**UNIVERSITY OF LONDON INTERNATIONAL PROGRAMMES**

**BSc Computer Science and Related Subjects**



**CM3070 PROJECT**

**PRELIMINARY PROJECT REPORT**

Fake News Detection

Author: Tay Yang Lin Marcus

Student Number : 180295106

Date of Submission: 28 November 2023

Supervisor : Chew Jee Loong

Contents

[CHAPTER 1: INTRODUCTION 3](#_Toc102463874)

[CHAPTER 2: LITERATURE REVIEW 3](#_Toc102463875)

[CHAPTER 3: PROJECT DESIGN 4](#_Toc102463876)

[CHAPTER 4: FEATURE PROTOTYPE 5](#_Toc102463877)

[CHAPTER 5: APPENDICES 5](#_Toc102463878)

[CHAPTER 6: REFERENCES 5](#_Toc102463880)

# INTRODUCTION

**Project Background: What is Fake News?**

With the constant advancement of technology, informative data has become more accessible, impacting daily lives in both a positive and negative manner. In the aspects of negative, increased accessibility has led to an increased spread of fake news. Fake news is generally fabricated to mislead its readers into believing misinformation of various forms such as propaganda, scandals, rumors. For example, Buzzfeed (https://www.buzzfeed.com/) compared and analyzed participation in 20 real news and 20 fake news articles (e.g., likes, comments, share activities) that spread the most on Facebook during the last three months of the 2016 US Presidential Election. According to the results, the participation rate of fake news (8.7 million) was higher than that of mainstream news (7.3 million), and 17 of the 20 fake news played an advantageous role in winning the election (Bogoan Kim, 2021). Another example, “There is much concern during this pandemic about the spread of misleading or inaccurate information. This article reports on a small study which attempted to identify the types and sources of COVID-19 misinformation. The authors identified and analyzed 1225 pieces of COVID-19 fake news stories taken from fact-checkers, myth-busters, and COVID-19 dashboards. The study is significant given the concern raised by the WHO Director-General that 'we are not just fighting the pandemic; we are also fighting infodemic'. The study concludes that the COVID-19 infodemic is full of false claims, half backed conspiracy theories and pseudoscientific therapies, regarding the diagnosis, treatment, prevention, origin and spread of the virus” (Salman Bin Naeem, 2021). Hence, fake news, with social media as a news medium, poses as a constant threat.

**AIM: Develop a Fake News Detection Model**

To tackle the spread of fake news, the idea for this project is to create a software module that can differentiate between real and fake news, embodied as a Jupyter Notebook, using Python. Some goals will be, can a machine-learning model identify the fake news, how accurate is the identification, what is considered an optimal model. Upon achieving each of these base goals for the project in order, the first goal will ensure that a machine-learning model will be created, the second goal will ensure the accuracy of the model result, the final goal, will ensure that that final model created, will be the most accurate model created during the project development process.

# LITERATURE REVIEW

In this section, literature review of past projects has been researched to aid in the development of this project for Fake News Detection.

Some Like it Hoax: Automated Fake News Detection in Social Networks (Eugenio Tacchini, 2017)

For this research, a dataset consisting of 15,500 Facebook Posts has been chosen and the aim is to classify these posts between hoax and non-hoax. The two types of models created were the Logistic Regression Model and the Harmonic Label-Crowdsourcing Model. The evaluation metric was conducted with average accuracy and standard deviation in mind and the sample data used was set between having ‘One-page out’ and ‘Half-page out’. After 50 runs, the Logistic Regression Model had an average of 71% (Half Page) – 79% (Full Page) whereas the Harmonic Label-Crowdsourcing Model had an overall average of 99% for both. In my opinion, even though a comparison was made, and the aim of the project has been met, this research is not very well done. The reason being that the dataset seems to be a little small and the number of runs conducted is less than 1% of the overall dataset, which impacts the overall accuracy of the research results. Hence, the takeaway from this research is that an appropriate dataset with a larger count in data is better to be used and that more runs must be conducted to ensure a fair data result for comparison.

Fake News Detection Using Naive Bayes Classifier (Mykhailo Granik, 2017)

For this research, a dataset consisting of 2282 Facebook Posts has been selected and the aim is to classify it been true and false. One thing of note is that only the Naïve Beyes Classifier Model has been developed for the classification. The research conducted has yielded results for the model accuracy having an overall score of 75%. Despite its decent accuracy level, this can be considered bad research. The dataset used was small, and some of the data were just previews of whole articles. Moreover, the author admitted that proper text-preprocessing steps had not been done stating that ways to improve accuracy included stopwords and stemming. Therefore, the takeaway for this research was that proper text-preprocessing steps must be done and a larger dataset must be used to avoid a skewed result for accuracy.

TI-CNN: Convolutional Neural Networks for Fake News Detection (Yang Yang, 2018)

For this research, a combination of two types of datasets were used, one being text while the other being images, the domain being news of the American Presidential Election. Different methods were used to conduct the experiment and their precision, recall and f-1 scores were used for evaluation. The methods were Convoluted Neural Network for images, Logistic Regression, Convoluted Neural Network, Long Short-Term Memory, Gated Recurrent Units were used for text data and finally a hybrid between Text and Image Convulted Neural Network. Viewing the comparison between the results between all the models, the hybrid model ended up having the highest scores by a large margin between almost all the other models. In conclusion, the research conducted was very well done and some methods to adopt is to involve the development of a hybrid model to ensure that the final product created will be highly accurate.

# PROJECT DESIGN

For the scope of this project, it will involve the users with access to Jupyter Notebook, and the domain of the project, will be the true and false news.

The plan for the project will be as follows:

A diagram of a project

Description automatically generated

Figure 1: Project Plan

For the methods, Python libraries such as “sklearn” and “keras” will be used to develop the models. Data pre-processing such as using stopwords and embeddings will be used in the process in improve the accuracy for the experiment. Some base models considered will be Naïve Beyes, Support Vector Machine while a final hybrid model will also be created, a Convoluted Neural Network hybrid model. For evaluation, precision, recall, f1-score and most importantly, accuracy will be included. For visual aid, confusion matrices, graphical results will be used for display and reading purposes.

**-Gantt Chart to be added at a later time-**

# FEATURE PROTOTYPE

For the first prototype, the model selected for development must be the most accurate of the various base models after comparison has been done between each of them.

# REFERENCES

Bogoan Kim, A. X. (2021, December 9). A systematic review on fake news research through the lens of news creation and consumption: Research efforts, challenges, and future directions. *PloS one, 16*. doi:10.1371/journal.pone.0260080

Eugenio Tacchini, G. B. (2017). Some Like it Hoax: Automated Fake News Detection in Social Networks. *CoRR, abs/1704.07506*. Retrieved from arXiv: https://arxiv.org/abs/1704.07506

Mykhailo Granik, V. M. (2017). *Fake news detection using naive Bayes classifier.* Vinnytsia, Ukraine : IEEE. doi:10.1109/UKRCON.2017.8100379

Salman Bin Naeem, R. B. (2021). An exploration of how fake news is taking over social media and putting public health at risk. *Health information and libraries journal*, 143-149. doi:10.1111/hir.12320

Yang Yang, L. Z. (2018). TI-CNN: Convolutional Neural Networks for Fake News Detection. *CoRR, abs/1806.00749*. Retrieved from http://arxiv.org/abs/1806.00749