COVID-19 Fake News Detection in Malaysia

Soo Mun Chong[1] and Vimala Balakrishnan[2]

[1]Faculty of Computer Science and Information Technology, Universiti Malaya, 50603 Kuala Lumpur, Malaysia
[2]Faculty of Computer Science and Information Technology, Universiti Malaya, 50603 Kuala Lumpur, Malaysia

***Abstract*—Social media has been flooded with enormous amounts of COVID-19-related information ever since the COVID-19 pandemic started back in 2020. Since then, Malaysian citizens have become more reliant than ever on social media for consumption of COVID-19 information. However, the lack of COVID-19 news regulations on social media platforms encouraged people to post unverified, fake and misleading COVID-19 related information. Because of the time-consuming nature of fact-checking, people often take these unverified COVID-19 news for granted. Consequently, people inadvertently spread these fake COVID-19 news to their families, friends and relatives on social messaging platforms like Whatsapp. The spread of COVID-19 fake news online in Malaysia can have severe sequences, causing widespread panic among fellow Malaysians. In this paper, we proposed a supervised learning approach to detect COVID-19 fake news. The fake news on COVID-19 were scraped from the website called Sebenarnya, and real news were scraped from The Star website. We applied a semantic model with different word representations which include Bag of Words (BOW), Term Frequency - Inverse Document Frequency (TF-IDF), Word2Vec and Global Vectors(GloVe). In the evaluation step, we applied 6 supervised machine learning algorithms such as Multinomial Naive Bayes, Support Vector Machines, Decision Tree, Random Forest, Logistic Regression and Adaboost. Afterward, 10-fold cross validation was used to train and evaluate the 6 supervised algorithms according to performance metrics such as accuracy, precision, recall, AUC-ROC, F1-score. The results showed that Random Forest with the word representation of TF-IDF performed the best with over 97% accuracy in contrast to numerous conventional supervised machine learning classifiers.**

***Index Terms*—Fake news detection, Covid-19, Malaysia**

I. Introduction

Fake news, defined as fabricated content that mimics legitimate sources of news in order to deceive its readers, is now a common social issue present on social media channels [1]. Social media, originally created as a powerful communication medium to build connections between friends, has become a breeding ground for the generation and dissemination of fake news in recent years. The most infamous case being the 2016 US presidential election campaign, where over 8 million shares, reactions, and comments were generated on the top twenty frequently discussed false election news [2]. The recent outbreak of Covid-19 worldwide pandemic has seen another surge in fake news on social media, with the WHO Director-General Tedros Adhanom Ghebreyesus calling it an "infodemic". It refers to a situation when there is an overload of information which might be true or false, making it difficult to decide what to believe [3].

In Malaysia, the inundation of Covid-19 fake news circulating the social media platforms is what was believed to have undermine the Ministry of Health's countrywide vaccination effort in the early phases, which led to the immediate action taken by the Malaysian government, to issue an Anti-Fake News ordinance on March 2021 which made it an offence to publish false covid-19 content in Malaysia [4]. Besides, the Malaysian Communications and Multimedia Commission (MCMC) has also taken the initiative to operate a third-party fact-checking website - sebenarnya.my portal to combat the spread of fake news. However, according to the Malaysian Communications and Multimedia Commission (MCMC) Internet Users Survey 2020, only 20.4% of the Internet users are aware of the existence of this portal even though the majority of Internet users (92.3%) consider fake news as a problem. Moreover, the spread of false information about Covid-19 pandemic can have dire social consequences, causing confusion and concerns among the public and undermining the government's efforts in combating Covid-19. Hence, researches have begun in recent years to address the fake news problems using machine learning detection algorithms on social media feeds, but there are a lack of studies related to Covid-19 fake news in Malaysia [5]. Thus, this study explores the application of different machine learning algorithms on Covid-19 fake news detection in Malaysia. This study extensively analyzes several machine learning algorithms and utilizes pre-trained word embeddings such as Word2Vec and GloVe. This paper is divided into several main sections: Section 2 describes the literature review and related works. Section 3 describes the problem formulation. Section 4 explains the proposed methodology. Section 5 shows the experimental results while Section 6 provides the discussion. Finally, Section 6 presents conclusions and plans for future work.

II. Literature Review

According to [6], extracting lexical and host-based features of associated uniform resource locators (URLs) for news articles could improve the efficiency and performance of Natural Language Processing classifiers. The dataset consisted of a collection of approximately 3, 350 news results, gathered from 2084 different URLs, which dated from January 20 to July 28, 2020 - the period where most countries experienced lockdowns. Since the data collection was conducted using search engine optimisation, several steps were proposed to reduce the potential bias which include using a VPN to be more consistent with the WSE domain inspected and its results, as well as using private/incognito window to prevent customised search results based on browsing histories. They used bag-of-words (BoW) and term frequency-inverse document frequency (TF-IDF) for mapping cleaned texts into

numeric representations. Further features (length, counting, and binary) were also extracted from URLs. Result shows that by incorporating URL features, Naive Bayes classifier with BoW achieved an F1-score of 81%, while SVM with TF-IDF got 79%, significantly exceeded results based on features built from lexical representations of the text (titles and descriptions) only.

