**ARTIFICIAL INTELLIGENCE LAB [CSL-411]**

**Project Name: Fake News Detection System**

**SEMESTER PROJECT**

|  |  |  |  |
| --- | --- | --- | --- |
| **Sr.no** | **Name** | **Enrolment** | **Semester** |
| **01** | **Muhammad Awais** | **02-134201-038** | **6-A** |
| **02** | **Imaan Shahid** | **02-134201-009** | **6-A** |
| **03** | **Manahil Shamim** | **02-134201-055** | **6-A** |

|  |  |
| --- | --- |
| **Name** | **Designation** |
| **Sir Tariq Siddique** | **Course Instructor** |
| **Ms. Salas Akbar** | **Lab Engineer** |

**Acknowledgement**

I would like to express my special thanks of gratitude to our instructor, Sir. Tariq Siddique as well as our lab instructor, Ms. Salas Akbar, who gave us the golden opportunity to do this wonderful project on the topic **Fake news detection**, which also helped us in doing a lot of Research and we came to know about so many new things, we are really thankful to them.

**Contents**

**1. Chapter 1**

**1.1. Problem Statement …………………………………………… 4**

**2. Chapter 2**

**2.2. Literature Review ……………………………………………... 4**

**3. Chapter 3**

**3.1. Methodology …………………………………………………… 6**

**4. Chapter 4**

**4.1. Code Snippet …………………………………………………. 8**

**5. Chapter 5**

**5.1. Conclusion & Comparative Analysis…………….……. 14**

**6. References ……………………………………………………….... 15**

**1. Chapter 1**

### 1.1. Problem Statement

To come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information and the ability to classify the news as fake or real and check the authenticity of the website publishing the news.

We aim to provide the user with the ability to classify the news as fake or real and check the authenticity of the website publishing the news.

1. **Chapter 2**

**2.1. Literature Review**

Recently, fake news has been incurring many problems to our society. As a result, many researchers have been working on identifying fake news. Most of the fake news detection systems utilize the linguistic feature of the news. However, they have difficulty in sensing highly ambiguous fake news which can be detected only after identifying meaning and latest related information. Fake news and hoaxes have been there since before the advent of the Internet. The widely accepted definition of Internet fake news is fictitious articles deliberately fabricated to deceive readers”. Social media and news outlets publish fake news to increase readership or as part of psychological warfare. In general, the goal is profiting through clickbaits. Clickbaits lure users and entice curiosity with flashy headlines or designs to click links to increase advertisements revenues. This exposition analyzes the prevalence of fake news considering the advances in communication made possible by the emergence of social networking sites. The purpose of the work is to come up with a solution that can be utilized by users to detect and filter out sites containing false and misleading information. We use simple and carefully selected features of the title and post to accurately identify fake posts.

The aim of the system testing process was to determine all defects in our project. The program was subjected to a set of test inputs and various observations were made and based on these observations it will be decided whether the program behaves as expected or not.

Diagram

Description automatically generated

**3. Chapter 3**

### 3.1. Methodology

Studying and training the different models with 4 different classifiers and choose the best classifier for final execution.

### Implementation

We can get online news from different sources like social media websites, search engine, homepage of news agency websites or the fact-checking websites. On the Internet, there are a few publicly available datasets for Fake news classification like Buzzfeed News, LIAR [15], BS Detector etc. These datasets have been widely used in different research papers for determining the veracity of news.

Then the dataset can be preprocessed, cleaned, tokenized, punctuation removal, vectorized etc.

### Algorithms used

Different classifiers can be investigated to predict the class of the text. Four different machine learning algorithms have been used to predict which one has the best accuracy – Multinomial Naïve Bayes Passive Aggressive Classifier and Logistic regression.

The implementations of these classifiers are done using Python library Sci-Kit Learn.

### Brief introduction to the algorithms

**1. Naïve Bayes Classifier:**

This classification technique is based on Bayes theorem, which assumes that the presence of a particular feature in a class is independent of the presence of any other feature. It provides way for calculating the posterior probability.

