Bangla Fake News Detection: Machine Learning Perspective

BY

Md. Ibrahim Khan

ID: 171-15-9155

AND

Md. Ashiqur Rahman

ID: 171-15-9169

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

Supervised By

Sheikh Abujar

Senior Lecturer

Department of CSE

Daffodil International University

Co-Supervised By

Ahmed Al Marouf

Lecturer

Department of CSE

Daffodil International University

DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

January 2021

APPROVAL

This Project/internship titled **"Bangla Fake News Detection: Machine Learning Perspective"**, submitted by **Md. Ibrahim Khan,** ID No: 171-15-9155 and **Md. Ashiqur Rahman**, ID No: 171-15-9169 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 28-01-2021.

BOARD OF EXAMINERS

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

Abdus Sattar — **Internal Examiner**
Assistant Professor
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Md. Jueal Mia — **Internal Examiner**
Senior Lecturer
Department of Computer Science and Engineering
Faculty of Science & Information Technology
Daffodil International University

Dr. Dewan Md. Farid — **External Examiner**
Associate Professor
Department of Computer Science and Engineering
United International University

Declaration

We here by declare that, this project has been done by us under the supervision of **MR. Sheikh Abujar, Senior Lecturer, Department of CSE Daffodil International University.** We declare that neither this project nor any part of this project has been submitted elsewhere for award of any degree or diploma.

Supervised By:

Mr. Sheikh Abujar
Senior Lecturer
Department of CSE
Daffodil International University

Co-Supervised By

Ahmed Al Marouf
Lecturer
Department of CSE
Daffodil International University

Submitted By:

Md. Ibrahim Khan
ID: 171-15-9155
Department of CSE
Daffodil International University

Md. Ashiqur Rahman
ID: 171-15-9169
Department of CSE
Daffodil International University

ACKNOWLEDGEMENT

First, we express our heartiest thanks and gratefulness to the Almighty for His divine blessings that make us possible to complete the final year thesis successfully.

We are very much grateful to **Mr. Sheikh Abujar**, Senior Lecturer, Department of CSE Daffodil International University, Dhaka. He was so helpful for the whole bunch of times. Whenever we needed his help no matter what he was always there for us. His motivation, inspiration, and direction helped us to fulfill our thesis work very effectively. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior drafts and correcting them at all stages have made it possible to complete this project.

We would also like to thank our co-supervisor **Mr. Ahmed Al Marouf, Lecturer,** Department of Computer Science and Engineering, Daffodil International University, Dhaka. He showed us many ideas on this human-computer interaction. He inspired and motivated us that we would be able to come up with this idea and would be able to make this work happen.

We would also like to express our heartiest gratitude to **Professor Dr. Touhid Bhuiyan**, Head of Computer Science and Engineering for all his appreciation, adoration and motivation.

We are also very much thankful to our parents and friends who were always there to criticize our work in a manner to improve that all the way long. So thank all of them from the core of our heart.

ABSTRACT

Detecting fake news is a challenging job and likewise has tremendous genuine political and social effects. Mostly it's spread over social news sites and social media. Fake news is written intentionally to mislead readers which is very bad for society, a country, and also for the whole of mankind. In our mother tongue, fake news spread a lot. Which makes the situation worst? To reduce fake news basically in the Bangla language and ensuring more online and social security we work on this thesis. Because of Fake news, people become misguided and make mistakes. That is why we discuss away in our research to detect Bangla fake news from online news using the advantages of "Machine Learning and Natural Language Processing (NLP)". We gather around nine thousand Bengali news to get the result. We utilize basic and diligently chose Bangla news to precisely recognize fake news. The test results show a 92.6% accuracy utilizing the SVM model.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER 1

Introduction

1.1 Introduction

Fake news can be as slithery to characterize all things considered to nail down. The news might be verifiably erroneous and promptly distributed to underscore a specific perspective or drive loads of guests to a site, or they could be part of the way evident however misrepresented or not completely actuality checked before distribution. Because of the absence of any administrative frameworks, this news can't be confirmed. Day by day we are depending on social media platforms online. People tend to seek out information but then some people get attracted to fake information which leads them towards the darkness.

The spread of fake news is anything but another idea. In any case, lately, it has become a genuine danger that can't be disregarded any longer. Simple admittance to the Internet and the hyperactivity of clients in different social news locales and web-based media stages have offered to ascend to the broad spread of fake news. The Internet has to a great extent supplanted conventional news media. Numerous individuals, particularly a gigantic bit of youth rely upon the Internet and online media as the essential hotspot for news utilization as a result of their simple access, minimal effort, and day in and day out accessibility. They have confidence in what they read on the web and spread the word that they thought to be valid. In this way, more often than not, parodies are not spread to misdirect. Be that as it may, some of the time a few people for their advantage exploit and advance the spread of parodies as genuine news. [1]

So much research has been done already to detect fake news by using English data. But Using Bangla data detecting Bengali fake news is still in the early stage. In this research, we are trying to detect fake news using the Bangla dataset. We try to get our desire result based on machine learning and NLP models.

1.2 Motivation

Presently a day, some non-trustworthy sources have been distributing fake and appealing reports. Consequently, these inconsistent sources can distribute anything they desire, and even now and again, it makes confusion in the public arena. Lately because of the straightforwardness in web accessibility and web-based media, the improper word can get out more rapidly than any time in recent memory. Sometimes, fake news is more alluring than the genuine one. In this way, individuals become confused and commit errors. That is the reason we will examine an approach to detect fake news in social media and online news sites using the advantages of “a Natural Language Processing (NLP)”.