Kannan et al.[7] states in their research that following the introduction of the COVID vaccine, there has been an inundation of hoax news articles. They presented an efficient way to detect fake news, using automation of supervised machine learning models. The models include Logistic Regression, XGBoost, RNN, and LSTM. Their findings show that linear SVC achieved superior performance compared to the other models, with over 95% accuracy and 0.96 F1-score. A noticeable result was the performance of the RNN and LSTM models, which both achieved 95% accuracy.

Priya & Kumar[8] states in their research that character-level features can be useful in a variety of machine learning applications because social media posts often contain grammatical errors, non-standard acronyms, and spelling errors. They explored the usability of character-level features with 8 different conventional machine learning classifiers and two different deep learning models. The implemented models are Support Vector Machines (SVM), Random Forest (RF), Logistic Regression (LR), K-Nearest Neighbor (KNN), Naive Bayes (NB), Gradient Boosting (GB), Decision Tree (DT), Adaboost, Dense Neural Network (DNN), and Convolutional Neural Network (CNN). To further improve the performance of machine learning models, they proposed a deep ensemble model that contains support vector machine (SVM), dense neural network (DNN), and convolutional neural network (CNN) in parallel to classify social media COVID-19 related posts into fake and real classes. The result shows that among all the implemented models, the deep ensemble model (SVM + DNN + CNN) with character-level features performed best with the weighted precision, recall and F1-score of 0.97. Their findings show that using character-level features with the classifiers performs better as compared to the ones with word-level features.

Bojjireddy et al.[9] describes in their research that with today's straightforward creation of social media posts, there has been an increasing amount of fake news, compared to traditional media in the past. They presented a machine learning approach to distinguish fake news from correct news (in tweets) based on content features. The result shows that the SVM model with TF-IDF feature representation performed the best with accuracy of 96%.

Isaakidou et al.[10] states in their research that a survey conducted by the European Commission on EU citizens shows that 40% come across fake news daily and 85% evaluate fake news as a problem. To overcome the problem, this study presented an Artificial Intelligence approach, the Decision Trees algorithm to identify COVID-19 fake news. The findings show that decision trees with entropy loss with best splitting criteria achieved 100% accuracy. However, the results also show that application of the decision tree algorithm on titles has an overall accuracy of 72.09%, which further conclude that titles are not a reliable part of an article to be used for fake news detection.

Bojireddy et al.[11] states in their research that fake news is false information about current events, intentionally created to mislead readers and the spread of such fake news has the potential to create a negative impact on individuals and society. They presented a machine learning approach for automating the detection of fake news and misleading contents and developed a fake news web application to make it easy for people to verify the trustworthiness of a COVID-19 news article.

Saghayan et al.[12] states in their research that one of the major complications of fake news detection lies in the fact that news in social media is multilingual, and therefore developing methods for each and every language in the world is impossible, especially for low resource languages like Persian. To solve this problem, researchers use machine translation to translate Persian tweets into English tweets. Since it requires lots of effort, time and expenses to develop a new set of Natural Language Processing tools for different languages, they proposed the method of transforming Persian tweets to English, which then allowed them to unify a low resource language to the prevalent language - English that has lots of Natural Language Processing tools available today. Results show that classification accuracy in original Persian tweets was 87%, while the accuracy dropped to 84% in machine translated dataset. The slight decrease in 4% accuracy was probably due to the fact that many deeper meanings of sentences are lost in the translation process.

Ul Hussna et al.[13] explored the deep learning based algorithm named distil BERT to accurately predict news about COVID-19 on social media posts. Distil BERT is a small and lightweight BERT's distillation transformer model and has 40% fewer parameters than Bert-Base-based. This study compared the performance of distil BERT to the other conventional traditional supervised machine learning models such as Multinomial Naive Bayes, Logistic Regression and Support Vector Machine. The results show that Naive Bayes, Logistic Regression and Support Vector Machine achieved 92%, 91%, and 93% accuracy respectively, while distil BERT achieved 97% accuracy, outperforming the traditional machine learning models.

Kaliyar et al.[14] states in their research that neural networks are powerful architectures for fake news classification. LSTMs could handle the problem of gradient exploding and memory access plaguing most neural network architectures. In this paper, they proposed a hybrid model using convolutional layers(CNN) with LSTM layers. Their findings show that the proposed C-LSTM model achieved exemplary results, with 92% accuracy, because it has the ability to capture both the temporal semantics as well as phrase-level representations.

Rajalakshmi et al.[15] explored three machine learning models which work well on text data such as Passive Aggressive Classifier, Multinomial Naive Bayes and Support Vector

Machine. Passive Aggressive Classifier is an 'online-learning algorithm' which works well on large scale data. In this study, the results show that the Passive Aggressive Classifier has the best performance with a 79% accuracy rate, while Naive Bayes and SVM achieved 76% and 77% respectively.

