

**REVIEW ANALYSIS OF RIDE-SHARING APPLICATION USING BILSTM
BASED RNN MODEL- BANGLADESH PERSPECTIVE**

BY

**Taminul Islam
ID: 181-15-11116**

**Rishalatun Jannat Lima
ID: 181-15-11120**

**Arindom Kundu
ID: 181-15-10557**

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

Supervised By

Most Hasna Hena
Assistant Professor
Department of CSE
Daffodil International University

Co-Supervised By

Mr. Abdus Sattar
Assistant Professor
Department of CSE
Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

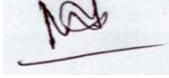
DHAKA, BANGLADESH

JANUARY 2022

APPROVAL

This Project/internship titled "**Review analysis of ride-sharing application using Bi-LSTM based RNN model-Bangladesh perspective**", submitted by Taminul Islam - ID No: 181-15-11116, Rishalatun Jannat Lima - ID No: 181-15-11120 and Arindom Kundu - ID No: 181-15-10557 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 January 02, 2022.

BOARD OF EXAMINERS

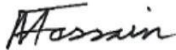


Chairman

Dr. Md. Ismail Jabiullah

Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

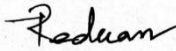


Internal Examiner

Dr. Md. Fokhray Hossain

Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



Internal Examiner

Md. Reduanul Haque

Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University



External Examiner

Dr. Mohammad Shorif Uddin

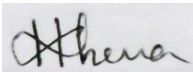
Professor

Department of Computer Science and Engineering
Jahangirnagar University

DECLARATION

We hereby declare that, this project has been done by us under the supervision of **Most Hasna Hena, Assistant Professor, Department of CSE** Daffodil International University. We also declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised by:

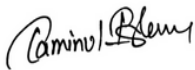


Most Hasna Hena

Assistant Professor

Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

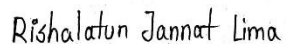
Submitted by:



Taminul Islam

ID: 181-15-11116

Department of Computer Science & Engineering
Daffodil International University



Rishalatun Jannat Lima

ID: 181-15-11120

Department of Computer Science & Engineering
Daffodil International University



Arindom Kundu

ID: 181-15-10557

Department of Computer Science & Engineering
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to almighty God for His divine blessing makes us possible to complete the final year project/internship successfully.

We really grateful and wish our profound our indebtedness to **Most Hasna Hena, Assistant Professor**, Department of CSE Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of "*Machine Learning*" to carry out this project. Her endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stage have made it possible to complete this project.

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

We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the course work.

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

Abstract

Technology and ride-sharing services have become more accessible and convenient as a result of the growth of the internet. Passengers increasingly focus on digital reviews to help them make purchasing decisions. Online reviews are incredibly inaccurate, as we've seen time and time again. False reviews were created to deceive customers for commercial purposes. A misleading review might have major repercussions for any organization. Providing good feedback to attract passengers and grow the market. It's possible that a bad review of an app would reduce interest in it. These false reviews endanger the reputation of a product. Because of this, it is critical to have a system in place for detecting fraudulent reviews. The goal of this research is to improve the performance of machine learning models that classify fake reviews. In this work Decision tree, Random Forest, Gradient Boosting, AdaBoost, and Bi-LSTM these five machine learning approaches have been implemented to get the best performance on our dataset. Data was collected from the current Bangladesh ride-sharing applications review section. After creating & running the model, Bidirectional Long Short-Term Memory (Bi-LSTM) achieved 85% best model accuracy and 89.0 F1-macro scores with training data rather than other machine learning algorithms.

LIST OF TABLES

TABLE	PAGE NO
Table 1: Research question criteria	3
Table 2: Summary of published research work	8
Table 3: Amount of collected individual data	12
Table 4: Amount of categorical data	12
Table 5: Descriptive category	12
Table 6: Example of data description	13
Table 7: Result comparison between five machine learning algorithms	23
Table 8: Accuracy report of DT	24
Table 9: Accuracy report of RF	25
Table 10: Accuracy report of GB	25
Table 11: Accuracy report of AdaBoost	26
Table 12: Classification report of Bi-LSTM	27
Table 13: Accuracy report of Bi-LSTM	27

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1.1: A Step-by-Step guide to detecting fake and real reviews	10
Figure 3.4.1: Statistics of platform wise data	12
Figure 3.5.1: Data Preprocessing steps	14
Figure 3.5.2: Example of punctuation removing from the text	15
Figure 3.5.3: Example of stopwords removing from the text	16
Figure 3.5.4: Example of stemming removal from the text	17
Figure 3.6.1: Proposed model workflow	17
Figure 3.7.2.1: Working sketch between Simple and Bi-SLTM models	21
Figure 4.1.1: Data Ratio	22
Figure 4.2.1: Confusion matrix of DT	24
Figure 4.2.2: Confusion matrix of RF	25
Figure 4.2.3: Confusion matrix of GB	26
Figure 4.2.4: Confusion matrix of AdaBoost	26
Figure 4.2.5: Accuracy and f1 macro average score of Bi-LSTM	27
Figure 4.2.6: Graphical representation of accuracy vs evaluation accuracy and loss vs evaluation loss	28
Figure 4.3.1: AUC comparison between classifiers	29

TABLE OF CONTENTS

CONTENTS	PAGE
Approval Page	i
Declaration	ii
Acknowledgement	iii
Abstract	iv
List of figures	v
List of tables	vi
CHAPTER	
CHAPTER 1: INTRODUCTION	1-5
1.1 Introduction	1
1.2 Motivation of the research	2
1.3 Rationale of the Study	3
1.4 Research Questions	3
1.5 Expected Output	4
1.6 Report Layout	5
CHAPTER 2: BACKGROUND	6-9
2.1 Introduction	6
2.2 Related Works	6
2.3 Research Summary	9

2.4 Scope of the Problem	9
2.5 Challenges	9
CHAPTER 3: RESEARCH METHODOLOGY	10-21
3.1 Introduction	10
3.2 Research Subject and Instrumentation	11
3.3 Data Collection Procedure	11
3.4 Statistical Analysis	11
3.5 Data Pre-Processing	13
3.6 Proposed Model Workflow	17
3.7 Proposed Methodology	17
3.8 Implementation Requirements	21
CHAPTER 4: EXPERIMENTAL RESULTS AND DISCUSSION	22-29
4.1 Experimental Setup	22
4.2 Experimental Results & Analysis	23
4.3 Discussion	28
CHAPTER 5: IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY	30-31
5.1 Impact on Society	30
5.2 Impact on Environment	30
5.3 Ethical Aspects	30

5.4 Sustainability Plan	31
CHAPTER 6: SUMMARY	32-33
6.1 Summary of the Study	32
6.2 Conclusions	32
6.3 Implication for Further Study	33
REFERENCES	33-35

CHAPTER 1

INTRODUCTION

1.1 Introduction

We have familiar with several ride-sharing applications like Uber, Patho, Obhai, Grab, etc. The quality of each company depends on their providing services. Users are able to submit their opinions by clicking on star & also doing comments in text which is known as review. We have seen there are a lot of reviews on the play store and other social platforms. Most of the time, the ranking and popularity of an app are determined by the reviews left by its users, making it one of the easiest methods for customers to express their opinions on the products or services they have purchased or used. As a result, the reviews have helped future passengers have a better understanding of the services they plan to use.

The use of machine learning techniques can make it easier to spot fake ride-sharing app reviews. Web mining techniques [1] employ a variety of machine learning algorithms to locate and collect interesting data from the internet. Content mining is one of the responsibilities involved in web mining. An interesting example of content mining is review mining [2], which involves using machine learning to train a classifier to assess review attributes and sentiments combined to discover the sentiment of text (positive or negative). Fake reviews are usually detected by looking at specific factors that aren't directly related to the content, such as the category in which the review appears. Natural language processing and text analysis are typically used to build review features. In order to generate false reviews that are related to the reviewer, extra details like the review time/date or the reviewer's writing style may be required. As a result, the key to finding false reviews is to build a system that extracts important characteristics from the reviewers.

We discovered a problem where reviewers manipulate reviews in order to disseminate false information. False information can be used to boost or degrade a company or application, depending on the intent. Fake reviews, review spams, and opinion spams are all terms used to describe this type of activity. In accordance with Jindal & Liu (2007, May), creating a false review is a form of opinion spamming. Instead of expressing their true ideas or

experiences, reviewers attempt to deceive readers or automated opinion mining and sentiment analysis algorithms, which is referred to as an unlawful behavior [3]. A product, service, or business body might benefit from favorable ratings from these people, but they can also publish negative reviews about other companies to harm their reputations. A fake review is one in which the reviewer knowingly provides untruthful or irrelevant information regarding the review item, whether it is all or part of the truth. Aside from being dubbed fake reviews, other terms for them include: bogus, scamming, misleading, and spam.

