Text Classification with an Efficient Preprocessing Technique for Cross-Language and Multilingual Data

by

Towhid Khan
18201035
David Dew Mallick
18201045
Md.Shakiful Islam Khan
18201198
Md Mahadi Hasan
18201062

Bachelor in Computer Science
Department of Computer Science and Engineering
Brac University
September 2022

Declaration

It is hereby declared that

1. The thesis submitted is my/our own original work while completing degree at Brac University.
2. The thesis does not contain material previously published or written by a third party, except where this is appropriately cited through full and accurate referencing.
3. The thesis does not contain material which has been accepted, or submitted, for any other degree or diploma at a university or other institution.
4. We have acknowledged all main sources of help.

Student's Full Name & Signature:

Towhid Khan	David Dew Mallick
18201035	18201045
Md.Shakiful Islam Khan	Md Mahadi Hasan
18201198	18201062

Approval

The thesis/project titled "Text Classification with an Efficient Preprocessing Technique for Cross-Language and Multilingual Data" submitted by

1. Towhid Khan (18201035)
2. David Dew Mallick (18201045)
3. Md.Shakiful Islam Khan(18201198)
4. Md Mahadi Hasan (18201062)

Of Summer, 2022 has been accepted as satisfactory in partial fulfillment of the requirement for the degree of B.Sc. in Computer Science on 22^{nd} *September* 2022.

Examining Committee:

Supervisor:
(Member)

Faisal Bin Ashraf

Lecturer
Department Of Computer Science and Engineering
Brac University

Program Coordinator:
(Member)

Md. Golam Rabiul Alam

Professor
Department Of Computer Science and Engineering
Brac University

Head of Department:
(Chair)

Ms. Sadia Hamid Kazi

Chairperson
Department of Computer Science and Engineering
Brac University

Abstract

The procedure of eradicating extraneous textual elements and preparing or processing the values to be fed into the classifier model is often indicates the concept of text-preprocessing. There are several preprocessing methods, however not all of them are effective when used with cross-language and multilingual datasets. Running a cross-lingual or multilingual dataset through a single pre-processing method and text classification model is rather challenging. What if a technique could be used to better classify data from multilingual and cross lingual datasets? In order to accelerate the process of improving accuracy, we tested various combinations of data pre-processing with text classification models on datasets in Bangla, English, and cross-lingual (Native language written in English letters). We may infer from our experiment that mLSTM functioned effectively for datasets in Bangla and English. Thus, mLSTM can be a helpful preprocessing method for datasets containing a variety of languages.

Keywords: NLP, Sentiment analysis, Information Retrieval, Review , LSTM , mLSTM, XGB, SVM, TF-IDF, Bag of Words, Logistic regression. Random Forest

Dedication

We dedicate this paper to all the Research Enthusiasts, who are looking for a better model to perform their pre-processing techniques in multilingual and cross lingual dataset.

Acknowledgement

(A blank page kept to acknowledge other's contributions and help in this paper; to be written in later phases)

Table of Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Text Classification

The process of evaluating the textual literature , documents, and records into two or multiple types of groups or types is commonly known as text classification . Text classification gives the privilege of working with a solid framework for learning and getting familiar with textual data processing[Figure 1.1].In addition, a text classification system requires several elements, such as acquiring documents and a hierarchy that describes the most relevant topic.The most crucial feature is that any type of text can be organized, structured, and classified using text classifiers.From

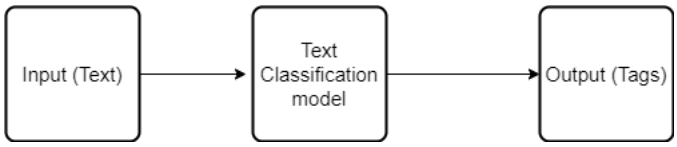

Figure 1.1: Text Classifier

[15] One area of artificial intelligence (AI) that offers robots the ability to read, comprehend, and convey meaning is natural language processing. One of the keys to good classification is understanding the data. However, most categorization tool developers are quite weak in this area. Text categorization systems are designed to make information more accessible and to make all newly found information available or applicable to help rational decision making.For the purpose of training a classification model, information needs to be extracted from raw text data in a variety of ways. Words and statements are the most frequent kinds of unstructured data. Although there is a lot of it, it might be challenging to extract relevant information. If not, mining the data would be time-consuming also.Both spoken and written language are rich in information. It's because writing and speech are the two main ways we communicate as sentiment creatures. Furthermore, NLP can do tasks like sentiment analysis, cognitive assistant, span filtering, spotting bogus news, and real-time language translation for us while analyzing this data. NLP is widely utilized in a wide range of products, including computers, cellphones, speakers, and websites. NLP-based machine translation is used by Google Translator. Google Translator

uses spoken and written natural language to translate languages requested by users. NLP enables Google Translate to recognize words in context, eliminate extraneous sounds, and enable CNN to comprehend native voice.

1.2 Necessity Of Text Classification

Natural language processing's core objective of text classification has several applications, including sentiment analysis and intent detection. It's a crucial technique for getting information or value out of unstructured data. With the use of text classification, we can quickly and efficiently evaluate thousands of texts to determine things like emotion or subject.For example: NLP Chatbot.

Text is among the most frequent kinds of large amounts of data, accounting for about 80 percent of all information. The chaotic nature of text makes it difficult and time-consuming to analyze, analyze, organize, and sort text data, which prevents conventional businesses from utilizing it to their maximum capabilities.Machine learning text classification is useful in this situation. Businesses can rapidly and inexpensively automatically arrange all types of pertinent text from emails, legal documents, social media, chatbots, polls, and more thanks to text classifiers. It enables businesses to analyze text data more quickly, automate business procedures, and make choices based on data.NLP is frequently used in chatbots. Because they eliminate the need for humans to ask customers what they need, chatbots are incredibly helpful. NLP chatbots are capable of asking a series of inquiries, such as what the user's issue entails and where to go for a solution. A capable chatbot is already built into Apple and Amazon's systems. The chatbot interprets the user's inquiries into language that can be understood in the internal system. Token then uses NLP to determine what questions users are posing. Information retrieval (IR) uses NLP . A software package called IR works with massive storage and information assessment from repositories' huge text documents. It will only pull out pertinent data. There are primarily three approaches of classifying texts: hybrid system, machine system, and rule-based system. To use a series of artisanal language rules, the rule-based technique divides texts into an organized group. Users are required to create a list of words that have been manually sorted into categories based on their linguistic characteristics. Donald Trump and Boris Johnson would be classified under the heading of politics. LeBron James and Cristiano Ronaldo are examples of athletes. A computer-based classifier learns to classify objects using previous data set observations. These words are pre-labeled: "user data" and "test data." It continuously learns by accumulating classification strategies from past inputs. For feature expansion, machine-based classifiers utilize a bag of words. The Hybrid Approach is the third classification approach for texts. In a hybrid technique, regulation and machine-based techniques are merged. Using a hybrid method, a regulation approach is utilized to establish a tag, while machine learning is used to train the system and construct a rule. A comparison is then made between the rule-based rule list and the machine-based rule list. When anything does not match a tag, the list is manually edited. It provides the most efficient method for categorizing text.

1.3 Importance of Efficient Pre-Processing

It is to be clear by now that data preparation is crucial.The consistency of the set is compromised by errors, redundancies, inconsistencies, and lost values. Therefore, we must address all of these problems for a more accurate result. For instance, using a flawed dataset to train a machine learning model to handle consumer purchases might be a bad idea. The likelihood that the system may exhibit distortions and aberrations that negatively impact the user experience is high. Thus, it is mandatory to organize and clean the data in as structured manner as possible. There are multiple ways to achieve an organized and structured dataset. Information preparation is broken down into four stages: cleansing, integration, diminishment, and change. Firstly, data cleaning, also abbreviated as "cleansing," is the method that involves making datasets more readable by taking into account for missing values, eradicating outliers, resolving contradictions between data points, and reducing noise in the data. Producing accurate and comprehensive samples for machine learning models is the primary goal of data cleaning. Secondly,data preparation must include data integration. Integration may result in several unnecessary and unreliable data points, which would eventually provide less accurate models.In addition,to decrease the expense of data mining or data analysis, data reduction is used to minimize the amount of information available. Finally,the process of transforming information from a particular format to yet another is referred to as data transformation. Fundamentally, it entails techniques for transforming information into acceptable representations that the computer can effectively learn from.However,an efficient preprocessing can generate good and valid results if it is implemented correctly into a model.A preprocessing implementation's main difficulty is that it would take longer to finish if the dataset was too big. So an efficient preprocessing technique might help in this case.Therefore, in this case, an effective preprocessing method might be useful.

1.4 Problem Statement

As previously discussed, finding a perfect model with an intelligent preprocessing combination that can run both on cross-language and multi language is indeed a challenging task. Although,cross language and multi language datasets have been implemented to detect the sentiment with traditional preprocessing techniques and yet there are distinct biases or flaws.We have seen times when preprocessing implementations are not truly able to transform the text from a dataset having non-english text.Since it is impractical to create a separate text processing program for every one of the world's languages, because some of them are less frequently spoken, it would require too much time and information. Thus,it is preferable to discover a means to standardize all of the terms in the observing text. So our aim is to classify and define such a text processing mechanism that can run both on datasets having cross-language and multi language.We propose a preprocessing technique which will work on multiple language and cross language datasets.See the following figure 1.2 to get idea about Our Proposed approach

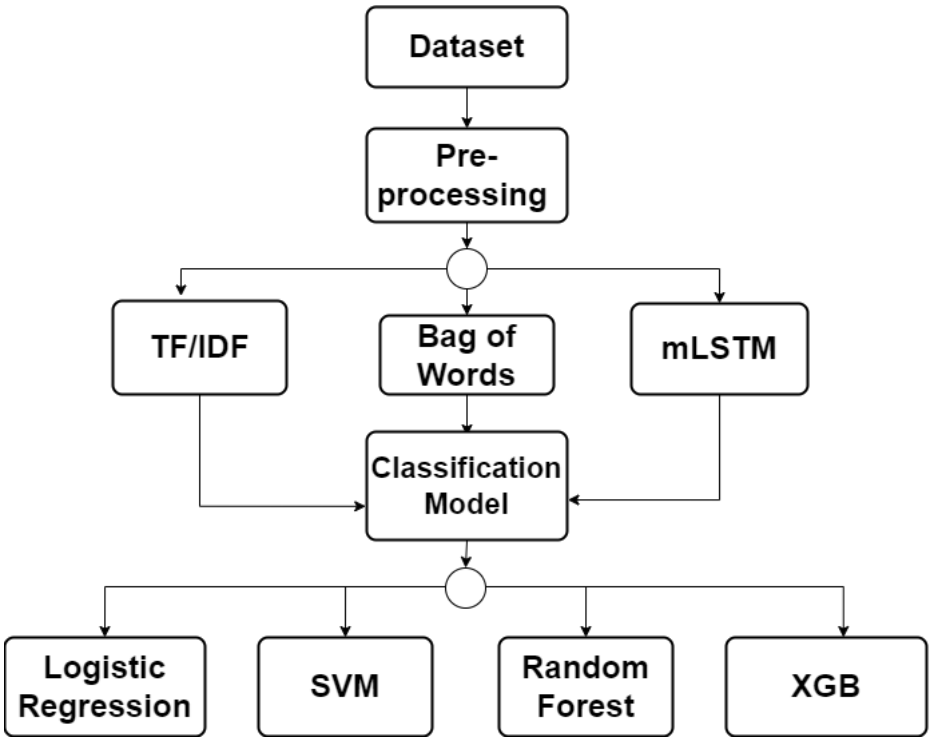

Figure 1.2: Proposed Model

1.5 Research Objectives

A precise text categorization that employs sophisticated preprocessing techniques can enhance not only the precision of the model but also the precision of the prediction.
This research aims to establish the following:

- Develop an NLP-based strategy in order to discover highly effective preprocessing methods.
- Collaborate with linguistically diverse and multilingual datasets.
- Concentrate on locating a model that can be used in conjunction with preprocessing procedures that will produce more accurate results and that can be combined with other models in a way that is dependable. combination.

Chapter 2

Literature Review

Text mining is the technique of discovering and retrieving pertinent and useful data from unstructured text [3]. This encompasses anything from information retrieval (such as document or website retrieval) through text classification, clustering, entity, relation, and event extraction (which has become more popular recently)[14]. Natural language processing is the endeavour to derive a deeper meaningful interpretation from unstructured text (NLP). This is much the same as discovering who did what to whom, when, how, and why. Parts of speech (nouns, verbs, adjectives, etc.) and grammatical structures are utilized often in NLP (expressed as phrases such as noun phrases and prefixes, or dependencies such as subject and object)[13]. It must address anaphora (a pronouns or other back-referential phrase) and ambiguity (in terms of words and grammatical structure, such as what is being modified by a specific word or prepositional phrase)[1]. To do this, it uses a number of knowledge representations, such as a lexicon of words along with their meanings, grammatical characteristics, and a set of grammatical rules, as well as a number of additional assets, such as an ontology of objects and activities, or a thesaurus of synonyms or abbreviations[1].

In the paper[7] they have explored the properties of the byte-level recurrent language model.They have explored that the sentiment unit of their byte mLSTM has a direct correlation on the generative process of the model. They found out that the multiplicative LSTMS converges faster than normal LSTM. Their proposed model works better for high density dataset. They experiment their byte mLSTM with different dataset such as MR, CR , SUBJ, MPQA and also compare it with others researcher model such as SKIPTHOUGHT, SKIPTHOUGHT(LN), SDAE, CNN, ADACENT. For CNN model the accuracy for MR is 83.1 but for byte mLSTM it's 86.9 on the other hand for CR dataset it is 86.3 and 91.4 for byte mLSTM.Their research emphasizes the sensitivity of learnt representations to the data distribution on which they were trained.Several promising future research directions are identified by their findings. Even on very identical domains, the observed performance plateau implies strengthening the representation model in terms of architecture and size. Due to the fact that their model functions at the byte level, hierarchical/multi-timescale extensions could enhance the quality of representations for lengthier documents. The sensitivity of learnt representations to their training domain could be mitigated by training on a broader variety of datasets with greater task coverage.The work of this study promotes additional research into language modeling because it proves that

learning high-quality representations is possible with only the conventional language modeling objective.

T.Mastan Rao et al. [8] proposed an algorithm to extract information from ratings and reviews provided by customers using Deep learning kits such as CNTK, KERAS. This paper firstly pre-processed the data into token level using NLP. Then it polarizes all the raw text into positive, negative, and neutral categories.

E. Suganya et al.[9] introduced a paper where they want to monitor customers' behavior and analyze the reviews. Moreover, Machine learning algorithms, eg. SVM, Random Forest, are used in this paper to get the result. This research is conducted on product review comments on online shopping platforms. A Graph representation using Hybrid SVM-CNN is the best approach for this paper.