III. Problem Formulation

Social media has been the primary source by which Malaysians obtain new information, with up to 91.7% of Malaysians staying updated on daily news on Facebook, according to Internet Users Survey 2020 by Malaysian Communications and Multimedia Commission (MCMC). With the huge amount of COVID-19 misinformation flooding the social media these days, it's important for Malaysians to discern fake news from real news. Therefore, it's become really important to come up with an automated machine learning system for Covid-19 fake news detection.

There has been tremendous work on various COVID-19 fake news approaches [16], but our methodology varies from those works in the sense that we only work with COVID-19 news in Malaysia. As for the dataset, we need a decent dataset to fulfill our research aims. A real news dataset must be collected from a trustworthy source and fake news collected must also be proved to be false by third-party fact checkers. We collected real news dataset from The Star website and scraped fake news from Sebenarnya website. Both the real news and fake news data dated from January 2020 to October 2021.

The main questions & objectives of this study are:

1. To explore the trend of fake news keywords from January 2020 to October 2021

- What are the trends of covid-19 fake news keywords from January 2020 to October 2021?

2. To detect fake news using supervised machine learning algorithms

- What is the proposed model to detect covid-19 fake news?

3. To assess the effectiveness of the fake news detector model

- What is the accuracy of the fake news model?

4. To develop a product for fake news detection system

- How to develop a fake news detection system?

IV. Methodology

We approached this task as a binary text classification problem in Natural Language Processing. Each news item belongs to either of the two distinct labels: "real" or "fake". Our proposed method consists of 5 main parts: (a) Data Collection, (b) Exploratory Data Analysis (EDA), (c) Text Pre-Processing, (d) Feature Extraction Techniques, (e) Model Training and Evaluation, and (f) Deployment. The overall flow diagram for the proposed model is in Figure 1 below.

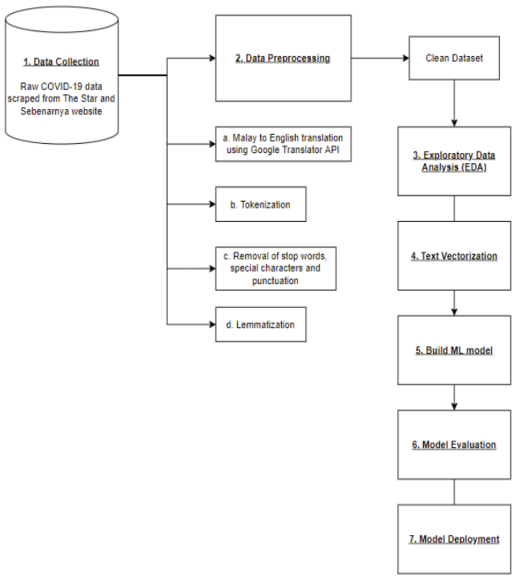

Figure 1: The proposed pipeline

4.1 Data Collection

The dataset contains real news(English) and fake news(Malay) in Malaysia, dated from January 2020 – October 2021. A total of 450 real news were scraped from "The Star" website, while a total of 436 fake news were scraped from "Sebenarnya" website. The dataset contains the following columns: "text": The body or content of the news article "label": "Fake" or "Real".

Table 1: Distribution of Real Fake COVID-19 news in the dataset

Label	Count
Real	450
Fake	436

Figure 2: SEBENARNYA.MY Website

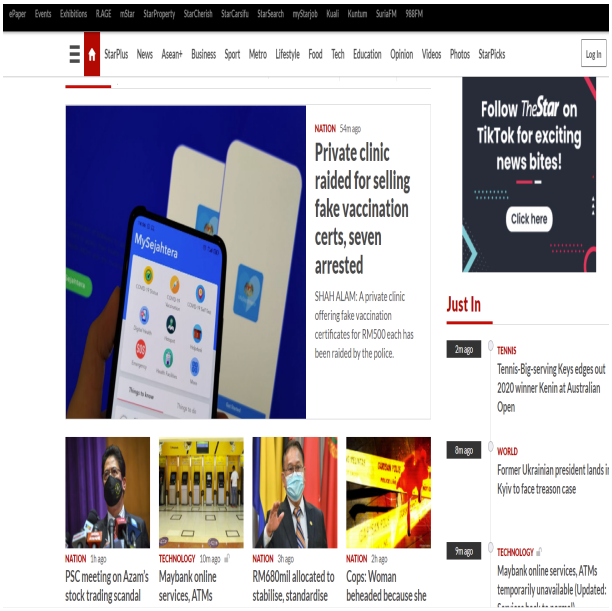

Figure 3: The Star Website

4.2 Exploratory Data Analysis (EDA)

Word cloud, visualizing the most popular keywords within fake news will be created across different timelines.

