

BENGALI FAKE NEWS DETECTION USING MACHINE LEARNING

BY

ADITI BALO

ID: 152-15-6064

AND

JAMIUL ISLAM

ID: 152-15-6147

AND

ABDULLAH AL BAKI

ID: 152-15-6169

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

Supervised By

Mr. Sheikh Abujar

Lecturer

Department of CSE

Daffodil International University



DAFFODIL INTERNATIONAL UNIVERSITY

DHAKA, BANGLADESH

MAY 2019

APPROVAL

This Project titled "**Bengali Fake News Detection Using Machine Learning**", submitted by Aditi Balo, ID No: 152-15-6064, Jamiul Islam, ID No: 152-15-6147, Abdullah Al Baki, ID No: 152-15-6169 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 3 may 2019.

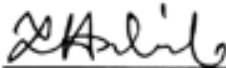
BOARD OF EXAMINERS



Dr. Syed Akhter Hossain
Professor and Head

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

Chairman



Md. Tarek Habib
Assistant Professor

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

Internal Examiner



Moushumi Zaman Bonny
Senior Lecturer

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

Internal Examiner



Dr. Swakkhar Shatabda
Associate Professor

Department of Computer Science and Engineering
United International University

External Examiner

DECLARATION

We hereby declare that, this project has been done by us under the supervision of Mr. Sheikh Abujar, Lecturer, 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:



Mr. Sheikh Abujar
Lecturer
Department of CSE
Daffodil International University

Submitted by:



Aditi Bala
ID: 152-15-6064
Department of CSE
Daffodil International University



Jamilul Islam
ID: 152-15-6147
Department of CSE
Daffodil International University



Abdullah Al Baki
ID: 152-15-6169
Department of CSE
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 successfully. We are really grateful and wish our profound indebtedness to **Mr. Sheikh Abujar, Lecturer**, Department of CSE, Daffodil International University, Dhaka. Deep Knowledge & keen interest of our supervisor in the field of “*Natural Language Processing*” to carry out this project. His endless patience, scholarly guidance, continual encouragement, constant and energetic supervision, constructive criticism, valuable advice, reading many inferior draft and correcting them at all stage have made it possible to complete this project. We would like to express our heartiest gratitude to the Almighty GOD and **Prof. Dr. Syed Akhter Hossain, Head**, Department of CSE, for his kind help to finish our project International and also to other faculty member and the staff of CSE department of Daffodil University. We would like to thank our entire course mate in Daffodil International University, who took part in this discuss while completing the coursework. Finally, we must acknowledge with due respect the constant support and patients so four parents.

ABSTRACT

This project based on NLP (Natural Languages Processing) techniques. The aim of this project is to identify the fake news of non-reputed news portal and to deliver the natural news or actual news to the news reader. As fake news misleads a great mishap among people so as a request of public demand we are going to run a model to detect all kinds of fake news. We collect data from reputed and non-reputed online news portal. This model build base on Bag of word (convert text format into vactorize format), tfidf matrix (extract facture) and RandomForestClassifier (for train and test dataset).By using this model we also create a predication where we test out site data is this data news fake or real. By using word tokenize we collect maximum number of keywords into our dataset for fake news and real news actually our model give result by comparing the dataset keywords. Our model achieves 86% accuracy.

TABLE OF CONTENTS

CONTENTS	Page
Board of examiners	I
Declaration	II
Acknowledgements	III
Abstract	IV
CHAPTER 1: INTRODUCTION	1-2
1.1 Introduction	1
1.2 Motivation	1
1.3 Research Question	2
1.4 Expected Output	2
1.5 Report Layout	2
CHAPTER 2: BACKGROUND	3-6
2.1 Introduction	3
2.2 Related Works	3
2.3 Research Summary	4
2.4 Scope of the Problem	5
2.5 Challenges	5
CHAPTER 3: RESEARCH METHODOLOGY	6-17
3.1 Introduction	6
3.2 Data Collection Procedure	7
3.3 Implementation Requirements	9

CHAPTER 4: EXPERIMENTAL RESULT AND DISCUSSION	18-21
4.1 Introduction	18
4.2 Experimental Result	18
4.3 Descriptive Analysis	20
4.4 Summary	20
CHAPTER 5: CONCLUSION AND FUTURE WORK	21-22
5.1 Summary	21
5.2 Conclusion	21
5.3 Future Work	22
REFERENCE	23
PLAGIARISM REPORT	24

LIST OF FIGURES

FIGURES	PAGE NO
Figure 3.1: Data preparation, train & testing flowchart	7
Figure 3.2: Data Collection Procedure	8
Figure 3.3: Sample data	10
Figure 3.4: Sample text to numeric format	12
Figure: 3.5 Random Forest tree	15
Figure 3.6: Confusion Matrix	16
Figure 3.7: Real News NLP Image	17
Figure 3.8: Fake News NLP Image	17
Figure 4.1: Real News Keywords Percentiles	19
Figure 4.2: Fake News Keywords Percentiles	19

LIST OF TABLES

TABLE	PAGE NO
Table 3.1: Data Category Table	11
Table 3.2: Textual data process using tf-idf	13
Table 3.3: World Could Analysis Table	16
Table 4.1: World Could Analysis Table	18
Table 5.1: Accuracy of Different Fake news Detection Model	21

CHAPTER 1

INTRODUCTION

1.1 Introduction

Now a day modern world provide different social media platforms, website because of upgrade our technology and from this site and platforms we get different news and also we consumes this news. Most of them people tend to seek out news from this site and platforms than traditional news organization. This traditional organization provide them real news but most of the time we get false news from different social platform and media. For this reason we can use machine learning techniques, natural language processing (NLP) to detect fake news. Through machine learning approach, we will learn machine which news is fake and which news is real. Then it can be capable to identify false news. And natural language processing is one of the most well-known fields that allows computer to process and manipulate human language. It is used for easy to learn and readable any language to machine. Fake news detection for bangle is more difficult and challenging than other language because of structure of bangle fake news is very confusing. Though some researchers are improve to detect fake news and use different method and algorithms but very few for bangle. There is still a lot of change to improve more.

1.2 Motivation

Fake is such a news which makes a humor among people. Sometimes people become too much frustrated by fake news. Mainly this type of news is like a social crime because it create a great mishap among the mass people. Nowadays it has become too difficult to identify news whether it is fake or real. For this reason we have research on some research model basis on fake news and afterword we have been able to build a model to detect any kind of news whether it I fake news or real news. Using our model one can be able to find news how much percent fake or how much the percent is real. Here many a keywords has been included into this system dataset to detect rapidly about the news fake or real. By this model it is hope that one can be able to remove his/her detection in using online news into his/her works or research. This must be helpful guide to anyone to ensure his doubt about news. In future this model would be more update to detect the online news fluffily. Finally it can be said that online news reader would be more benefited by using this model.