1.3 Rationale of the study

The number of people who use the Bengali language as their local language is tremendous. As we already knew that spreading fake news is now a common concept. In Bangla, it is also a common matter. But it is very hard to detect which news is fake and which one is correct. Additionally in this advanced time, the devices and innovation of the Bengali language are not as rich as different languages. Therefore, we need to assemble the developments for this language. The greater part of the content-related issues can be desired by NLP gadgets and strategies. Most and significant NLP strategies as of now work for different languages, for example, English, French, Chines and so on. But for Bengali content, a couple of models have been constructed which isn't sufficient. In this manner, the exploration region of Bengali NLP should be expanded. Doing any research on Bangla fake news detection data the main obstacle is Bangla fake news dataset. In this exploration work, we attempt to show how to detect Bangla fake news from thousands of authentic data.

1.4 Research Questions

- What is fake news?
- How Bangla fake news spread?
- How does Bangla fake news detection work?
- What are the advantages of Bangla fake news detection?
- What is the dissimilation among Bangla and English fake news?
- How to preprocess Bangla text data in Machine learning and NLP?
- What are the future works of Bangla Fake News Detection?
- How does Bangla fake news detection Model work?

1.5 Expected Output

Since this is an exploration project, our key concern was to convey an assessment paper in an associated field. Exploration works reliably a constant cycle. Various individuals dissect express examination subjects to locate a helpful arrangement. By then, the engineer intensifies the instruments for the end-clients. The most noteworthy number of exploration work and apparatuses are prospered in English utilizing natural language processing, deep learning, neural networks, etc. But using the Bengali language is still in its early stage. Some researchers and developers are trying to make a Bangla fake news dataset and resources for everyone. In the Bengali language, fake news detection is still a new research topic. A very little amount of experiments are done previously for Bangla fake news detection. In this research, we applied Machine learning strategies for detecting Bangla fake news and show vital strides on the most ideal approach to build a diagram for Fake news identification. Our main expected output is to detect Bangla fake news from social media news.

1.6 Report Layout

This report has a total of 5 chapters.

- Chapter 1 contains an outline of the entire research work. It has a few segments, for example, 1.1 Introductions of the work, 1.2 Motivation of this examination, 1.3 Rational Study of the pursuit, 1.4 Research Questions, 1.5 Expected Output, and 1.6 Reports Layout of the exploration.
- In Chapter 2 we have talked about Background Studies of the exploration and their subsections are 2.1 Introductions, 2.2 Related works, 2.3 Research Summary, 2.4 Scope of the Problem, 2.5 Challenges.
- In Chapter 3 we have talked about the entire Research Methodology with subsections 3.1 Introduction, 3.2 Research Subject, and Instrumentation, 3.3 Data assortment strategy, 3.4 Statistical Analysis of Datasets, 3.5 Implementation Requirements.
- In Chapter 4 Experiment and Results of the examination are talked about and the subsection is 4.1 Introduction, 4.2 Experimental Results, 4.3 Descriptive Analysis, 4.4 Summary.
- Chapter 5 contains the Conclusion and future works of the research with subsections 5.1 Summary of the Study, 5.2 Conclusion, 5.4 Implication for Further Study.

End of all areas given the references which encouraged us in our exploration work.

CHAPTER 2

BACKGROUND

2.1 Introduction

fake news is the very much arranged spread of deception by means of standard news media or online media. fake news disperses strikingly quickly. This is shown by the way that, when one fake news site is brought down, another will in no time have its spot. Individuals can peruse articles from various destinations, share the data, re-share from others and before the day's over, the fake news has gone so distant from its key site that it gets indistinct. Data control is certifiably not a surely known theme and by and large not at the forefront of anybody's thoughts, particularly when fake news is being shared by a companion or close one.

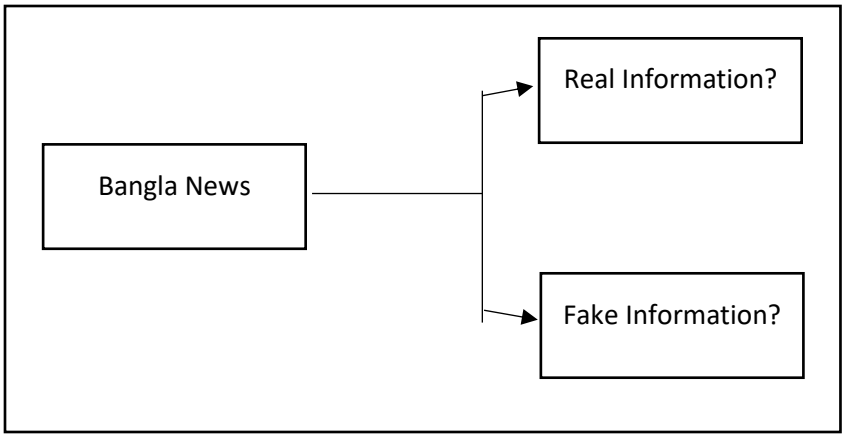

Figure 2.1.1: Bangla Fake news detection

A large portion of the user will in general allow their watchman to down via online media and conceivably assimilate all the bogus data as though it were reality data. This is likewise considerably more damaging thinking about how clients will in general depend via online media to educate them regarding governmental issues, significant occasions, and breaking news.