Figure 4: Word count of COVID-19 fake news in the first half of 2020

Figure 4 shows that words like “people”, “hospital”, “covid”, “virus” and “positive” are most used and often seen in the first half of 2020. This is the period where covid-19 epidemic outbreak emerged in Malaysia. Words like “march”, “state” and “close” showed up frequently in fake news during the same period because Malaysia had its first country wide lock-down from March 2020 to June 2020, whereby the Malaysian citizens were prohibited from travelling due to the strict lockdown measures implemented by the government.

Figure 5: Word count of COVID-19 fake news in the second half of 2020

Figure 5 shows that during the second half of 2020, covid-19 fake news in Malaysia often contain words like “covid”, “close”, “school”, “face”, “mask”, “hand”. Malaysians became more aware of the dangers posed by the coronavirus back then, where we had seen the school closures in every state in Malaysia. Besides, the Malaysian government had also made it compulsory for people in Malaysia to wear masks in public. Those failing to do so would receive a fine up to RM 1,000 under the Prevention and Control of Infectious Disease Act.

Figure 6: Word count of COVID-19 fake news in the first half of 2021

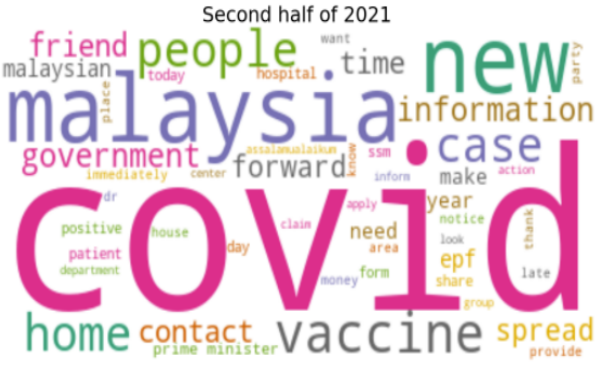

Figure 7: Word count of COVID-19 fake news in the second half of 2021

Word cloud in Figure 6 shows that words like “covid”, “sop”, and “allow” appeared frequently in covid fake news during the first half of 2021. While Figure 7 shows words like “vaccine”

started to appear in covid fake news during the second half of 2021, where people started spreading all sorts of false information regarding the side effects of covid vaccine.

N-grams of texts are extensively employed in text mining and natural language processing. An n-gram is a contiguous sequence of n items from a given sample of text or speech. An n-gram of size 1 refers to as “unigram”; size 2 is a “bigram”; size 3 is a “trigram”. The frequently occurring unigrams, bigrams and trigrams of the dataset will be visualized in colorful charts.

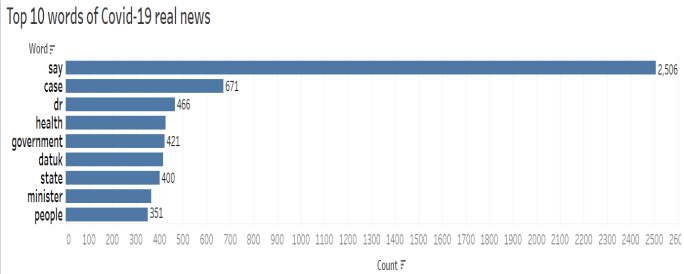

Figure 8: Top 10 most popular words in COVID-19 real news

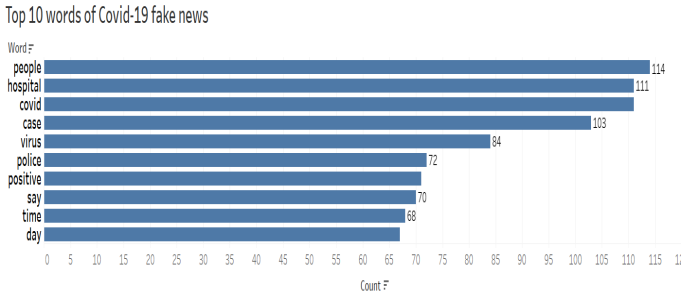

Figure 9: Top 10 most popular words in COVID-19 fake news

Figure 8 shows words like “say”, “case”, “dr”, “health”, and “government” are most used in covid real news, while Figure 9 shows words like “people”, “hospital”, “covid”, “virus”, and “positive” are most often used in covid fake news.