1.3 Research Question

What is the effect of a supervised system to improve the accuracy level of real and fake news detection? RandomForestClassifier is the supervised system? The purpose of this study is to find out the performance level of a supervised system in the area of news detection. Supervised system performs well in this field, where unsupervised system still has to improve.

- How collect fresh data and validated data.
- How store data and arrange data?
- How remove regular expiration?
- How remove data noise and html extra tag?
- How remove punctuation and stop words?

1.4 Expected Output

In this system news title, category, description and link will be submitted as input. Then using some build in function remove all type of noise and test the news using our model. This model gives result to compare the key word of fake news and real news and it stores some keyword for fake news and real news. It also stores this key word by training all fake news and real news dataset. So when any news test our model that is compared with the key words of stores dataset. When it finds more key word of fake news, it finally declares that “It’s a fake news” on the contrary when it finds the real news keyword it reveals that “It’s a real news” from fake news and real news dataset. Thus, detecting the keywords of fake news or real news our model gives a final decision or result whether the news is real or fake.

1.5 Report Layout

We divide this report into five sections. This is the first section where we talk about motivation for our work and the expected outcome. In the second section (CHAPTER2) We discuss about related works in this field, scope of the problems, challenges etc. In the third section (CHAPTER 3) we discuss about data collection procedure and implementation. Section four (CHAPTER 4) is for experimental result and analysis. Conclusion and future work are discussed in CHAPTER5.

CHAPTER 2

BACKGROUND

2.1 Introduction

By using the linguistic rule or stochastic rule or both, many fake news detection model was developed for different languages. In this part we discuss about those research paper which paper we used our reference.

2.2 Related Works

Distorted news and “alternate facts” were not a problem in society two years ago, despite the long-term deep changes in the news market [1]. The social concern about these kinds of news has been rather deeply accelerated by the term “fake news”, coined by the US elected President, Donald Trump, conveying its origins in the political arena. For example, among other fake news that emerged during the Trump campaign one of the most popular ones consisted on the Pope Francis reported endorsement of Donald Trump for president of the US. The news piece was advanced by the website “Ending The Fed”, managed by a Romanian youngster. BBC [4] also refers to the advancement of particular (often extreme) political causes as one of the main sources of fake news, defining them as false information deliberately circulated by those who have scant regard for the truth and act under the motivation of fostering political causes or obtaining revenue out of the online traffic. In this domain, Facebook has faced an increasing criticism over its role in the 2016 US presidential election because it allowed the propagation of fake news disguised as news stories coming from unchecked websites. This spreading of false information during the election cycle was so severe that Facebook was labelled as “dust cloud of nonsense.”[7] The fact is that the presidential election year has shown how the lines have blurred between facts and speculation, with people profiting off the spread of fake news. There were more than 100 news sites that made up pro-Trump content traced to Macedonia, according to a BuzzFeed News investigation⁸. Then again, reality checking approaches depend on computerized confirmation of recommendations made in the news articles [9] to survey the honesty of their cases [11]. Learning databases, for example, DBpedia 2 have been utilized to question the Web in an organized way. The consequences of such inquiries would then be able to be utilized to test whether unique sources additionally contain data affirming the news guarantee [15]. Different works have utilized interpersonal organization movement [10] on a particular news thing to evaluate its believability, for example by distinguishing tweets voicing wariness about the honesty of a case made in a news article [13]. Despite the fact that reality checking approaches are getting to be progressively ground-breaking, a noteworthy downside is that they are based on the reason that the data can be confirmed utilizing outside sources, for example FakeCheck.org and Snopes.com. In any case, this isn't a direct assignment, as outer sources probably won't be accessible, especially for simply distributed news things. Along these lines, the reality checking approach is prevalently helpful for the discovery of misdirection in writings for which outer, obvious data is accessible. Besides, likewise identified with the present paper is take a shot at the programmed ID of beguiling content, which has investigated spaces, for example, discussions, buyer audits sites, internet promoting, internet dating, and crowdfunding stages [12]. While counterfeit news recognition is firmly identified with trickery discovery [5], there are essential contrasts between the two errands. To start with, counterfeit news makers for the most part look for political or monetary profit just as self-

promotion while double crossers have inspirations that are all the more socially determined, for example, self-security, strife or on the other hand hurt shirking, impression the executives or personality covering. Second, they vary fundamentally in their objective and in the structure they spread: counterfeit news things are typically scattered at bigger scale through the Internet and online networking while trickery is all the more explicitly focused at people. Be that as it may, since the two undertakings manage tricky substance, we conjecture that there are phonetic perspectives that may be shared between these assignments. In this way, we center on the etymological methodology and expand upon a rising assemblage of research on PC robotized verbal trickiness recognition

2.3 Research Summary

A supervised system with KNN, Decision Tree, Naïve Bayes, Logistic Regression, SVM performs satisfactorily for detect fake news. But due to the huge dataset requirement, it will be very laborious to develop. Its performance level depends on the size of training data. Moreover, the neural network has done this job pretty well, but still hard to develop. Now it is the term for an unsupervised system. A limited work has done in this section for detection purposes, which performs fairly well. But all of them have a number of limitations; those make a scope for improvement. So, we develop a supervised detecting system, powered by and modified version of previously used algorithm [14]. In this system, each word is inspected for appropriate keywords. There are also some rules that help to make decisions to get an appropriate keyword.

2.4 Scope of the problem

The problem is the part of an experiment. There have a number of scopes for occurring problems.

- **Manually collect data:** Collect data without any software and facing problem to detect data which data is fake and which data is real and it was herculean task for us.
- **Arrange data category:** some data collect without category. In this data set category depending data title and description. Then all dataset arrange category wise. To arrange data into category wise
- **Detect Data:** All data seems look like same for this detect reason detect data is so default which is real and which is fake
- **Reducing noise data:** Reducing data noise use some build in function but some stop words are not reduce and this stop words is so effect for dataset. Because stop word key are common for real news data and fake news data.

2.5 Challenge

As this is the first time we are going to detecting the Bangla fake news on Bengali language so this may called a great challenge for us to this project. This is only reason is we didn't find any full research paper in detecting Bangla news. As the grammar of Bengali language is too different to English grammar so to identify the Bangla fake news we have to build a new model that can detect Bangla news easily. This task is not so easy like to detecting English news. Besides, collection of data for this project was so difficult to us because the fake and real news has no other exceptional identity to detect them. So, we have to research more to collect all types of data.

CHAPTER 3

RESEARCH METHODOLOGY

3.1 Introduction

The goal of this study is to find out the performance of a supervised system in the purpose of fake news detection. To successfully conduct the thesis below steps were taken.