2.2 Related Work

Fake news is one of the most analysis-able themes in NLP. Much investigation work has been done in this field for various languages. Significant exploration was held on fake news identification. This segment will be examined about some respectable investigation in these fields.

A glance at synchronous academic work exhibit that the exposure of fake news has been a significant worry among researchers from different foundations. For example, a few creators have seen that fake news is not, at this point a safeguard of the promoting and publicizing workplaces.[2]

They propose a model that combines three characteristics for a more accurate and automated prediction. Propose model called CSI which is composed of three modules: Capture, Score and Integrate. CSI model consists of two main parts, a module for extracting temporal representation of news articles, and a module for representing and scoring the behavior of users. They proposed a model using Neural Network. But the availability of labeled examples of true and fake news may be limited. It has lacked user labels. CSI is based on deep neural networks. [3]

This paper gives analysts a guide of the current scene of veracity appraisal strategies, their significant classes, and objectives, all to propose a hybrid approach to deal with the framework plan. These strategies have risen out of isolated advancement streams, using dissimilar methods. They have utilized two significant classes of strategies: 1. Linguistic Approach, 2. Network Approach. Additionally utilized profound punctuation examination, Semantic Analysis, Rhetorical Structure, and Discourse Analysis for better performance. [4]

In this paper, they have used machine learning methods to get the output of their research. Their best performing models by in general ROC AUC are Stochastic Gradient Descent models prepared on the TF-IDF include set as it were. They see that PCFGs don't add a lot of prescient worth, however balance the Recall for our top-performing model. They needed to demonstrates that PCFGs are useful for a Fake-News Filter type execution versus, state, focusing on phony news destinations for survey. [5]

This paper shows a straightforward methodology for fake news identification utilizing a naive Bayes classifier. They utilized this methodology as a product framework and tried it against an

informational collection of Facebook news posts. The primary objective of the exploration is to look at how this specific strategy functions for this specific issue given a manually marked news dataset and to help to utilize computerized reasoning for counterfeit news discovery. [6]

2.3 Research Summary

In this research, we used machine learning methods for Bangla Fake News Detection. We used various machine learning models. To implement this model, we have used our dataset and also used a dataset from Kaggle. Dataset has collected form social media and different news sites also. At first collect Bengali news articles from different media. At that point make a summary of every piece of information. Consequently, the data set contains 8 sections, which are article id, domain, date, category, source, relation, headline, content, and label. The total number of around nine thousand data in the dataset that we used. Before applying the machine learning algorithms, we preprocessed our Bangla dataset. In this preprocessing stage, we check all noises and try to remove all of them. We used Count Vectorizer then. Which is used to change a given text into a vector-based on the frequency (count) of each word that happens in the whole content. After using the count vectorizer and characterize all functions and library then we train the model for more than 5 hours. At that point, we discovered a decent reaction from the machine.

2.4 Scope of the problem

Since Fake news detection is a new examination in Bengali NLP distinctive procedure is developed step by step. This exploration work utilizes Machine learning techniques for identification. In our dataset, so many noises were found. So, from the early stage, we didn't get the expected results. Also, fake news data is not sufficient enough so we had to use little authentic data to balance the dataset. For this research purpose, we used Machine Learning models. Also, we used Supervised learning algorithms for our labeled data then we get a good result.

2.5 Challenges

First of all, finding a Bangla dataset is not an easy task. All information is available in an unstructured manner. Thusly, information assortment is a test for this research. We couldn't find a reliable dataset to use. Along these lines, we need another dataset to complete this exploration work. Since the assortment of the dataset, labeling all of the data is another challenging work. Therefore, the preprocessing step needs fresh coding to set up the content as a contribution of a model. We have faced so many problems to run the dataset. Another problem is to run the Bangla dataset as a string. For other languages like English have a well maintain dataset. But for the Bengali language, it's still in an early stage, so we had to do it on our own. Finding a large amount of exact fake news is another test is in this examination. On the other hand, if the dataset has countless fake information, it would give assists with creating a more precise result.

CHAPTER 3

Research Methodology

3.1 Introduction

Here, we will examine the entire strategy of the exploration action. Each examination action has an interesting tackling strategy. Applying all methodologies is remembered for the technique part. Here give an itemized conversation of applying to utilize a model with a short depiction of each piece of the technique.

In our examination, we have utilized AI models for fake news. Machine learning algorithms are used to includes computers learning from information provided so that they do certain tasks. The Machine Learning approach is partitioned into three classes:

- Supervised learning
- Unsupervised learning
- Reinforcement learning

For this research purpose, we used a supervised algorithm. Each Machine learning model requires a decent dataset to locate a precise outcome. To apply the calculation, the dataset should be gathered and preprocessed. Then, each segment of the methodology is examined separately. Given all areas are followed when the examination work is finishing. A superior clarification of technique builds the proficiency of work and gives honorability. Numerical conditions and a graphical perspective on the model with their depiction is assisting with understanding the entire work. Thusly, further examination and expanding the exploration field great clarification of the strategy is required. The entire work resembles a structure. All means of the procedure are momentarily talked about in this part. A subsection of some center segments is assisting with understanding the essence of the model with its motivation of utilization. The working progress of the entire examination work is given beneath which gives a short perspective on the absolute exploration work.

