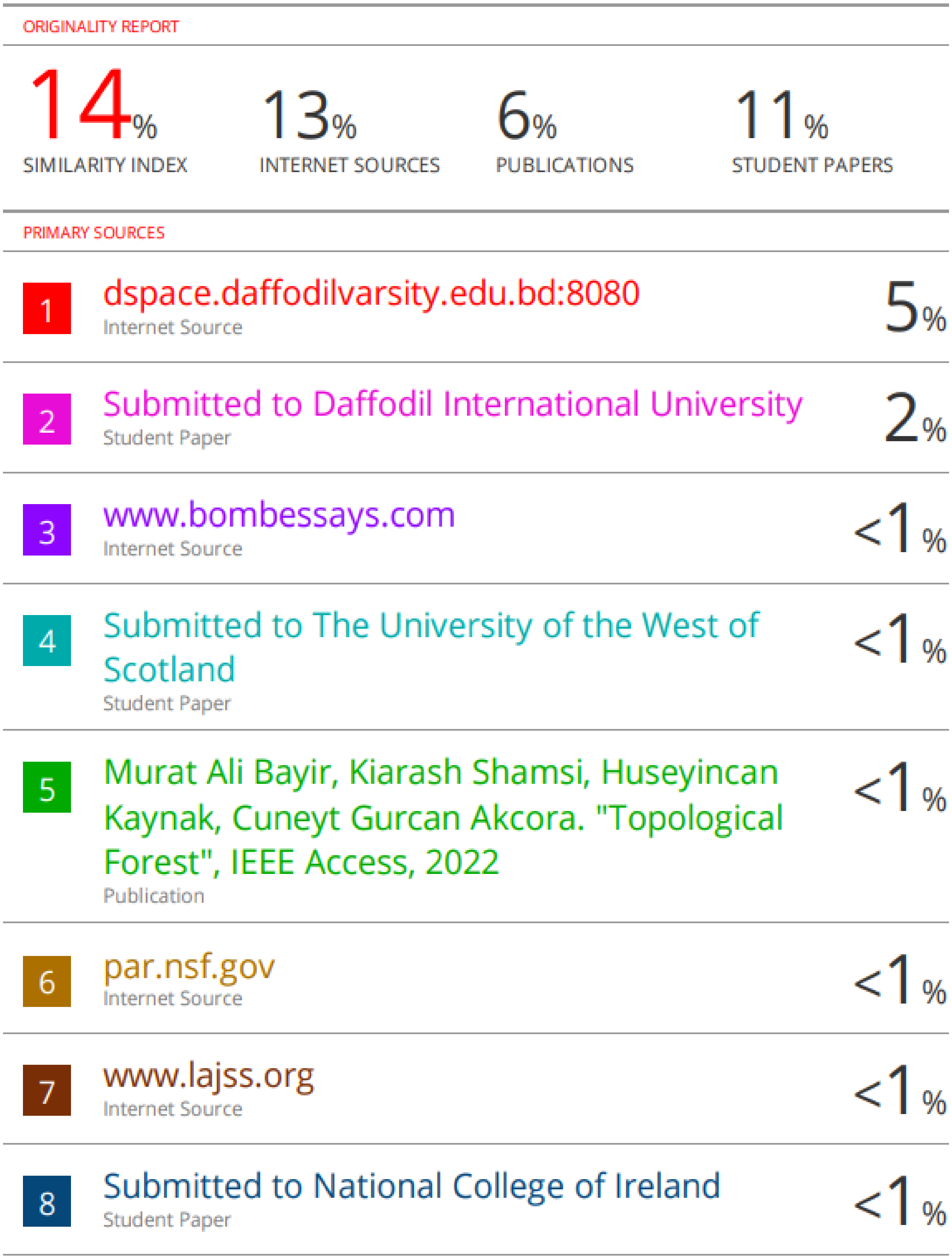
[13] Petkovic, Dragutin, Russ Altman, Mike Wong, and Arthur Vigil. "Improving the explainability of Random Forest classifier–user-centered approach." In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2018: Proceedings of the Pacific Symposium*, pp. 204-215. 2018.

[14] Zhu, Ji, and Trevor Hastie. "Classification of gene microarrays by penalized logistic regression." *Biostatistics* 5, no. 3 (2004): 427-443.

[15] Jimenez, Matthieu, Cordy Maxime, Yves Le Traon, and Mike Papadakis. "On the impact of tokenizer and parameters on n-gram based code analysis." In *2018 IEEE International Conference on Software Maintenance and Evolution (ICSME)*, pp. 437-448. IEEE, 2018.

[16] Webster, Jonathan J., and Chunyu Kit. "Tokenization as the initial phase in NLP." In *COLING 1992 volume 4: The 14th international conference on computational linguistics*. 1992.

[17] Webster, Jonathan J., and Chunyu Kit. "Tokenization as the initial phase in NLP." In *COLING 1992 volume 4: The 14th international conference on computational linguistics*. 1992.



ORIGINALITY REPORT

14%	13%	6%	11%
SIMILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	5%
2	Submitted to Daffodil International University Student Paper	2%
3	www.bombessays.com Internet Source	<1%
4	Submitted to The University of the West of Scotland Student Paper	<1%
5	Murat Ali Bayir, Kiarash Shamsi, Huseyincan Kaynak, Cuneyt Gurcan Akcora. "Topological Forest", IEEE Access, 2022 Publication	<1%
6	par.nsf.gov Internet Source	<1%
7	www.lajss.org Internet Source	<1%
8	Submitted to National College of Ireland Student Paper	<1%