- To build dataset collect data from different news portal reputed and non-reputed news portal.
- Arrange data into category wise and every data set one class.
- Prepare data for train and testing by using re, string and beautiful soup function (remove regular expiration and html extra tag).
- Data read by pd.read_csv() function and show sample data using data.head() function.
- Data.shape to show rows and columns and data.columns to show all dataset columns.
- Data.isnull().sum tho check missing data
- data['Class'].value_counts() to show real data and fake data of dataset.
- data['Category'].value_counts() to show all category and calculate all category.
- data.dropna(inplace=True) by using this function prepare data fully and if any data is missing then this function trace this field and make it as a true value.
- Then reducing nosing and normalization remove all type of noise.
- Target variable encoding to create data shape and partition data.
- By using bag of word convert text documents into corresponding numerical features.
Count Vectorizer: The most straightforward one, it counts the number of times a token shows up in the document and uses this value as its weight. Finally, an unsupervised Bangla POS tagger based on suffix analysis is proposed to increase the accuracy level.
- By using tf-idf convert numeric data into matrix format and Finally Construct training and testing sets then model train using Randomforestclassifier.

Flow Chart: All dataset preparation for training and testing flow chart

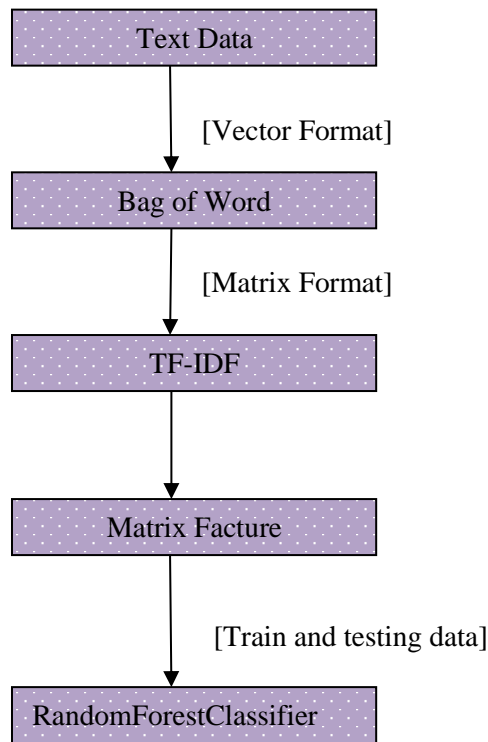


Figure 3.1: Data preparation, train & testing flowchart

Figure 3.1 we shown how to process data. In our process, at first we collect text data that contain fake and real news from different reputed and non-reputed news platform. Then this data convert into CSV form for readable to machine using panda's library of python. Then this text data convert into numeric format using bag of word process. In this process, we collect frequency of each word and we assign maximum feature and document frequency. From this method we get numeric format of text data. Then we calculate TF and IDF value. Bag of word calculated frequency in specific document. For this reason we calculate TF and IDF value. Because TF and IDF calculate frequency calculated all document. Then it convert into matrix and create train and test data. In our work, 80 news we use for train and 20 news we use for test. At we use RandomForestClassifier algorithms. It is a supervised algorithms it will classify fake and real news from train data. From this algorithms we get our expected result.

3.2 Data collection Procedure:

Our collection of data divided into five parts (Title, Category, Description, Link, Class) and classes divided into two parts (Real & Fake), Real class (256) Fake class (244) our total dataset is 500 in this data 80% training used and 20% testing used. Real class data collect Bangladeshi reputed online news portal site such as Daily Prothom alo, Bangladesh Pratidin, Ittefaq, Daily KalerKantho Daily NayaDiganta bdnews24.com etc.

The Fake class data collect some non-reputed online news source such as Dhaka Channel Khbor24.com etc. Afterwards, we check all sorts of data by Google scraping whether it is fake or real. For Google scraping we have to search by news title and made a result for that news and to define that how much the percent of this news is shown in fake news portal or real news portal. Then we made a Google form where we took the public opinion on all kinds of news to define how much people support it as fake news and how much people consider it as a real news. Depending all above description we made two classes of news, one is fake news class and other is real news. In figure 3.2 we shown some main topic and we show flowchart how we collect data and how to detect data. Figure 3.2 we also show which site give real data and which site give fake data for this reason this flowchart we divided tow part for news site one is reputed and another is non-reputed. For collecting data we also noticed news category that's means which published for joke and which news published for entitlement. Some time we get some news this news title is so attractive but when we read this news then we show anther things. So to detect news properly should be analysis all news and tis called read beyond.