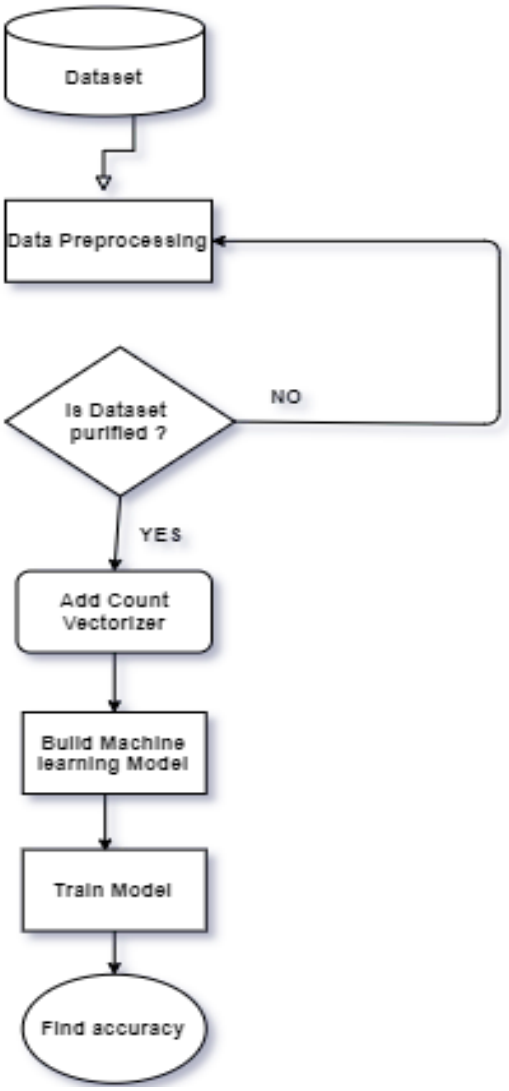

Figure 3.1.1: Workflow for fake news detection

3.2 Research Subject and Instrumentation

Our research topic name is “Bangla Fake News Detection Using Machine learning and NLP (Natural Language Processing)”. This is a significant examination zone in Bengali NLP. We have examined the way toward making a system of fake news detection in Bangla with the applied and hypothetical cycle first to now. To evaluate the hole work needs a high setup Computer with GPU and other instruments. A catalog is given beneath the necessary apparatus for this work.

Hardware and Software:

- Intel Core i3 / Core i5 including minimum 8GB RAM
- 1 TB HDD
- Google Colab including 12GB GPU and 350 GB RAM

Advancement Tools:

- Windows 10
- Python 3.7
- TensorFlow Backend Engine
- NLTK
- Pandas
- NumPy

3.3 Data collection and Data preprocessing

We utilized properly gathered information for the exploration reason. All information is gathered from online media and online news sites such as bd-pratidin, jugantor, bd24live, somoynews, etc. There is some unpredictability to gather information from online sites for security issues. For that, we gather information utilizing manual methodologies. Around 9000 data is being aggregated from online sites. After aggregating data, the dataset needs to preprocess and create a correct dataset to implement codes. The important strides of the preprocessing period in figure 3.3.1 are examined area shrewd beneath.

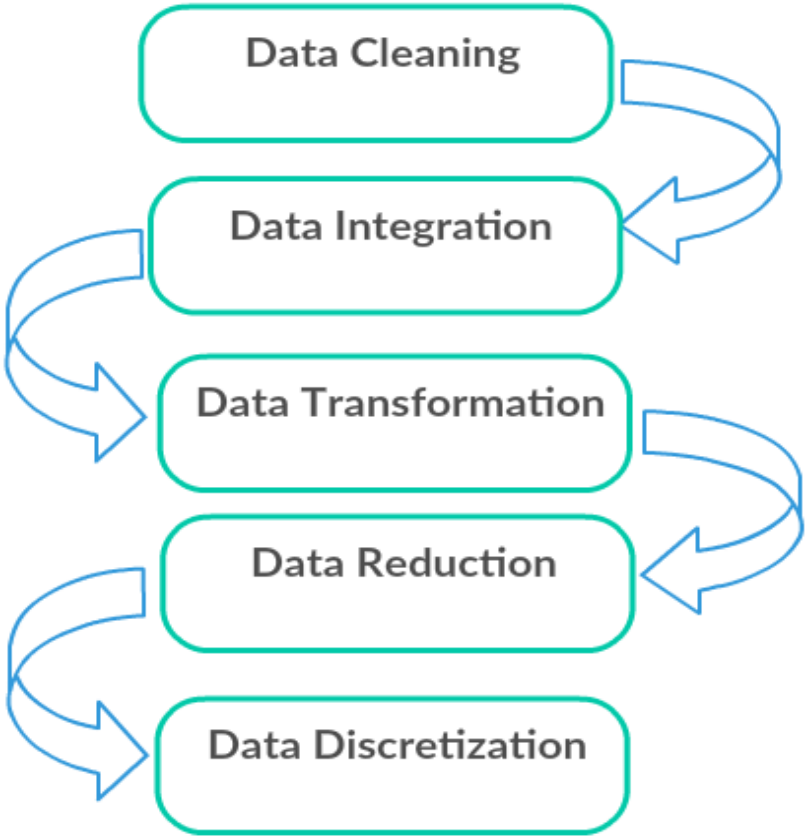

Figure 3.3.1 Steps of data preprocessing

3.3.a Data cleaning

It is the way toward getting ready information for examination by eliminating or modifying data that is incorrect, incomplete, irrelevant, copied, or inappropriately organized. We can easily clean data by dropping or imputing missing data, removing unwanted data, fixing structural errors and so many ways to clean a full dataset. In Bengali languages, it is a very common problem of missing data when we are making a dataset so we need to clean it for better performance.

