AL-Balqa Applied University

Faculty of Engineering

Computer Engineering Department

COVID-19 Fake News Detection Using Natural Language Processing

A Project Report Submitted in Partial Fulfillment of B.Sc. Degree in Computer Engineering

Prepared By:

Sarah Mahmoud Bani Issa

Rua'a Yahia Eid

Supervised By:

Dr. Abeer Al-Hyari

Al-Salt, Jordan

AUGUST 2020

EXAMINATION COMMITTEE SIGNATURE

We certify that we have read the project entitled “COVID-19 Fake News Detection Using Natural Language Processing” and as an examining committee, examined the students in its context and that in our opinion it is adequate with_____ standing as a graduation project for the degree of B.Sc. in Computer Engineering.

Supervisor Name

Dr. Abeer Al-Hyari

Committee Member Name

_____

Committee Member Name

_____

ABSTRACT

The misinformation related to Covid-19 is considered a huge threat, since it's becoming more popular, and people are confused about what should they believe or not, and how exactly they can have an idea of what can fake news look like.

Therefore, this project is about analyzing both fake and real news related to COVID-19 virus using word visualization tool [word cloud], and NLP techniques[word sentiment and topic modeling];to train a model that is able to discriminate between real and fake news, and to predict whether a given news is fake or real. This project compares the performance of three machine learning algorithms, namely: logistic regression, multinomial Naive Bayes, and support vector machine. The results of the three algorithms are compared to select the best algorithm among them.

ACKNOWLEDGEMENT

In this short acknowledgment, we would like to thank our instructors in Engineering Department for this rich long educational journey that improved both our knowledge and personality, specifically Dr. Abeer who helped us in this project, we were grateful to have you.

Dedication

To all those who helped us when we needed them most.

Our families and friends

Our Teachers

Chapter 1 Contents

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Chapter 1

1.1 Introduction

This project is about detecting news about Corona virus that spread like wildfire around the world using machine learning and Natural Language Processing (NLP) techniques for analyzing the data and building a model that can distinguish whether certain news is real or fake.

1.2 Project idea

Fake news has flooded the social media highly impacting our society; it reshapes people's opinions, and confuses their knowledge, sometimes fake news can affect voting behavior like it did in 2016 US elections. Unfortunately, people tend to blindly share news before checking them, or some entities that intentionally spread it for their own political or economic goals.

When it comes to corona virus, fake news and misinformation led people to disobey the safety rules which caused more infections and deaths, they thought that their government wanted to direct them by a virus conspiracy, or it had something to do with the five G technology, or they don't need a mask in public gatherings, and other misinformation that were spread online.[6]

1.3 Project importance

After we witness the huge consequences of corona virus fake news in the whole world, and how dangerous it could be, we have decided to contribute in preventing it; this project is presenting an idea to handle this problem, for helping the society, so it won't be a victim of false information, reducing the effect of fake news and marking it.

1.4 Objectives of this project

The aims of this project:

- Distinguish whether given news is fake or real.
- Find the patterns of both fake and real news.

CHAPTER TWO

BACKGROUND

Chapter 2

2.1 Introduction:

In this chapter, machine learning algorithms and natural language processing will be explained. Also, some of the main research that has been done in this domain is represented in the literature review section.

Machine Learning:

Basically, it is Field of study that gives computers the ability to learn without being explicitly programmed. It is about giving a training data to a specific learning algorithm, let the computer analyze it and give an output.[7]
There are three general Types of machine learning:

- **Supervised Learning**: where we give the computer labeled data with a desired output, and then the algorithm can predict the output for other unlabeled data, the accuracy of the model when given new unseen data.

- **Unsupervised Learning**: the given data is not labeled, and the machine is asked to identify patterns and classify the data.

- **Reinforcement learning**: where the learning algorithm interacts with dynamic environment that gives it feedback in terms of rewards and punishments, like self-driving cars being rewarded to stay on the road.