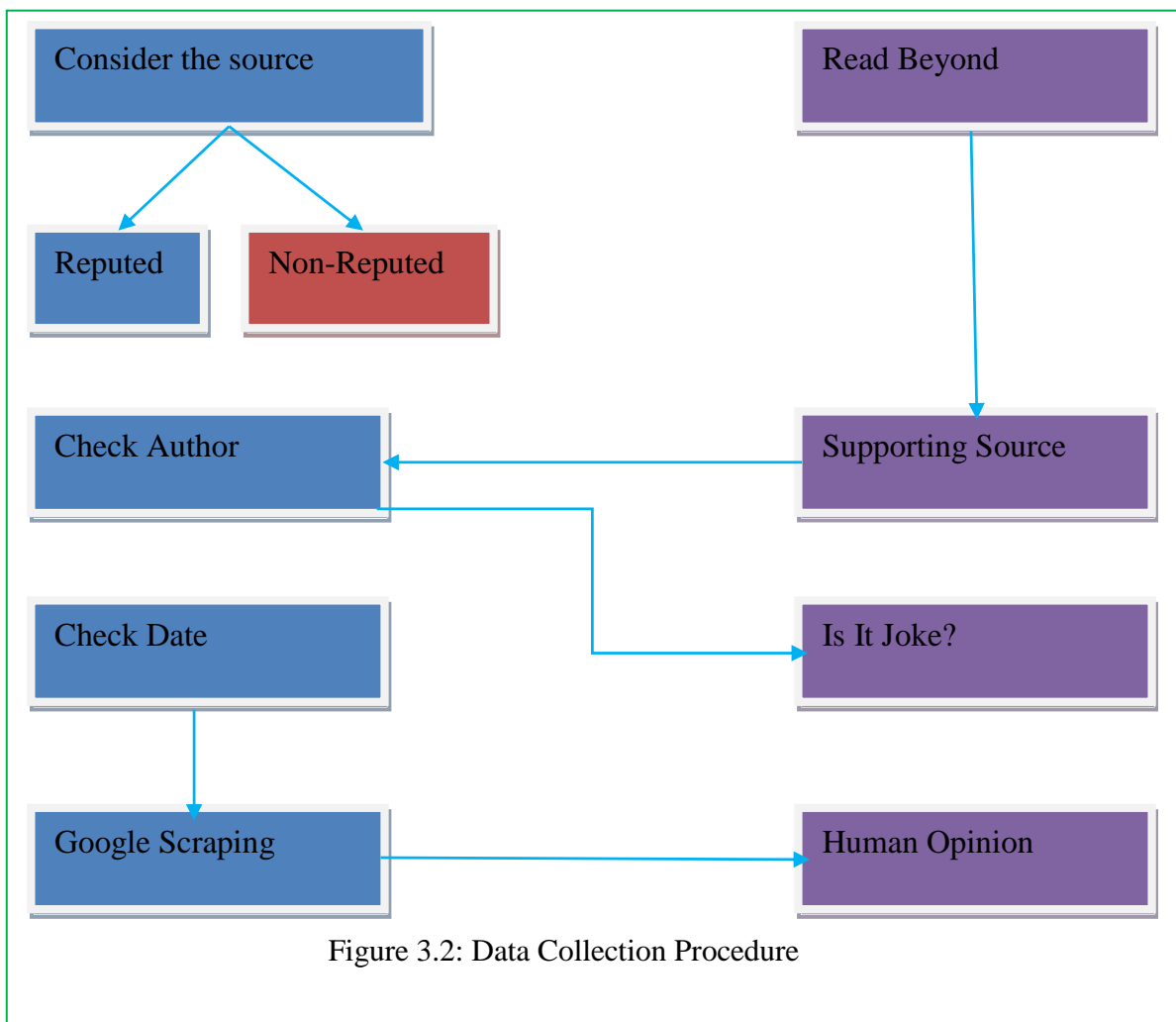


Figure 3.2: Data Collection Procedure

- **Consider the source:** To collect valeted data must be consider data source so we collect real data consider the reputed source and collect fake data from non-reputed source.
- **Read Beyond:** To collect data must be read full data only see headline not comment about full so collect valeted data must be read or check full potion of data
- **Check Author:** To identify data that`s the data is real or fake must be check author because some non-reputed author publish fake news and reputed author always publish real news so it`s so important to detect data and collect valeted data.
- **Supporting source:** Some reputed site always publish real news and this site are many famous to using this site we detect some non-reputed site news for this at first we collect the non-reputed news data then we search this data into Google and find that this data of news published the reputed site if this type of news published some reputed site then its data might be consider as a real data and it get better performance for collect valeted data
- **Check Date:** To collect news we focused about news date that`s means that news we collect those news publish which date
- It`s a joke: To collecting data we checked this data is type of joke or funny because when we arrange data into category wise then it`s be helpful to arrange data into category.
- **Google Scraping:** To collect valeted data we scrape Google by hold news title by use news title we search Google and find some result if this news is real then show some reputed news site and if this news is fake then show non-reputed site or only show one site because actually fake news published non-reputed site and if is news is fake then it published individual site.
- **Human Option:** To detect fake news and real news he create a Google from where we show our all dataset fake news and real news and we create 2 radio button for counting human option.

3.3 Implementation Requirement

Bangla is very inflectional language, where each word may have more than one meaning based on inflection. For detecting every word and remove noise we use some library function by using this library function we remove regular expression, html tag, punctuation. For removing this we use some requirement.

Requirement

- **Python:** Python is a programming language. For machine learning python language is the best language of the world for this reason we use python for implement our model.
- **Pandas:** Pandas is a library function of python language. We use pandas as pd for data manipulation and analysis.
- **Numpy:** Numpy is a library function of python language. We use numpy for provides a high performance multidimensional array and basic tools to compute with and manipulated these array.
- **Itertools:** The Python itertools module is a collection of tools for handling iterators. Simply put, iterators are data types that can be used in a for loop.
- **Matplotlib:** Matplotlib is a plotting library for the Python programming language and its numerical mathematics extension NumPy. It provides an object-oriented API for embedding plots into applications using general-purpose GUI toolkits like Tkinter, wxPython, Qt, or GTK+.
- **Sklearn:** Learning and predicting. ... In scikit-learn, an estimator for classification is a Python object that implements the methods fit(X, y) and predict (T). An example of an estimator is the class sklearn.svm. SVC, which implements support vector classification.

	Title	Category	Description	Link	Class
0	ডাকসুর ভিপি'র দায়িত্ব নিচ্ছেন নুরুল	রাজনীতি	ঢাকা বিশ্ববিদ্যালয় কেন্দ্রীয় ছাত্র সংসদে (ডাকসু...	www.prothomalo.com/bangladesh/article/1584785/...	Real
1	অস্ত্রোপচারের পর ভালো আছেন কাদের	রাজনীতি	বাংলাদেশ আওয়ামী লীগের সাধারণ সম্পাদক এবং সড়ক প...	www.prothomalo.com/bangladesh/article/1584632/...	Real
2	প্রধানমন্ত্রীর সাক্ষাৎ চান আন্দোলনরত শিক্ষকরা	রাজধানী	জাতীয় প্রেসক্লাবের সামনে দুদিন ধরে অবস্থান নিয়...	www.ittefaq.com.bd/capital/39030/প্রধানমন্ত্রী...	Real
3	অফিসকক্ষ বুঝে নিতে চিঠি, দায়িত্ব নিচ্ছেন ভিপি নুর	শিক্ষাঙ্গন	প্রায় তিন দশক পর ডাকসু নির্বাচনের দরজা খুলেও ...	www.ittefaq.com.bd/education/39020/অফিসকক্ষ-বু...	Real
4	ঐক্যফ্রন্ট থেকে নির্বাচিতরাও শপথ নেন: সুলতান...	রাজনীতি	ঐক্যফ্রন্ট থেকে বহিষ্কৃত ডাকসুর সাবেক ভিপি ও ছ...	www.ittefaq.com.bd/politics/39038/ঐক্যফ্রন্ট-থ...	Real

Figure 3.3: Sample data

Sample data Table: Figure 3.3 show dataset sample using shape.head () function. We arrange 500 dataset here we just show five dataset as a sample data. In our dataset we keep news title news category news description news link and class (class means which news is real and which news is fake here we declare tow class fake (for fake news) and real (for real news).

Table 3.1: Data Category Table

Category Name	Category Value
খেলাধুলা	86
রাজনীতি	68
বিনোদন	46
আন্তর্জাতিক	55
জাতীয়	73
শিক্ষাঙ্গন	11
বিজ্ঞান-প্রযুক্তি	29
অর্থনীতি	8
স্বাস্থ্য	5
তথ্যপ্রযুক্তি	2

Data Category: In table 3.1 we have shown all category of our dataset. Here we represent ten types of category and how much times each of category is present like রাজনীতি, খেলাধুলা, আন্তর্জাতিক, জাতীয় etc. To show this category table we call `data['Category'].value_counts()`.