A picture containing text

Description automatically generated

### 2. Random Forest

The random forest is a classification algorithm consisting of many decisions trees. It uses bagging and feature randomness when building each individual tree to try to create an uncorrelated forest of trees whose prediction by committee is more accurate than that of any individual tree.

### 3. Logistic Regression

It is a classification not a regression algorithm. It is used to estimate discrete values (Binary values like 0/1, yes/no, true/false) based on given set of independent variable(s). In simple words, it predicts the probability of occurrence of an event by fitting data to a logit function. Hence, it is also known as logit regression.

### 4. SVC

SVC, or Support Vector Classifier, is a supervised machine learning algorithm typically used for classification tasks. SVC works by mapping data points to a high-dimensional space and then finding the optimal hyperplane that divides the data into two classes.

### Evaluation Matrices

Evaluate the performance of algorithms for fake news detection problem; various evaluation metrics have. These metrics are commonly used in the machine learning community and enable us to evaluate the performance of a classifier from different perspectives. Specifically, accuracy measures the similarity between predicted fake news and real fake news.

**4. Chapter 4**

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Graphical user interface, text, application, email

Description automatically generated

Text

Description automatically generated

Graphical user interface, text, application

Description automatically generated

Graphical user interface, text

Description automatically generated

# 5. Chapter 5

## 5.1. Conclusion and Comparative Analysis

All the algorithms that we trained gave accuracy above 0.95. Thus, on our dataset all the algorithms performed well but to compare the accuracy scores the best was linear SVM with 0.999 and the least accuracy was Multinomial naïve bayes 0.95. Logistic regression performed well as 0.990 and Random Forest accuracy was 0.996. According to research as well Linear SVM is the best algorithm considered for text classification problems because SVM is more powerful to address non-linear classification tasks. SVM generalizes well in high dimensional spaces like those corresponding to texts. It is effective with more dimensions than samples. It works well when classes are well separated. And in our case too SVM gave the best accuracy out of all. The next best was Random Forest, Random Forest (RF) is one of the best classifiers widely used for regression and classification tasks. Algorithmic simplicity makes it an attractive choice for text classification. In addition, its capability to handle high dimensional data and high performance under imbalanced datasets are significant advantages over other machine learning models. However, in our case the dataset was balanced so not all benefits of random forest algorithm were utilized but again it handles high dimensional data such as text well. Logistic regression work fine for text classification but it isn’t always the best approach because Logistic regression models tend to overfit the data, particularly in high-dimensional settings (which is the clever way of saying cases with lots of predictors). For this reason, it’s common to use some kind of regularization method to prevent the model from fitting too closely to the training data. Naïve bayes gave relatively less accuracy compared to other examples Naïve Bayes classifier is based on the hypothesis that each attribute is mutually independent; thus, it is widely used for its easiness and high efficiency. But because of the text redundant features and rough parameter estimation, the performance of Naïve Bayes in text classification tasks is not good.

**6. Chapter 6**

# 6.1. References

<http://www.cs.utoronto.ca/~strider/publications/Chapter9.pdf>

[https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detectionalgorithms-36d53571365e](https://towardsdatascience.com/r-cnn-fast-r-cnn-faster-r-cnn-yolo-object-detection-algorithms-36d53571365e)

[https://www.researchgate.net/publication/334850808\_Improvement\_of\_MaskRCNN\_Object\_Segmentation\_Algorithm](https://www.researchgate.net/publication/334850808_Improvement_of_Mask-RCNN_Object_Segmentation_Algorithm) <https://towardsdatascience.com/image-segmentation-part-1-9f3db1ac1c50>[https://www.analyticsvidhya.com/blog/2019/07/computer-vision-implementing-maskr-cnn-image-segmentation/](https://www.analyticsvidhya.com/blog/2019/07/computer-vision-implementing-mask-r-cnn-image-segmentation/)

☺