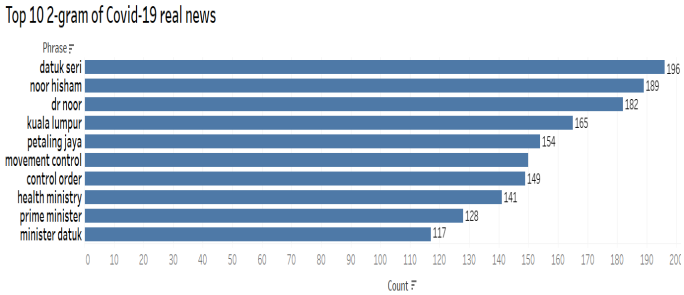

Figure 10: Top 10 most popular 2-gram in COVID-19 real news

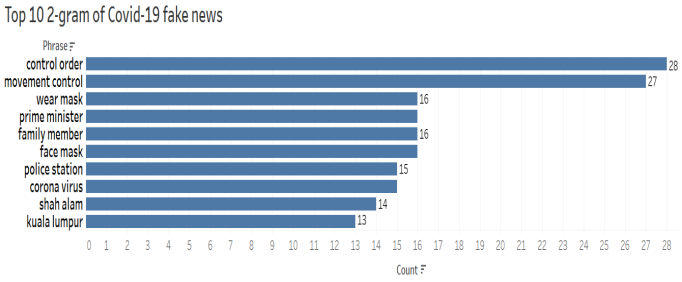

Figure 11: Top 10 most popular 2-gram in COVID-19 fake news

Figure 10 shows bigram like “datuk seri”, “noor hisham”, “dr noor” are frequently used in covid real news, whereas bigram like “control order”, “movement control”, “wear mask” are often used in covid fake news.

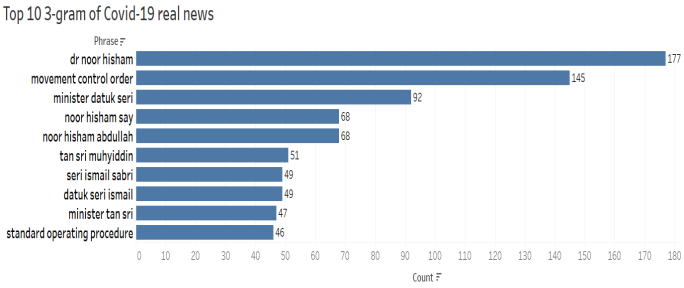

Figure 12: Top 10 most popular 3-gram in COVID-19 real news

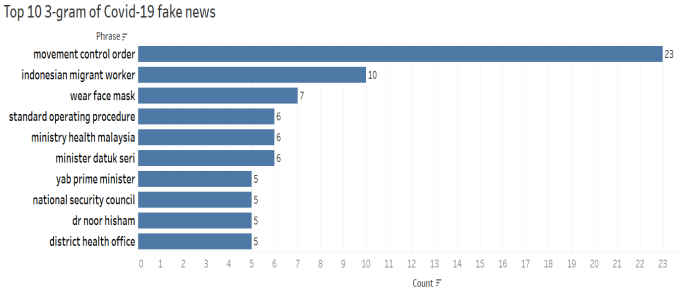

Figure 13: Top 10 most popular 3-gram in COVID-19 fake news

As for trigram, phrases like ‘dr noor hisham“, “movement control order”, and “standard operating procedure” show up frequently in both covid real news as well as covid fake news, as shown in Figure 12 and Figure 13.

4.3 Text Pre-Processing

Raw texts have to be pre-processed into machine-readable form. Pre-processing steps are listed in sequence as follows:

a. Convert all the texts to English using Googletrans API

b. Converting all letters to lowercase

c. Remove punctuation

e. Remove stop words : some common words in English like "the", "a", "on", "is" do not carry important meaning and need to be removed from texts.

f. Stemming/Lemmatization : Reduce words to their root form.

4.4 Feature Extraction Techniques

Next, the cleaned texts have to be given a word vector representation. Word embedding is a methodology in NLP to map words or phrases from vocabulary to a corresponding vector of real numbers which can be fed into machine learning algorithms. The techniques that will be used are Bag-of-Words model, TF-IDF, Word2Vec and GloVe.

A. Bag of Words (BoW)

A bag-of-words is a representation of text that describes the occurrence of words within a document. It involves two things:

- A vocabulary of known words
- A measure of the presence of known words

B. Term Frequency - Inverse Document Frequency (TF-IDF)

- Term frequency: raw count of instances a word appears in a document
- Inverse document frequency: The logarithm of the number of documents divided by the number of documents that contain the word

Multiplying these two numbers results in the TF-IDF score of a word in a document. The higher the score, the more relevant that word is in that particular document.

C. Word2Vec

Word2Vec is a classical method that creates word embeddings in the field of Natural Language Processing (NLP), developed by Tomas Mikolov and his team at Google in 2013. Word2Vec takes in words from a large corpus of texts as input and learns the vector representation of words. The resulting word vector can be used as features in many natural language processing and machine learning applications. In this study, we used the pre trained Word2Vec vector file developed by Google.

D. Global Vectors (GloVe)

A matrix factorization technique on aggregate global word-word co-occurrence statistics from a corpus that obtains vector representations for words which share interesting linear substructures of the word vector space.

4.5 Model Training

Supervised machine learning is the construction of algorithms that are able to produce general patterns hypotheses by using externally supplied instances to predict the fate of future instances [17]. Supervised machine learning classification algorithms aim at categorizing data from prior information [17]. In this study, six supervised machine learning algorithms are used, as listed below:

Multinomial Naive Bayes

Naive Bayes classifier is a core technique used for text retrieval and classification, focussing on the distributional assumptions made about word occurrences in documents [18]. An alternative approach to modelling term frequencies is to treat the bag of words for a length f document as resulting from f draws on a d-valued multinomial variable X. The naive bayes assumption is that the draws on X are independent - each word of the document is generated independently from every other.