Spammers may easily build excitement for a product or service by generating good reviews in large numbers. This is the fundamental concern with review spams. False reviews now have a significant impact on how customers perceive a brand [4]. For organizations, positive reviews can result in large financial gains; on the other hand, poor evaluations can quickly destroy a company's good name. Automated systems or hired reviewers can create reviews. Fake positive evaluations for a company's products or services can be written by people or third-party groups hired by the companies or merchants. Since anybody can simply create and submit a review on the internet, the practice of spamming ride-sharing applications with fake reviews has grown in popularity.

The work introduced and also applied some machine learning approaches such as Decision Tree, Random Forest, Gradient Boosting, AdaBoost & Bidirectional Long Short-Term Memory (Bi-LSTM). And we got the best accuracy on Bi-LSTM model.

1.2 Motivation of research

Reviews have substituted other sources of information for customers looking to make purchasing decisions regarding services or products. So, when a passenger decides to schedule a ride, they can check out what other people have to say about it by reading reviews. They determine whether or not to schedule a ride based on the feedback they receive from the reviews. If the reaction from the reviews was favorable, they'll almost certainly book the trip. As a result, historical analyses gained a lot of trust among many online users. Due to the fact that reviews are seen as a way to authentically share customers'

experiences with both good and bad services, any attempt to alter such reviews through the use of false or misleading information is disapproved upon and characterized as fake. As a result, this study will focus the find the true reviews from the huge number of fake reviews. As part of our study, some famous machine learning models should be implemented to get the accurate result on our gathered dataset.

1.3 Rationale of the study

This research will make way for individuals and society as a whole. Though identifying fake reviews is a difficult task, it's important. Our team collects this data from a variety of ride sharing application sources before analyzing it. There are other approaches for review categorization there that we may learn about. Other sources also discuss various machine learning techniques and how to improve review analysis accuracy by using them.

1.4 Research Questions

A suitable research question is critical for uncovering related works in machine learning and approaches for ride sharing application. (B. Kitchenham et al, 2010) outlines the steps necessary to answer the appropriate research questions, such as population, intervention, outcomes, and context [5]. Table 1 shows the research topic criteria.

TABLE 1: RESEARCH QUESTION CRITERIA

Criteria	Details
Population	Bangladeshi Ride-Sharing application users
Intervention	Machine Learning & Deep Learning approaches for prediction
Outcome	Important attributes, Accuracy & Classification
Context	Ride-Sharing application's review section

According to the findings, the following research questions should be pursued:

Q1: How can we fetch reviews from the ride sharing applications?

Q2: What are the approaches to find the real review?

Q3: What are the approaches of the preprocessing data?

Q4: What is the market value of ride sharing application?

Q5: What are the key machine learning approaches for review analysis?

Q6: What is the performance of the present proposed models?

1.5 Expected Output

This proposed solution is to get the proper justification on several ride-sharing applications based on review. This solution will help to contribute make a genuine dashboard of a review section for any application. From this proposed idea we will get real and false review from the review dataset. This work will provide good accuracy on this particular sector. By implementing this work, we will able to know the real market value of ride sharing apps in Bangladesh. It will also categorize the originality of review. This work will help to find the most popular ride sharing application in Bangladesh. Hopefully this research work will produce a machine learning model which will detect real, fake or partially fake news from the inputted data with high accuracy.

1.6 Report Layout

The report has total 6 Chapters which will be followed given by instructions:

This study is summarized in Chapter 1 of the report. This chapter's primary focus is on introduction and discussion. This chapter does a good job of explaining why people get motivated. This sensible study is also essential since it shows what the research questions will be and what the expected results of the investigation will be, as explained in the previous section.

In Chapter 2, What has been done before this issue was researched? What exactly were the researchers trying to achieve with this study, and how did they go about it? What are the research's issues, and how will it solve them? As a conclusion, the study design has been provided in the final section.

This work's statistical methodologies are discussed in Chapter 3's theoretical explanation of research. These procedures have been illustrated in this chapter, and the final section describes how the model is evaluated through the use of machine learning models.

In Chapter 4, the findings of this study are detailed and discussed. Some research-related images have made it easier to grasp the work's criteria.

Chapter 5 contains the Ethical aspects of this research including impact of this research to our society and environment. This chapter also includes the sustainability goal of our project.

This section contains the conclusions from Chapter 6. It's important to the achievement of the entire segment. The idea of a substantial research study was presented. In addition, what are the restrictions on conducting this research, which will be useful to other scholars in the future.

CHAPTER 2

BACKGROUND

2.1 Introduction

Related work relates to the previous researcher's work on Fake Review detection. Though there are a few works in ride sharing application, therefore we considered the best research on fake review detections in this chapter. In this section, we'll talk about several research methodologies, as well as limitations and inaccuracies that were discovered over the course of the study. This piece was inspired by a number of different publications and journals. Everything we've done has been summarized into a single report, which outlines our objectives and the steps we took to achieve them. As a result of the problem scope, we will have a complicated issue with our research. The difficulties we have are almost exclusively the result of unexpected technical issues.

2.2 Related Work

More than 15 million evaluations from more than 3.5 million users from three major travel sites were included in that study by Jindal, N., & Liu, B [6]. There have been three main motions in their work. To begin with, they developed brand-new tools for detecting disparities across many sites. Second, they carried out the first comprehensive research of cross-site variations using real data and produced a technique with a 93% accuracy. The TrueView score was then presented. 20% of hotels appear to have a low trustworthiness score, based on the results.

In this study, Hyadri et al. [7] primarily analyze and categorize the models that mostly identify spam in reviews. This project will continue to improve in terms of accuracy and output. It's important to note that each sort of detection method has various strengths and drawbacks. The paper's limitation is that they can't achieve more precise results without going through the process of systematic analysis.

M. Crawford et al [8], Most of these studies have one thing in common: they turn reviews into word vectors, which can provide tens of thousands of unique characteristics. However, little research has been done on how to appropriately reduce the size of the feature subset

to a tolerable quantity. There were two ways to reduce the size of a feature subset in their paper: filter-based feature rankers and word-frequency-based feature selection. These approaches are used in the review spam domain. These results illustrate that there are no one-size fits-all method to feature selection, and the optimum technique to minimize the size of the feature subset depends on the classifier employed and the intended size of the feature subset. They have used Decision tree, Logistic Regression, Naïve Bayes, SVM, Multinomial Naïve Bayes to find the best accuracy on their proposed model & got the best accuracy 83% in Decision Tree.

Review spam detection is no different in that finding labeled datasets is always a difficulty for machine learning researchers. Using Amazon Mechanical Turk (AMT) to produce fake reviews for their dataset and combining them with "true" TripAdvisor ratings, Ott et al [9] developed a unique technique. To come up with their final dataset, they gathered a total of 400 false and 400 true reviews. These classifiers were tested on a variety of different data sets, including unigrams, bigrams, and trigrams. There was no statistical analysis done to see if the difference between SVM and bigrams in terms of performance was significant since the dataset was rather small. Some published Fake news detecting works have been summarized in the table below:

TABLE 2: SUMMARY OF PUBLISHED RESEARCH WORK

Ref	Year	Contribution	Dataset	Models	Accuracy
[10]	2021	Developed spam recognition framework which highlights display review illuminating lists as metadata structures to configure spam ID approach into a collection of issues.	HIN Resource Dataset	Naïve Bayes, Decision Tree	DT achieved the best 92.06% accuracy
[11]	2021	Developed graph model to detect the fake reviews & the fake reviewer.	Yelp Dataset	Convolutional Neural Network.	75%
[12]	2021	Developed model to detect the fake review from hotel & restaurant section.	Crowdsourcing, Amazon Mechanical Turk, TripAdvisor	Random Forest, AdaBoost, SVM, CNN, LSTM	RF achieved the best 90% accuracy
[13]	2021	Proposed a data sampling technique that improves the accuracy of the fake review class.	Yelp Dataset	Logistic Regression, SVM, Multilayer Perceptron (MLP), Bagging Predictor (BP), Random Forest, AdaBoost	SVM achieved the best 89.74% accuracy
[14]	2020	Proposed a method to identify false reviews using multiple feature fusion and collaborative rolling training.	YelpCHI	Random Forest, Logistic Regression, Latent Dirichlet Allocation, KNN, DT, Naïve Bayes, SVM	SVM achieved the best 84.45% accuracy
[15]	2018	Developed model according to behavior feature of reviewer to detect fake & true review.	Yelp Dataset	Random Forest, SVM	RF achieved the best 91.396% accuracy

2.3 Research Summary

Work in knowledge discovery such as fake review identification can be extremely hard to execute. For over a decade, researchers have studied fraud in review data from several aspects. The focus of our study is on approaches and classification models for identifying individual fake reviews from various review data sources. This study shows how to spot fake reviews using machine learning and deep learning techniques. The spread of false information may be prevented, though, if we don't allow it to be communicated in the first place. To place limits on a service, we need the assistance of the appropriate state. Here, we're looking for ways to reduce the number of reviews we utilize for machine learning and deep learning while also preventing false material from being sent out into the wider world. In order to decrease the spreading of falsehood, one of our main objectives is to restrict the circulation of fake ride-sharing app reviews. Finding high-quality data, such as samples of both fake and real reviews over a wide range of topics, is the most difficult part of our study. No matter how challenging it may be, our new algorithm has shown to be quite accurate in spotting fake news.