Su Su Htay et al. [2] proposed a model to summarize the customer's Comment through adjective, verb, adverb, and noun using pattern knowledge and opinion lexicon. Pattern knowledge. In this article, we will analyze sentences using part-of-speech tagging and opinion lexicon models. They expected to get results from the extraction of opinion mining. They showed comparisons and obtained results by extending both explicit and implicit functionality for future work. Both features help provide more accurate results when determining polarity.[2].

Chau Vo et al . [11] introduced a paper where a useful review search task has been developed using elastic net normalization techniques. Multiple type linear regression models are used in this research paper to get the optimized result. This approach obtains 83% accuracy with the amazon dataset. However, They are willing to improve their regression model using advanced algorithms such as deep learning and want to improve the prediction model's effectiveness.
Oscar Romero [5] defines sentiment analysis as the problem in which a machine learns to predict and interpret the sentiment conveyed by a person or a contextual opinion about anything. He explains that the modifications needed to improve the classifier's accuracy are determined by the data and the language it uses. The machine learning approach learns more efficiently and generalizes better when transformations and filtering of the least essential data are used.

Ammar Mars et al . [6] proposed a method to extract product features' opinions of customers from social networks using text analysis techniques. This research contributes to lexical ontology, product presentation, opinion mining, and visualization of their result. MapReduce, Drvad are used for extracting data. Afterward, OSPM, FE, and POS methods process the extracted data. Using these methods, the polarity percentage of product features is shown with the classified portion of the product. But, the approach of POS tagging wasn't much categorized and not used much.

Bangla is the 7th most spoken language in the world and people cyber bullies in bangla everyday in the social media which is why this paper [16] fully focuses on detecting the cyberbullying using deep neural network. They have labeled their dataset into 4 categories as Non-bully,Sexual,Threat,Troll,Religious. Next they pre-processed the dataset in three parts: Stop words Removal, Tokenization of String

and Padded sequence conversion [16]. Binary classification and multi-class classification models were used and the predicted results from both the models were applied ensemble method in order to improve their accuracy. In binary classification they have shown that they have achieved validation accuracy of 87.91% and precision of 90%, recall of 75% and F1-score of 82%. Their model can successfully predict 95% of the 'not bully' comments and 75% 'bully' comments. And for the multi-class classification they have used Random Forest, SVM, KNN, Naïve Bayes etc. classifiers amidst them SVM has the better accuracy of 85%. They have successfully predicted 91% of 'Not bully' comments, 85% of 'Religious' comments, 81% of 'Sexual' comments 50% of 'Threat' and 84% of 'Troll' comments. 87.91% accuracy was acquired using the binary classification and 79.29% accuracy using the multi-class.classification, which they then combined with the binary classification using the composite approach in order to categorize the incidents of harassment into a variety of subcategories. Their model required significantly more time to train, and on occasion, it provided a false positive result when applied to lengthy words. They have plans to address their weaknesses in the near future.

In the paper[4] polarity detection from all the online news articles was the main goal if they are negative or positive or even neutral. First, they recommended selecting a News Article from which they would then extract sentences and words according to their sentence type (simple sentence, compound sentence, complicated phrase, or compound-complex sentence). Finally, they would sum all the polarities to obtain the article's polarity. They have selected three categories of newspaper circulation based on Google Analytics: The Independent, The Telegraph, and The Daily Star. They've acquired 91.07% accuracy with their recommended method. When they have a small number of sentences in their paper, which makes it harder to discern the genuine polarity, they have experienced challenges with this method. Their primary goal was to identify the polarity of a newspaper article. They believe that algorithm analysis would aid big corporations in identifying the positives and downsides of their product reviews.

2.1 Major Findings and Scope of Research

From the paper, we have studied and we have acquired some major points regarding sentimental analysis, Different types of machine learning approaches, etc. In addition, the technique of extracting interesting and important information from unstructured or free text is called text mining[3]. However Natural language processing is the process of extracting a more comprehensive meaning representation from the free text (NLP). This is almost the same as figuring out who did what to whom, when, why, how, and why. With the help of NLP, T.Mastan Rao et al. [8] suggested an algorithm to extract information from customer ratings and reviews using Deep learning kits such as CNTK and KERAS. On the other hand, Ngoc-Bao-

Van Le et al. [17] established a technique to evaluate a customer's sentimental view by extracting their emotional perspective from their social media or microblogging platform comment or review. Furthermore, customers' reviews are also categorized using POS, NEG, and NUE algorithms based on product cost, shipment, quality,

design, and satisfaction. Thus customers reviews and behaviors were important also which was introduced by E. Suganya et al.[9]. They used machine learning algorithms to demonstrate their work and came up with the best solution which was Hybrid SVM-CNN. Moreover, Su Su Htay et al.[2] proposed a model to summarize the customer's comment through adjective, verb, adverb, and noun using pattern knowledge and opinion lexicon. They have also used POS tagging, and Opinion Lexicon models to parse the sentence. We have also learned that Chau Vo et al. [11] introduced a paper where a helpful review retrieval task has been developed by using the elastic net regularization method.This approach obtains 83% accuracy with the amazon dataset. Besides the machine learning approach learns more efficiently and generalizes better when transformations and filtering of the least essential data are used. However, the modifications needed to improve the classifier's accuracy are determined by the data and the language it uses which was explained by Oscar Romero [5]. Lastly, Ammar Mars et al. [6] proposed applying text analysis techniques to obtain customer opinions on product attributes from social networks. In addition, this research contributes to lexical ontology, product presentation, opinion mining, and visualization.

Chapter 3

The Dataset

3.1 Dataset

In our research , we will be working with multiple Binary Datasets.In addition , we have gathered 3 categories of datasets based upon cross-language and multi language which are English, Bangla and Mixed(Native Language Written In English).

3.1.1 ebay Dataset

eBay Inc. is mostly popular for top Selling Categories like Video Games, Health and Beauty etc.eBay is a multi-billion dollar e-commerce institute operating in approximately 32 countries.The company operates through a website named after its company eBay, an online shopping and auction platform where customers can buy products and businesses around the world conducts buy and sell operations with a wide variety of goods and services.There are sales data available of eBay.eBay dataset was introduced by Wojtek Bonicki with the help of python web scraping technique.
Example: Table 3.1

Label	Sentence
1	all good all good product described
0	not working asked return havent heard not working
1	nice gary good comfy headset
1	wont go wireless xbox one theyre good
0	bad story not worth 60

Table 3.1: eBay Binary Dataset

Dataset Collection: From the beginning , we needed feedback from the customers.So we thought eBay might be the right way to collect data from them.In addition, eBay is a very vast E-commerce companies have varieties of customers from all over the world which will likely have less chance of biased feedback.Nevertheless, our main goal was to search for unbiased datasets which will help us in accurate sentiment results.Lastly, we have taken the dataset from a open-source platform.
Source: eBay Dataset

Dataset Analysis: We collected around 44311 user, and from that, around 2026 comments were labeled as not recommended comments and the rest 42285 comments were recommended comments.However, we only considered the positive and negative feedbacks and not neutral feedbacks.Initially, we spliced the dataset to 80% for our training purpose and the rest of the remaining 20% for the testing purpose.There were two major column in the datasets:1.rating 2.review.It is good to keep in mind that , the following dataset was initially preprocessed for the model.In addition , we are taking reviews as feature from the dataset and output as the labeled data.The visual representation of the dataset "Ebay" is shown in figure 3.1 -

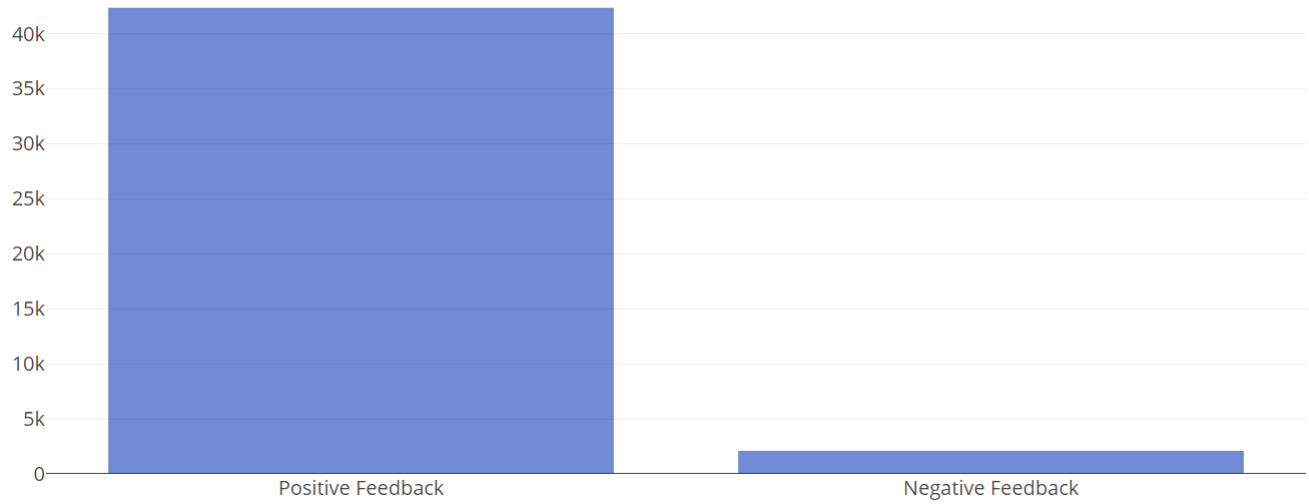

Figure 3.1: eBay Dataset Representation

3.1.2 Steam Reviews Dataset

Steam is basically a digital entertainment and video game distribution platform with a massive worldwide gaming community. There are many players who provide honest or biased reviews on the game's blog page or website and can choose whether or not it is recommended to the game for others. Although, this dynamic classification of sentiment from text would help Steam autonomously to tag the reviews which were pulled from various websites around the internet, allowing them to determine the popularity of titles.

Example : Table 3.2

Dataset Collection: We initially intended to work with product reviews from different e-commerce sites but we took a different approach. Why not steam game reviews ? Then we came to know that the steam dataset[12] contained about 17495 unique reviews which can help us a lot for measuring our model accuracy as it the review as totally game specific.There are 64 game titles reviewed, each having a review text, user recommendations, and more details.Similarly, we considered two necessary

Label	Sentence
1	This game is a breath of fresh air! If you can ignore the F2P business style. The gameplay is sound and teamwork prevails. If you have a groupof people thinking about playing I would definitely give it a go!
0	no matter how many times i re arrange my disk space, it never lets me download it again! there are none of the game files on my computer and it isnt in my library, although it says it is.
1	Even all the efforts of the devs it still has some major bugs, but getting better every week.If u wanna shout battleships to pieces and release the steam then its for u. Free to play, nothing to loose.
0	I. Really. Don't. Care.
0	I currently wouldn't reccomend this game atm, as the stronger zombies take more than 100 bullets to kill and if you've just started off that will mean instant death- which happens alot.

Table 3.2: Steam Binary Dataset

parts for our research that is review text and user recommendation.Lastly,we use review text as our main corpus for the model and user recommendation as labeled data.

Dataset Analysis: In this dataset there were around 17495 user feedbacks, and from that, we took around 7527 comments which were not recommended and the rest 9969 comments were recommended by the users . However, we only considered the positive and negative feedback and not neutral feedback as they might not work with our model. Initially, we sliced the dataset to 80% for our training purpose and the remaining 20% for the testing purpose.There were two major columns in the datasets: *user_review* and *user_suggestions*. However, the following dataset was initially preprocessed for the model.In addition , we took reviews as feature from the dataset and output as the labeled data.The visual representation of the dataset "Steam Review Dataset" is shown in figure 3.2 -

3.1.3 IMDB Dataset

A well-liked and well known dataset for text- and language-related machine learning lessons is the IMDB collection. Additionally, it is easily included in the Keras library, and Keras has a few built-in algorithms for pre-processing and data loading.

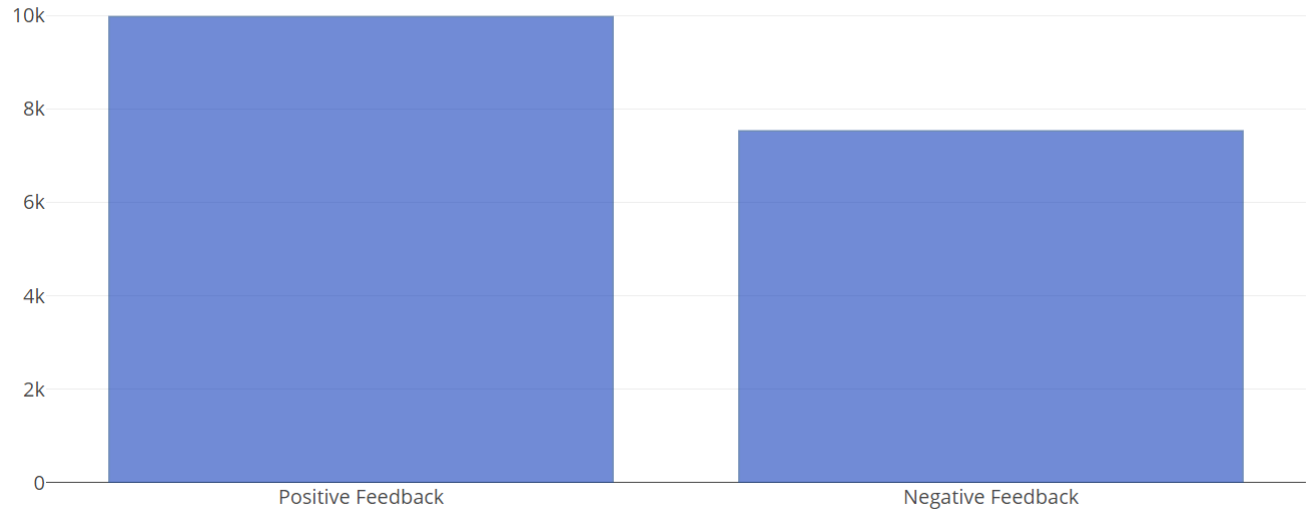

Figure 3.2: Steam Dataset Representation

Example : Figure 3.3

label	Sentence
0	The film is strictly routine.
1	This is a stunning film, a one-of-a-kind tour de force.
0	Final verdict: You've seen it all before.
1	A slick, engrossing melodrama.
1	A fun ride.

Table 3.3: IMDb Binary Dataset

Data Collection: IMDB dataset is known for its good integrity as it holds the raw emotion text of the user for most of the movies.As discussed before our main goal was to fulfill the requirements of our binary sentiment classification model.It could generate more accurate result for our model because of the varieties of sentiment included in the dataset[7].