Support Vector Machines (SVM)

SVM is a machine learning classifier that separates the training data into classes by learning a separating hyperplane [19]. It also works best for non-linear problems due to its kernel functions that take original vector space to high dimensional space where it is much easier to find a hyperplane.

Random Forest

Random forest, proposed by Leo Beriman and Adele Cutler in 2001, is an ensemble learning method for classification and regression that constructs a number of decision trees at training time and delivers the class that is the mode of the classes output by individual trees [20].

Logistic Regression

Logistic regression model apprehends a vector of variables and evaluates coefficients or weights for each input variable and then predicts the class of stated fake news on COVID-19 in the form of a word vector [21]. Logistic Regression can be used only when the dependent variable is dichotomous (binary).

Decision Tree

Decision Trees are one the powerful methods commonly used in various fields, such as machine learning, image processing and identification of patterns. Decision Tree chooses the best characteristic which produces the highest information gain at each node and this process continuous recursively until all the leaf nodes become pure or when no additional classification is required [22].

Adaboost

The AdaBoost algorithm, introduced in 1995 by Freund and Schapire, is a variant of the earlier boosting algorithms. Adaboost calls a given weak or based learning algorithm that has large bias repeatedly in sequence to be combined to generate a strong learner with low bias [23]. In general, Adaboost is a specific boosting algorithm that turns a weak classifier into a strong classifier.

4.6 Model Deployment

I developed a final product for this project by using HTML, CSS and Bootstrap as frontend and Django framework as the backend. After that, the website was deployed and hosted using Heroku. The screenshot of the website homepage is shown in Figure 15 below. The documentation of the website can be found here click here.

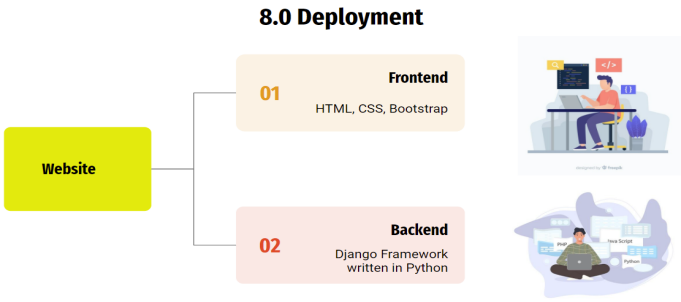

Figure 14: Deployment

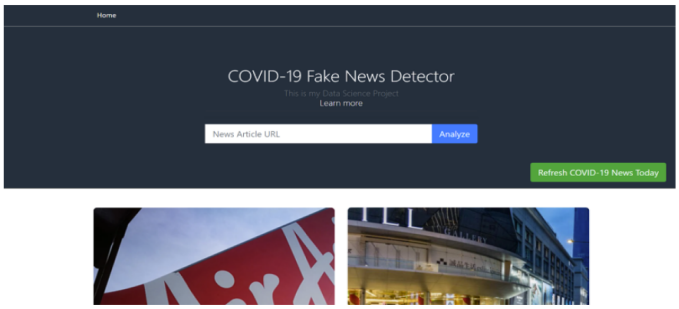

Figure 15: Website

V. Results

Abbreviations Used in Table 2-5

MNB	Multinomial Naive Bayes
SVM	Support Vector Machines
DT	Decision Tree
RF	Random Forest
LR	Logistic Regression
Adaboost	Adaboost

Table 2: Accuracy, Precision, Recall, AUC, F1-score of BoW

Bag of Words(BoW)					
Model	Accuracy	Precision	Recall	AUC	F1-score
MNB	0.92	0.96	0.87	0.92	0.91
SVM	0.96	0.96	0.96	0.96	0.96
DT	0.93	0.94	0.93	0.93	0.93
RF	0.97	0.98	0.95	0.97	0.97
LR	0.96	0.95	0.97	0.96	0.96
Adaboost	0.96	0.96	0.96	0.96	0.96

Table 3: Accuracy, Precision, Recall, AUC, F1-score of TF-IDF

TF-IDF					
Model	Accuracy	Precision	Recall	AUC	F1-score
MNB	0.88	0.96	0.79	0.88	0.87
SVM	0.95	0.98	0.92	0.95	0.95
DT	0.94	0.96	0.92	0.94	0.94
RF	0.97	0.98	0.96	0.97	0.97
LR	0.95	0.95	0.96	0.95	0.95
Adaboost	0.96	0.96	0.96	0.96	0.96

