Green University of Bangladesh
Department of Computer Science and Engineering (CSE)
Faculty of Sciences and Engineering
Semester: (Spring, Year:2021), B.Sc. in CSE (Day/Eve)

Course Title: Pattern Recognition Lab
Course Code: CSE 422 Section: DA

Lab Project Name: _ _ _ _ _ _ _ _ Fake news detection _ _ _ _ _ _ _ _ _ _ _ _ _ _ _ _

Student Details

	Name	ID
1.	Sium Hossain	181002033
2.	Salauddin	181002034
3.	Sumaiya akhter	181002174
4.	Md. Alamgir Hossen	181002049

Submission Date : _ _ _ _ _ _ _ _ _ _ 11-Sep-20121 _ _ _ _ _ _ _ _ _ _ _ _
Course Teacher's Name : _ _ _ _ _ _ _ Md. Mamunur Rahman_ _ _ _ _ _ _ _

[For Teachers use only: Don't Write Anything inside this box]

Lab Project Status

Marks: …… **Signature: .....**

Comments: ..... **Date: .....**

Table of Contents

Chapter 1

Introduction

1.1 Introduction

Fake news is false or misleading information presented as news. It often has the aim of damaging the reputation of a person or entity, or making money through advertising revenue.So Detection of fake news online is important in today's society as fresh news content is rapidly being produced as a result of the abundance of available technology.There are several methods to identify fake news.Linguistic feature extraction is one of them.And this approach is detect fake news by catching the information manipulators in the writing style of the news content.

1.2 Design Goals/Objective

The aim is not only to detect fake news, but to also achieve the highest possible accuracy levels in the detection.The purpose of the work is to come up with a solution that can be utilized by users to detect and filter out containing false and misleading information.Our project is based on supervised learning technique.Supervised learning, also known as supervised machine learning, is a subcategory of machine learning and artificial intelligence. It is defined by its use of labeled datasets to train algorithms that to classify data or predict outcomes accurately.Once the relevant attributes are selected after the data cleaning and exploration phase, the next step involves extraction of the linguistic features. Linguistic features involve certain textual characteristics converted into a numerical form such that they can be used as an input for the training models.The experimental results show a 97-99% accuracy using different kinds of classifiers.And also we have a option for manual testing and and we can test manually by input news headline.Then classifier algorithm will give us boolean output whether the news is true or false.

Chapter 2

Design/Development/Implementation of the Project

2.1 Section (Data preprocessing)

At first we collected two datasets (true_news.csv & fake_news.csv) for training,testing and evaluating performance of our selected classifier algorithm.

Inserting fake and real dataset

```
df_fake = pd.read_csv("Fake.csv")
df_true = pd.read_csv("True.csv")
```

Then we add a class in two dataset which was identified as true and false news.In false news dataset we put 0 value and 1 for true news.Here is example:

False_news.csv:

	text	subject	date	class
	21st Century Wire says This week, the historic...	Middle-east	January 20, 2016	0
	By Dady Chery and Gilbert MercierAll writers ...	Middle-east	January 19, 2016	0
	Vic Bishop Waking TimesOur reality is carefull...	Middle-east	January 19, 2016	0
	Paul Craig RobertsIn the last years of the 20t...	Middle-east	January 19, 2016	0
	Robert Fantina CounterpunchAlthough the United...	Middle-east	January 18, 2016	0