Reducing noise on dataset: It depends how you characterize the "clamor" and how it is caused. Since you didn't give much data about your case, I'll accept your inquiry as "how to make the bend smooth". Kalman channel can do this, however it's excessively perplexing, I'd incline toward basic IIR channel.

Fake news and real has some mutual word, bracket, tag because of this item to detect fake news is very difficult .For this reason we use normalization function and import some library of python.

At first Beautiful Soup (`text.strip()`, "lxml") using this method we call for text script .its return returnsoup.get_text() through this method that give only text and remove others. And it will be store on soup. Next method returnre. Sub ("`\\[[^]]*\\]`", "", text) it is remove different bracket then return only text. At last through this method we get noise free text. After reducing noise data have fully prepared for training and testing without reducing noise data machine can't detect keywords for this reason data reducing is necessary for training and testing.

Bag of Words

Bag of words model is one of a series of techniques from a field of computer science known as Natural Language Processing or NLP to extract features from text. The way it does this is by counting the frequency of words in a document.

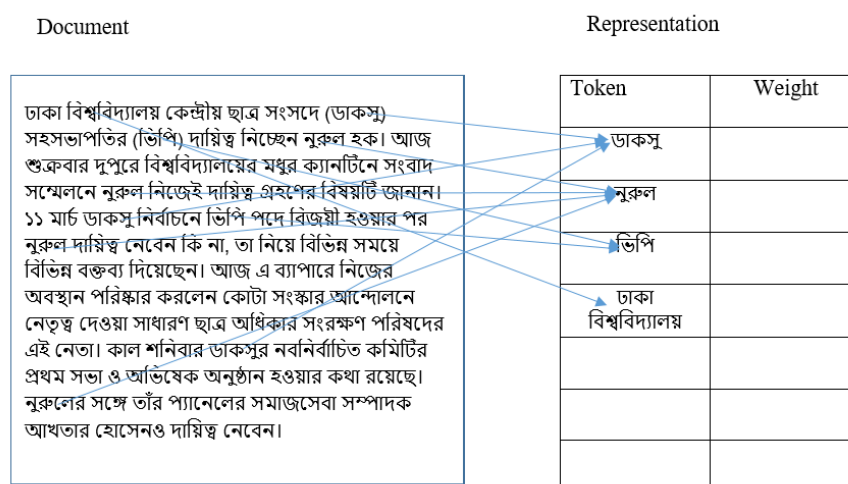


Figure: 3.4 Simple text numeric format

Mainly bag of word used for numeric format. In one single line same word occur many time. We can count their frequency that's called numeric format. In figure 3.4 ডাকসু occur 3 time from this text script, then নুরুল occur 4 time, ভিপি occur 2 time, ঢাকাবিশ্ববিদ্যালয় 1 time occur. Finally using of bag of word process, we get frequency of word. And we can simulate this text script and represent numeric format. In this process we import pickle library. This library used for vectorization. CountVectorizer(max_features=1500, min_df=5, max_df=0.7) this method we use for frequency define. Here df means that document frequency, max-df remove too much frequency. Here max df =.7 means that more than 70 percent frequency it will remove. And min frequency remove number of low frequency. Here min_df=5 that means less than 5 frequency it will remove. [6]

TF-IDF

The bag of words approach works fine for converting text to numbers. However, it has one drawback. It assigns a score to a word based on its occurrence in a particular document. It doesn't take into account the fact that the word might also be having a high frequency of occurrence in other documents as well. TFIDF resolves this issue by multiplying the term frequency of a word by the inverse document frequency. The TF stands for "Term Frequency" while IDF stands for "Inverse Document Frequency [6]".

The term frequency is calculated as:

$$\text{Term frequency} = X/Z \dots \dots \dots \text{(I)}$$

$X = \text{Number of Occurrences of a word}$

$Z = \text{Total words in the document}$

And the Inverse Document Frequency is calculated as:

$$\text{IDF} = [V/N] \dots \dots \dots \text{(II)}$$

$V = \text{Total number of documents}$

$N = \text{Number of documents contain in g the word}$

Table 3.2: Textual data process using tf-idf

WORD	TF	IDF
The	1/7	Log (2/2) = 0
Car	1/7	Log (2/1) = 0.3
Truck	0	Log (2/1) = 0.3
Is	1/7	Log (2/2) = 0
Driven	1/7	Log (2/2) = 0
On	1/7	Log (2/2) = 0
The	1/7	Log (2/2) = 0
Road	1/7	Log (2/1) = 0.3
Highway	1/7	Log (2/1) = 0.3

In table 3.2, sentence the car truck is driven on the road highway from this sentence we calculate TF and IDF. At first ‘The’ value calculated TF=1/7 that means in this sentence ‘The’ occur 1 time and here total number of word 7. So TF gives result 1/7. And TDF calculated log (2/2) that means here total number of document =2 and word containing number of document=2. Every word calculated in same way. In this method we get tf and idf value then it is convert array through this method and using this method we create our dataset to apply this method we use `tfidfconverter.fit_transform(X).toarray ()`.

Classification Algorithm

- Logistic Regression
- Decision trees
- Support vector machine (SVM)
- Naïve Bayes
- Random forest
- Linear regression
- Polynomial regression
- SVM for regression

All classification and regression algorithm come under supervised learning so we can use any types of above algorithm. We use above all algorithm but we get better result for Random Forest Classifier algorithm so we use random forest for train and testing dataset.

Random Forest Classifier:

Precondition: A training set $X := (a_1, b_1), \dots, (a_n, b_n)$, features E , and number of trees in forest T .

```
1 function RandomForest(X, T)
2  $G \leftarrow \emptyset$ 
3 for  $i \in 1, \dots, T$  do
4  $X(i) \leftarrow$  A bootstrap sample from  $X$ 
5  $rt \leftarrow$  RandomizedTreeLearn( $X(i)$ ,  $E$ )
6  $G \leftarrow G \cup \{rt\}$ 
7 end for
8 return  $G$ 
9 end function

10 function RandomizedTreeLearn( $X$ ,  $E$ )
11 At each node:
```