Data Analysis: Similarly, the IMDB dataset contained about 8742 unique values of reviews of different movies.However,we spliced the dataset into 80% for training purpose , 20% testing purpose.We took 7793 unique reviews for training the model Whereas 1748 for the testing and rest of the datas for the validation. From 8742 unique reviews 4519 was positive feedback and 4222 was negative feedback. This IMDB dataset was pre-made for binary text classification models so it saved our time in the process.The visual representation of the "IMDB Review Dataset" is shown in figure 3.3 -

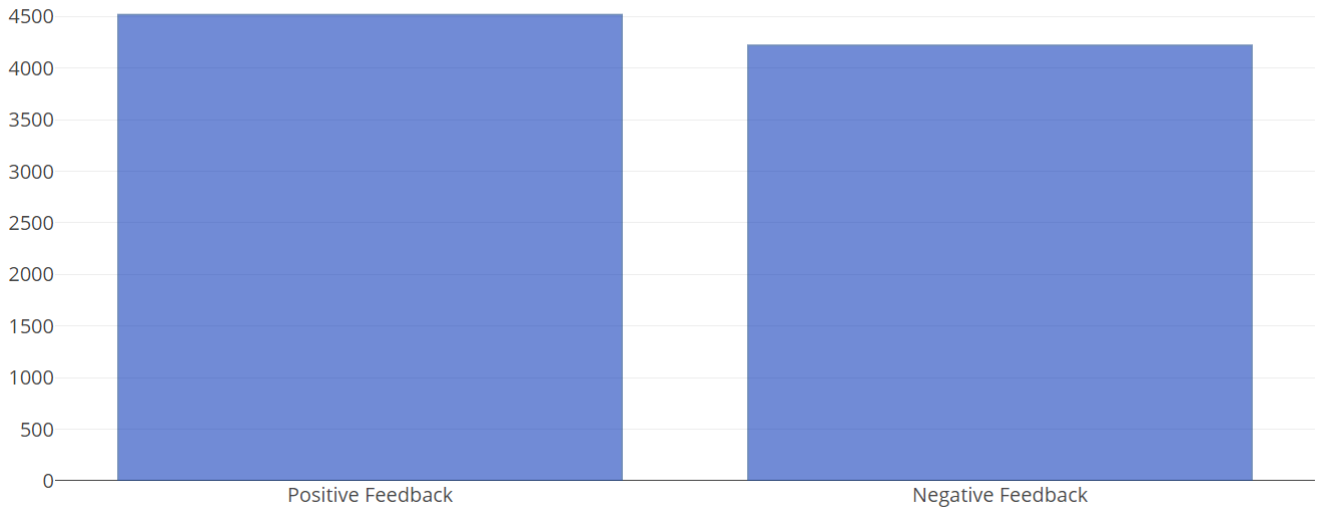

Figure 3.3: IMDb Dataset Representation

3.1.4 Bangla Language Dataset

Among the top 100 most spoken languages worldwide, Bangla is placed 5th. More than 210 million people speak Bengali as a primary or second language, with the majority of them living in the Indian states of West Bengal, Assam, and Tripura. There are also substantial immigrant populations throughout the United Kingdom, the United States, and the Middle East.
Example: Figure 3.4

Sl.	Label	Sentence
1	0	বাংলার প্রধান শত্রু রে
2	0	চুক্তি পাইছি পানির ন্যায্য হিস্
3	1	কসাই হোক এটা ভাল নাম
4	1	চিন্তার কি আছে? "রক্ত কথা কয়"..... যেমন রক্ত তেমন ক
5	1	চমৎকার গীতিকার ভাইজান।

Figure 3.4: Bangla Dataset

Data Collection: During our research and gathering multi language dataset we did found Bangla Dataset for sentiment analysis.Keeping in mind, there is not much of good Bangla Dataset around us.So we took a dataset that has been worked on similar purpose like us[10].

Data Analysis: We have chosen the bangla dataset that has been used in several researches.This dataset consists of raw bangla text related to sports and peoples feedback regarding sports events.The dataset contained around 8498 feedbacks . We split the whole dataset by following the normal convention of machine learning.So we considered around 6498 for training the model and 2000 for testing purposes. From 8498 unique reviews 4079 was positive feedback and 4419 was negative feedback The visual representation of the dataset is given in the following figure 3.5:

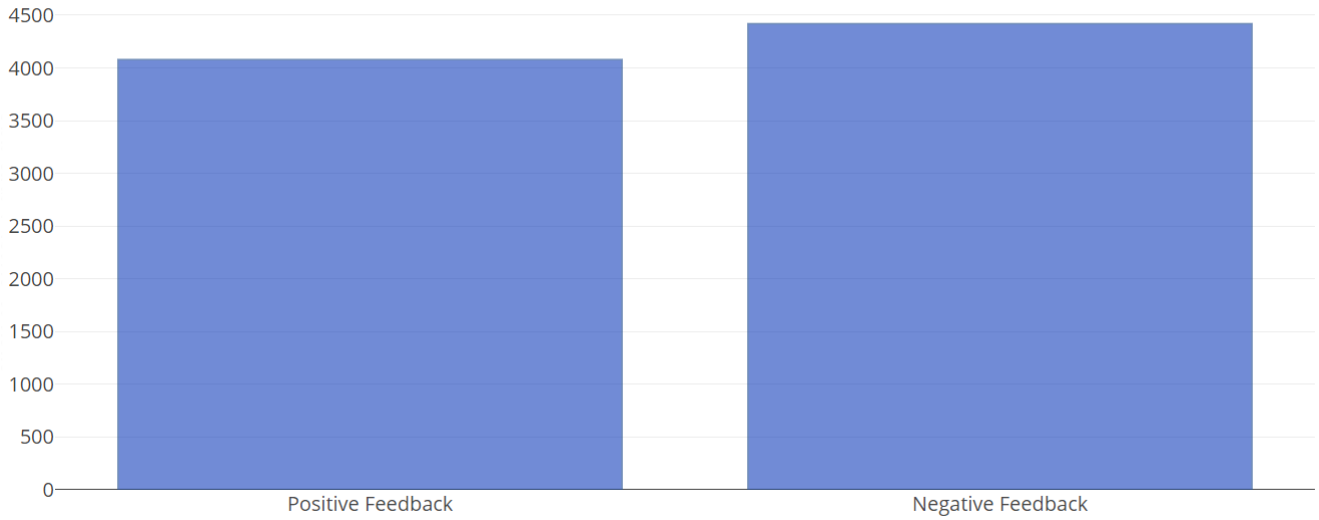

Figure 3.5: Bangla Dataset Representation

3.1.5 Cross-Language Dataset(Hate Speech)

Cross-Language is the written representation of Native language in English letters. Many people use their native language and use English letters to express what they are trying to say in social media.The most common and casual way of expressing feeling and feedback online relies on Cross Language.

Example: Table 3.4

Label	Sentence
1	Sobai sundorer pujari
1	Jana Galo Pori moner akta cheler khoj
1	Amader ei eisob party er sponser korbe abar
0	Kanki ki der Allah sob samoy valo rake
0	Erkom bokachoda News channel thakle entertainment er ovab hobe na..

Table 3.4: Cross Language Binary Dataset

Data Collection: Throughout the research we came across many cross language datasets.However we have chosen a particular dataset that is worked upon for Hate

Speech Detection using machine learning approach.Why Hate Speech you may ask? Because people tend to express feelings more freely and openly online without knowing the full consent.So the whole expression thing is wrapped with cross language[18].

Data Analysis: This cross language hate speech dataset had been implemented to many to research.The dataset consists of raw emotion and thoughts categorized as not hatred and hatred.The dataset contained about 5000 speeches where around 2836 speeches recognized as hatred and around 2164 speeches as not hatred.Again we follow the normal conventions for splitting the dataset.Lastly,the visual representation of the dataset is given in the following figure 3.6:

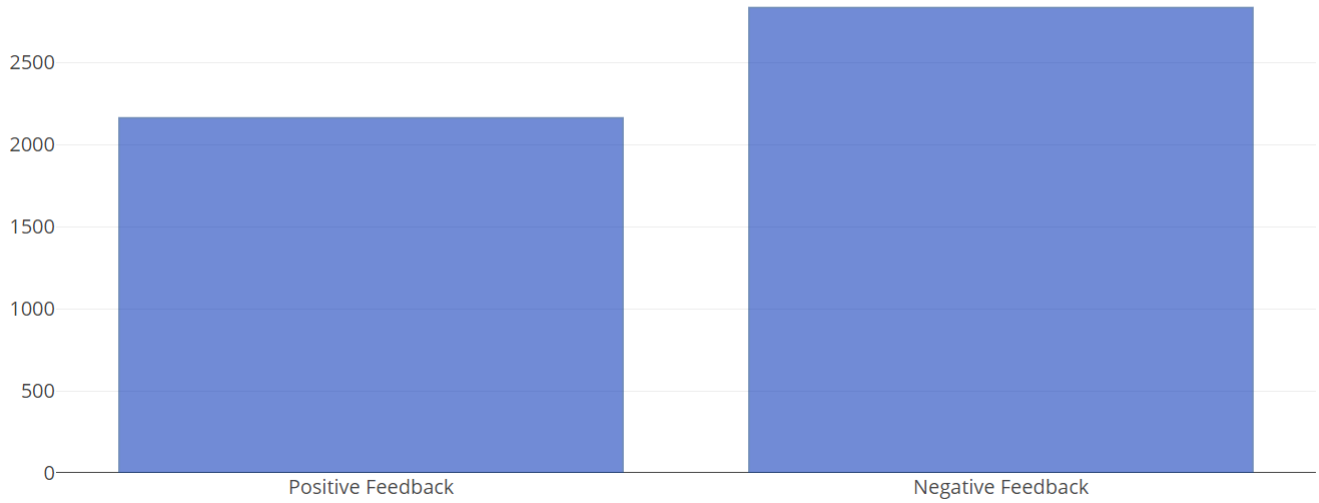

Figure 3.6: Cross language Dataset Representation

3.2 Pre-Processing

A step in the machine learning process called data preprocessing which changes or encodes data in order to make it easier for computers to analyze it. In other words, features in the data may now be easily interpreted by algorithms.

3.2.1 Bag Of Words

The NLP tool Bag of Words is a text modeling tool. Features are extracted from text data using this method. Feature extraction from documents is simplified and adapted in this approach.Word frequency in a document can be represented by its "bag of words," a text structure that stores words in the order in which they appear. We just tally up the total quantity of words and pay no mind to punctuation or sentence structure. Since there is no indication of how the words are organized, the document is called a Bag Of Words. The model cares only about the presence or absence of certain phrases, not their precise placement. Text's potential for chaos

and disorganization is a major limitation. Using the Bag-of-Words method, we can convert free-form sentences into deterministic vectors. The inputs to machine learning algorithms perform best when they are well-structured and uniform in length, and we have the means to transform texts of varying lengths into such inputs. The first step is preparing the data to be analyzed. All capitalization and punctuation within words must be altered to lowercase, and all other characters must be removed. Next, we'll look for the words that appear most frequently in the text. A vocabulary list should be created, phrases should be tokenized into individual words, and the frequency of use of each word should be calculated. The model is built afterward. Counting how often a word appears in text requires creating a vector. If it is a common word, it gets a 1, but else it gets a 0.

Sentence 1 : The Book is in front of the table.

Sentence 2 : The pen is under the table.

Model Vocabulary [7] = [book,front ,in,of,table,pen,under]

Vectorized Sentence1 = [1,1,1,1,1,0,0]

Vectorized Sentence2 = [0,0,0,0,1,1,1]

3.2.2 mLSTM

mLSTM is a multiplicative recurrent neural network architecture that incorporates with the long short term memory (LSTM).It is to be believed that its mLSTM is more expressive for estimating autoregressive densities because it has many recurrent transition function for each potential input.This architecture aims to combine the extended time delayed and overall functioning of LSTMs with the customizable input-dependent transformation of mRNNs. It might be simpler to control or avoid the complex transitions that arise from the factorized concealed weight matrix due to the LSTMs' gated units.

Even more flexible input-dependent transformation functions than in standard mRNNs are possible thanks to the additional sigmoid input and forget gates present in LSTM units. The overall performance of the low-tier type models after one data set analysis to help with problem solving.In addition , this preprocessing was selected for the mass scale experiment which is single layer working with multiplicative LSTM.In simpler word this preprocessing works good and more accurate when the data sentences are long.

3.2.3 TF/IDF

The TF, or Term Frequency, and the IDF, or Inverse Document Frequency The statistical significance of a word in a collection of documents is determined. In order to calculate how often a word appears in a given document, we multiply its inverse document frequency by itself.

In machine learning algorithms, it is most beneficial for scoring text analysis; this is especially true in NLP and a number of other applications, the most prominent of which is automated text analysis.

TF-IDF was invented to increase the efficiency of the retrieval system. It increases the frequency with which a word appears in a document, but this impact is neutralized by the quantity of papers that include the phrase. Even though they exist often in all writings, words like this, what, and if are unclassified since they have little to do with the specific content.

In contrast, if the term Bug appears frequently in one document but not in others, it is considered to be extremely significant. When classifying NPS(Net Promoter Score) answers into topics, the word Bug is likely to be related with the subject Reliability since most responses including the word Bug will be about the issue of reliability.

The frequency of a word or phrase is the number of times it appears in a given text. One way to quantify this is by tallying the occurrences of a key phrase in the original text. Depending on the page's length and the frequency of the most frequent phrases, the individual can modify the frequency. Change the direction of the document frequency graph for a given term over a collection of documents. This represents how common or rare a term is over the full text corpus. Words and phrases that are closer to zero in frequency are more common. Logarithmically dividing the total number of documents by the number of documents that include a specific phrase yields this statistic. As a result, if the term is widely used and appears in numerous publications, its count will become close to 0. Unless otherwise specified, the value will be 1. The TF-IDF score of a word in a document is calculated by multiplying its frequency of occurrence by its lexical similarity to a specified term. A higher weighting indicates greater significance in the text. It can also be written as the formula for determining the TF-IDF score of the word t in document d from set D which is given below :

$$TF\left(i,j\right) = \frac{n\left(i,j\right)}{\sum\ n\left(i,j\right)}$$

$$IDF\ =\ 1 + log\left(\frac{n}{dn}\right)$$

$$TF - IDF\ =\ TF\ \times IDF$$

Chapter 4

Methodology

4.1 Logistic Regression(LR)

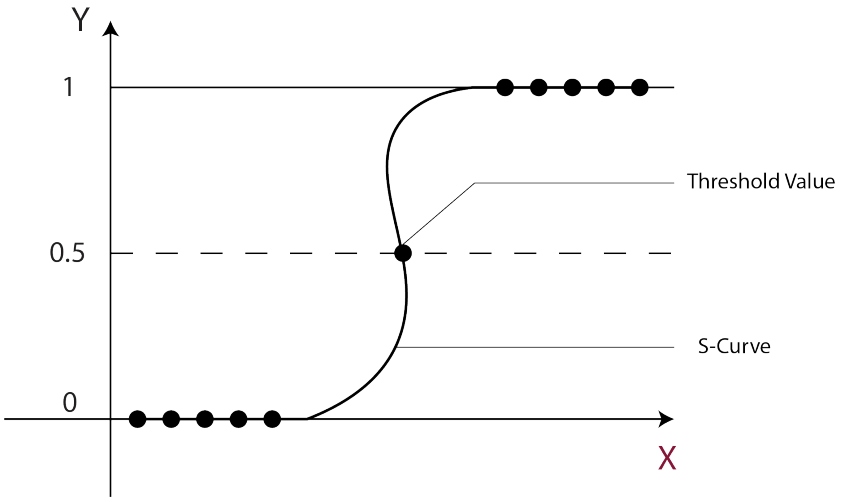

Figure 4.1: Logistic Regression Classifier

In the Supervised Learning framework, one of the most often employed Machine Learning algorithms is logistic regression. It is employed to make forecasts about a categorical dependent variable using a number of independent variables. The goal of logistic regression is to predict the value of a categorical dependent variable. Therefore, the solution must be a single, deterministic number. It's possible to get probabilistic values between 0 and 1 instead of absolute ones like Yes or No otherwise 0 or 1, true or false, etc.