Table 4: Accuracy, Precision, Recall, AUC, F1-score of Word2Vec

Word2Vec					
Model	Accuracy	Precision	Recall	AUC	F1-score
MNB	0.90	0.90	0.92	0.90	0.90
SVM	0.92	0.96	0.88	0.92	0.92
DT	0.85	0.85	0.84	0.85	0.84
RF	0.92	0.97	0.85	0.92	0.91
LR	0.94	0.96	0.92	0.94	0.94
Adaboost	0.93	0.95	0.91	0.93	0.93

Table 5: Accuracy, Precision, Recall, AUC, F1-score of GloVe

GloVe					
Model	Accuracy	Precision	Recall	AUC	F1-score
MNB	0.90	0.89	0.91	0.90	0.90
SVM	0.88	0.98	0.78	0.88	0.86
DT	0.90	0.90	0.90	0.90	0.90
RF	0.92	0.97	0.87	0.92	0.92
LR	0.95	0.96	0.94	0.95	0.95
Adaboost	0.96	0.98	0.93	0.96	0.95

For evaluation purposes of performance, we have computed five metrics - Accuracy, Precision, Recall, Area Under the ROC curve, and F1-score. According to the results obtained in this study, to analyze the fake news on COVID-19 using machine learning techniques, we used 6 different classifiers: Multinomial Naive Bayes(MNB), Support Vector Machines (SVM), Random Forest (RF), Logistic Regression (LR), Decision Tree (DT), and Adaboost. For each of these classifiers, we trained it using 10-fold cross validation on 4 different word embeddings: Bag of Words (BoW), Term Frequency - Inverse Document Frequency (TF-IDF), Word2Vec, and Global Vectors (GloVe). From the comparison of the different measures, we find that Random Forest performed the best on Bag of Words, TF-IDF, and Word2Vec, while Adaboost performed the best on GloVe. Overall, the best performing classifier is Random Forest on TF-IDF. Consequently, we find that bagging algorithms like Random Forest and boosting

algorithms like Adaboost are generally efficient in COVID-19 fake news classification problems. For all the experiments in this section, the performance shown is based on the mean results of the Out of Bag(OOB) data used in 10-fold cross validation.

VI. Discussion

The findings of all machine learning models can be seen in the aforementioned figures and tables. All of the classifiers performed greatly, even though certain classifiers outperformed others in terms of accuracy, precision, recall, AUC, and F1-score. We have evaluated the trained models using 10-fold cross validation on the entire dataset using 4 different word embeddings. The performance of each of the models on Out of Bag (OOB) data was recorded and compared. Random Forest outshined all the other classifiers on 3 word embeddings: Bag of Words, TF-IDF and Word2Vec, while Adaboost reigns supreme on Global Vectors (GloVe) embedding. Taking into account the accuracy, precision, recall, AUC, and F1-score, Random Forest on TF-IDF came out best with accuracy, precision, recall, AUC, and F1-score of 0.97, 0.98, 0.96, 0.97 and 0.97 respectively. The second best classifier is Adaboost on TF-IDF with the scores of 0.96 on accuracy, precision, recall, AUC and F1-score.

VII. Conclusion

In our research, six machine learning algorithms - Naive Bayes, Support Vector Machines, Random Forest, Logistic Regression, Decision Tree, and Adaboost, are used to analyze and detect COVID-19 fake news in Malaysia. We conducted different techniques on the Malaysian COVID-19 dataset that we scraped from 2 reliable sources, namely The Star website and Sebenarnya.my website. Furthermore, the models have been analyzed in the aspects of accuracy, precision, recall, AUC, and F1-score.

In the future, we plan to use a large dataset, and include COVID-19 news from various websites, social messaging platforms and social media like Twitter.

VIII. Acknowledgement

We gratefully thank Associate Prof. Ts. Dr. Vimala Balakrishnan & Associate Prof. Dr. Salimah Mokhtar for their supervision in this data science project. I learned a lot from this project as it improved my research skills from conducting this research project under their supervision.

References

[1] Domenico, G. D., Sit, J., Ishizaka, A., Nunan, D. (2021). Fake news, social media and marketing: A systematic review. Journal of Business Research, 124, 329–341. doi:10.1016/j.jbusres.2020.11.037

[2] Zhou, Xinyi Zafarani, Reza Shu, Kai Liu, Huan. (2019). Fake News: Fundamental Theories, Detection Strategies and Challenges. 10.1145/3289600.3291382.

[3] Hao, K. and Basu, T. (2020) The coronavirus is the first true social-media 'infodemic', MIT Technology Review. Available at: https://www.technologyreview.com/s/615184/thecoronavirus-is-the-first-true-social-media-infodemic/ (Accessed: 15 April 2020).