3.3.b Data Integration

Data Integration is a technique of integrating the data which resides in different sources. The goal is to provide users with a holistic view of the data. It can be viewed more as a practice of consolidating data from various disparate sources. This is viewed as one of the most important steps in Data preprocessing. It has some techniques like data replication, data virtualization, streaming data integration, etc.

3.3.c Data transformation

In our dataset, we used data transformation also for joining two datasets (authentic data and Fake data) and changing column names. Data transformation entails some very general tasks, such as joining datasets or changing column names.

3.3.d Data reduction

Data reduction refers to the technique of reducing the dimension of a data feature set. In our datasets which contain hundreds of columns (i.e., features) or an array of points, creating a massive sphere in a three-dimensional space.

3.3.e Data Discretization

Data discretization is the way toward changing over persistent information into discrete containers by gathering it. Discretization is likewise known for the simple viability of the information. Preparing a model with discrete information turns out to be quicker and more viable than while endeavoring the equivalent with consistent information. Albeit nonstop esteemed information contains more data, tremendous measures of information can back the model off. Here, discretization can help us find some kind of harmony between both.

3.4 Statistical Analysis

1. Complete amount of information 8502. All information of our dataset was collected from online news sites. A short perspective on our dataset is given underneath in table 3.4.1.

Table 3.4.1: Sample of the dataset

ArticleID	Domain	Date	Category	Source	Relation	Headline	Lebel
1	bd-pratidin.com	08.09.2019	Politics	Reporter	Related	আওয়ামী লীগ ও ঐক্যফ্রন্ট দুই দলের নির্বাচনী ইশতেহারেই যা নেই	1
2	jugantor.com	25.11.2020	Lifestyle	Reporter	Unrelated	"এতো মেকআপ করো কেন?" এই প্রশ্নে বিরক্ত হয়ে প্রেমিকের মাথা ফাটিয়ে দিলো তরুণী!	0
3	bd24live.com	05.09.2020	Miscellaneous	Reporter	Unrelated	নোয়াখালী ছাড়া দেশের কোথাও ফকির মিসকিন নেই	0
4	channeldhaka.news	19.05.2015	Politics	Reporter	Related	মাশরাফি নমিনেশন পেলে লেংটা হয়ে দৌড়ানোর প্রতিশ্রুতি দিলেন এক সাকিবভক্ত	0

5	somoynews.tv	06.08.2013	Politics	Reporter	Unrelated	এনজেলা মার্কেলের পর ওবামার শিকার এম কে আনোয়ার | দৈনিক মতিকণ্ঠ	0
6	banglanews24.com	05.02.2018	Miscellaneous	Reporter	Related	বিশ্বের পবিত্র ৭ গাছের তালিকায় রয়েছে গাঁজা!	0
7	earki.com	01.14.2019	Miscellaneous	reporter	Unrelated	এবার মশার বিরুদ্ধে গর্জে উঠলো ছাত্রলীগের কামান	0

2. The total number of columns is 8.

3. The complete amount of identical words in the dataset is 4392k.

4. 95% of the word we utilized in our model.

5. Dataset is stored in a CSV document which augmentation is .csv.

3.5 Implementation Requirements

3.5.a. Problem discussion

In our dataset, we have used 5 machine learning models. We have used classification and regression models by using a split method in our research.

3.5.b. Split method

The train-test split is a procedure for assessing the exhibition of a machine learning algorithm. We utilized this strategy for grouping or relapse issues. The system includes taking a dataset and separating it into two subsets. We divided our dataset into two subsets. One subset is 80% and one subset has 20% of all data. Data will be taken randomly from datasets. We used two-column for the final implementation: Headline and label. Lebel is divided into two categories. 0 and 1 which represents fake as 0 and true as 1.

3.5.c. Decision tree

The decision tree algorithm falls into the classification of directed learning. They can be utilized to tackle both relapse and characterization issues. Here in our work we used the decision tree to find out either the news is fake or true and what is the probability of the news is to be fake.

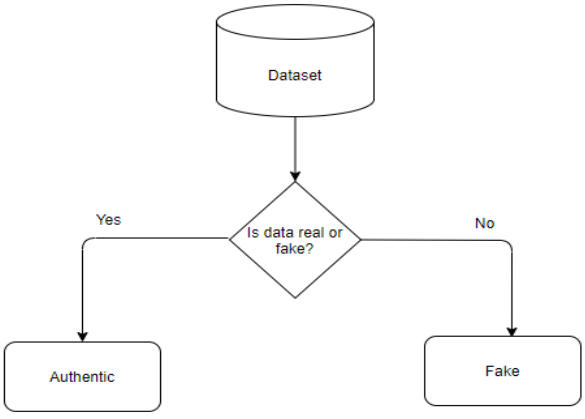

Figure 3.5. c.1. Decision tree for fake news detection.