2.4 Scope of the Problem

Due to the fact that we must work with a fresh dataset, that's why it was a big struggling part to gather all the new reviews from the authentic sources. Fake review analysis also faces the challenge of detecting language patterns. For the fresh dataset, model accuracy must in concern. That's why we have faced many problems to take the accuracy in higher scale.

2.5 Challenges

There are some challenges we have faced during this work -

- Collecting quality data.
- Data preprocessing.
- Data purification and source authentication.
- Preparation and publication of a data collection for research study.
- Getting accuracy above 80%.
- Making a decision based on the results of tests.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

We go through our research approach in this chapter. An outline of our study process will be seen in Figure 3.1. This chapter examines the dataset's source and features. In addition, contextual aspects are addressed here. Some classification models and evaluation procedures are briefly explored in the later portion of this chapter. The steps in our study process are as follows:

Dataset Creation: The dataset has been created from the review section of Bangladeshi Ride-Sharing application. There are four domains has been analyzed. These are: Uber, Pathao, Obhai & Shohoz.

Preprocessing: To deal with noisy and inconsistent data, preprocessing techniques are employed. Many different preprocessing procedures have been used to improve the quality of the final product. Main techniques that were applied include tokenization, lemmatization, Punctuation removal, Stopwards removal and others.

Feature Extraction: A feature set for the classification model is built using attributes that were retrieved after preprocessing the data in the review database.

Train the Model: Several classification algorithms are then trained for experiments associated to our study.

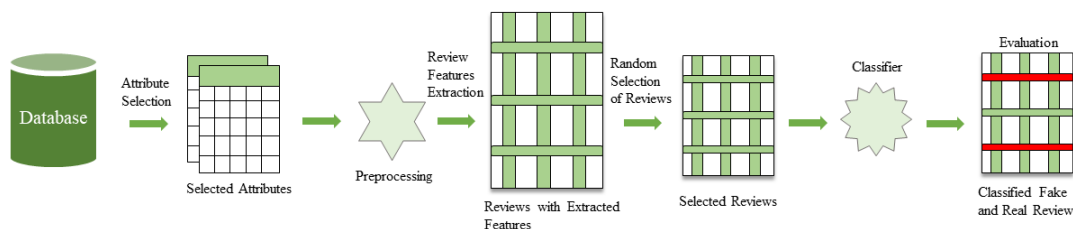


Figure 3.1.1: A Step-by-Step guide to detecting fake and real reviews

3.2 Research Subject & Instrumentation

Fake reviews may be detected using both supervised and unsupervised learning algorithms, as mentioned in related work. As far as we're aware, supervised learning techniques topped unsupervised ones. As a result, supervised learning has an advantage over unsupervised learning, this is why it is implemented. Fake reviews may be detected by looking at contextual features in reviews as well as the platforms themselves. For this experiment, we pulled on user reviews from four different ride-sharing applications in Bangladesh. The review dataset is mined for contextual features. There are many instrumentations has been used to complete this research work. To write the research paper we have used Latex & Microsoft word document. To develop the & train the model we have used Python, Google Colab, SPSS, Tensorflow and so on. To draw the figures & charts we have used Microsoft PowerPoint.

3.3 Data Collection Procedure

Data collection is always being a challenging part of a research study. It was a tough task to collect all the fresh data within a short time from all the authorized sites. Our team members have done a great job in this task. We have collected online user reviews from four (4) Bangladeshi ride sharing applications review section. These are- Uber, Pathao, Obhai and Shohoz. These data were collected from their individual websites & social platforms & google play store review section. All the data were collected manually by our researchers.

3.4 Statistical Analysis

This work has gathered total 2042 online reviews from individual 4 platforms. Data was collected from individual website & social platform of four (4) Bangladeshi ride sharing application. These are- Uber, Pathao, Obhai, Shohoz. These data were collected manually by our researchers. Table 3 illustrates the amount of collected data from individual apps section.

TABLE 3: AMOUNT OF COLLECTED INDIVIDUAL DATA

Applications	Data Size
Uber	1000
Pathao	583
Obhai	411
Shohoz	98
Total	2092

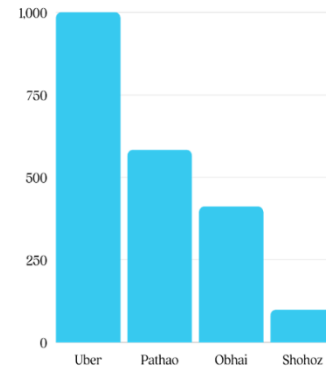


Figure 3.4.1: Statistics of platform wise data

In the bellow table 4 states the amount of total True reviews, Partially False reviews and False reviews. The exact number of this statistics is – True (1263), Partially False (385), False (444). All the data was collected manually by our researchers.

TABLE 4: AMOUNT OF CATEGORICAL DATA

Category	Amount of data
True	1263
False	444
Partially False	385

Here is the confusion for the description of collected category. These data were classified into three category – True, False & Partially False. Table 5 state the description of categories that classified.

TABLE 5: DESCRIPTIVE CATEGORY

Category	Description
True	The given text includes contents that are clearly apparent or capable of being logically proved.
Partially False	Main claim in given text might be True but also contain false or misleading information, not surely true and not certainly false.
False	The main content of given text is fake.

Our researchers have categorized data followed by above description table. In the below table 6 illustrates the examples of reviews was collected according to its rating.

TABLE 6: EXAMPLE OF DATA DESCRIPTION

<i>Number</i>	<i>Review Text</i>	<i>Platform (Apps)</i>	<i>Rating</i>
01	The app is good at showing the data user the direction of the driver. However, yesterday I was disappointed when ordering a ride. I was ordering the Chap Chap option and at the time, I was forced to give out my payment details such as PayPal, Credit or Debit Card details. The app didn't provide for M-PESA or cash option. I was forced to give out my credit or debit details before ordering a ride which is kind of disappointing. This compromises my data security	UBER	True
02	I hate this app now , gps map is too poor , it won't show us all the map direction like google map , need some update but who cares , all need money without struggle , like; easy money!!! very very very dis-appointed app ever.	Pathao	Partially False
03	Very organized and on time delivery service	Shohoz	False

3.5 Data Preprocessing

Pre-processing data is the initial step in doing research. Processing the data is the first stage in data mining. There are several platforms which we obtain our reviews for this reason. We start by preparing the data to fix this. These data sets are split down into a wide range of numerical values. This data is processed one at a time. Machine learning and deep learning models cannot be applied to data which is only in the form of text. It is able to handle numerical data only. On the text and rating dimensions, we used the following data preparation approaches in this research work. Figure 3.5.1 shows the steps of data preprocessing.

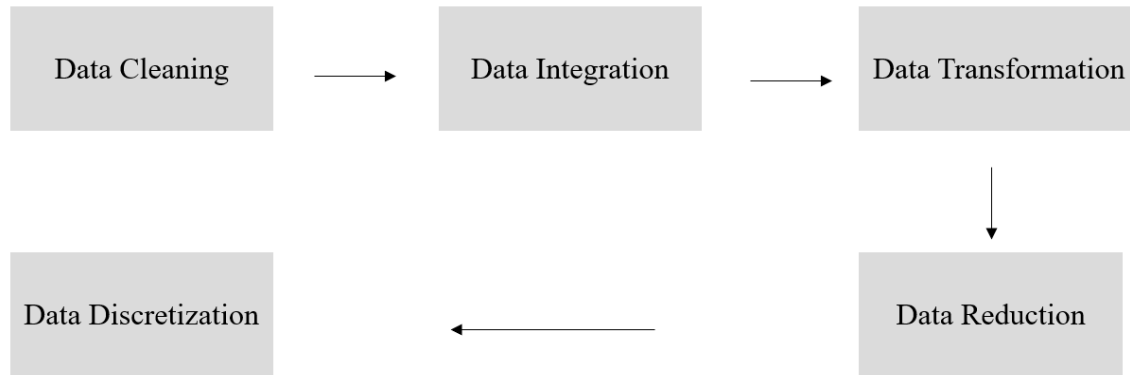


Figure 3.5.1: Data Preprocessing steps

Reduce of Dimension

Unnecessary qualities are responsible for increasing the length of time it takes to complete an operation. The public id and title characteristics are deleted from the dataset prior to data being entered into the model. The text column contains the input data, while the target column has the rating.