The two forms of regression analysis, Logistic Regression and Linear Regression, are relatively similar, with the exception of their separate uses. For Regression-related difficulties, statisticians turn to Linear Regression, while for Classification issues, they use Logistic Regression.

A logistic function, shaped like an S, is fitted in logistic regression to predict two maximum values instead of a single one (0 or 1).

If cells are malignant, whether a mouse is obese based on weight, etc., are all represented by curves on the logistic function.

Logistic Regression is an essential machine learning method due to its ability to create probabilities and classify fresh data from both continuous and discrete datasets.

Logistic Regression can be used to classify observations based on a wide range of data types and can easily identify the most effective classification factors. The **figure** 4.1 portrays the representation of Logistic Regression Classifier.

4.2 Support vector machines (SVM)

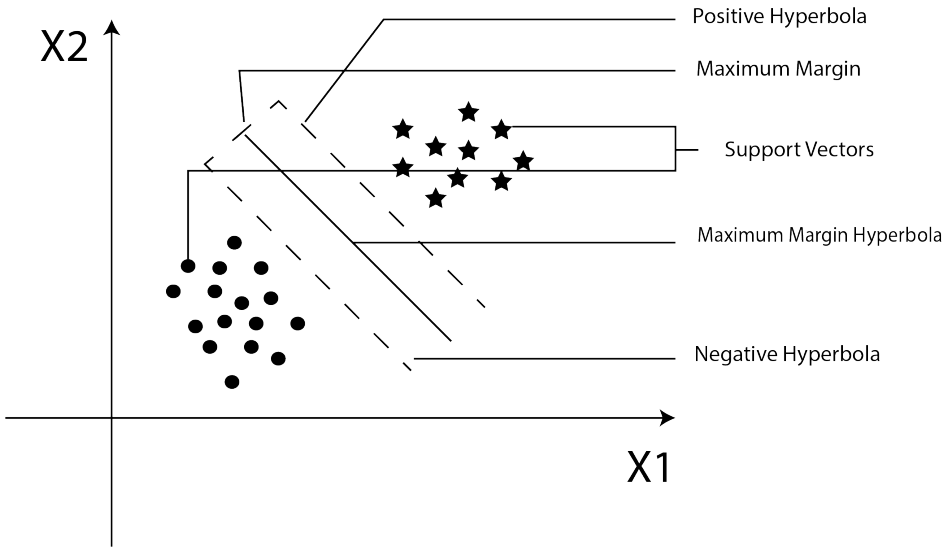

Figure 4.2: SVM Classifier

Support Vector Machine, or SVM, is one of the most popular classification and regression techniques in Supervised Learning. In Machine Learning, it is primarily used for Classification problems.

The goal of the SVM algorithm is to generate the optimal line or decision boundary that divides n-dimensional space into classes, thereby facilitating the classification of subsequent data points. This optimal decision limit is known as a hyperplane.

SVM chooses the extreme points or vectors that help make the hyperplane. Support vectors are these extreme cases, and the technique that uses them is called the Support Vector Machine. A simple linear SVM classifier makes a straight line between the two groups. This means that all of the data points on one side of the line will represent one category, and all of the data points on the other side will represent a different category. This means that there are no limits on how many lines can be used.

The linear SVM algorithm is superior to others, such as k-nearest neighbors, because it classifies your data points using the optimal line. It chooses the line that separates the data that is farthest from the data points that are closest.

A two-dimensional illustration clarifies the terminology of machine learning. You have essentially a grid of data points. You are separating these data points by the category to which they belong, but none of them should fall into the wrong category. This indicates that you are attempting to identify the line connecting the two closest points that maintains the distance between the remaining data points.

Consequently, the two nearest data points provide the necessary support vectors for locating the line. This is referred to as the decision limit.The **figure** 4.2 is the representation of SVM Classifier.

4.3 Random Forest(RF)

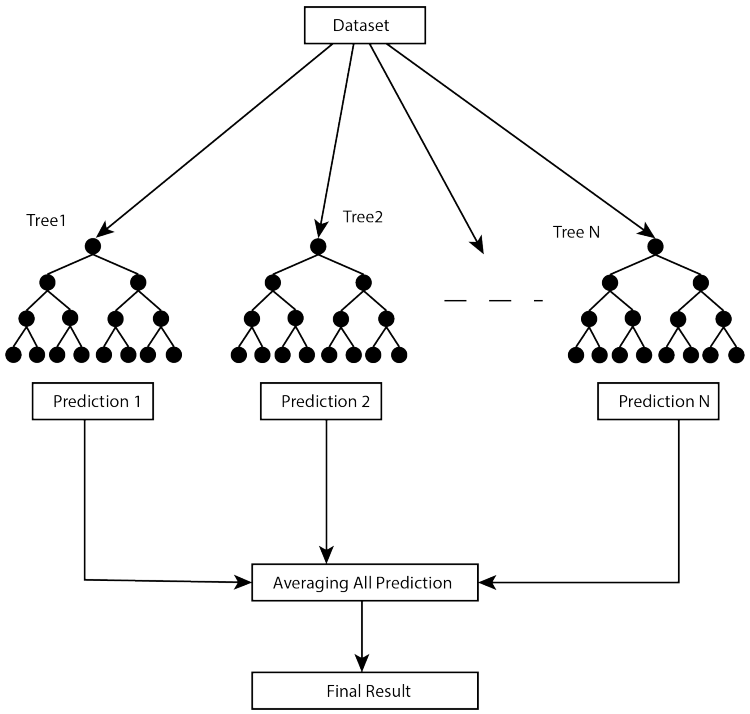

Figure 4.3: Random Forest Classifier

Supervised machine learning techniques, like random forest, are commonly employed in classification and regression issues. Different from regression, where the sample vote average is used, classification uses the sample with the greatest vote count. The Random Forest Algorithm's flexibility in handling data sets with both continuous and categorical variables, for use in tasks such as regression and classification, is one of its most valuable qualities. The results for problems with classifying things are enhanced.

Ensemble refers to the combination of multiple models. Consequently, a collection of models, as opposed to a single model, is used to generate predictions. Ensemble utilizes two distinct methodologies.

1.Bagging - In bagging, training data samples are replaced randomly to generate a new training subset, and the final result is determined by a majority vote of the

training data. The Random Forest algorithm is an example.

2.Boosting - Boosting transforms ineffective learners into effective ones through the creation of sequential models with increasing precision. For example, ADA BOOST, XG BOOST.

When constructing a tree, not all attributes/variables/features are considered because each tree is unique. Due to the fact that each tree does not consider all features, the feature representation is minimized. Each tree is constructed uniquely using unique data and attributes. This means we can utilize the CPU to its maximum capacity when constructing random forests. We do not need to separate the data for testing and training in random forest because 30

Random Forest is widely used in many industries because it works so well. It can handle data that is binary, continuous, or categorized. One of the best things about the random forest is that it can handle missing values. This makes it a great choice for anyone who wants to build a model quickly and easily. Random forest is a quick, simple, versatile, and robust modeling technique, despite its limitations.The **figure** 4.3 is the representation of Random Forest classifier.

4.4 Extreme Gradient Boosting(XGB)

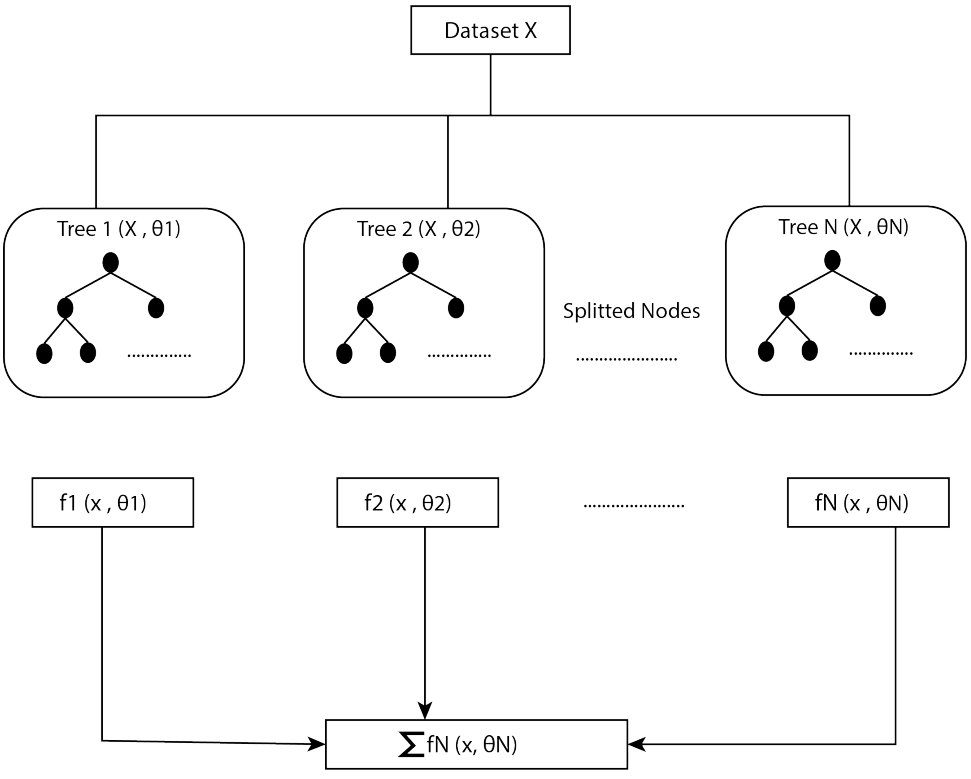

Figure 4.4: XGB Classifier

XGBoost is a machine learning approach. If you have previously predicted data, you can anticipate any form of data more accurately. You can classify any kind of

data. Additionally, it may be used to categorize text.

Gradient Boosting, the same as Random Forest (another decision tree methodology), is a technique for supervised classification problems including classification (male, female) and regression (expected value). Gradient Boosting machines (abbreviated GBM) and XGBoost are the most popular terms for the implementations of this technique. Numerous recent Kaggle tournaments have been won by XGBoost, which has led to its broad adoption.

Gradient Algorithm is an optimization learner like Random Forest approach. This signifies that a final model will be produced by combining many models. Individually, these models are incapable of accurate prediction and prone to errors, but combining a large number of them into an ensemble might yield a superior outcome. Similarly to Random Forests, decision trees are the most popular weak model in Gradient Boosting machines.

The elegance of this potent method lies in its flexibility, which enables quick learning via simultaneous and distributed processing and provides optimal memory use.

It is therefore not unexpected that CERN found it the most effective way for categorizing Large Hadron Collider signals. CERN required a scalable system capable of processing 3 petabytes of data per year and successfully isolating an extremely uncommon signal from background noise through a sophisticated physical process in order to solve this particular challenge. XGBoost emerged as the most efficient, straightforward, and resilient alternative.

Chapter 5

Result Analysis

5.1 Precision, Recall and F1 Score

It is essential to assess the efficiency of categorization models in order to make effective use of them in production for the purpose of finding solutions to actual problems. Because of this, the performance metrics of effective classification models are used to evaluate how well the method performed in relation to a set of predetermined criteria. In addition, we utilize four distinct kinds of evaluation criteria, which comprise F1-score, Recall, Precision, and Accuracy respectively. With the aid of these four performance indicators, we are able to assess the capabilities of the algorithm and the constraints of its applicability when producing predictions. Despite the fact that these four performance criteria are dependent on certain measurements that are being produced when predictions are being made.

True Positive(TP) : The true positive indicates the extent to which the classifer properly predicts the positive class. In other terms, the occurrence is positive, just as the model anticipated. When determining how often true positives our model predicts properly, real positives are crucial. True positives are significant because they demonstrate the model's performance when confronted with of positive data.

False Positive(FP): False positives arise when the algorithm is inaccurately assumes that an instance belongs to a class when it actually does not. False positives can be problematic since they can lead to poor conclusions. In simpler language are so many methods for determining false positives, including false positive rates. The false positive rate refers to the proportion of all negative instances that are wrongly regarded as positive.

True Negative(TN): The negatives are actually the occurrences for which the correct identification model is negative. True negatives are a measurement used to evaluate a classification effectiveness of the algorithm. A significant proportion of negative cases is frequently indicative of a well-performing model.

False Negative(FN): A false negative occurs when a model incorrectly interprets a good outcome as negative. Particularly in the health care field, false negative results may be tremendously costly. False negatives are often more destructive than false - positive, thus they must always be taken into consideration when assessing

the effectiveness of a classification algorithm.