Let X is a set of instances, A is an attribute (example: label, headline), Xv is the subset of X with A = v, and Values (A) is the set of all possible values of A, then

Gain (X, A) = Entropy(X) - \sum_ {v \epsilon Values(A)} \frac {\left | X_{v} \right |} {\left | X \right |}. Entropy(X_{v}).

3.5.d. Naive Bayes

Naive Bayes classifiers are an assortment of classification algorithms dependent on Bayes' Theorem. It's anything but a solitary of algorithms yet a group of calculations where every one of them shares a typical guideline, for example, each pair of features being grouped is free of one another.

The dataset is separated into two sections, specifically, the feature matrix and the response vector.

- The feature matrix contains all the vectors(rows) of the dataset in which every vector comprises the estimation of ward features. In our dataset, features are 'headline', and 'content'.

- Response vector contains the estimation of the class variable(prediction or yield) for each line of feature matrix. In our dataset, the class variable name is 'label'.

Bayes' Theorem finds the likelihood of an occasion happening given the likelihood of another occasion that has just happened. Bayes' hypothesis is expressed numerically as the accompanying condition:

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

- Fundamentally, we are attempting to discover the likelihood of occasion A, given the occasion B is valid. Occasion B is likewise named as proof.

- P(A) is the priority of A (the earlier likelihood, for example, Likelihood of occasion before a proof is seen). The proof is a property estimation of an obscure instance(here, it is occasion B).

- P(A|B) is a posteriori likelihood of B, for example, the likelihood of occasion after proof is seen.

Presently, with respect to our dataset, we can apply Bayes' hypothesis in after way:

$$P(y|X) = \frac{P(X|y)P(y)}{P(X)}$$

where y is class variable and X is a reliant element vector (of size n) where:

$$X = (x_1, x_2, x_3, ..., x_n)$$

3.5.e. Random Forest

Settling on more number choice trees to make the backwoods we won't utilize a similar apache of building the choice with data gain or Gini file approach. In our exploration, handles the missing qualities and gives a clear-cut result.

Random Forest works in two-stage initially is to make the random forest by consolidating N choice tree, and second is to make expectations for each tree made in the principal stage.

The Working cycle can be clarified in the underneath steps and chart:

Step-1:

Select arbitrary K information focuses from the training set.

Step-2:

Build the decision trees related to the chosen data focuses (Subsets).

Step-3:

Choose the number N for decision trees that we need to construct.

Step-4:

Repeat Steps 1 and 2.

Step-5:

For new information focuses, discover the predictions of every decision tree, and allow the new information to focus on the classification that wins the majority votes.

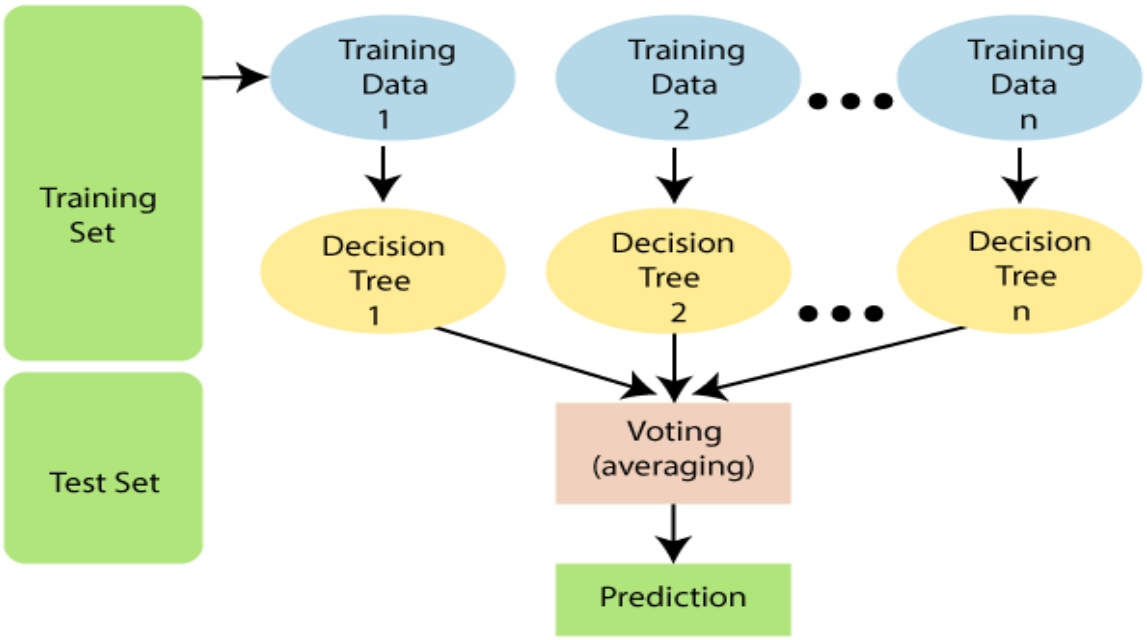

Figure 3.5.e.1: Random Forest Algorithm working procedure

3.5.f. K-nearest neighbors (KNN)

K-nearest neighbors (KNN) computation uses 'incorporate likeness' to predict the assessments of new datapoints which further suggests that the new data point will be given out a value reliant on how eagerly it organizes the concentrations in the arrangement set. We used KNN algorithms in our research with the help of the following steps −

Step 1

We need to pick the value of K for example the closest information focuses. K can be any number.