True_news.csv:

	text	subject	date	class
	SAO PAULO (Reuters) - Cesar Mata Pires, the ow...	worldnews	August 22, 2017	1
	GENEVA (Reuters) - North Korea and the United ...	worldnews	August 22, 2017	1
	GENEVA (Reuters) - North Korea and the United ...	worldnews	August 22, 2017	1
	COPENHAGEN (Reuters) - Danish police said on T...	worldnews	August 22, 2017	1
	UNITED NATIONS (Reuters) - Two North Korean sh...	worldnews	August 21, 2017	1
	BRUSSELS (Reuters) - NATO allies on Tuesday we...	worldnews	August 22, 2017	1
	LONDON (Reuters) - LexisNexis, a provider of l...	worldnews	August 22, 2017	1
	MINSK (Reuters) - In the shadow of disused Sov...	worldnews	August 22, 2017	1
	MOSCOW (Reuters) - Vatican Secretary of State ...	worldnews	August 22, 2017	1
	JAKARTA (Reuters) - Indonesia will buy 11 Sukh...	worldnews	August 22, 2017	1

Then we merge two dataset into a single dataset name as manual.csv for our further classifier work.Then we drop unnecessary columns (subject,date,index_number) which are not needed for the classifier.

Finally for more optimization,we declared a function which is take care of making uppercase letter to lowercase,special character (@,$,*,& etc)

```
def wordopt(text):
    text = text.lower()
    text = re.sub('\[.*?\]', '', text)
    text = re.sub("\\W"," ",text)
    text = re.sub('https?://\S+|www\.\S+', '', text)
    text = re.sub('<.*?>+', '', text)
    text = re.sub('[%s]' % re.escape(string.punctuation), '', text)
    text = re.sub('\n', '', text)
    text = re.sub('\w*\d\w*', '', text)
    return text
```

```
df["text"] = df["text"].apply(wordopt)
```

2.2 Section (Splitting the dataset into training set and testing set)

After splitting the dataset into training and testing,we converted text to vectors.Vectorization or word embedding is the process of converting text data to numerical vectors. Later those vectors are used to build various machine learning models.

```
x_train, x_test, y_train, y_test = train_test_split(x, y, test_size=0.25)
```

Convert text to vectors

```
from sklearn.feature_extraction.text import TfidfVectorizer
```

```
vectorization = TfidfVectorizer()
xv_train = vectorization.fit_transform(x_train)
xv_test = vectorization.transform(x_test)
```

We used 75% data for training and 25% data for testing purposes.

2.3 Section (Classifier)

We used four different Classification Algorithms for making differentiate between fake news and true news which are :

- **Logistic Regression**
  In statistics, the logistic model (or logit model) is used to model the probability of a certain class or event.
- **Decision Tree Classification**
  Decision tree learning or induction of decision trees is one of the predictive modelling approaches used in statistics.
- **Gradient Boosting Classifier**
  Gradient boosting classifiers are a group of machine learning algorithms that combine many weak learning models together to create a strong predictive model.
- **Random Forest Classifier**
  A random forest is a machine learning technique that's used to solve regression and classification problems. It utilizes ensemble learning, which is a technique that combines many classifiers to provide solutions to complex problems. A random forest algorithm consists of many decision trees.

2.4 Section (Manual Testing)

We define another function for manual testing of those news from the dataset.If we put a news headline into this function,it will return the result based on four different classifiers which we discussed in the previous section.

Example:

```
LR Prediction: Not A Fake News
DT Prediction: Not A Fake News
GBC Prediction: Not A Fake News
RFC Prediction: Not A Fake News
```

Chapter 3

Performance Evaluation

3.1 Performance

3.1.2 Performance of Logistic Regression

```
In [33]: LR.score(xv_test, y_test)
Out[33]: 0.9867260579064588
```

```
In [34]: print(classification_report(y_test, pred_lr))
```

	precision	recall	f1-score	support
0	0.99	0.99	0.99	5842
1	0.99	0.99	0.99	5383
accuracy			0.99	11225
macro avg	0.99	0.99	0.99	11225
weighted avg	0.99	0.99	0.99	11225

3.1.3 Performance of Decision Tree Classification

```
In [38]: DT.score(xv_test, y_test)
Out[38]: 0.995456570155902

In [39]: print(classification_report(y_test, pred_dt))
```

	precision	recall	f1-score	support
0	0.99	1.00	1.00	5842
1	1.00	0.99	1.00	5383
accuracy			1.00	11225
macro avg	1.00	1.00	1.00	11225
weighted avg	1.00	1.00	1.00	11225