Accuracy: The precision of a classifier is a measurable statistic for learning algorithms. In addition, calculated ratio of accurate positive data to accurate negative observations. In simpler words, accuracy is the percentage of score that our model correctly guessed an outcome out of all the cases in which it made a judgment.

$$AccuracyScore = \frac{TP + TN}{TN + FN + TP + FP} \tag{5.1}$$

Recall (RS): The model's ability to accurately estimate positive instances is assessed by the algorithm recall score. This differs significantly from precision, which evaluates the fraction of correct positive predictions made by models that are accurate. Recall grades demonstrate how effectively a machine-learning algorithm can differentiate between the positive and negative data. Greater recall scores are preferable. Other terms for recall include sensitivity or true positive rate. The higher recognition score indicates whether effectively the model can identify instances of achievement.

$$RecallScore(RS) = \frac{TP}{FN + TP} \tag{5.2}$$

Precision Score (PS): The proportion of categories for which positive predictions were produced effectively is reflected by a specific classifiers accuracy score. Positive predictive ability is an equivalent term for accuracy. When classes are imbalanced, the score of precision is a useful measure of the forecast's accuracy. It mathematically represents the fraction of accurate positives towards the total amount of genuine positives and misleading positives.

$$PrecisionScore(PS) = \frac{TP}{FP + TP} \tag{5.3}$$

F1 Score (FS): This score portrays the classifiers score as a function of its recall and precision scores. F-score is an alternative to Accuracy metrics that provides equal merit to both Recall and Precision when evaluating the effectiveness of a deep learning classifier with respect to accuracy. It is frequently employed as a single number that offers high-level information on the quality and performance of the model.

$$F1Score(FS) = \frac{2 * PS * RS}{PS + RS} \tag{5.4}$$

5.2 Result

We collected five datasets from different sources. The datasets are eBay, Steam, IMDb, Bangla Language, Cross Language (Hate speech). Initially we categorized the datasets as binary datasets.In addition, we considered only two mandatory columns

as “label ” and “sentence”. Furthermore, we worked on three different preprocessing algorithms such as TF/IDF,Bag of words, mLSTM. Afterwards, we used four different classification algorithms with individual preprocessing techniques to find a better combination which will work on multilingual datasets such as Bangla, English and Cross language. For example:We took each dataset and ran it through with each preprocessing algorithm and classification models. Moreover, after each completion we noted the four performance metrics. Additionally, we computed the precision, recall, and F1-score to every binary dataset. In a nutshell,the five dataset were run through in each possible combination of preprocessing and classification algorithms.The final outcome is organized by dataset type and tokenization procedure.The result of classifiers are listed in the following table 5.1

		TF/IDF				Bag of Words				mLSTM			
Dataset		**LR**	**SVM**	**RF**	**XGB**	**LR**	**SVM**	**RF**	**XGB**	**LR**	**SVM**	**RF**	**XGB**
eBay	**Precision**	0.998	0.999	0.998	0.994	0.985	0.977	0.997	0.996	0.986	0.983	0.980	0.994
	Recall	0.975	0.977	0.973	0.982	0.991	0.998	0.974	0.945	0.993	0.986	0.998	0.988
	F1	0.987	0.988	0.985	0.988	0.988	0.987	0.986	0.970	0.989	0.984	0.989	0.991
	Accuracy	0.973	0.9761	0.971	0.977	0.977	0.975	0.972	0.942	0.979	0.97	0.977	0.982
Steam	**Precision**	0.865	0.865	0.919	0.854	0.825	0.543	0.908	0.549	0.768	0.835	0.723	0.885
	Recall	0.699	0.714	0.592	0.697	0.691	0.927	0.594	0.910	0.862	0.741	0.891	0.788
	F1	0.773	0.782	0.720	0.768	0.752	0.684	0.718	0.685	0.813	0.785	0.798	0.834
	Accuracy	0.791	0.802	0.707	0.787	0.777	0.648	0.707	0.656	0.837	0.812	0.814	0.855
IMDb	**Precision**	0.826	0.846	0.771	0.778	0.820	0.812	0.760	0.850	0.906	0.906	0.905	0.921
	Recall	0.777	0.799	0.727	0.736	0.795	0.762	0.732	0.618	0.922	0.890	0.917	0.918
	F1	0.801	0.822	0.749	0.757	0.807	0.786	0.746	0.716	0.914	0.898	0.911	0.919
	Accuracy	0.795	0.816	0.741	0.750	0.805	0.779	0.741	0.663	0.913	0.897	0.910	0.919
Bangla-Language	**Precision**	0.713	0.820	0.885	0.797	0.782	0.645	0.887	0.438	0.753	0.765	0.934	0.653
	Recall	0.757	0.908	0.938	0.856	0.736	0.854	0.931	0.728	0.678	0.826	0.920	0.780
	F1	0.735	0.861	0.911	0.825	0.758	0.735	0.908	0.547	0.714	0.795	0.927	0.711
	Accuracy	0.7655	0.880	0.921	0.847	0.787	0.789	0.918	0.670	0.752	0.819	0.934	0.758
Cross -Language	**Precision**	0.680	0.680	0.727	0.657	0.567	0.653	0.653	0.387	0.439	0.613	0.389	0.497
	Recall	0.622	0.622	0.552	0.525	0.673	0.609	0.570	0.637	0.630	0.417	0.420	0.419
	F1	0.650	0.650	0.627	0.584	0.616	0.630	0.609	0.481	0.517	0.497	0.404	0.454
	Accuracy	0.780	0.780	0.741	0.719	0.748	0.770	0.748	0.750	0.647	0.627	0.628	0.642

Table 5.1: Classification report with accuracy

5.2.1 Bag of Words Tokenized Binary Dataset

		Bag of Words			
Dataset		**LR**	**SVM**	**RF**	**XGB**
eBay	**Precision**	0.985	0.977	0.997	0.996
	Recall	0.991	0.998	0.974	0.945
	F1	0.988	0.987	0.986	0.970
	Accuracy	0.977	0.975	0.972	0.942
Steam	**Precision**	0.825	0.543	0.908	0.549
	Recall	0.691	0.927	0.594	0.910
	F1	0.752	0.684	0.718	0.685
	Accuracy	0.777	0.648	0.707	0.656
IMDb	**Precision**	0.820	0.812	0.760	0.850
	Recall	0.795	0.762	0.732	0.618
	F1	0.807	0.786	0.746	0.716
	Accuracy	0.805	0.779	0.741	0.663
Bangla-Language	**Precision**	0.782	0.645	0.887	0.438
	Recall	0.736	0.854	0.931	0.728
	F1	0.758	0.735	0.908	0.547
	Accuracy	0.787	0.789	0.918	0.670
Cross -Language	**Precision**	0.567	0.653	0.653	0.387
	Recall	0.673	0.609	0.570	0.637
	F1	0.616	0.630	0.609	0.481
	Accuracy	0.748	0.770	0.748	0.750

Table 5.2: Bag of words tokenized Binary Dataset

From the table 5.2, we are witnessing that when we used Bag of Words as our pre-processing algorithm for all the datasets we get some promising result.From all the results we can see that Logistic Regression outperforms in most of the datasets except for Bangla Language and Cross-language dataset having the accuracy of 0.977, 0.777, 0.805 which is from eBay, steam, IMDb datasets respectively.However,for the Bangla Language and Cross-Language Datasets Random Forest and SVM stands out from all other text classifiers with the accuracy of 0.918 and 0.770 respectively. Now we will discuss the performance of each model using F1 Score , Precision and Recall Firstly, Random Forest has the best precision value from 3 of the dataset out of 5. Which means Random Forest has predicted correctly relative to total positive predictions from 3 of the dataset very accurately. In the other 2 dataset which are IMDb and Cross-Language, has the precision of 0.850 and 0.653 for XGB and SVM respectively. As for the recall we can see that in the dataset eBay Where the text classifier SVM has the highest value of 0.998 and for the other datasets it's 0.927, 0.762, 0.854, 0.609 respectively. For IMDb and Cross-Language dataset Logistic Regression has the highest value of 0.795 and 0.673 respectively out of all other classifiers. On the other hand in Bangla dataset Random Forest has the highest recall value of 0.931 which is very high compared to other classifiers and it means that predicted correct positives are related to total actual positives. In Case of F1 score we found that Logistic Regression(LR) has the highest value of 0.988, 0.752

and 0.807 F1 score for 3 datasets which are eBay, Steam and IMDb. On the other hand for Bangla Language and Cross-Language the value of Random Forest and SVM respectively were much higher than other classifiers. Random Forest(RF) had a much higher value of 0.908 F1 score on Bangla Language Dataset and on Cross-Language Dataset SVM was much better than other classifiers with the F1 score of 0.630. In conclusion, while using Bag of Words as a pre-processing algorithm we have come acrossed and have seen that LR, SVM and RF were the models Which were giving us excellent accuracy and accurate results.So, We can say that These are the best classifiers for Bag of Words Pre-processing algorithm.

5.2.2 mLSTM Tokenized Binary Dataset

		mLSTM			
Dataset		**LR**	**SVM**	**RF**	**XGB**
eBay	**Precision**	0.986	0.983	0.980	0.994
	Recall	0.993	0.986	0.998	0.988
	F1	0.989	0.984	0.989	0.991
	Accuracy	0.979	0.97	0.977	0.982
Steam	**Precision**	0.768	0.835	0.723	0.885
	Recall	0.862	0.741	0.891	0.788
	F1	0.813	0.785	0.798	0.834
	Accuracy	0.837	0.812	0.814	0.855
IMDb	**Precision**	0.906	0.906	0.905	0.921
	Recall	0.922	0.890	0.917	0.918
	F1	0.914	0.898	0.911	0.919
	Accuracy	0.913	0.897	0.910	0.919
Bangla-Language	**Precision**	0.753	0.765	0.934	0.653
	Recall	0.678	0.826	0.920	0.780
	F1	0.714	0.795	0.927	0.711
	Accuracy	0.752	0.819	0.934	0.758
Cross -Language	**Precision**	0.439	0.613	0.389	0.497
	Recall	0.630	0.417	0.420	0.419
	F1	0.517	0.497	0.404	0.454
	Accuracy	0.647	0.627	0.628	0.642

Table 5.3: mLSTM Tokenized Binary Dataset

In the table 5.3,we can see the four performance metrics of mLSTM with four different classification models in which the highest accuracy is 0.982 for XGB in eBay datset. It clearly indicates that XGB performed very well comparing to other classification model. Furthermore,in Steam we also witness that XGB classification model scores are outperforming the other classification model which was 0.837, 0.812, 0.814, 0.855. So it clarifies that XGB is again performing well in the Steam dataset in which the accuracy is 0.855. Afterwards, we can see that the IMDb dataset with the XGB classification model performed well, having a score of 0.919. But keeping in mind that ,Logistic Regression(LR) and Random Forest(RF) scored a similar kind of value ,having 0.913 and 0.910.On the other hand,in bangla language dataset

shows unique result in the Random Forest classifier. Random Forest classifier perform very well in bangla dataset , having a accuracy of 0.934. It is unique in a sense that compared to the whole table only Random Forest (RF) scored this good in Bangla language dataset.Moreover for Cross-language dataset Logistic Regression(LR) performed well from most of the classifiers as it scored 0.647. The rest of the classifiers such as SVM,RF,XGB scored 0.627, 0.628 ,0.642 respectively. From the whole table, eBay dataset has a very high precision value of 0.986, 0.983, 0.980, 0.994 from which XGB scored the highest precision value. Furthermore, Random Forest scored 0.989 recall value in eBay dataset.Both having quite similar accuracy and precision value but it does not necessary mean that the model classifier model are equal. From Steam dataset,XGB again performed well in precision score from the rest of the classifier model in which each classifier model scores 0.768 , 0.7835 , 0.723 , 0.885 . On the other hand, Random Forest(RF) scored 0.891 in recall score which is higher among other recall value of classifiers. Furthermore, in IMDb dataset Logistic Regression(LR) and SVM scores similar kind of value in precision score, having value of 0.906 which indicates both classifier worked well. But XGB is the classifier which stands out the most in precision score, having 0.921.However,Logistic Regression (LR) scored higher in recall value which is 0.922. It means a high recall algorithm returns the majority of the pertinent results. In the Bangla-Language dataset ,Random Forest outperformed most of the classification models in precision value which is 0.934. No other classification algorithm scored this well throughout our research.In Cross-Language dataset ,SVM has the higher precision value which is 0.613 compared to other scores . On the other hand, Logistic Regression (LR) scores higher in recall value , having 0.630 . To conclude , after analyzing the table we can state that mLSTM preprocess works well with Random Forest , XGB classifiers.These two classifier scores very well in each dataset.

5.2.3 TF/IDF Tokenized Binary Dataset

		TF/IDF			
Dataset		**LR**	**SVM**	**RF**	**XGB**
eBay	**Precision**	0.998	0.999	0.998	0.994
	Recall	0.975	0.977	0.973	0.982
	F1	0.987	0.988	0.985	0.988
	Accuracy	0.973	0.9761	0.971	0.977
Steam	**Precision**	0.865	0.865	0.919	0.854
	Recall	0.699	0.714	0.592	0.697
	F1	0.773	0.782	0.720	0.768
	Accuracy	0.791	0.802	0.707	0.787
IMDb	**Precision**	0.826	0.846	0.771	0.778
	Recall	0.777	0.799	0.727	0.736
	F1	0.801	0.822	0.749	0.757
	Accuracy	0.795	0.816	0.741	0.750
Bangla-Language	**Precision**	0.713	0.820	0.885	0.797
	Recall	0.757	0.908	0.938	0.856
	F1	0.735	0.861	0.911	0.825
	Accuracy	0.7655	0.880	0.921	0.847

Table 5.4 continued from previous page

Cross-Language	**Precision**	0.680	0.680	0.727	0.657
	Recall	0.622	0.622	0.552	0.525
	F1	0.650	0.650	0.627	0.584
	Accuracy	0.780	0.780	0.741	0.719

Table 5.4: TF/IDF tokenized Binary Dataset

In the following table 5.4, we have denoted the recall, F1 score and precision for each dataset and classification models in TF/IDF preprocessing. From analyzing the above table , we can find out the most consistent classification model for TF/IDF algorithm which is Support Vector Machine (SVM). SVM has the accuracy from each dataset having 0.976 , 0.802, 0.816, 0.880, 0.780 in eBay, Steam, IMDb, Bangla-Language, Cross-Language respectively. Compared to other results, the eBay review dataset is the most consistent dataset we came upon because its accuracy in each classification algorithm is quite similar and solid. From our conducted research it performed well throughout the whole process. Apart from that, rest of the datasets have their own distinct value from where Steam,IMDb,Bangla-Language,Cross-Language(Hate-Speech) have accuracy of 0.791, 0.795, 0.765, 0.780 in Logistic Regression. Logistic Regression provided the highest accuracy of 0.973 for eBay dataset but could not outperform the results of SVM which was 0.976. On the other hand,Random Forest classifiers scored very closely compared to SVM model scores. Random Forest scored 0.971 in the eBay dataset.

However, other datasets such as Steam,Bangla-Language,Cross-language(Hate-Speech), IMDb scored 0.707,0.921,0.741,0.741 respectively. We can see that the IMDb and Cross lingual(Hate-Speech) dataset has the same accuracy of 0.741 in Random Forest classification. But it does have different F1-scores of 0.627 and 0.749 in Cross-Language, IMDb respectively.However random forest has a accuracy that is similar to SVM but SVM has the highest precision value in table which is 0.999 . This indicate how well the SVM successfully predicted each of the test outcomes.However , XGB has the highest Recall value of 0.982 in eBay dataset which is slightly greater than SVM. A high recall number indicates that there were less false negatives and suggests that the classifier's conditions for categorizing some as positive were more liberal. In the steam dataset Random Forest(RF) classifier generates the second highest precision value of 0.919 which clarifies that random forest can also predict well as SVM. Lastly, SVM and XGB have the same highest F1-score in eBay dataset but not compared to the value of precision score. In spite of having similar F1-score doesn't necessary it would indicate that both models are equal. In conclusion, SVM, RF, and XGB are the best classification models for Binary TF/IDF. Each of these classifiers delivered a good precision, recall, and accuracy score.