Punctuation Removing

There are many punctuation marks, links, numbers, and other special characters used in news articles, none of which have any influence on whether the review is true or incorrect in the vast majority of cases. In addition, punctuation appears often and has a substantial influence on the measurements for punctuation, but it has no effect on the classification of the text, which is a mixed bag [16]. In the below figure 3.5.2 it is shown the example of punctuation removing.

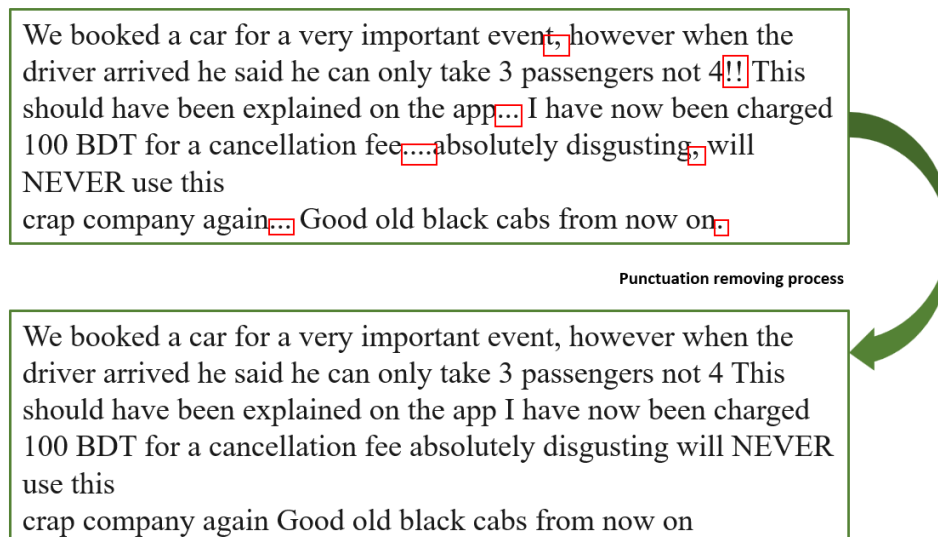


Figure 3.5.2: Example of punctuation removing from the text

Noise removal

All data must be clear and free of noise. Otherwise, there might be some unanticipated issues that need to be dealt with. Words that are unnecessary to tokenize and vectorize must be removed from the input sequence. Tokenization is improved by converting uppercase characters to lowercase ones.

Tokenization

Tokenization is the process of breaking down a review text into individual words (tokens). For example, review material is tokenized and converted into tokens. To calculate RCS and capital variety, tokenization is a critical step since it allows each word in the review to be separated [17]. For word tokenization, we made use of the NLTK library. For example, let's consider "greatest". Here Character tokens: g-r-e-a-t-e-s-t and Sub word tokens: great-est.

Removing Stopwords

Due to the fact that stopwords are widespread in natural language and do not convey any unique meaning, they are less significant in a phrase [18]. Stopwords may increase the amount of time it takes to process data in data analysis. Because of this, it is important to

remove stopwords from the phrase. For the purpose of deleting stopwords from the phrase, we employed the NLTK library. We must remove all of the text and strings from the data in order to make it trainable. Because of this, we convert all of our text into numbers so that it may be utilized as a teaching aid. Figure 3.5.3 is an example of stopwords being removed from a text.

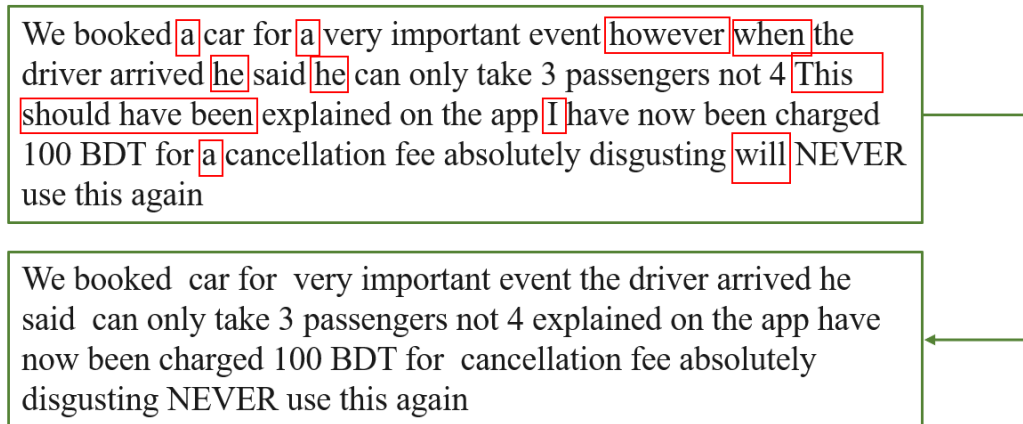


Figure 3.5.3: Example of stopwards removing from the text

Capitalization

In a computational model, it is ideal to use the same register level regardless of whether upper- or lowercase characters are used [19]. It doesn't matter what kind of register level you use when it comes to digits. Lowercase letters were used in this study.

Stemming Removal

Eliminating suffixes and prefixes from a word is known as stemming. Using the stemming method, we can get a word back to its root structure. Figure 3.5.4 shows the way of stemming.

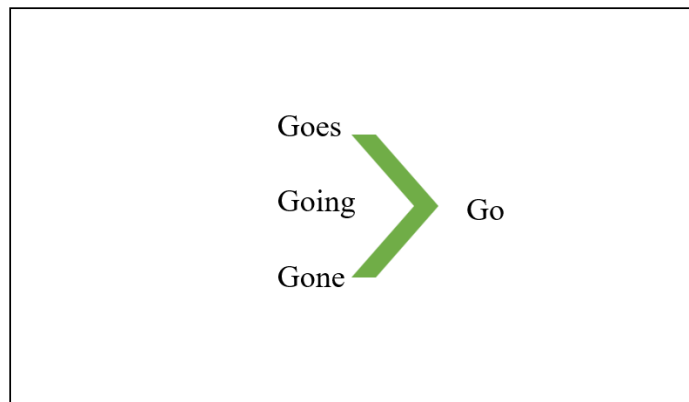


Figure 3.5.4: Example of stemming removal from the text

3.6 Proposed Model Workflow

The figure 3.6.1 shows the workflow of our proposed research methodology.

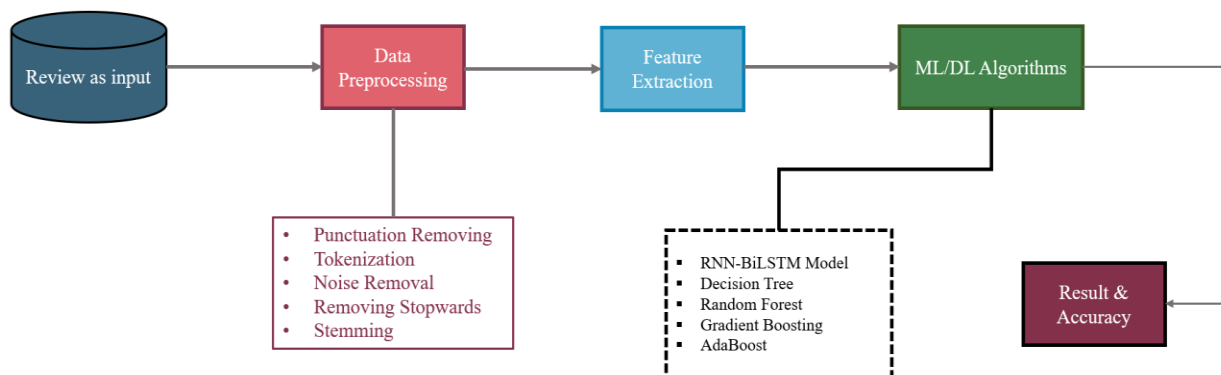


Figure 3.6.1: Proposed model workflow

3.7 Proposed Methodology

We have collected data from four different ride sharing applications. Our main goal was to find the best machine learning model that can detect fake, true or partially fake review smoothly. We applied five machine learning algorithms to find the best accuracy on this dataset. These are Decision tree, Random Forest, Gradient Boosting, AdaBoost and Bi-LSTM. We got the best accuracy 85% from the Bi-LSTM model.