3.1.4 Performance of Gradient Boosting Classifier

```
In [43]: GBC.score(xv_test, y_test)
Out[43]: 0.9955456570155902

In [44]: print(classification_report(y_test, pred_gbc))
```

	precision	recall	f1-score	support
0	1.00	0.99	1.00	5842
1	0.99	1.00	1.00	5383
accuracy			1.00	11225
macro avg	1.00	1.00	1.00	11225
weighted avg	1.00	1.00	1.00	11225

3.1.5 Performance of Random Forest Classifier

```
In [48]: RFC.score(xv_test, y_test)
Out[48]: 0.987706013363029

In [49]: print(classification_report(y_test, pred_rfc))
```

	precision	recall	f1-score	support
0	0.99	0.99	0.99	5842
1	0.99	0.98	0.99	5383
accuracy			0.99	11225
macro avg	0.99	0.99	0.99	11225
weighted avg	0.99	0.99	0.99	11225

3.2 Results and Discussions

We can see that four different classifier give us over 98% accuracy on average which is a very good number in this kind of situation.But Gradient Boosting Classifier takes too much time from others classifier.And we can also see the accuracy result for both true and false news.

And for manual testing we also get almost accurate result form = ((23481, 5), (21417, 5)) this kind of big data frame.

Chapter 4

Conclusion

4.1 Introduction

The purpose of the work is to come up with a solution than can be utilized by the users and social media to detect and filter out false news.Because false news can make a big impact on social life as well as personal life.But this is not a very easy task because the world internet has not any boundaries. Linguistics problem is another difficulty for this kind of work.

4.1 Practical Implications

Social media for news consumption is a double-edged sword. On the one hand, its low cost, easy access, and rapid dissemination of information allow users to consume and share the news. On the other hand, it can make viral “fake news”, i.e., low-quality news with intentionally false information. The quick spread of fake news has the potential for calamitous impacts on individuals and society. This kind of project will prevent it from spreading fake news all over the world.

References

1. sklearn.linear_model.LogisticRegression - Scikit-learn [https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjk8O-B7fbyAhVjlOYKHRPOCMkQFnoECAQQAQ&url=http%3A%2F%2Fscikit-learn.org%2Fstable%2Fmodules%2Fgenerated%2Fsklearn.linear_model.LogisticRegression.html&usg=AOvVaw3jQzWkNHz4EsVdoj30jRb3]
2. 1.10. Decision Trees — scikit-learn [https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwiozaWh7fbyAhU463MBHUSOAWIQFnoECAQQAQ&url=http%3A%2F%2Fscikit-learn.org%2Fstable%2Fmodules%2Ftree.html&usg=AOvVaw0lSaFzDc1nnQhPpTXeI-LR]
3. sklearn.ensemble.GradientBoostingClassifier - Scikit-learn [https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwi8yuvB7fbyAhWe8HMBHczWAdAQFnoECAQQAQ&url=http%3A%2F%2Fscikit-learn.org%2Fstable%2Fmodules%2Fgenerated%2Fsklearn.ensemble.GradientBoostingClassifier.html&usg=AOvVaw2Di5F4WfMEDdgwCFwiB_zU]
4. sklearn.ensemble.RandomForestClassifier - Scikit-learn [https://www.google.com/url?sa=t&rct=j&q=&esrc=s&source=web&cd=&cad=rja&uact=8&ved=2ahUKEwjIz_Xa7fbyAhVCjOYKHYPLDu0QFnoECAUQAQ&url=http%3A%2F%2Fscikit-learn.org%2Fstable%2Fmodules%2Fgenerated%2Fsklearn.ensemble.RandomForestClassifier.html&usg=AOvVaw2H-u90wd4bJLLgLmGGWK0Y]