5.3 Confusion Matrix

The findings of a prediction task involving a classification difficulty are summarized in a confusion matrix. Count values are used to calculate the sum of accurate and inaccurate predictions, which are then separated by class. This is the key to understanding the confusion matrix. It gives a glimpse into not just the errors your classifier is producing, but also the error categories. This separation eliminates the

disadvantage of depending entirely on categorization precision. This is an extremely effective way for calculating Recall, Precision, and Accuracy as well. Figure 5.1 is given to get a better understanding of Confusion Matrix

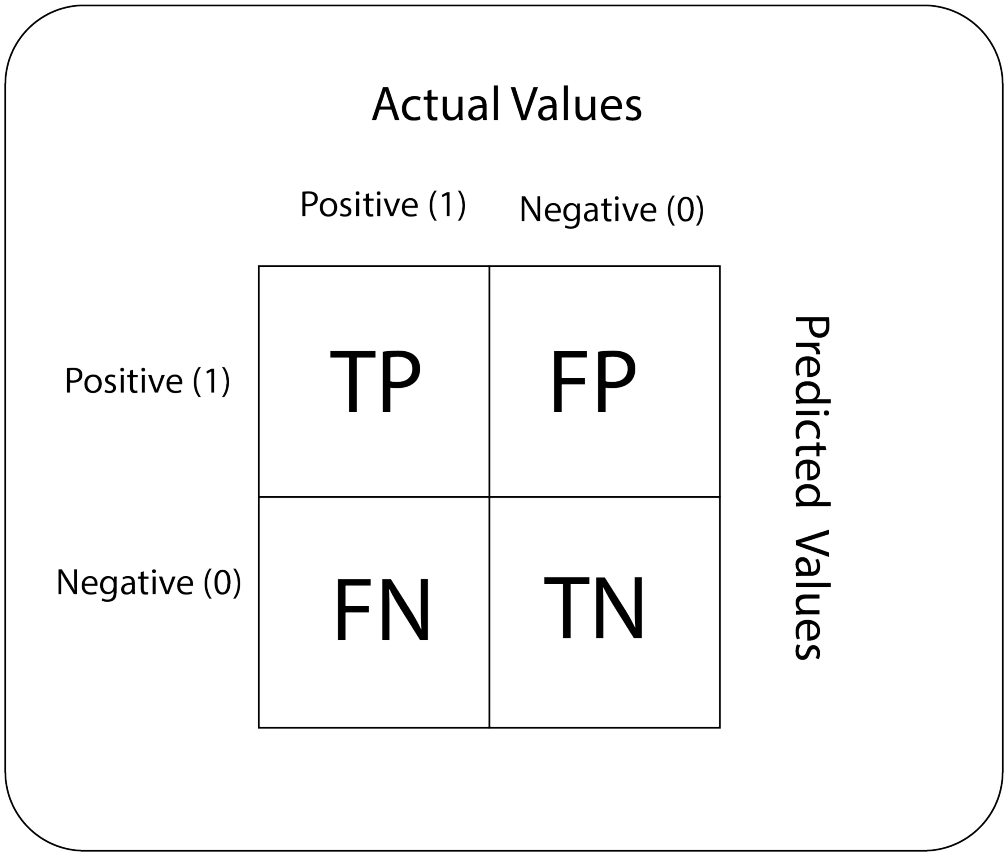

Figure 5.1: Confusion Matrix

On the following page are the Confusion Matrices for the Binary Datasets utilizing Bag of Words, mLSTM, and TF/IDF, which were done on several classifiers such as Logistic Regression, Support vector machines (SVM), Random Forest, and Extreme Gradient Boosting (XGB).

5.3.1 Confusion Matrix Based Result

The algorithms that were applied to detect if the feedback or comments were hate or non-hate can be easily determined by Confusion Matrix. It essentially demonstrates the differences between the true positives, false positives, true negatives, and false negatives. We have only used one type of Dataset which is Binary Dataset and it is labeled as hate and non-hate. Bag of words, mLSTM and TF-IDF have been used as our tokenization process. As we have 5 Datasets for which we will have 12 confusion matrices for each Datasets.