12 $e \leftarrow$ very small subset of E

13 Split on best feature in E

14 return the learned tree

15 end function

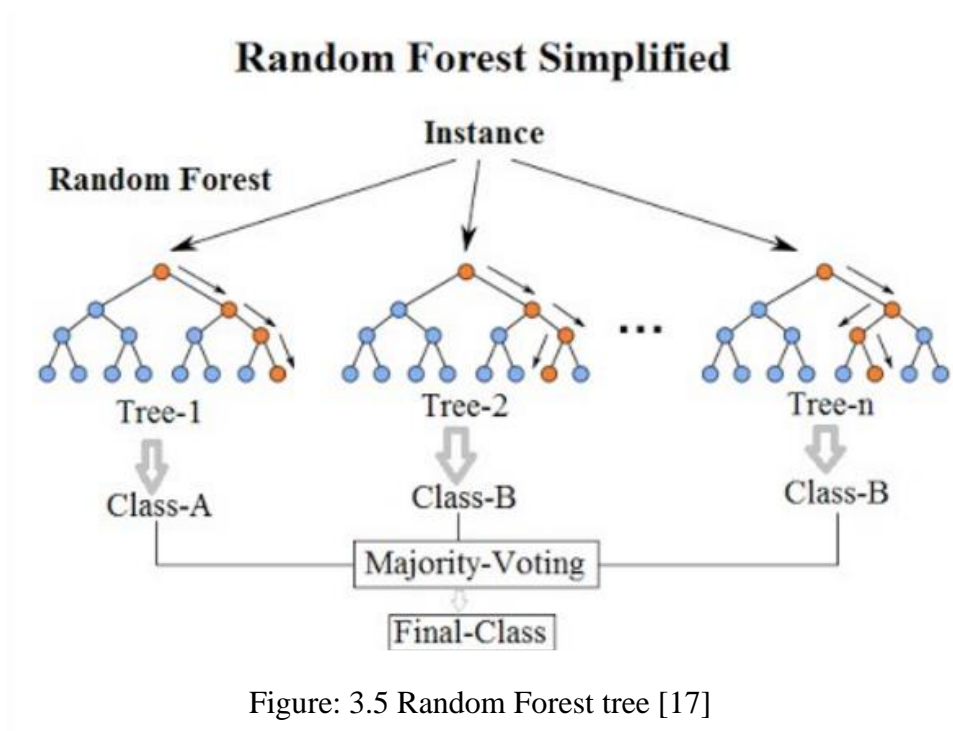


Figure: 3.5 Random Forest tree [17]

Figure 3.5 we shown random forest classifier tree and above we define random forest algorithm pseudocode. Figure 3.5 we show many tree (tree-n that's means here belong many trees) actually random forest algorithm create many tree using data its called vectorisiton. This algorithm create mane vector tree and we know that which algorithm is best that algorithm create many tree. Figure 3.5 we show some tree and every tee gabe a class such that class A, class B, etc. At last we get final class using this to majority and voting those above class. And in pseudocode, at first we represented our training set, variable (a, b) that contain fake news and real news. And total number of tree in forest T that contain 500 news. Firstly we assign a variable G and store null value. Then we start loop.it will continue until T that means it travers 500 news. Then variable rt is storing randomized news and feature of news. Then its value and G's mutual value store on variable G and return this value. Then small subset of feature such as real of fake news will be store on variable (e). Finaly from this training set function will return specific feature of tree. That means which new is fake and which news is real.at last end of the function.

Confusion Matrix: Figure 3.6 shown confusion matrix represent true positive and negative value. Here total fake news is 37 and total real news is 49. Real and fake news combination is 12 and 2. Here true positive value represent real news and false positive value represent total fake news. And false positive and false negative value represent fake and real news combination. From this confusion matrix, we get accuracy sum of TP and TN divided by sum of positive and negative.

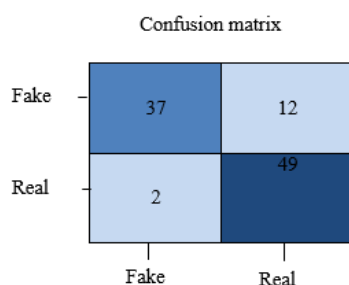


Figure 3.6: Confusion Matrix

Table 3.3: Word Cloud Analysis Table

Step	Real	Fake
word tokenize	5826	11025
Keyword	20	19

Word Cloud Analysis table: Table 3.3 shown our machine able to collect how much token for fake news and real news. Our machine collect 5826 token for real news and 11025 token for fake news. And our Machen collect some keywords (keywords is maximum number of token which token is store into our machine) for real news collect 20 kay words and for fake news collect 19 key words . Using this key words our machine draw two images for real news nlp (show figure 3.3.7) and fake news nlp(show figure 3.3.8)

NLP Images: Here we show the dataset out put keyworeds drawing image. Draw this image we use word could analysis library function. At first we collect keywords using word_tokenize function then arrange keywords their parity using sorted function and at draw image use plt.figure, plt.imshow, plt.axis. plt.tight_layout function and showing this image use plt.show() function.

CHAPTER 4

EXPERIMENTAL RESULT AND DISCUSSION

4.1 Introduction

In this part we discuss about our experimental result and show accuracy table of our model. In this part we also show the key words bar chart and discuss the model descriptive analysis and the summary.

4.2 Experimental results

The efficiency of a system can be measured from its accuracy level. Our proposed algorithm is applied to the testing dataset, which is collected from different popular online newspaper. There are almost 100 dataset available in the testing dataset. Accuracy is measured from the ratio of the number of correctly data the total number of data. Our system can detect 86 news out of 100 news. Our system micro average precision 0.86, recall 0.86, fl-score 0.86, support 100 and macro average precision 0.88, recall 0.86, fl-score 0.86, support 100 and weighted average precision 0.87, recall 0.86, fl-score 0.86 support 100. Final recall and fl-score 0.86 it means, our system obtains 86% accuracy, which is not a bad figure. Our system detects Fake news and Real News. The result is shown in Table 4.1

Table 4.1: Experiment Result

Total Dataset	Train	Test	Accuracy %
500	80%	20%	86%

Experiential Result: Table 4.1 shown the accuracy level of our model. This model will collect some keywords by train all dataset the number of fake news keywords is 5026 and real news is 11026. There are two bar chart will be shown below on the basics of these keywords presenters.

Keywords Chart: For fake news and real news our system collect some keywords. For fake news system detect some keywords and calculate this keywords that's means which key word in more for fake news and real news system sore number of keyword and using this keywords system detect fake news and real news. In below figure 4.1 and 4.2 we shown real news and fake news keyword bar chart.