3.7.1 Feature Selection and Extraction

The Keras library was used to create our BiLSTM model, which was then tested. It is possible to make a model using a glove embedding of 100d. The sequential model is used as the foundation for this experiment's analysis. A number of different techniques are used, including embedding, dropout layers, and a layer with 256 neurons that is totally connected to the rest of the network. We have a dataset with many classes. That's why, soft-max activation is used to apply the output layer to the final layer. It is consistent with other algorithms, such as Random Forest, Gradient Boosting, and AdaBoost. N-gram features, such as uni gram, bi gram and tri gram were employed in all machine learning techniques. The model is trained using 20 epochs and 128 batch sizes of training data to achieve optimal performance. The accuracy of this model was determined to be 85%, while the F1-macro score was found 89%

3.7.2 Machine Learning models

There are four ride-sharing apps that we've collected data from. Our major aim was to create the best machine learning model that can recognize fake, real, or partially fake reviews without any difficulty. On this dataset, we used five different machine learning methods to determine the one with the greatest accuracy. These include the Decision Tree, Random Forest, Gradient Boosting, AdaBoost, and Bi-LSTM models. The models have emphasized in this section.

Decision tree

Classification and regression models may be built at regular intervals using a decision tree. In terms of categorization and predictions, it's the most effective and widely used technology available today. There are many different types of decision trees; the most common is the flowchart-like tree structure, in which each internal node symbolizes a test on a certain characteristic, and each branch reflects a conclusion of the test [20]. The last word is a node in a tree having nodes for decisions and nodes for leaves. Alternate nodes either a few or a large number of branches. Decisions or classifications are represented by a leaf node. The root node of a passing tree, which corresponds to the highest successful

predictor, is the simplest decision node in the tree. Decision trees can deal with any type of data, whether it's numerical or categorical [21].

Random Forest

The choice tree is the basic component of random forest classifications. The choice tree is littered with living trees including a variety of elements at each node. The entropy of a specified collection of characteristics is supported by the nodes. In the random forest, a collection of decision trees is linked to a collection of bootstrap samples derived from the source data set. Trees are the building blocks of a forest, and the more trees there are, the more stable it will be [22]. By creating call trees out of data samples, the random forest algorithm receives the forecast for every one of these trees, then votes on which is best. Breiman's articles include extensive information on random forest classifiers. At times while using the quality random forest strategy, the bootstrapping technique is used to help create an appropriate random forest with the requisite number of decision trees thus boosting classification accuracy using the notion of overlap dilution as described. To train and optimize, random forests are often like growing trees, making it easier. As a result, random forest is a good method for numerous packets [23].

Gradient Boosting

Many machine learning methods are combined into Gradient Boosting Classifiers (GBCs) in order to create a strong predictive model. Gradient boosting is a technique in which decision trees are occasionally employed. Gradient boosting models have lately been used to win multiple Kaggle informatics challenges due to their success in categorizing large data sets. The main goal is to lower the amount of error in the next model by aligning the desired outcomes. How are the goals determined? There are many different methods to build gradient boosting classifier in the Python machine learning, Scikit-Learn. This article examines the theory underlying gradient boosting models and looks at two distinct techniques to construct gradient boosting models in Scikit-Learn [24-25].

AdaBoost

Multi-learner approaches to problem solving are referred to as "ensemble learning" [26]. When it comes to learning, ensemble techniques are a popular choice because of their

superior capacity to generalize. Due to its strong theoretical foundation, precise prediction, tremendous simplicity (Schapire noted it required only "only 10 lines" of code), and extensive and successful use cases, the AdaBoost algorithm [27] created by Yoav Freund and Robert Schapire was among the most significant ensemble techniques. Because AdaBoost is the most widely used ensemble algorithms, its huge influence is not surprising. The theoretical and practical aspects of these two topics are briefly discussed in this article. Because of AdaBoost, there has been an abundance of theoretical research on ensemble approaches, which is readily available in the machine learning and statistical literatures.

Bi-LSTM

In comparison to Long Short-Term Memory, Bi-Directional Long Short-Term Memory (Bi-LSTM) excels at categorizing sequences (LSTM). It's the process of creating a neural network that can process information in both forward and reverse orientations. The Bi-LSTM is composed of two LSTMS, one for forward and one for reverse input. It is feasible to communicate data in both directions using disguised states. Each time step, the outputs of two LSTMs are merged to generate one [28]. The Bi-LSTM technique contributes to the reduction of the restrictions associated with traditional RNNs. The context is more easily comprehended as a result of Bi-LSTM's high degree of accuracy. However, with bi-directional input, we can ensure that both the future and the past are preserve [29].

Based on the previous validated reviews, an NLI model is built using Bidirectional LSTM neural networks in this phase to assess the validity of each individual claim. Here is a contrast of the simple model and the Bi-LSTM model. We started by training a simple machine learning model with simply the assertions (hypotheses) as input. The NLI-based model is then trained to infer the accuracy of the claim based on previous information (premises). We test the suggested NLI-based strategy to identifying false reviews by comparing the outcomes of these two models. Figure 3.7.2.1 shows the process of this method.

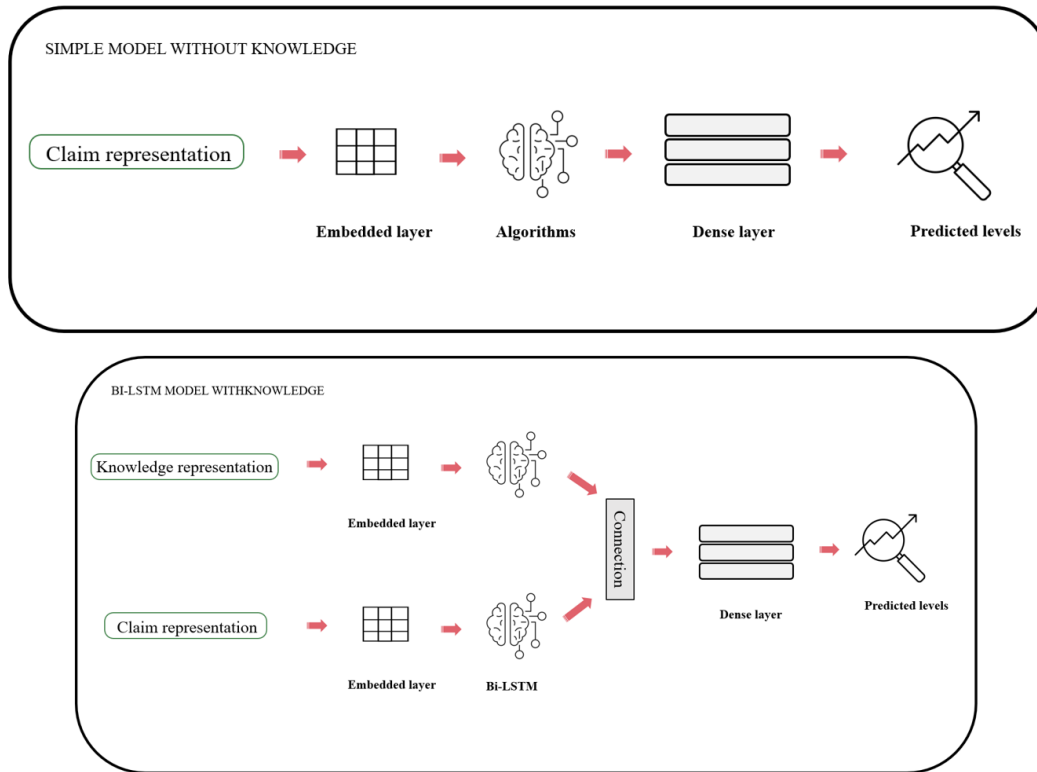


Figure 3.7.2.1: Working sketch between Simple and Bi-SLTM models

3.8 Implementation Requirements

These are some requirements needed to complete this project.