5.3.2 Confusion matrix of eBay Dataset

Confusion matrix of Cross-Language Dataset Shown in Figure 5.2

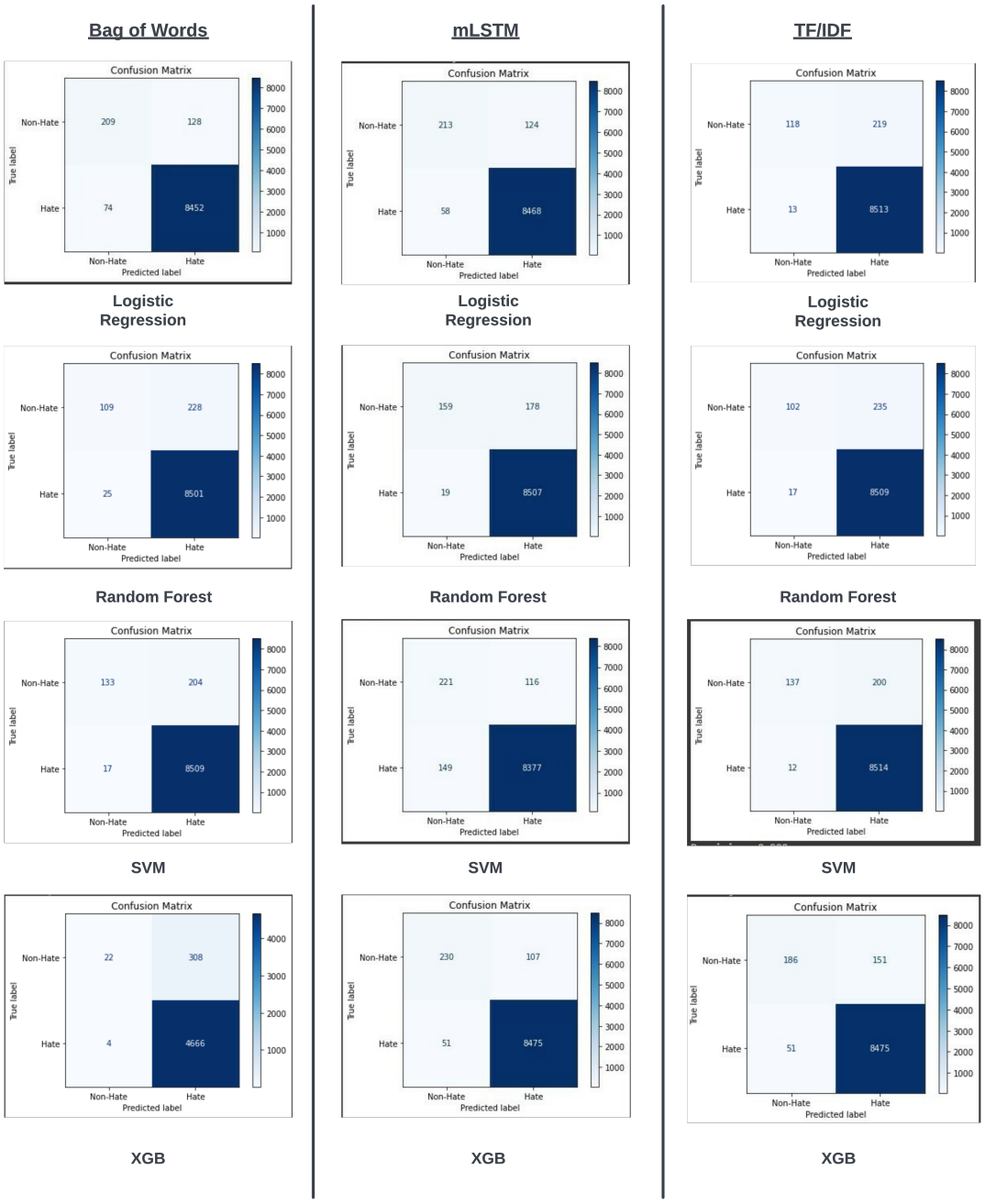

Figure 5.2: Confusion matrix of eBay Dataset

5.3.3 Confusion matrix of Steam Dataset

Confusion matrix of Cross-Language Dataset Shown in Figure 5.3

5.3.4 Confusion matrix of IMDb Dataset

Confusion matrix of Cross-Language Dataset Shown in Figure 5.4

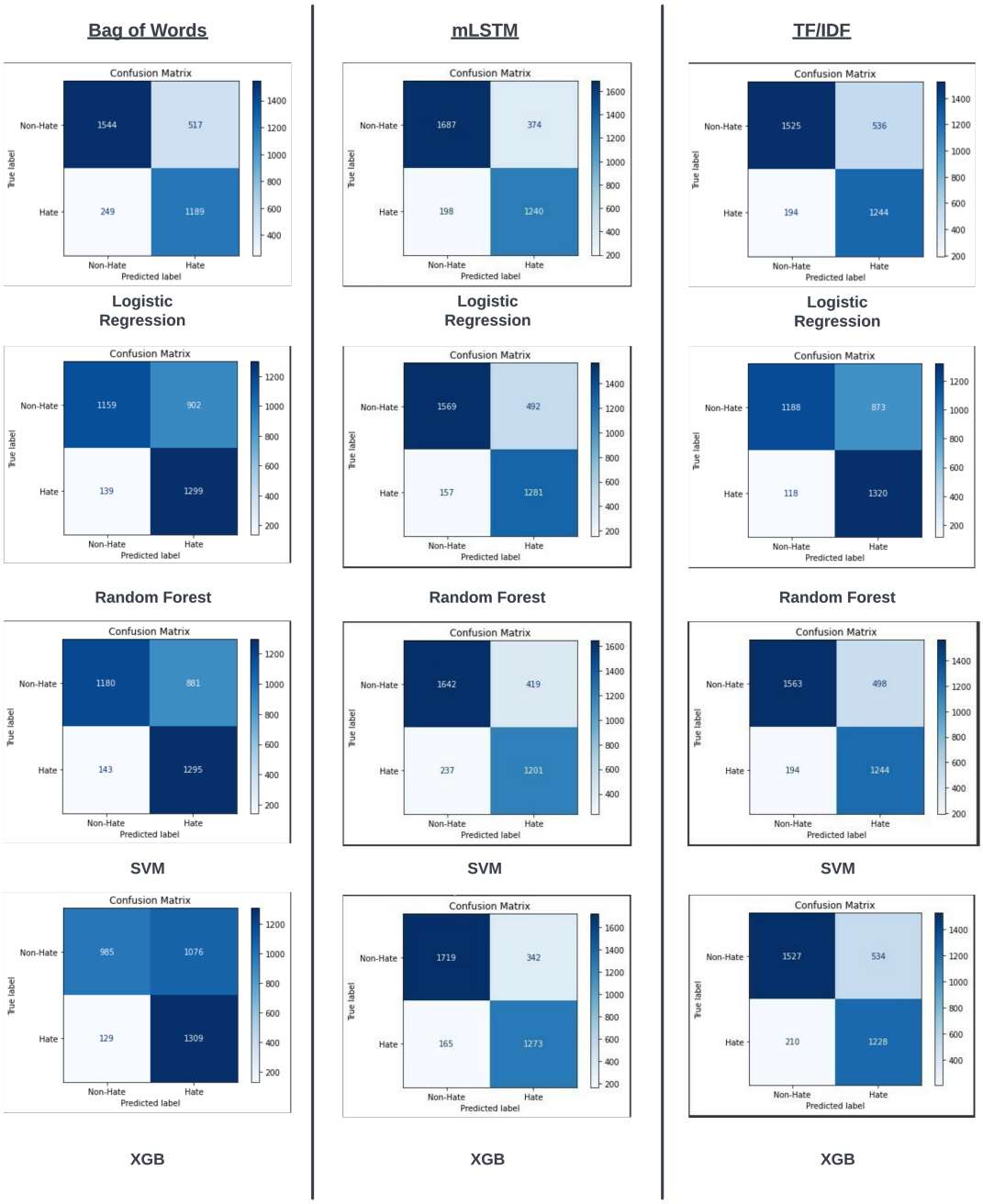

Figure 5.3: Confusion matrix of Steam Dataset

5.3.5 Confusion matrix of Bangla Dataset

Confusion matrix of Cross-Language Dataset Shown in Figure 5.5

5.3.6 Confusion matrix of Cross-Language Dataset

Confusion matrix of Cross-Language Dataset Shown in Figure 5.6

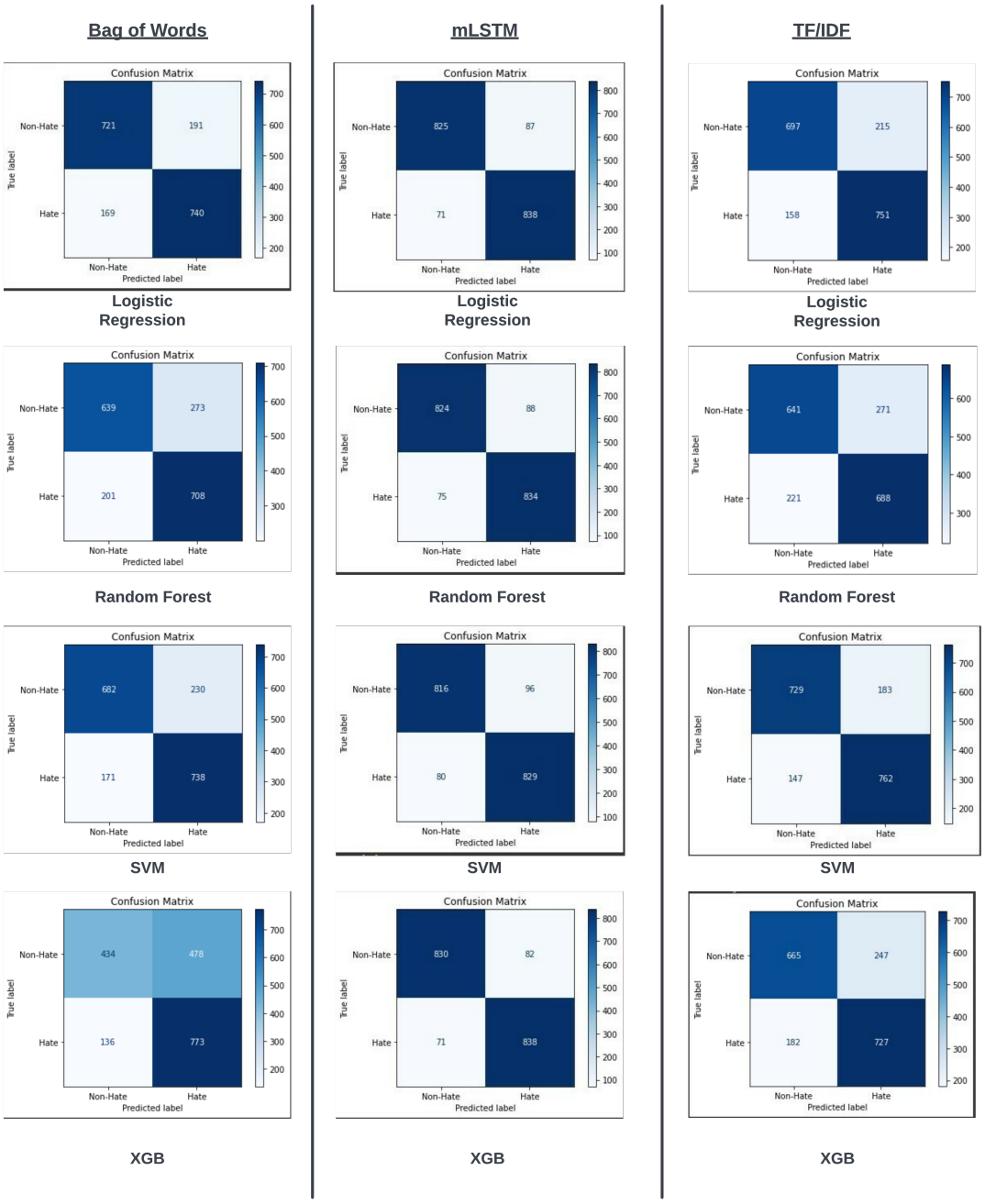

Figure 5.4: Confusion matrix of IMDb Dataset

5.4 Result Comparison

5.4.1 eBay Dataset

To find greater accuracy, we applied 12 different approaches of preprocessing and classifier. The figure 5.7 shows the XGB,RF,SVM,LR where the average result is 0.967, 0.973, 0.969 and 0.976 respectively. This indicates that both RF and LR, which have scores of 0.97 and 0.97 respectively, are doing better. In addition, LR is performing better for eBay.Moreover, eBay dataset is a very consistent dataset because overall, it performed well in every possible combination.But mLSTM with

Figure 5.5: Confusion matrix of Bangla Dataset

XGB has the highest accuracy result then rest of the classifier model, having a value of 0.982.

The dataset was produced for a project purpose at a data science bootcamp (Source). The project's objective was to create a sentiment analysis model. The author used his own Python web scraping programs to produce the dataset.However ,the predictions were presented by an author named "Bruno A. T. Freitas" ,having an accuracy of 88.2% . Whereas our mLSTM predicted an accuracy of 0.982.

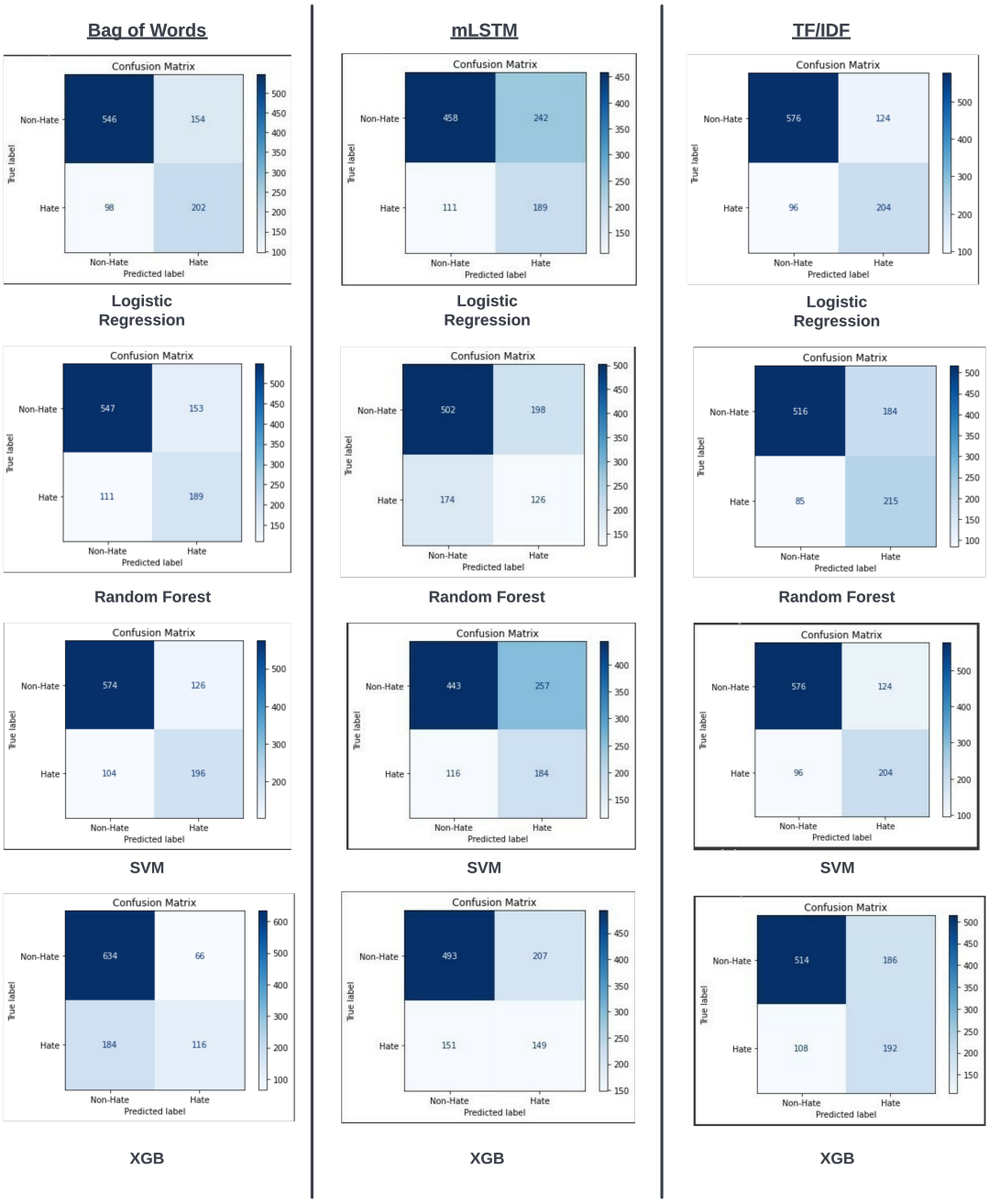

Figure 5.6: Confusion matrix of Cross-Language Dataset

5.4.2 Steam Dataset

We performed the methodology for the Steam dataset in all conceivable configurations, as shown in the figure 5.8. We first applied three preprocessing techniques to each classifier. For each preprocess and classifier model, we obtain average results. First, we receive an average LR score of 0.801. After that, we obtained an SVM average of 0.754 for each preprocessing technique. Additionally, we obtained average scores of 0.742 and 0.766 for Random Forest and XGB, respectively.

A similar study was carried out by author named as Zhen Zuo in 2018[12].The re-

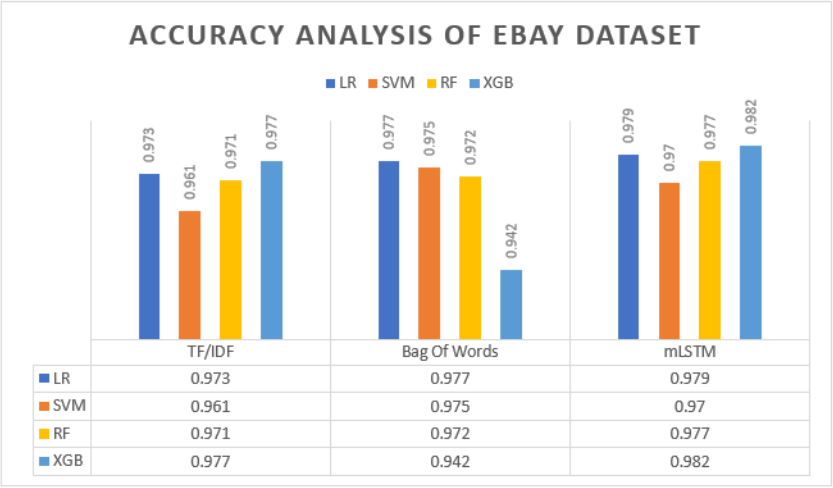

Figure 5.7: Accuracy Analysis for eBay Dataset

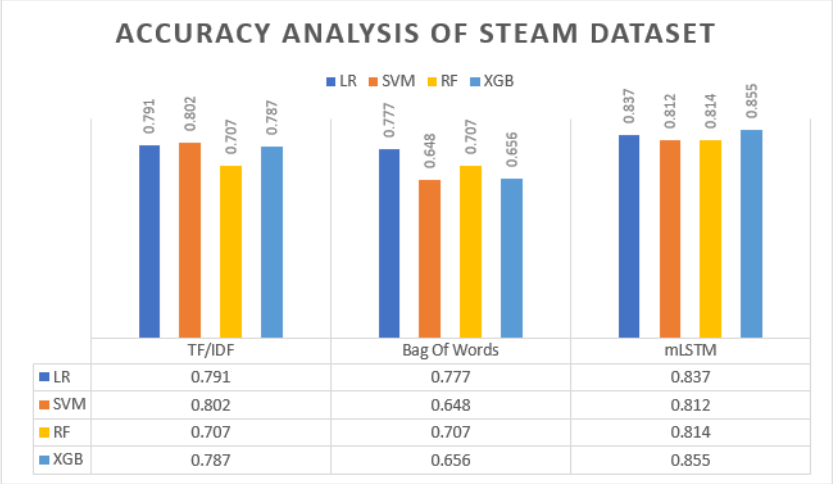

Figure 5.8: Accuracy Analysis for Steam Dataset

search was on "Sentiment Analysis of Steam Review Datasets using Naive Bayes and Decision Tree Classifier". In their research, they worked with a similar steam dataset in which they used Naive Bayes and Decision Tree Classifier. They conclude by stating that Decision Tree achieved accuracy of 0.75 in the Steam Dataset. Whereas , we got the highest accuracy of 0.855 with the same dataset.

5.4.3 IMDb Dataset

From the figure 5.9 of IMDb accuracy analysis , we take similar approach to find average results for classification model which are LR 0.837 , SVM 0.830 , for RF 0.797 for XGB 0.777 i.e LR is working better here and for pre processing model mLSTM it is 0.909, bag of words 0.747 and TF/IDF 0.775 so we can say that for all mLSTM is providing us better accuracy for all of this 4 pre-processing tech-

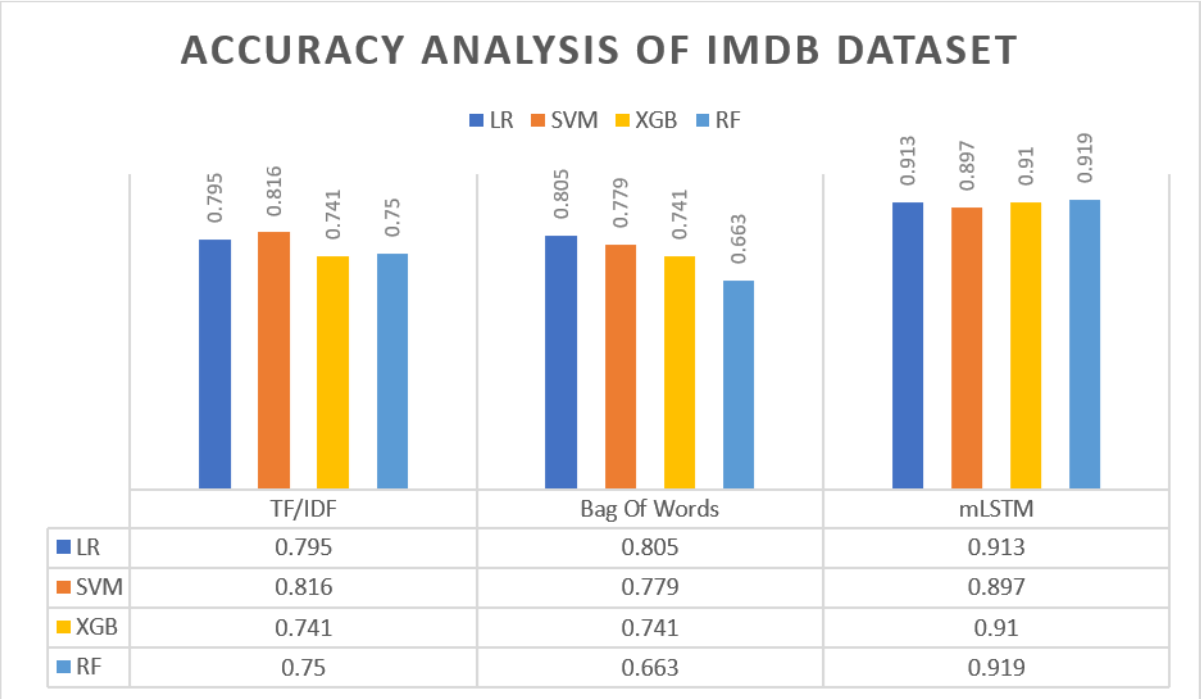

Figure 5.9: Accuracy Analysis for IMDb Dataset

niques and the classification model LR is providing higher accuracy. So we can say that, for the IMDB dataset the combination could be mLSTM. But here we got 0.919 using mLSTM and XGB. A similar type of research was conducted with this IMDB dataset.The research was on "Learning to Generate Reviews and Discovering Sentiment" conducted by authors named as Alec Radford; Rafal Jozefowicz; Ilya Sutskever [9].In their paper they conducted their research with byte mLSTM from which they got a highest accuracy result of 86.9 in this IMDB. Whereas,we got 0.919 in xgb using mLSTM.