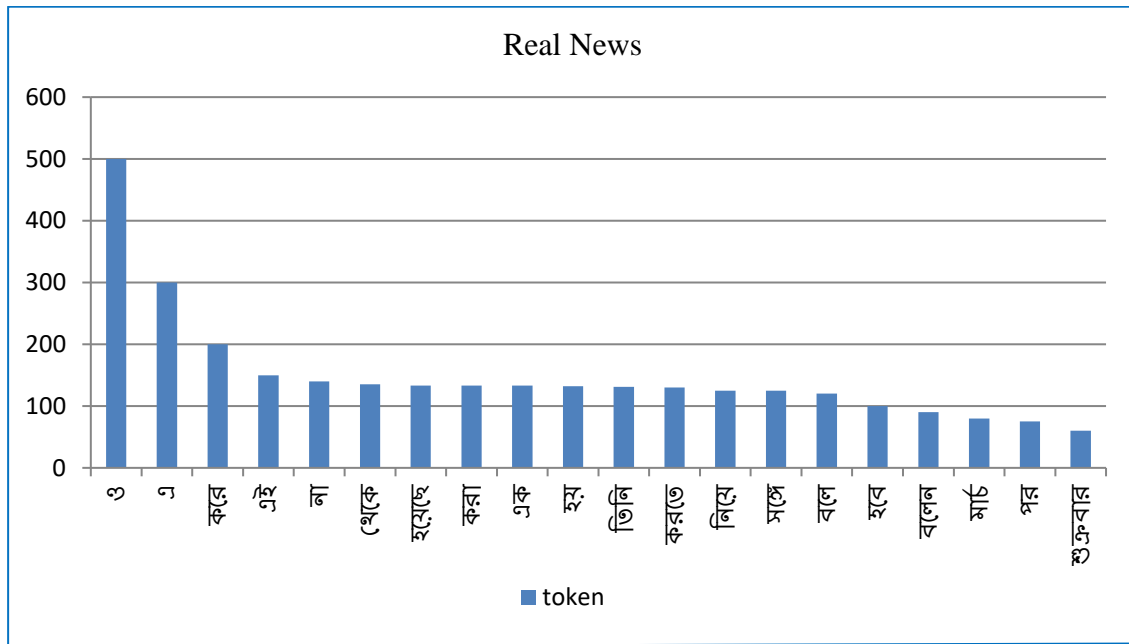


Figure 4.1: Real News Keywords Percentiles

Real News Key Words Chart: Figure 4.1 shown about the keywords of real news. This keyword collect our machine as a percentiles for this using percentiles we dare this bar chart her. For real news we total collect 5026 keywords. For that ও keyword get 500 times, এ (300), করে, (200), এই(150) those for 20 keyword draw bar chart.

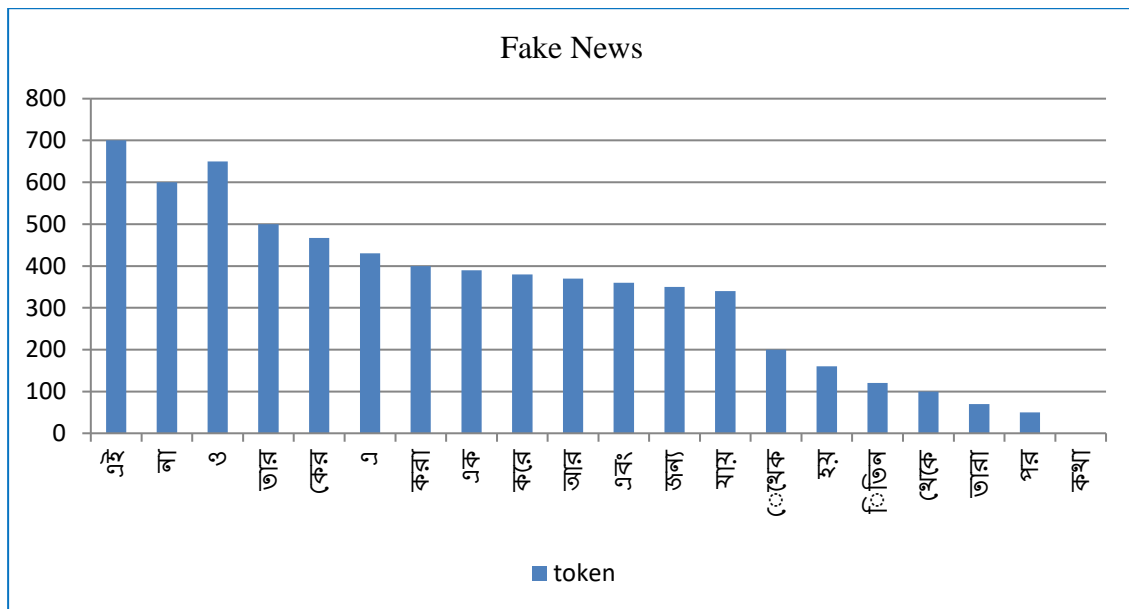


Figure 4.2: Fake News Keywords Percentiles

Fake News Keywords Chart: Figure 4.2 shown about the keywords of fake news. This keyword collect our machine as a percentiles for this using percentiles we dare this bar chart her.

For real news we total collect 11025 keywords. For that ও keyword get 500 times, এই (700), না (600), তার(500) thus for 10 keyword draw bar chart.

4.3 Descriptive Analysis

By analyzing the result, we identify some constraints. Those are mentioned below-Here Firstly, we collect some news without category for this news we set category to depended news title and description for this reason some news category are not properly set. A news result depended her category so when any news category is not set properly for this reason model get different result that's means if nay news category change and test this news then those news result could be change

Secondly, this model manly works news description. The collect keywords from description and compare ta model stories key words. So when any news are testing and give some potion or news then it get one result and if give maximum potion or full potion then it's could get another result, many grammatical rules are not applicable to them.

4.4 Summary

After result experiment it is seen that the model declares result on the basis of description and category. To show a result the model search category wise keyword from description to show a final result. To find the key words, the model compare with the stored keywords which is included into the dataset. After that it detect the keyword of fake news and real news portal into the input news portal and then it fixed a decision on the basis of similarity of dissimilarity of the stored keyword sand finally declare that the news is fake or the news is real news.

CHAPTER 5

CONCLUSION AND FUTURE WORK

5.1 Summary

After result experiment it is seen that the model declares result on the basis of description and category. To show a result the model search category wise keyword from description to show a final result. To find the key words, the model compare with the stored keywords which is included into the dataset. After that it detect the keyword of fake news and real news portal into the input news portal and then it fixed a decision on the basis of similarity of dissimilarity of the stored keyword sand finally declare that the news is fake or the news is real news.

Table 5.1: Accuracy of Different Fake news Detection Model

Model	Rada Mihalcea [11]	Hadeer Ahmed [6]	Kia Shu [1]	Our System
Accuracy %	76%	92%	80%	86%

Comparing Accuracy: Table 5.1 shown some research paper model accuracy and compare those model accuracy to our model. All model build English language fake news but we are the first team who word bangle fake news detection. And we got 86% accuracy it's a great archive for our team.

5.2 Conclusion

We have been able to build our model successfully. This model is now active to identify fake news and real news. To build this model about 500 length dataset has been made in which 256 is real and 244 is fake data. Then this raw data was turned into numeric format using bag of words, tfidf matrix has been used to transfer numeric data into matrix feature. This matrix feature has been trained by using RandomForestClassifier. Of the total data 80% is for train and rest for test. After testing the rest of 20% data we got 86% accuracy. So this model accuracy is about 86% which is more than others research work. As this is the first research on Bengali fake news detection, so it can be said that the accuracy by this model has got is like a successful work. Our model have some limitation such that our model get result to calculate data keywords for this our model get some keyword for real data and fake data. And arrange this data into category wise for this reason when we test out site news or our model then if just change news category then our model could get different result.

