FACULTY OF INFORMATION & COMMUNICATION TECHNOLOGY

BITI 2513

INTRODUCTION TO DATA SCIENCE

CAPSTONE PROJECT

PROJECT TITLE : FAKE AND REAL NEWS CLASSIFIER

GROUP : DS GENG

Name:	No Matric :
Muhammad Naim Syahmi bin Roslan	B031810312
Muhamed Hussain Bin Hithayatullah	B031810392
Muhammad Fikrun Amin	B031810404
Ramanan Gobalakrishnan	B031810334

Lecturer Name : AP DR SHARIFAH SAKINAH SYED AHMAD

Introduction

As the world progresses more and more, it keeps getting harder to identify a given public information as truthful or fake. The reputation and state of journalism these days were damaged by the rise of media sensationalism and fake news. The worrying rise of low-quality hack writers also makes it difficult to find authentic and good articles.

The project aims to solve this problem by creating a classifier that is able to label a given news as either fake or real based on their attributes such as the text and subject of the article. It involves natural language processing and text analysis procedures. It is believed that the data science techniques used when developing this project can be applied to other similar domains of text analysis.

Objective

1. To distinguish between real and fake news

2. To use text analysis on labelled data

3. Help create more trustworthy flow of news

Goals

Our main goal in this project is being able differentiate between fake and reliable news. This mainly because of the abundance fake or inaccurate news on the internet. As students who are learning on becoming a computer specialist we have to make sure the news given out by the media is always trustworthy and reliable. Fake news can sometimes cause great danger not only economically but also physically. Physical attacks may occur just because of a piece of fake news and we as students think such think are inacceptable.

Not only that but our goal is also use artificial intelligence in being able to differentiate between fake and real news. Most of the news online is considered inaccurate, thus using Artificial Intelligence is distinguishing between real and fake news is extremely efficient and useful.

Questions

How can we help the public to distinguish between real public information and fake public information using sentiment analysis based on the title, text, subject of the information, and date on which the information was posted?

Success And Measurements

"Success" And Its Measurements

The project will involve performing text analysis on labelled data to generate a useful model. In our case, "Success" occurs when the model generated by supervised learning of the data is able to classify a given news item as fake or true according to its attributes. We measure our success by looking at the performance of the model by using measures such as accuracy or the confusion matrix. We can consider the project a success if the model has good performance and can be run without any errors.

Measurable Outcomes

- ➢ The model can classify news as either fake or real news.
- ➢ The model generated can predict with a high accuracy.

Data Source

The dataset used of this project is "fake and real news dataset" from kaggle website and article. This dataset contains two types of news that is fake and real news. In article, this dataset was gathered from real world sources: the truthful article were collected from reuters.com(news website) and for the fake news article were collected from unreliable websites that were flagged by politifact (a fact-checking organization in the USA) and wikipedia . In kaggle website the dataset consists of two csv files. The first file named "true.csv" contains 21417 data and the file second files named "fake.csv" contains 23502 data and this dataset also contains 4 attributes, which are:

1. Title
2. Text
3. Subject
4. Date

Dataset url : https://www.kaggle.com/clmentbisaillon/fake-and-real-news-dataset?select=True.csv