5.4.4 Bangla Dataset

The accuracy analysis of the Bangla Language dataset, which we ran through 12 various preprocessing and classifier model combinations, is shown in the figure 5.10. At first, we used one preprocessing technique to run four classification algorithms. Furthermore, we select TF/IDF with Logistic Regression (LR), Standard Vector Machine (SVM), Random Forest (RF), and Extreme Gradient Boosting (XGB). Then, we repeat this procedure for every preprocessing algorithm. Additionally, the average LR score is 0.768. By carrying out this procedure further, we arrive at an average SVM score of 0.829. Additionally, RF has a 0.924 accuracy score on average. The final result is the XGB average score, which is 0.758. mLSTM with Random Forest, on the other hand, gets the highest accuracy score, coming in at 0.934. A similar type of research was conducted with this Bangla-Language dataset. The research was on "N-Gram Based Sentiment Mining for Bangla Text Using Support Vector Machine" conducted by authors named as SM Abu Taher;

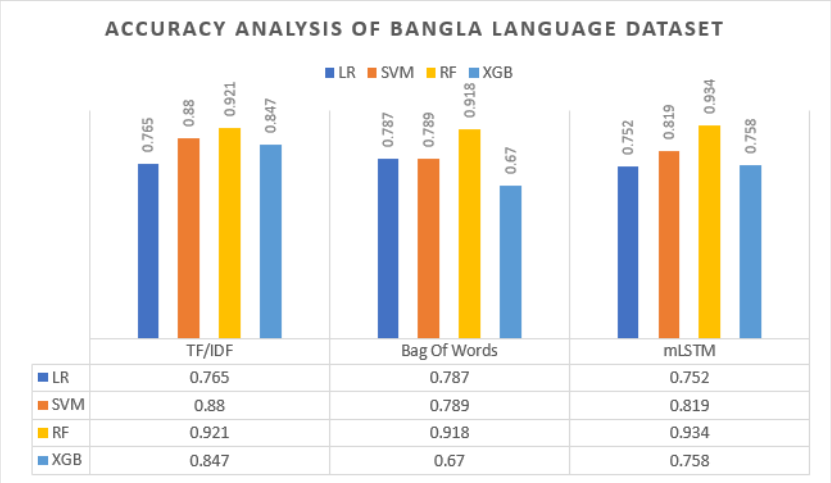

Figure 5.10: Accuracy Analysis for Bangla Dataset

Kazi Afsana Akhter; K.M. Azharul Hasan in the year 2018 [10].In their paper they conducted their research with N-Grames and SVM from which they got a highest accuracy result of 89.271 in this Bangla dataset. While they achieved 89.271, we only achieved 0.88 using SVM, but we did achieve a substantially better score of 0.934 in mLSTM using Random Forest.

5.4.5 Cross-Language Dataset

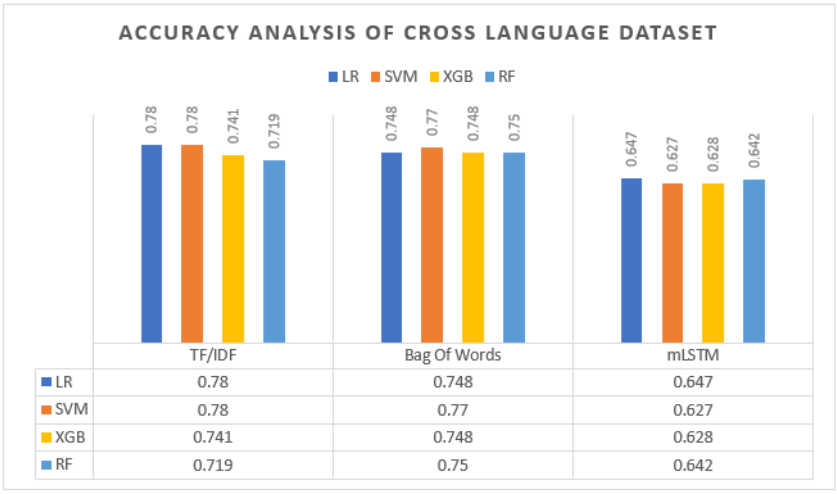

Figure 5.11: Accuracy Analysis for Cross-Language Dataset

The accuracy analysis of a cross-language dataset is shown in the figure 5.11, and as can be seen, TF-IDF performs well. Similar to SVM, Logistic Regression has almost the same accuracy. However, compared to other datasets, this one has the lowest score for both data pre-processing and the data analysis model. We obtain

an average LR of 0.725 from this. Additionally, the average for the SVM is 0.726. Additionally, we obtain an XGB accuracy average of 0.706. Last but not least, the average accuracy score for RF is 0.704. A very similar type of work was done on this Cross-Language dataset "Hate Speech Detection Using Machine Learning Techniques" which was conducted by the following author's Tahbib Manzoor, Md. Wahidur Rahmank, Monjurul Sharker Omi, Arpan Das Abir, Tanvir Ahmed Abir in the year 2022 [18]. In the paper they used 2 pre-processing algorithms and 8 classifiers and their best classifier for binary dataset in TF-IDF and in Bag of Words was Multinomial NB with the accuracy of 0.740 for both case which is the highest accuracy of from all other classifiers. Similarly they also applied SVM in which they got an accuracy result of 0.729. However, for our case we got our highest accuracy result in TF/IDF pre-processing with LR and SVM, having a similar score of 0.78.

5.4.6 Overview of the Comparison

An overview of the comparison between others has been shown in the table-5.5

Paper Name	**Datasets**	**Accuracy**	**Our Accuracy**
(**eBay Dataset**)	eBay	0.88	0.98
[12]	Steam	0.75	0.85
[7]	IMDb	0.86	0.91
[10]	Bangla	0.89	0.93
[18]	Cross Language	0.72	0.78

Table 5.5: Comparing with previous work

5.5 Overall Discussion

From our conducted research , we ran all possible combinations of preprocessing techniques with four different classification models on five datasets.Initially, we see from our accuracy analysis figure (5.7) that from eBay dataset where we got highest accuracy in mLSTM with XGB , having accuracy of 0.982. Among all preprocessing mLSTM worked well in the eBay dataset.After that, we continued our process in accuracy analysis figure (5.8) of Steam dataset where we got a similar result in which mLSTM with XGB has the highest accuracy , having a value of 0.855.In addition, in figure (5.9) of IMDb dataset, we got the highest value in accuracy with mLSTM and XGB where the accuracy score was 0.919. However,in figure (5.10) of Bangla dataset we got something different from the previous outcomes. The following illustration (5.12) provides a graphical representation of an accuracy curve for mLSTM preprocessing across five datasets.
We got a better accuracy for Bangla dataset with mLSTM and Random Forest(RF),having an accuracy score of 0.934.Lastly, we did the similar process for figure (5.11) of cross-language dataset where we got similar accuracy in both TF/IDF

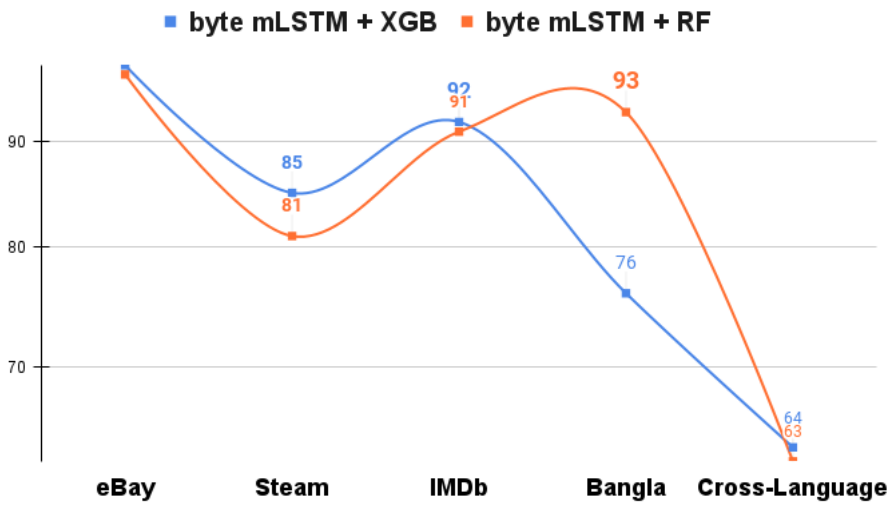

Figure 5.12: byte mLSTM accuracy curve

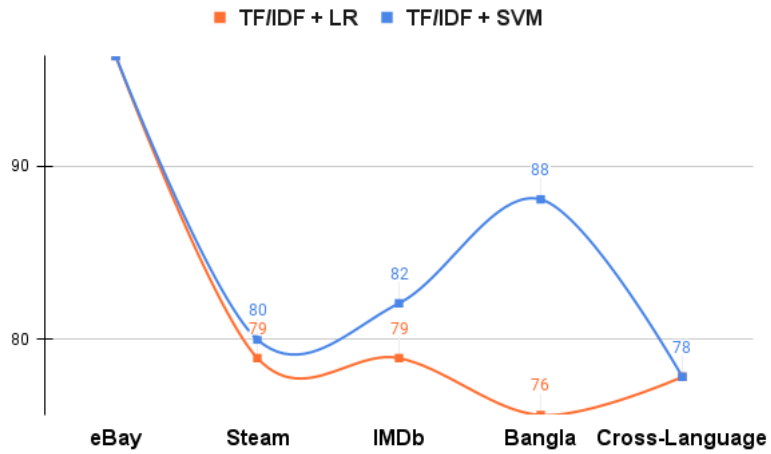

Figure 5.13: TF/IDF accuracy curve

preprocessing along with LR and SVM classifier model, having a accuracy of 0.78. Therefore, it is clear that mLSTM preprocessing performed better in four distinct datasets: Bangla-Language, IMDb, Steam, and eBay. English is used in the first three datasets, whereas Bangla is used in the last dataset. As a result, we can see that mLSTM performed well for both Bangla and English datasets. mLSTM can thus be a useful preprocessing approach for datasets that contain many languages. On the other hand, we note that TF/IDF generally performed well on cross-linguistic datasets. A graphical representation of TF/IDF accuracy curve is shown in the following figure (5.13). However, two classifier models—RF and XGB—stand out above the others when it comes to mSLTM preprocessing procedures. With English-based datasets, XGB with mLSTM performed better, but RF with mLSTM performed well with a dataset in Bangla. For the LR and SVM classification models, the TF/IDF technique also worked well. With TF/IDF in the Cross-Language dataset, both LR and SVM performed better.

Chapter 6

Conclusion

Different methodologies are required to accurately detect the sentiments from various types and categories of dataset. The majority of the time, datasets based on the English language are taken into account while building a classifier model. Only a small fraction of classifiers were trained using multilingual datasets, including Bangla datasets. Furthermore, our main objective was to carry out an experiment to discover a better preprocessing method for a classifier model. In response, we implemented mLSTM, a preprocessing method that performs well on datasets in both Bangla and English. As opposed to mLSTM, TF/IDF outperformed it in cross-lingual datasets. Additionally, whereas TF/IDF works well with LR and SVM, mLSTM performs better with RF and XGB.In contrast to TF/IDF, mLSTM and BOW performed only moderately well in cross-lingual datasets, which presented us with some challenges. However, due to the short phrases in the cross-language dataset, mLSTM was unable to function at its peak level. In addition, there are challenges we experienced while doing the research, which we want to overcome this in the future. To sum up, future researchers may observe and decide on the optimum pre-processing and classifier model combination to acquire the best possible results for multilingual and cross-language datasets.

Bibliography

[1] A. Kao and S. R. Poteet, *Natural language processing and text mining*. Springer Science & Business Media, 2007.

[2] S. S. Htay and K. T. Lynn, "Extracting product features and opinion words using pattern knowledge in customer reviews," *The Scientific World Journal*, vol. 2013, 2013.

[3] G. Nunnari and S. Nunnari, "Clustering and prediction of solar radiation daily patterns," in *Proceedings of the International Conference on Data Science (ICDATA)*, The Steering Committee of The World Congress in Computer Science, Computer …, 2016, p. 3.

[4] M. U. Islam, F. B. Ashraf, A. I. Abir, and M. A. Mottalib, "Polarity detection of online news articles based on sentence structure and dynamic dictionary," in *2017 20th International Conference of Computer and Information Technology (ICCIT)*, 2017, pp. 1–5. DOI: 10.1109/ICCITECHN.2017.8281777.

[5] O. R. Llombart, "Using machine learning techniques for sentiment analysis," *published in June of*, 2017.

[6] A. Mars and M. S. Gouider, "Big data analysis to features opinions extraction of customer," *Procedia computer science*, vol. 112, pp. 906–916, 2017.

[7] A. Radford, R. Jozefowicz, and I. Sutskever, "Learning to generate reviews and discovering sentiment," *arXiv preprint arXiv:1704.01444*, 2017.

[8] T. M. Rao, N. Mounika, K. H. Chowdary, and T. Sudhir, "A framework for generating rankings to e-commerce products based on reviews using nlp," *International Journal of Mechanical Engineering and Technology (IJMET) Volume*, vol. 9, 2018.

[9] E. Suganya and S. Vijayarani, "Sentiment analysis for scraping of product reviews from multiple web pages using machine learning algorithms," in *International Conference on Intelligent Systems Design and Applications*, Springer, 2018, pp. 677–685.

[10] S. Taher, K. Akhter, and K. M. Hasan, "Bangla dataset for opinionmining," Sep. 2018. DOI: 10.13140/RG.2.2.20214.96327.

[11] C. Vo, D. Duong, D. Nguyen, and T. Cao, "From helpfulness prediction to helpful review retrieval for online product reviews," in *Proceedings of the Ninth International Symposium on Information and Communication Technology*, 2018, pp. 38–45.

[12] Z. Zuo, "Sentiment analysis of steam review datasets using naive bayes and decision tree classifier," 2018.

[13] B. A. E. García, "Site-specific rules extraction in precision agriculture," Ph.D. dissertation, Universidad de Zaragoza, 2019.

[14] J. Igelbrink, "Empirical research and method approach," in *Perceived Brand Localness: An Empirical Study of the German Fashion Market*. Wiesbaden: Springer Fachmedien Wiesbaden, 2020, pp. 67–152, ISBN: 978-3-658-28767-2. DOI: 10.1007/978-3-658-28767-2_3. [Online]. Available: https://doi.org/10.1007/978-3-658-28767-2_3.

[15] P. Neupane. "Understanding text classification in nlp with movie review example example." (2020), [Online]. Available: https://www.analyticsvidhya.com/blog/2020/12/understanding-text-classification-in-nlp-with-movie-review-example-example/ (visited on 12/11/2020).

[16] M. F. Ahmed, Z. Mahmud, Z. Biash, A. Ryen, A. Hossain, and F. Ashraf, "Cyberbullying detection using deep neural network from social media comments in bangla language," Jun. 2021.

[17] N.-B.-V. Le, J.-H. Huh, *et al.*, "Applying sentiment product reviews and visualization for bi systems in vietnamese e-commerce website: Focusing on vietnamese context," *Electronics*, vol. 10, no. 20, p. 2481, 2021.

[18] T. Manzoor, M. Omi, A. Abir, M. Araf, T. Abir, and F. Ashraf, "Banglish (bengali in english letter) hate speech dataset," Jun. 2022. DOI: 10.17632/58sff5bdxd.1.