Hardware/Software Requirements

- Operating System (Windows 7 or above)
- Hard Disk (minimum 250 GB)
- Ram (minimum 4 GB)
- Web Browser (preferably chrome)

Developing Tools

- Python with Tensor Flow
- Google Colab
- Notepad++
- Orange

Writing & Designing tools

- Microsoft Word & Latex
- Microsoft Power point

CHAPTER 4

EXPERIMENTAL RESULT & DISCUSSION

4.1 Experimental Setup

We have collected data from four different ride sharing applications. Our main goal was to find the best machine learning model that can detect fake, true or partially fake review smoothly. We applied five machine learning algorithms to find the best accuracy on this dataset. These are Decision tree, Random Forest, Gradient Boosting, AdaBoost and Bi-LSTM. Data collection is always being a challenging part of a research study. It was a tough task to collect all the fresh data within a short time from all the authorized sites. Our team members have done a great job in this task. This work has gathered total 2042 online reviews from individual 4 platforms. We collected 100 reviews from Uber, 583 reviews from Pathao, 411 reviews from Shohoz and 98 reviews from Obhai. Our data was collected with three categories. Where True data contains 1263, False contains 444 and Partially False contains 385 data.

To begin the modeling procedure, we divided our dataset in two parts.

- Dataset for Training
- Dataset for Testing

80% of the data have utilized for training, and 20% for testing. And this is also what we expect to observe in our model. Figure 4.1 illustrates the data ration for the training & testing.

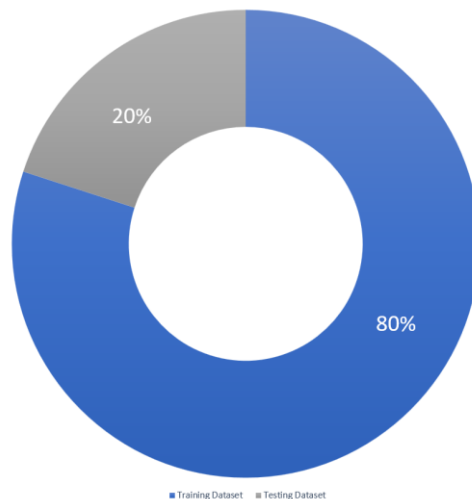


Figure 4.1.1: Data Ratio

DT, RF, GB, and AdaBoost are used to classify the extracted features. Python is used to implement all of the classifiers. Machine learning libraries in Python are extensive and feature-rich. There are two popular machine learning software programs called 'SKLEARN' and 'PANDAS'. Precision, recall, F1-measure, and accuracy are used to evaluate the classification model's performance. Besides, The Keras library was used to create our Bi-LSTM model, which was then tested. It is possible to make a model using a glove embedding of 100d. The sequential model is used as the foundation for this experiment's analysis. A number of different techniques are used, including embedding, dropout layers, and a layer with 256 neurons that is totally connected to the rest of the network. The model is trained using 20 epochs and 128 batch sizes of training data to achieve optimal performance and also precision, recall, F1-measure, and accuracy are used to evaluate this classification model's performance.

4.2 Experimental Results & Analysis

After the experimental setup, getting the result is the only remaining work of this research. We have applied total five machine learning algorithms on our fresh dataset. There is a tight comparison between one to another algorithms. In between five algorithms Bidirectional LSTM (Bi-LSTM) achieved the best accuracy, precision, recall and F1 score. Bi-LSTM achieved the best 85% accuracy where Random Forest & AdaBoost works well and achieved 83% accuracy both. We found 80% accuracy in Gradient Boosting & 79% accuracy on Decision Tree algorithm. Below Table 7 illustrates the result comparison between five machine learning algorithms

TABLE 7: RESULT COMPARISON BETWEEN FIVE MACHINE LEARNING ALGORITHMS

ALGORITHMS	AUC	PRECISION	RECALL	F1 SCORE
Decision Tree	0.798	0.691	0.723	0.691
Random Forest	0.834	0.690	0.723	0.693
Gradient Boosting	0.80	0.703	0.718	0.629
AdaBoost	0.830	0.706	0.731	0.702
Bi-LSTM	0.85	0.86	0.91	0.89

Let's have a look the individual result of five different algorithms, then we can evaluate the result from this end.

Decision Tree

We used the decision tree approach to train our model and acquired an accuracy of 79 % and a f1 -macro average score of 69%. The following table 7 summarizes the accuracy report, and also the confusion matrix seen in Figure 4.2.1.

TABLE 8: ACCURACY REPORT OF DT

Algorithm	Accuracy	Precision	Recall	F1 Macro Score
Decision Tree	79%	0.69	0.72	69%

		Predicted			Σ
		FALSE	Partially FALSE	TRUE	
Actual	FALSE	28	3	63	94
	Partially FALSE	6	17	42	65
	TRUE	20	12	337	369
Σ		54	32	442	528

Figure 4.2.1: Confusion matrix of DT

Random Forest

We also used the Random Forest approach to train our model and acquired an accuracy of 83% which is better than DT and a f1 -macro average score of 69% that is similar with DT. The following table 8 summarizes the accuracy report, and also the confusion matrix seen in Figure 4.2.2. If we compare with DT, the we find the difference on average accuracy it increases 4%.

TABLE 9: ACCURACY REPORT OF RF

Algorithm	Accuracy	Precision	Recall	F1 Macro Score
Random Forest	83%	0.69	0.72	69%

		Predicted			Σ
		FALSE	Partially FALSE	TRUE	
Actual	FALSE	22	6	66	94
	Partially FALSE	1	14	50	65
	TRUE	13	8	348	369
Σ		36	28	464	528

Figure 4.2.2: Confusion matrix of RF

Gradient Boosting

In the Gradient Boosting approach after train our model achieved accuracy of 80% which performs relatively better than DT and worse than RF and it scores f1 -macro average sof 62% that performs not good. The following table 9 summarizes the accuracy report, and also the confusion matrix seen in Figure 4.2.3.

TABLE 10: ACCURACY REPORT OF GB

Algorithm	Accuracy	Precision	Recall	F1 Macro Score
Gradient Boosting	80%	0.70	0.71	62%

		Predicted			Σ
		FALSE	Partially FALSE	TRUE	
Actual	FALSE	9	2	83	94
	Partially FALSE	1	5	59	65
	TRUE	3	1	365	369
Σ		13	8	507	528

Figure 4.2.3: Confusion matrix of GB

AdaBoost

In the AdaBoost approach after train our model achieved accuracy of 83% which performs relatively better than DT and GB and it scores f1 -macro average of 70% that performs good. The following table 10 summarizes the accuracy report, and also the confusion matrix seen in Figure 4.2.4.

TABLE 11: ACCURACY REPORT OF AdaBoost

Algorithm	Accuracy	Precision	Recall	F1 Macro Score
AdaBoost	83%	0.70	0.73	70%

		Predicted			Σ
		FALSE	Partially FALSE	TRUE	
Actual	FALSE	33	1	60	94
	Partially FALSE	3	17	45	65
	TRUE	22	11	336	369
Σ		58	29	441	528

Figure 4.2.4: Confusion matrix of AdaBoost

Bi-LSTM

The proposed model's output identifies the review item presented. The review is either true, false, or partially false. The RNN model can't handle text, that's why true is considered to

be 2, false is 0, and partially false is 1. We applied 20 epochs to train our model and got the best accuracy of 85% and the best f1 macro average score of 89%. We found the best performance in Bi-LSTM method relatively all other algorithms. Table 11 shows the model's accuracy and f1 macro average score, as well as the classification report. Table 12 summarizes the accuracy report and the following figure 4.2.5 shows the accuracy and f1 macro average score of this model.

TABLE 12: CLASSIFICATION REPORT OF BI-LSTM

Class	Precision	Recall	F1-macro score
True	0.86	0.91	0.89
False	0.91	0.89	0.90
Partially False	0.49	0.37	0.42

TABLE 13: ACCURACY REPORT OF Bi-LSTM

Algorithm	Accuracy	Precision	Recall	F1 Macro Score
Bi-LSTM	85%	0.86	0.91	89%

Evaluating Model ...

	precision	recall	f1-score	support
0	0.91	0.89	0.90	173
1	0.49	0.37	0.42	65
2	0.86	0.91	0.89	357
accuracy			0.85	595
macro avg	0.75	0.72	0.74	595
weighted avg	0.84	0.85	0.84	595

Figure 4.2.5: Accuracy and f1 macro average score of Bi-LSTM

The following figure 4.6.2 demonstrates the connection between our proposed model's accuracy and evaluation accuracy and loss and evaluation loss. Both of which are achieved using Bi-LSTM model. These graphics demonstrate that our suggested model is accumulating knowledge from its predecessors.