Step 2

For each point in the test, information do the following −

I. Figure the distance between test information and each line of preparing information with the assistance of the strategies in particular: Euclidean, Manhattan, or Hamming distance. Euclidean is the most regularly utilized technique to ascertain distance.
II. Presently, in view of the distance esteem, sort them in rising request.
III. Then, it will pick the top K lines from the arranged array.
IV. Presently, it will set a class to the test point dependent on the most successful class of these columns.

Step 3

End

3.5.g. Support Vector Machine (SVM)

SVM is a supervised AI calculation that is generally utilized for characterization and regression difficulties.

For a dataset comprising of features set and labels set, an SVM classifier fabricates a model to anticipate classes for new models. It doles out new model/information focuses to one of the classes. In the event that there are just 2 classes, at that point, it tends to be called a Binary SVM Classifier.

There are 2 kinds of SVM classifiers:

1. Linear SVM Classifier
2. Non-Linear SVM Classifier

SVM Linear Classifier:

In the linear classifier model, we accepted that preparation models plotted in space. These information focuses are relied upon to be isolated by a clear gap. It predicts a straight hyperplane partitioning 2 classes. The essential concentration while drawing the hyperplane is on augmenting the separation from hyperplane to the closest information purpose of one or the other class. The drawn hyperplane is called a maximum-margin hyperplane.

SVM Non-Linear Classifier:

In reality, our dataset is for the most part scattered up somewhat. To tackle this difficult partition of information into various classes based on a straight linear hyperplane can't be viewed as a decent decision. For this Vapnik proposed making Non-Linear Classifiers by applying the kernel trick to maximum-margin hyperplanes. In Non-Linear SVM Classification, data focus plotted in a higher-dimensional space.

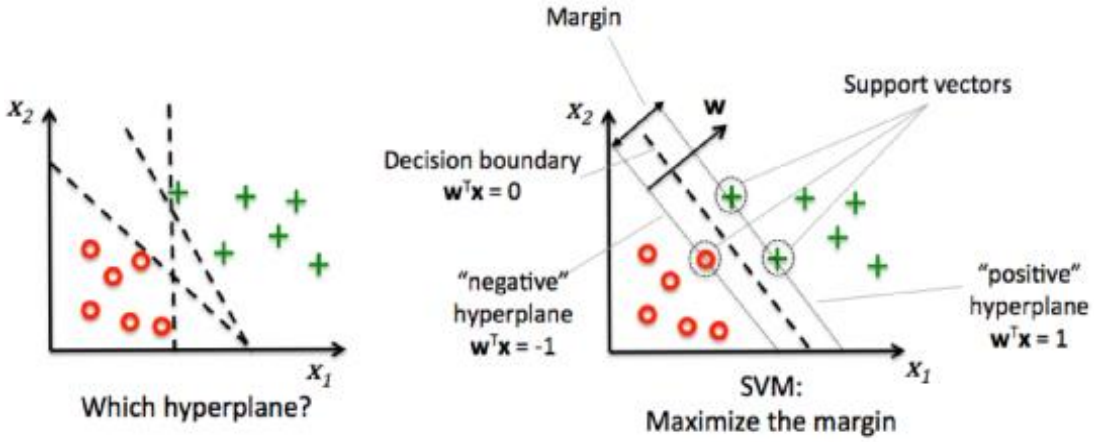

Figure 3.5.g.1: SVM Classifier

CHAPTER 4
Experimental Results and Discussion

4.1 Introduction

Fake news detection is not an easy task. The machine can detect fake news automatically but trace an exact result is very hard. Accuracy prediction is more significant for this fake news detection. Whereas the machine gives the result based on the highest feasibility which we get different models in our research. After collecting data from different online news sites at first we need preprocessing to remove errors and missing values from the dataset. At that point, the dataset needs to prepare for the model to learn the machine. We use the split method to train the dataset.

As we all know that an all-around prepared model can send the best result in the trial period. Lofty design computer needs to prepare the model for quick results. We train our model right off the bat on the direct computer using Anaconda software. That sets aside some additional effort to prepare the model. At last, we train the model utilizing Google Colab. Which gives free GPU administration to a client. That is absolutely the first and decreases trial period. Presently the accuracy of utilizing various models in our research is given below.

Table 4.1.1: Accuracy of models

Model	Accuracy
Decision Tree	88.5%
Random Forest	91%
Naïve Bayes	85%
K Nearest Neighbors	87%
Support Vector Machine(SVM)	92.6%

The highest accuracy we get **92.6%** from **Support Vector Machine Algorithm**.

4.2 Experimental Results

The machine gives result almost the genuine result. Everyone realizes that no machine gives a 100% exact result. Also, our prepared model gives a decent outcome however not for all qualities. Once in a while it reactions to wrong information relating to the dataset. Yet, the greatest number of the reaction is indicating our ideal outcome.

We train our model using the split method. We slit our dataset 80/20. We have used 5 different models of machine learning algorithms to get the most accurate result. And finally, we get 92.6% accurate results using the machine learning algorithm for fake news detection.

