NITTE MEENAKSHI INSTITUTE OF TECHNOLOGY

(AN AUTONOMOUS INSTITUTION, AFFILIATED TO VISVESVARAYA TECHNOLOGICAL UNIVERSITY,
BELGAUM, APPROVED BY AICTE & GOVT.OF KARNATAKA

Introduction to Machine Learning (18CSE751)

Learning Assessment
Laboratory based Exercise

Submitted in partial fulfillment of the requirement for the award of Degree of

Bachelor of Engineering

in

Computer Science and Engineering

Submitted by:

K. Viddya	1NT18CS065
Krithika Devadiga	1NT18CS080
Jasmine	1NT18CS061

Under the Guidance of
Dr. Vani Vasudevan
Professor, Dept. of CS&E, NMIT

Department of Computer Science and Engineering
(Accredited by NBA Tier-1)

2020-2021

ABSTRACT

The proliferation of misleading information in everyday access media outlets such as social media feeds, news blogs and online newspapers have made it challenging to identify trustworthy news sources. The spread of fake news and misinformation is causing serious problems to society, partly due to the fact that more and more people only read headlines or highlights of news assuming that everything is reliable, instead of carefully analyzing whether it can contain distorted or false information. Specifically, the headline of a correctly designed news item must correspond to a summary of the main information of that news item. This increases the need for a computational tool that is able to provide insight to the reliability of the content. In this project, we have used natural language processing techniques and machine learning algorithms to classify whether the news is fake or not. Using a benchmark dataset and machine learning models like SVM, Naive-bayes and Logistic Regression, we are predicting the possibility of the news to be true or false.

CONTENTS

INTRODUCTION

Fake news denotes a type of yellow press which intentionally presents misinformation or hoaxes spreading through both traditional print news media and recent online social media. Fake news has existed for a long time. In recent years, due to the booming developments of online social networks, fake news for various commercial and political purposes has been appearing in large numbers and widespread in the online world. With deceptive words, online social network users can get infected by this online fake news easily, which has brought about tremendous effects on the offline society already.

Fake news has significant differences compared with traditional suspicious information, like spam in various aspects:

- Impact on society: spams usually exist in personal emails or specific review websites and merely have a local impact on a small number of audiences, while the impact fake news in online social networks can be tremendous due to the massive user numbers globally, which is further boosted by the extensive information sharing and propagation among these users

- Audiences' initiative: instead of receiving spam emails passively, users in online social networks may seek for, receive and share news information actively with no sense about its correctness.

- Identification difficulty: via comparisons with abundant regular messages (in emails or review websites), spams are usually easier to be distinguished; meanwhile, identifying fake news with erroneous information is incredibly challenging, since it requires both tedious evidence-collecting and careful fact checking due to the lack of other comparative news articles available.

Fake news paves the way for deceiving others and promoting ideologies. These people who produce the wrong information benefit by earning money with the number of interactions on their publications. Spreading disinformation holds various intentions, in particular, to gain favour in political elections, for business and products, done out of spite or revenge. Humans can be gullible and fake news is challenging to differentiate from the normal news. Most are easily influenced especially by the sharing of friends and family due to relations and trust

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. Due to this false information is reaching a large number of people.

DATASET

The LIAR dataset is a new benchmark dataset for fake new detection that was published by William Yang in July 2017. He in turn retrieved the data from PolitiFact's API, which provides detailed analysis reports and links to source documents for each case. This website collects statements made by US 'speakers' and assigns a truth value to them ranging from 'True' to 'Pants on Fire'. The statements that Yang retrieved primarily date from between 2007 and 2016. LIAR is a publicly available dataset for fake news detection. A decade-long of 12.8K manually labelled short statements were collected in various contexts from Politifact.com. The LIAR dataset4 includes 12.8K human labelled short statements from Polifact.com's API, and each statement is evaluated by a Polifact.com editor for its truthfulness.