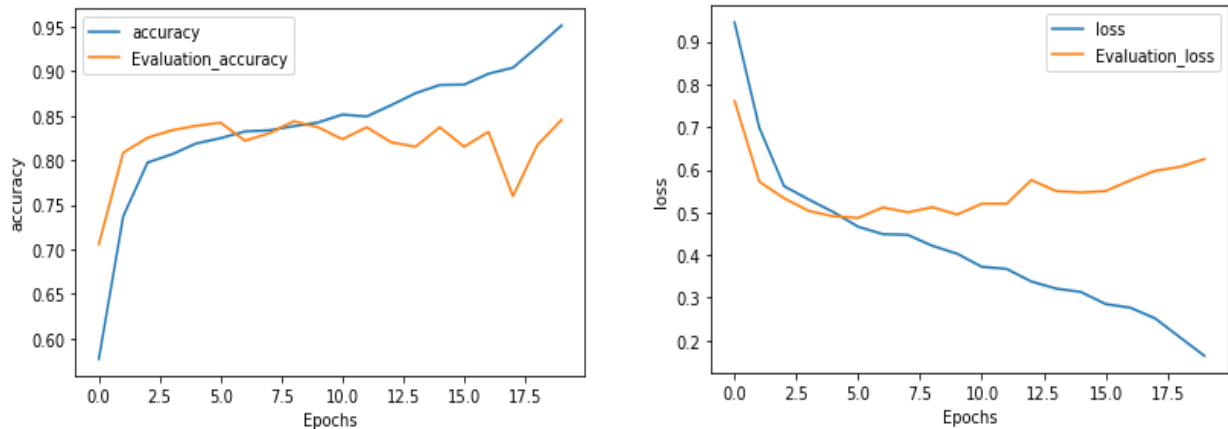


Figure 4.2.6: Graphical representation of accuracy vs evaluation accuracy and loss vs evaluation loss

4.3 Discussion

To compare classifiers, all methods behave similarly. However, each classifier is notably different from the others, with Bi-LSTM, AdaBoost, and Random Forest being the best and Decision tree and Gradient Boosting being the worst. Among all other algorithms, Bi-LSTM produces the best figure. Figure 4.3.1 illustrates the sharp difference more clearly. It is critical to notice that this figure depicts the confidence interval for the mean AUC score across all subset sizes, indicating that RF and AdaBoost are not mutually exclusive.

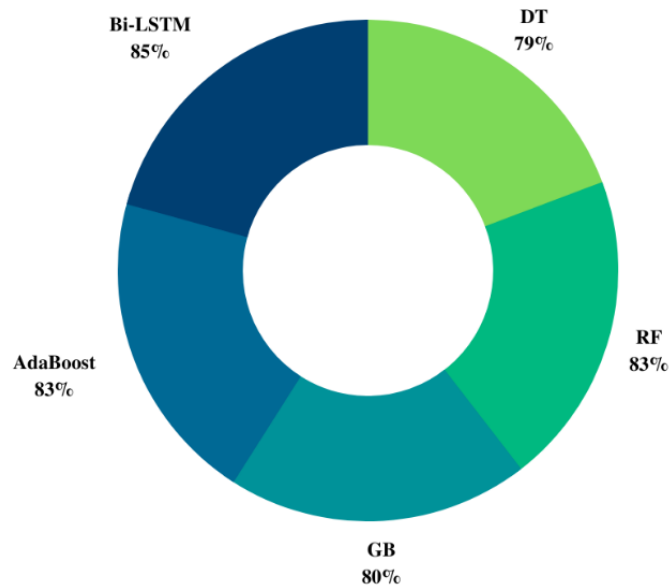


Figure 4.3.1: AUC comparison between classifiers

We have collected data manually by our researches. It was a challenging task to collect all quality fresh reviews from different platforms. The result satisfies us in terms of the new dataset. If we work hard to collect the best possible quality data, then our model should achieve more accuracy in every sector.

CHAPTER 5

IMPACT ON SOCIETY, ENVIRONMENT AND SUSTAINABILITY

5.1 Impact on Society

Ride-sharing applications are heavily impacted by online consumer and passenger reviews. That is why it has such a profound effect on our society. Passengers contemplating the use of a service may benefit from reading internet evaluations prior to making their final selection. Fake internet reviews may impact users' purchase decisions. Fake reviews are used to promote or disparage services on ride-sharing applications. It has the potential to destroy the reputation of a reputable service, resulting in financial loss for a well-known firm. This research has given Bangladesh a new perspective. Our findings will have a profound effect on society. This will assist us in obtaining accurate reviews for our ride sharing application.

5.2 Impact on Environment

This research is totally based on computer technology. Our study has no adverse effect on the environment. We collected online user reviews from the review sections of four (4) Bangladeshi ride sharing programs. Uber, Pathao, Obhai, and Shohoz are among them. These statistics were gathered from their respective websites, social media platforms, and the Google Play store review area. Our researchers gathered all data manually. Then, we used machine learning techniques to determine if the reviews were fraudulent or genuine. We obtained some accuracy as a result of this investigation. Therefore, there are no impact on the environment through this work.

5.3 Ethical Aspects

This research work must respect every user's security & privacy. All the data have collected manually by our researchers. There is no previously published data have used in this research. The results we have provided are correct. There is no false information regarding this research. No animals were harmed during our research. All the members of our group helped equally and we walked with our supervisor's guidance.

5.4 Sustainability Plan

We know fake review detection have a wide range. We have just picked a new area for exploring our model. There are more plans to work in future with this domain. We will cover the other applications through our model. We will collect more quality data to develop our accuracy in this research. Besides, we will implement more machine learning & NLP algorithms to detect fake reviews from different applications of Bangladesh. We will consider other regions for analyzing their perspective. So, we have a long sustainability plan regarding this research work.

CHAPTER 6

SUMMARY

6.1 Summary of the study

Ride-sharing applications are largely influenced by customer and passenger evaluations found online. Passengers who are considering using a service might benefit from reading online reviews before making their final choice. Users' purchasing decisions might be influenced by fake online reviews. On ride-sharing apps, fake reviews are used to promote or degrade services. It can tarnish a good service's reputation, resulting in financial loss for a well-known business. In rare circumstances, a corporation might benefit from falsely positive evaluations. Customers and companies alike are harmed by fake reviews, which are detrimental to both parties. Since 2007, researchers have been focusing on the identification of bogus reviews.

Fake reviews, individual spammers, and spammer groups are the focus of the majority of the current investigation work being done in these fields. The purpose of this research was to determine whether or not some machine learning algorithms were being used to detect bogus reviews. Four Bangladeshi ride-sharing apps have been analyzed by our researchers. The classification methods used in this study performed well. Our research has effectively revealed a previously unknown aspect of the fake review detecting field.

6.2 Conclusions

Nowadays, ride sharing applications are an important factor of Bangladesh's infrastructure. Fake reviews significantly impair consumers' ability to obtain authentic information. Our research focuses on identifying fraudulent reviews through the use of well-known machine learning techniques. Our researchers collected data from Uber, Pathao, Sohoz, and Obhai, four of Bangladesh's most popular ride-sharing applications. Following feature extraction and model construction, Bi-LSTM attained the highest accuracy of 85 %. Additionally, Random Forest & AdaBoost perform well, with an accuracy of 83 %. The model was tested using newly collected data. Collecting high-quality data can help enhance accuracy. In the future, we hope to test our proposed technique on a bigger, more varied dataset in order to get over our existing limitations.

6.3 Implication for Further Study

We know fake review detection have a wide range. We have just picked a new area for exploring our model. There are some plans to work in future with this domain.

- The dataset we will make with more quality & information.
- We will explore other machine learning approaches with the new dataset.
- There will be variety on application sector for considering the fake review analysis.
- We will consider other regions for analyzing their perspective.

REFERENCES

- [1] Kosala, R., & Blockeel, H. (2000). Web mining research: A survey. *ACM Sigkdd Explorations Newsletter*, 2(1), 1-15.
- [2] Johnson, F., & Gupta, S. K. (2012). Web content mining techniques: a survey. *International journal of computer applications*, 47(11).
- [3] Jindal, N., & Liu, B. (2007, May). Review spam detection. *Proceedings of the 16th international conference on World Wide Web*, 1189-1190.
- [4] E. D. Wahyuni , A. Djunaidy,” Fake Review Detection From a Product Review Using Modified Method of Iterative Computation Framework”, InProceeding MATEC Web of Conferences, 2016
- [5] B. Kitchenham et al., "Systematic literature reviews in software engineering—a tertiary study," *Information and software technology*, vol. 52, no. 8, pp. 792-805, 2010.
- [6] A. j. Minnich, N. Chavoshi, A. Mueen, S. Luan, and M. Faloutsos, “True view: Harnessing the power of multiple review sites”,In *ACM WWW*, 2015.
- [7] A. Heydari, M. A. Tavakoli, N. Salim, and Z. Heydari,” Detection of review spam: a survey”, *Expert Systems with Applications*, vol. 42, no.7, pp. 3634–3642, 2015.
- [8] M. Crawford, T. M. Khoshgoftaar, and J. D. Prusa,” Reducing Feature SetExplosion to Faciliate Real-World Review Sapg Detection”, In *Proceeding of 29th International Florida Artificial Intelligence Research Society Conference*, 2016.
- [9] Ott, M., Choi, Y., Cardie, C., & Hancock, J. T. (2011). Finding deceptive opinion spam by any stretch of the imagination. *arXiv preprint arXiv:1107.4557*.