Natural Language Processing:

The branch of AI which deals with the interactions between computers and humans using Natural languages is NLP, it processes text data to extract the meaning of these texts, for example: if you are a client and wrote a feedback comment to the company that provides you with a specific service, NLP should eventually know whether your feedback is positive or negative; therefore NLP is a basic element of virtual assists like Siri.[8]

NLP is loaded with different techniques that serve different objectives; the following is the basic techniques:

- **Topic Modeling:** is a complicated method for identifying the natural topics in a text.
- **Sentiment Analysis (polarity):** is a technique for finding the polarity of a text between -1 (negative) and 1(positive).
- **Sentiment Analysis (subjectivity):** a technique for finding if a certain text is subjective or opinioned in range of 0 fact to 1 very opiniened.
- **Text Generation:** The process of transforming structured data into natural language.
- **Text Summarization:** is a method for extracting the important words from the text, removing all unnecessary words with less meaning.
- **Named Entity Recognition (NER):** Identifies entities such as people, locations, organizations, dates, etc. from the text.

2.2 Software requirements

Programming language used in this project: Python 3.7

Table2.1: list of python basic libraries used in this project

Basic libraries used	Library functions
pandas	data manipulation and analysis. [10]
sklearn	various classification, regression, and clustering algorithms. [9]
nltk	a suite of libraries and programs for symbolic and statistical natural language processing for English written. [14]
matplotlib	creating statics, animated, and interactive visualizations in Python. [11]

numpy	high-level mathematical functions to operate on arrays. [12]
wordcloud	performing visual representation of text data. [13]
string	manipulation on text data.
csv	reading and writing on csv files. [15]

2.3 Literature Review

Analysis and Detection of Health-Related Misinformation on Chinese Social Media

The data set was collected from both reliable and unreliable health-related articles from multiple Chinese online social media , then analyzed the differences with respect to writing style, text topic and feature distribution by both intuitive and statistical analysis, selected manually 104 linguistic and statistical features that are useful for machine learning classifiers lastly proposed a Health-related Misinformation Detection framework (HMD) that includes a feature-based method and a text-based method.[1]

Facebook Company

Facebook already applies machine learning algorithms to detect sensitive content. Though fallible, this software goes a long way toward ensuring that photos and videos containing violence and sexual content are flagged and removed as swiftly as possible.
the company is set to use similar technologies to identify false news and take own action on a bigger scale.[2]

Detecting Fake News in Social Networks

A Dataset collection of approximate seventeen thousand news articles extracted from online, also social media platforms
The project consists of Rumor Detection / Clickbait Detection / Email Spam Detection / Truth Discovery and Hot Topic Detection. [3]

Fake News Analysis Modeling Using Quote Retweet

This study proposes a fake news analysis modeling method by identifying a variety of features and collecting various data from Twitter, The method proposed in this study can increase the accuracy of fake news analysis by acquiring more potential information from the Quote Retweet feature added to Twitter in 2015, compared to the more conventional and common Retweet only. Furthermore, fake news was analyzed through neural network-based classification modeling by using the preprocessed data and the identified best features in the learning data [4]

Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator

In this project, the results of a fake news identification study that documents the performance of a fake news classifier are presented. The Textblob, Natural Language, and SciPy Toolkits were used to develop a novel fake news detector that uses quoted attribution in a Bayesian machine learning system as a key feature to estimate the likelihood that a news article is fake. The resultant process precision is 63.333% effective at assessing the likelihood that an article with quotes is fake.[5]

CHAPTER THREE

METHODOLOGY

Chapter 3

3.1 Flow Chart

We have two flow charts: one for data analysis and the second for the machine learning algorithms.