5.3 Future Work

In future the aim of this model to create a database where all sorts of news keyword will be stored into database category wise. A news alarm will be include in this model to define clearly about applied news. Nevertheless, it will show how much people consider this model as fake and how much people is on the behalf of this news is real. It will also logically define by comparing or showing the client with strong online news portal who has already submitted this news on their webpage.

REFERENCES

- [1] Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, Huan Liu, "Fake News Net: A Data Repository with News Content, Social Context and Spatial temporal Information for Studying Fake News on Social Media" The 9th International Conference on Emerging Ubiquitous Systems and Pervasive Networks (Submitted on 5 Sep 2018 (v1), last revised 27 Mar 2019 (this version, v3)).
- [2] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, Huan Liu, "Fake News Detection on Social Media A Data Mining Perspective" ACM SIGKDD Explorations Newsletter, Volume 19 Issue 1, June 2017.
- [3] Naman Singh ; Tushar Sharma ; Abha Thakral ; Tanupriya Choudhury "Detection of Fake Profile in Online Social Networks Using Machine Learning" 2018 International Conference on Advances in Computing and Communication Engineering (ICACCE)
- [4] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, RadaMihalcea, "Automatic Detection of Fake News", Computation and Language (cs.CL) Submitted on 23 Aug 2017.
- [5] Benjamin Riedel, Isabelle Augenstein, Georgios P. Spithourakis, Sebastian Riedel, "A simple but tough-to-beat baseline for the Fake News Challenge stance detection task" (Submitted on 11 Jul 2017 ([v1](#)), last revised 21 May 2018 (this version, v2)).
- [6] Hadeer Ahmed, authorIss, TraoreSherifSaad, "Detection of Online Fake News Using N-Gram Analysis and Machine Learning Techniques, Conference paper, First Online: 11 October 2017.
- [7] Minyoung Huh, Andrew Liu, Andrew Owens, Alexei A. Efros, "Fighting Fake News Image Splice Detection via Learned Self-Consistency" Computer Vision and Pattern Recognition (cs.CV).
- [8] Benjamin D. Horne, SibelAdali, "Fake News Packs a Lot in Title, Uses Simpler, Repetitive Content in Text Body, More Similar to Satire than Real News" Published at The 2nd International Workshop on News and Public Opinion at ICWSM.
- [9] Emerson F.Cardoso Renato, M.SilvaTiago, A.Almeida, "Towards automatic filtering of fake reviews" Volume 309, 2 October 2018.
- [10] W. KenRedekop, "Fake news, big data, and the opportunities and threats of targeted actions" [Health Policy and Technology](#), 7(2), 113-114, 2018.
- [11] Veronica P ´ erez-Rosas ´ 1 , Bennett Kleinberg2 , Alexandra Lefevre1 Rada Mihalcea1, "Automatic Detection of Fake News" University of Michigan 2Department of Psychology, University of Amsterdam.
- [12] A.Peters, E.Tartari, N.Lotfinej, P.Parneix, D.Pittet, "Fighting the good fight: the fallout of fake news in infection prevention and why context matters" 2018 Published by Elsevier Ltd on behalf of The Healthcare Infection Society.
- [13] S.MoJangPhD, TiemingGeng, Jo-YunQueenie, LiaRuofanXia, Chin-TserHuangPhD, HwalbinKimPhD, JijunTangPhDb, "A computational approach for examining the roots and spreading patterns of fake news: Evolution tree analysis" 2018 Elsevier Ltd.
- [14]Mauridhi Hery Purnomo, Surya Sumpeno, Esther IrawatiSetiawan, DianaPurwitasaria, "Keynote Speaker II: Biomedical Engineering Research in the Social Network" Analysis Era: Stance Classification for Analysis of Hoax Medical News in Social Media", 2017 Published by Elsevier B.
- [15] Monther Aldwairi, Ali Alwahedi, "Detecting Fake News in Social Media Networks" [Volume 141](#), 2018, Pages 215-222.
- [16] Avaro Figueira, Luciana Oliveira, "The current state of fake news: challenges and opportunities" CENTERIS / ProjMAN / HCist 2017, 8-10 November 2017, Barcelona, Spain [Volume 121](#), Pages 817-825.
- [17] Sholk Gilda "Evaluating machine learning algorithms for fake news detection" 2017 IEEE 15th Student Conference on Research and Development (SCORED).

Turnitin Originality Report

Processed on: 01-Apr-2019 23:21 +06
 ID: 1103902488
 Word Count: 6493
 Submitted: 1

Similarity Index

20%

Similarity by Source

Internet Sources: 18%
 Publications: 5%
 Student Papers: 12%

**FAKE NEWS DETECTION
 USING MACHINE LEARNING** By
 Aditi Balo

4% match (Internet from 24-Oct-2018)

<http://web.eecs.umich.edu/~mihalcea/papers/perezrosas.coling18.pdf>

2% match (Internet from 03-Feb-2019)

<https://pyplanet.herokuapp.com/pyplanet/1224/>

2% match (student papers from 05-Apr-2018)

[Submitted to Daffodil International University on 2018-04-05](#)

2% match (student papers from 10-Apr-2018)

[Submitted to Daffodil International University on 2018-04-10](#)

1% match (student papers from 12-Apr-2018)

[Submitted to Daffodil International University on 2018-04-12](#)

1% match (student papers from 07-Apr-2018)

[Submitted to Daffodil International University on 2018-04-07](#)

1% match (student papers from 03-Apr-2018)

[Submitted to Daffodil International University on 2018-04-03](#)

1% match (student papers from 20-Jun-2017)

[Submitted to Bridgepoint Education on 2017-06-20](#)

< 1% match (Internet from 21-Sep-2018)

<http://dspace.library.daffodilvarsity.edu.bd:8080/bitstream/handle/20.500.11948/2730/152-15-5887.pdf?isAllowed=y&sequence=1>

< 1% match (Internet from 14-Jan-2019)

http://www.public.asu.edu/~skai2/papers/wsdm_fake_news_tutorial.pdf

< 1% match (Internet from 01-Apr-2019)

<https://arxiv.org/abs/1707.03264>

< 1% match (Internet from 16-Dec-2018)

<http://resits.its.ac.id/expert/details/594>

< 1% match (Internet from 21-Jul-2018)

<https://dblp.uni-trier.de/search?q=external+feature>

< 1% match (Internet from 03-Feb-2019)