[4] Emergency Ordinance not aimed at restricting freedom of speech, protects affected parties — MCMC: Malaysian Communications and Multimedia Commission (MCMC), Malaysian Communications and Multimedia Commission (MCMC) — Suruhanjaya Komunikasi Dan Multimedia Malaysia (SKMM), 2021 accessed January 19, 2021, https://www.mcmc.gov.my/en/media/press-clippings/emergency-ordinance-not-aimed-at-restricting-freed

[5] Balakrishnan, V., Ng, K. S., Rahim, H. A. (2021). To share or not to share – The underlying motives of sharing fake news amidst the COVID-19 pandemic in Malaysia. Technology in Society, 66, 101676. doi:10.1016/j.techsoc.2021.101676

[6] Mazzeo, V., Rapisarda, A., Giuffrida, G. (2021). Detection of Fake News on COVID-19 on Web Search Engines. Frontiers in Physics, 9, 351. doi:10.3389/fphy.2021.685730.

[7] S. Kannan, S. Saravanan, P. Chandirasekeran and S. Rani Patra, "Detection of Fake News related to COVID-19 using Natural Language Processing," 2021 Asian Conference on Innovation in Technology (ASIANCON), 2021, pp. 1-6, doi: 10.1109/ASIANCON51346.2021.9545047.

[8] A. Priya and A. Kumar, "Deep Ensemble Approach for COVID-19 Fake News Detection from Social Media," 2021 8th International Conference on Signal Processing and Integrated Networks (SPIN), 2021, pp. 396-401, doi: 10.1109/SPIN52536.2021.9565958.

[9] Bojjireddy, S., Chun, S. A., Geller, J. (2021). Machine Learning Approach to Detect Fake News, Misinformation in COVID-19 Pandemic. DG.O2021: The 22nd Annual International Conference on Digital Government Research, 575–578. Presented at the Omaha, NE, USA. doi:10.1145/3463677.3463762

[10] Isaakidou, Marianna Zoulias, Emmanouil Diomidous, Marianna. (2021). Machine Learning to Identify Fake News for COVID-19. 10.3233/SHTI210130.

[11] Bojjireddy, Sirisha Chun, Soon Geller, James. (2021). Machine Learning Approach to Detect Fake News, Misinformation in COVID-19 Pandemic. 575-578. 10.1145/3463677.3463762.

[12] M. H. Saghayan, S. F. Ebrahimi and M. Bahrani, "Exploring the Impact of Machine Translation on Fake News Detection: A Case Study on Persian Tweets about COVID-19," 2021 29th Iranian Conference on Electrical Engineering (ICEE), 2021, pp. 540-544, doi: 10.1109/ICEE52715.2021.9544409.

[13] A. Ul Hussna, I. I. Trisha, M. S. Karim and M. G. R. Alam, "COVID-19 Fake News Prediction On Social Media Data," 2021 IEEE Region 10 Symposium (TENSYMP), 2021, pp. 1-5, doi: 10.1109/TENSYMP52854.2021.9550957.

[14] Kaliyar, Rohit Goswami, Anurag Narang, Pratik. (2021). A Hybrid Model for Effective Fake News Detection with a Novel COVID-19 Dataset. 1066-1072. 10.5220/0010316010661072.

[15] C. Rajalakshmi, T. Subika, K. Vaishali and J. Shana, "Fake News Prediction On COVID Dataset Using Machine Learning," 2021 12th International Conference on Computing Communication and Networking Technologies (ICCCNT), 2021, pp. 1-7, doi: 10.1109/ICCCNT51525.2021.9579543.

[16] Ul Hussna, I. I. Trisha, M. S. Karim and M. G. R. Alam, "COVID-19 Fake News Prediction On Social Media Data," 2021 IEEE Region 10 Symposium (TENSYMP), 2021, pp. 1-5, doi: 10.1109/TENSYMP52854.2021.9550957.

[17] A. Singh, N. Thakur and A. Sharma, "A review of supervised machine learning algorithms," 2016 3rd International Conference on Computing for Sustainable Global Development (INDIACom), 2016, pp. 1310-1315.

[18] D. Lewis, "Naıve (Bayes) at forty: The independence assumption in information retrieval", Journal of Machine Learning, pp. 4–15, 1998.

[19] M. A. Chandra and S. S. Bedi, "Survey on SVM and their application in image classification," International Journal of Information Technology, vol. 13, 2020.

[20] Y. Al Amrani, M. Lazaar, and K. E. El Kadirp, "Random forest and support vector machine based hybrid approach to sentiment analysis," Procedia Computer Science, vol. 127, pp. 511– 520, 2018.

[21] A. Prabhat and V. Khullar, "Sentiment classification on big data using Naïve bayes and logistic regression," 2017 International Conference on Computer Communication and Informatics (ICCCI), 2017, pp. 1-5, doi: 10.1109/ICCCI.2017.8117734.

[22] S. Umadevi and K. S. J. Marseline, "A survey on data mining classification algorithms," in 2017 International Conference on Signal Processing and Communication (ICSPC), pp. 264– 268, Coimbatore, India, 2018.

[23] Freund, Yoav, and Robert E. Schapire. "A Short Introduction to Boost-

ing." In Proceedings of the Sixteenth International Joint Conference on Artificial Intelligence, Morgan Kaufmann, 1999, pp. 1401–06.