The dataset comes pre-divided into training, validation and testing files. For our purposes, we will use the files as follows:

- **Training dataset:** 5-fold cross-validation of our models, using an 80/20 split
- **Validation dataset:** evaluate our model results and choose our model
- **Testing dataset:** judge the final model

This dataset can be used for fact-checking research as well. The liar dataset has the following attributes:

Columns	**Description**
Id	The json ID of the statement
Label	Truth value of the statement; 6 categories from 'true' to 'pants on fire'
Statement	Title of the PolitiFact article, often but not always the actual statement
Subject	The subject(s) of the statement
Speaker	The source of the statement
Speaker_job Speaker_us_state speaker_affiliation	The speaker's job title, US state where they're based, and party affiliation where they are available.
Speaker_bt Speaker_pof (5 features)	Total count of truth values to the speaker(truth credit history), excluding truth count and including truth statements.
context	The context (value/location of the speech statement)

There are several possible reasons for the models' poor performance:

- The lack of dates translate to **lack of historical information.**

- There is a prevalence of **missing speaker jobs and affiliations**, which means that these features may not have been very useful for determining which piece of news was fake**.**
- Some of the articles in the LIAR dataset are **from the wrong set of data** (PolitiFact's Flip-o-Meter rather than its Truth-o-Meter), and yet are tagged with a truth value. As a result, those data points are not useful for training the model because they are mislabelled.

MACHINE LEARNING METHODS

Classifiers will be used to predict fake news. Firstly, feature selection will be performed on the dataset. The features extracted here will be fed into different classifiers.The classifiers used will be Naive-bayes, Logistic Regression and Linear SVM.
Naive-bayes classifier- It is a probabilistic classifier based on Bayes theorem. It uses features to make a prediction on a target variable and assumes that features are independent of each other and there is no correlation between features.

SVM- SVM algorithm creates the best line or decision boundary that can segregate n-dimensional space into classes so that we can easily put the new data point in the correct category. This best boundary or region is called a hyperplane. Algorithm finds the closest point of the lines from both the classes. These points are called support vectors.

Logistic Regression-It is used for predicting the categorical dependent variable using a given set of independent variables. Logistic regression is used for solving the classification problems. Instead of fitting a regression line, we fit an "S" shaped logistic function, which predicts two maximum values (0 or 1).Logistic Regression provides probabilities and classifies new data using continuous and discrete datasets.

Certain terms have higher importance while detecting fakeness of the news, while common words like the, an, a do not hold any importance. So, to determine the most important word term frequency tfidf vectorizer will be used.
Comparison of different models will be required to select the best out of all. So, to find out the best model f1 score and confusion matrix will be considered.

The goal of the machine learning models is to detect fake news with high accuracy and inform the user.

Assessment
F1 score metric usually tells us how precise and robust our classifier is.It is a harmonic mean between recall and precision. The Better the F1 score better will be performance.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

F1 score of all the models will be compared for the dataset. The best performing model will be selected for fake news detection.
A Precision-Recall curve is a graph with Precision values on the y-axis and Recall values on the x-axis. A good Precision-Recall curve will have greater area under the curve.
We can use the Precision-Recall curve to see how training and test sets perform when we increase the amount of data in our classifiers.

PRESENTATION AND VISUALISATION

First we will have to train, test and validate data files and then perform some pre-processing techniques like image feature extraction and selection methods to make it easier to use in the further stages. These extracted features can be fed into different classifiers. we have to build all the classifiers for detection of fake news. We will be using Naive-bayes, Logistic Regression and Linear SVM models to implement this idea. We will aim to extract as many features as possible from the term-frequency TF-IDF vectorizer to see what words are mostly used and are important in each of the classes. The application must ultimately take a news article as input from the user and then the article is predicted to be genuine or fake.

ROLES

Group Member	**Task**
Jasmine	Preprocessing of the dataset and implementation of Logistic Regression model.
Viddya	Preprocessing of the dataset and implementation of SVM.
Krithika	Preprocessing of the dataset and implementation of Naive-bayes classifier.

SCHEDULE

Date	**Task to be completed**
20/12/21	Submission of proposal
05/01/22	System design and implementation
13/01/22	Testing performance and accuracy
17/01/22	Final report submission
18/01/22	Final presentation

BIBLIOGRAPHY

https://towardsdatascience.com/identifying-fake-news-the-liar-dataset-713eca8af6ac

https://sites.cs.ucsb.edu/~william/papers/acl2017.pdf “Liar, Liar Pants on Fire”: A New Benchmark Dataset for Fake News Detection”

A. Jain and A. Kasbe, "Fake News Detection," 2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS), 2018, pp. 1-5, doi: 10.1109/SCEECS.2018.8546944.

https://data-flair.training/blogs/advanced-python-project-detecting-fake-news/

https://www.hindawi.com/journals/complexity/2020/8885861/