4.3 Descriptive Analysis

Before applying models on the Bangla dataset, we apply these models for the English dataset for detecting fake news. We have tested our system in both noisy and noise-free environments and our result is quite satisfactory. We have tested the accuracy of the system by testing the dataset using a different model. But we have found some problems in a noisy environment. As in a noisy environment, the system was getting confused with the other noisy commands. So the accuracy level in a noise-free environment is quite praiseworthy in terms of the accuracy level of the system in a noisy environment. So above all, this system can cope up with all the surrounding problems and we are working to solve the noisy issue.

4.4 Summary

In a nutshell, we can say that we have implemented our idea that we had dreamt of. We wanted to detect Bangla fake news and finally, we did it. And we are pretty much successful in that attempt.

CHAPTER 5
Conclusion and Future Work

5.1 Summary of the Study

Our entire research work is identified with the Bengali NLP. In this task, we have utilized machine learning models for detecting Bangla fake news. That is useful for making a programmed Bengali fake news detector. We have finished this research in a half year. The entire cycle of work is partitioned into certain parts. The entire synopsis of the research is given beneath bit by bit.

Step 1: Planning

Step 2: Problem Analysis

Step 3: Model design

Step 4: Data collection form online news sites

Step 5: Summarize the collected data

Step 6: Labeling all the data

Step 7: Data preprocessing

Step 8: Data transformation

Step 9: Train dataset

Step 10: Train models

Step 11: Check the outcome and examination of the reaction of the machine

This research will aid our Bengali NLP exploration local area to fabricate a completely reliant programmed fake news detector and further examination of Bengali fake news discovery.

Presently we will talk about the future work and finish of this exploration work.

5.2 Conclusion

The fundamental worry of this exploration work is creating and expanding the Bengali NLP research territory. We have utilized the Bengali information as the contribution of our model. From the outset, we construct the model for the English dataset then we make this model for the Bengali dataset. Our dataset is not large. Because Bangla data is not available so much. However, the machine gives superb reactions to this dataset. This research is for Bangla fake news detection. Finding fake news is like looking for a needle in a haystack.

All things considered, no machine gives a completely exact outcome. Each machine has a few restrictions in its working field. Additionally, our research work likewise has a few constraints. In any case, the primary concern is that the model can detect Bangla fake news for the Bengali language. This is an accomplishment for our Bengali NLP recorded which accommodating for future exploration work.

5.3 Recommendations

In the following phase of our work, we will expand the dataset and their rundown for improving the model execution. We will attempt to fabricate another model that will assist us with finding the best entertainer for Bengal fake news detection. We work only using some machine learning models. Some recommendations for Bangla fake news detection is given below,

- Make a big dataset of Bangla fake news
- Understand the pattern of fake news
- Understand which sites are using mostly for publishing fake news.
- Make a better model
- Try to get more accuracy.

5.4 Implication for Further Study

Some restriction is introduced in our research work, for example, work using only the machine learning model, the dataset isn't sufficient. In any case, the model is worked for the future turn of events. Whereas any exploration work is a consistent cycle. Hence, this model will be created step by step for the Bengali language. To locate a legitimate arrangement any works, need more examination. At that point, all exploration finds a legitimate answer for a particular issue. Along these lines, research work needs future execution or advancement. Future usage is subject to the constraints of the past work. Settling the restrictions of the past work assists with making an effective framework. In this work, the destiny work will be expanding the dataset of the Bengali fake data and we will use neural network and deep learning models for better performance

Subsequent to finishing exploration, the framework needs to send. Consequently, making an application like web and portable application is significantly dependent on the fate of computerized reasoning. Such an application can easily give the output of an input data that the data is either fake or authentic.

References

[1] M. A. M. M. S. I. Arnab Sen Sharma, "Automatic Detection of Satire in Bangla Documents: A CNN Approach Based on Hybrid Feature Extraction Model," *International Conference on Bangla Speech and Language Processing(ICBSLP),* pp. 27-28, 2019.

[2] Y. C. a. N. J. C. Victoria L. Rubin, "Deception Detection for News: Three Types of Fakes," *ASSIST,* pp. 1-4, 2015.

[3] S. S. Y. L. Natali Ruchansky, "CSI: A Hybrid Deep Model for Fake News Detection," *International Conference on Information and Knowledge Management,* pp. 797-806, 2017.

[4] V. L. R. a. Y. C. Nadia K. Conroy, "Automatic Deception Detection: Methods for Finding Fake," *ASSIST,* vol. 52, no. 1, pp. 1-4, 2015.

[5] S. Gilda, "Evaluating Machine Learning Algorithms for Fake News Detection," *2017 IEEE 15th Student Conference on Research and Development (SCOReD),* pp. 110-115, 2017.

[6] V. M. Mykhailo Granik, "Fake News Detection Using Naive Bayes Classifier," *IEEE First Ukraine Conference on Electrical and Computer Engineering (UKRCON),* 2017.

ORIGINALITY REPORT

20%	18%	7%	10%
MILARITY INDEX	INTERNET SOURCES	PUBLICATIONS	STUDENT PAPERS

PRIMARY SOURCES

1	dspace.daffodilvarsity.edu.bd:8080 Internet Source	10%
2	www.ukessays.com Internet Source	2%
3	Submitted to K. J. Somaiya College of Engineering Vidyavihar, Mumbai Student Paper	1%
4	arxiv.org Internet Source	1%
5	www.learnbay.co Internet Source	1%