- [10] Vachane, D. (2021). Online Products Fake Reviews Detection System Using Machine Learning. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(1S), 29-39.
- [11] Manaskasemsak, B., Tantisuwankul, J., & Rungsawang, A. (2021). Fake review and reviewer detection through behavioral graph partitioning integrating deep neural network. *Neural Computing and Applications*, 1-14.
- [12] Yao, J., Zheng, Y., & Jiang, H. (2021). An Ensemble Model for Fake Online Review Detection Based on Data Resampling, Feature Pruning, and Parameter Optimization. *IEEE Access*, 9, 16914-16927.
- [13] Budhi, G. S., Chiong, R., & Wang, Z. (2021). Resampling imbalanced data to detect fake reviews using machine learning classifiers and textual-based features. *Multimedia Tools and Applications*, 80(9), 13079-13097.
- [14] Wang, J., Kan, H., Meng, F., Mu, Q., Shi, G., & Xiao, X. (2020). Fake review detection based on multiple feature fusion and rolling collaborative training. *IEEE Access*, 8, 182625-182639.
- [15] Kumar, J. (2020). Fake Review Detection Using Behavioral and Contextual Features. *arXiv preprint arXiv:2003.00807*.
- [16] Pradha, S., Halgamuge, M. N., & Vinh, N. T. Q. (2019, October). Effective text data preprocessing technique for sentiment analysis in social media data. In *2019 11th international conference on knowledge and systems engineering (KSE)* (pp. 1-8). IEEE.
- [17] Webster, J. J., & Kit, C. (1992). Tokenization as the initial phase in NLP. In *COLING 1992 Volume 4: The 14th International Conference on Computational Linguistics*
- [18] Gerlach, M., Shi, H., & Amaral, L. A. N. (2019). A universal information theoretic approach to the identification of stopwords. *Nature Machine Intelligence*, 1(12), 606-612.
- [19] Lev, B., & Sougiannis, T. (1996). The capitalization, amortization, and value-relevance of R&D. *Journal of accounting and economics*, 21(1), 107-138.
- [20] Myles, A. J., Feudale, R. N., Liu, Y., Woody, N. A., & Brown, S. D. (2004). An introduction to decision tree modeling. *Journal of Chemometrics: A Journal of the Chemometrics Society*, 18(6), 275-285.
- [21] Fletcher, S., & Islam, M. Z. (2019). Decision tree classification with differential privacy: A survey. *ACM Computing Surveys (CSUR)*, 52(4), 1-33.

- [22] Pal, M. (2005). Random forest classifier for remote sensing classification. *International journal of remote sensing*, 26(1), 217-222.
- [23] Ali, J., Khan, R., Ahmad, N., & Maqsood, I. (2012). Random forests and decision trees. *International Journal of Computer Science Issues (IJCSI)*, 9(5), 272.
- [24] Bahad, P., & Saxena, P. (2020). Study of adaboost and gradient boosting algorithms for predictive analytics. In *International Conference on Intelligent Computing and Smart Communication 2019* (pp. 235-244). Springer, Singapore.
- [25] Chakrabarty, N., Kundu, T., Dandapat, S., Sarkar, A., & Kole, D. K. (2019). Flight arrival delay prediction using gradient boosting classifier. In *Emerging technologies in data mining and information security* (pp. 651-659). Springer, Singapore.
- [26] Freund, Y., Schapire, R., & Abe, N. (1999). A short introduction to boosting. *Journal-Japanese Society For Artificial Intelligence*, 14(771-780), 1612.
- [27] Murthy, S. K., Kasif, S., & Salzberg, S. (1994). A system for induction of oblique decision trees. *Journal of artificial intelligence research*, 2, 1-32.
- [28] Liu, Y., Wang, L., Shi, T., & Li, J. (2022). Detection of spam reviews through a hierarchical attention architecture with N-gram CNN and Bi-LSTM. *Information Systems*, 103, 101865.
- [29] Braşoveanu, A. M., & Andonie, R. (2019, June). Semantic fake news detection: a machine learning perspective. In *International Work-Conference on Artificial Neural Networks* (pp. 656-667). Springer, Cham.

Turnitin Originality Report

Processed on: 04-Dec-2021 16:04 +06

ID: 1720338018

Word Count: 7355

Submitted: 1

tamim By Most. Hena

< 1% match (Internet from 06-Aug-2020)

Similarity by Source	
Similarity Index	
3%	
Internet Sources:	2%
Publications:	2%
Student Papers:	0%

http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/4095/P15433%20%2829_%29_.pdf?isAllowed=y&sequence=1

< 1% match (Internet from 25-Mar-2021)

http://dspace.daffodilvarsity.edu.bd:8080/bitstream/handle/123456789/5250/162-15-7955%20%287_%29.pdf?isAllowed=y&sequence=1

< 1% match (publications)

["Proceedings of International Conference on Intelligent Computing, Information and Control Systems", Springer Science and Business Media LLC, 2021](#)

< 1% match (publications)

[Danny Joel Devarapalli. "Chapter 19 Classification Method to Predict Chances of Students' Admission in a Particular College", Springer Science and Business Media LLC, 2021](#)

< 1% match (publications)

[Shirina Samreen. "Memory-Efficient, Accurate and Early Diagnosis of Diabetes Through a Machine Learning Pipeline Employing Crow Search-Based Feature Engineering and a Stacking Ensemble", IEEE Access, 2021](#)

< 1% match (Internet from 19-Jul-2021)

<https://aclanthology.org/2019.icon-1.2.pdf>

< 1% match (publications)

[Wael Etaoui, Ghazi Naymat. "The Impact of applying Different Preprocessing Steps on Review Spam Detection", Procedia Computer Science, 2017](#)

< 1% match (Internet from 13-Oct-2021)

[https://akademik.arel.edu.tr/arastirma-birimleri/scopus-bildiri-koleksiyonu/a-data-driven-approach-to-kinematic-analytics-of-spinal-motion\(50a4d293-c173-4297-b062-2c89ed2cb60c\)/tum-oge-kaydi](https://akademik.arel.edu.tr/arastirma-birimleri/scopus-bildiri-koleksiyonu/a-data-driven-approach-to-kinematic-analytics-of-spinal-motion(50a4d293-c173-4297-b062-2c89ed2cb60c)/tum-oge-kaydi)

< 1% match (Internet from 30-Sep-2021)

<https://coek.info/pdf-an-ensemble-approach-for-spam-detection-in-arabic-opinion->

< 1% match (publications)

[Amirah Mohamed Shahiri, Wahidah Husain, Nur'aini Abdul Rashid. "A Review on Predicting Student's Performance Using Data Mining Techniques", Procedia Computer Science, 2015](#)

< 1% match (publications)

[Jake Vasilakes, Sicheng Zhou, Rui Zhang. "Natural language processing", Elsevier BV, 2021](#)

< 1% match (publications)

["Pattern Recognition. ICPR International Workshops and Challenges", Springer Science and Business Media LLC, 2021](#)

< 1% match (publications)

[Ting-Ying Chien, Hsien-Wei Ting, Chih-Fang Chen, Cheng-Zen Yang, Chong-Yi Chen. "A Clinical Decision Support System for Diabetes Patients with Deep Learning: Experience of a Taiwan Medical Center", Research Square Platform LLC, 2021](#)

< 1% match (Internet from 21-Oct-2021)

<https://dokumen.pub/intelligent-sustainable-systems-proceedings-of-iciss-2021-lecture-notes-in-networks-and-systems-213-1st-ed-2022-9811624216-9789811624216.html>

< 1% match ()

[Canlin Zhang, Daniel Biś, Xiuwen Liu, Zhe He. "Biomedical word sense disambiguation with bidirectional long short-term memory and attention-based neural networks", BMC Bioinformatics](#)

< 1% match (publications)

[Ramesh Kumar Huda, Haider Banka. "Efficient feature selection methods using PSO with fuzzy rough set as fitness function", Soft Computing, 2021](#)