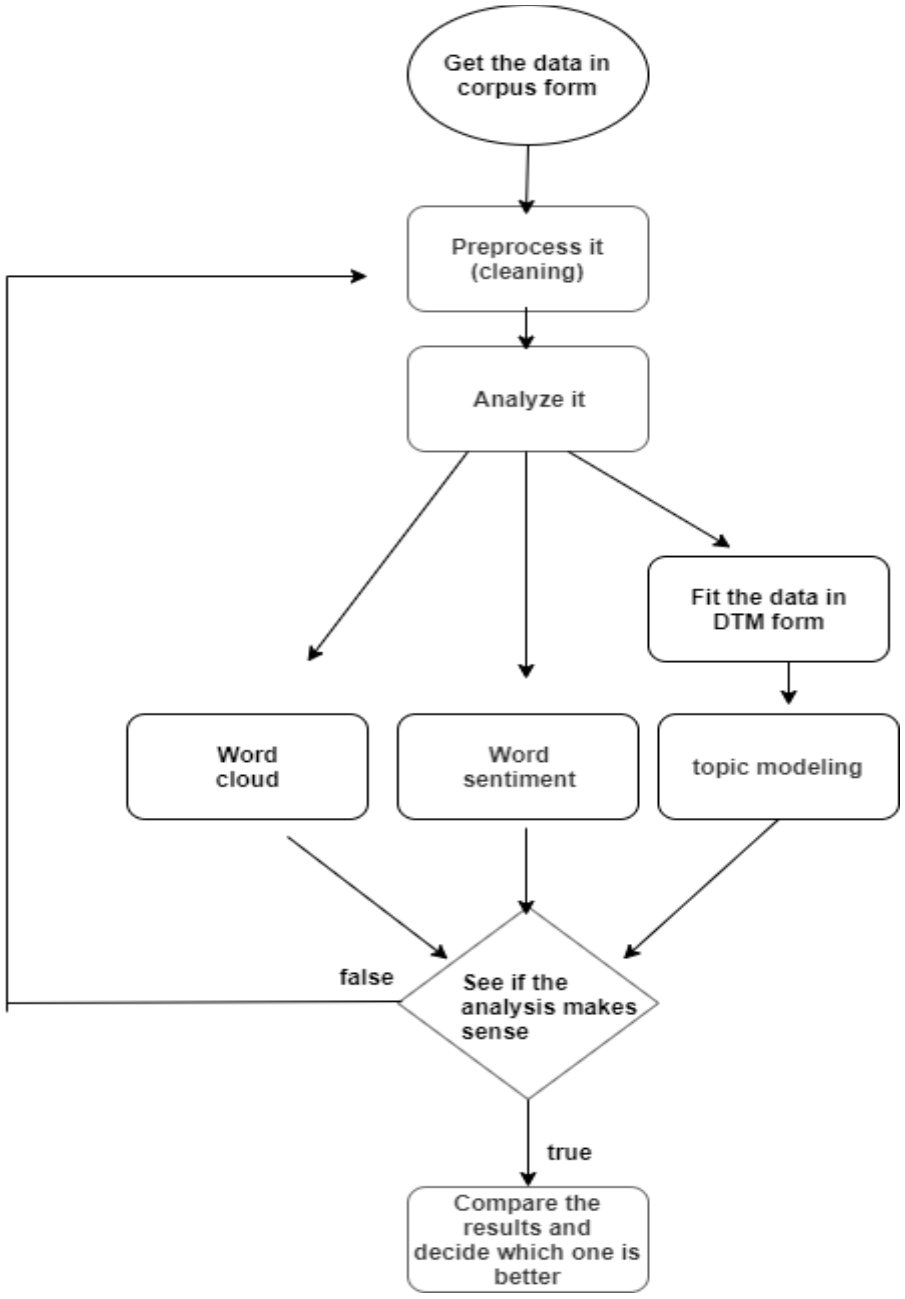

Figure 3-1: data analysis Flow chart

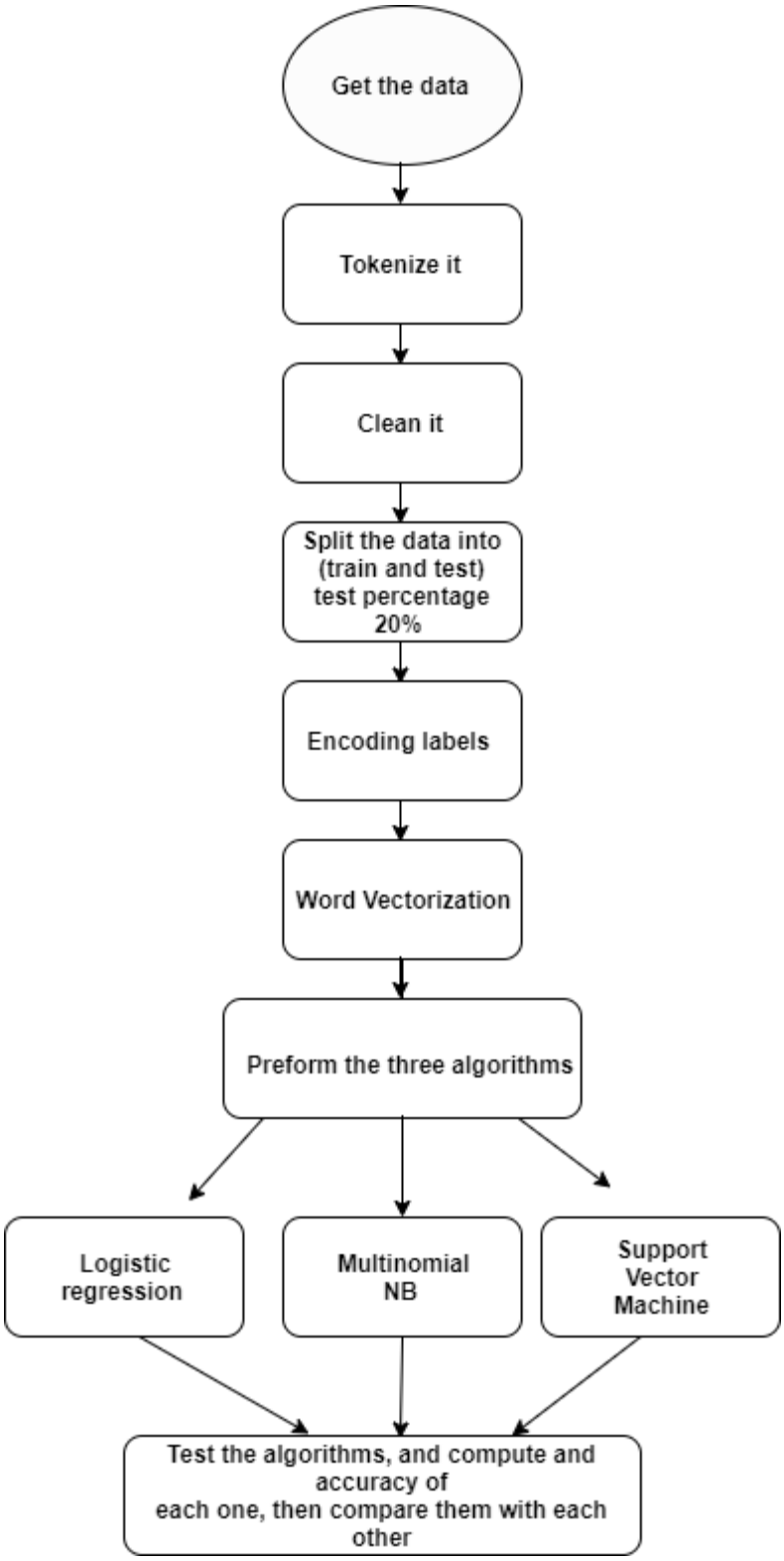

Figure 3-2: Machine learning algorithms Flow chart

3.2 The Data:

The data was obtained from a GitHub project [16], with a size of 42KB 74 x 4 for training models classified into two labels (fake and real), and 36KB of corpus data for analysis tools.

3.3 Data Preprocessing:

The data on its ordinary form is not suitable for machine learning algorithms, its need to be organized and preprocessed.
And there are two standard data forms:

- **Corpus**: a collection of text.
- **Document-Term Matrix** - word counts in matrix format.

Texts are really the important when analyzing the news, corpus form will only include [texts and labels].

The data must not have any kind of duplication, punctuation marks, numbers, stop words, words with similar meanings like [play, playing],the data must also be in lower case to become suitable for analyzing.[19]

For data analysis, cleaning the data and put it in corpus form was enough, as for machine learning algorithms there are more steps yet, since the machine don't understand natural language, Text data must go through a two operations so it can be fitted in the classifiers:

- **Encoding label column:** converting text data into numerical values.
- **Word vectorization for texts column**: or **word embedding** is a set of techniques that aims at extracting information from a text corpus and associating to each one of its word a vector like [-1,2, -4].

3.4 Analysis Tools:

- **Word Cloud:**

  Is a technique for visualizing frequent words in a text where the size of the words represents their frequency as a compressed image of words with different sizes and colors, most repeated words appears in bigger sizes. It is an easy way to show what a text mostly talks about. In this project word cloud was used to

explore what both fake and real news usually discuss, by viewing most frequented words in both labels.

- **Word Sentiment:**
  Is a NLP machine learning technique that detects polarity and subjectivity like a positive or negative opinion within text. Its mostly used in customers feedbacks about a certain product or service, identifying whether this feedback is a positive or negative one, also for further analysis subjectivity is used, to identify how much emotions are in the text, for example: if you are reading comments in a post on facebook, it's likely pinions, whereas reading a news article from a news website, or some president speech about improving the country's economy, it's likely a subjective article have much less emotions. Word sentiment was chosen in this project to discover how polarity and subjectivity the news can be, to see the differences between the results for fake and real Covid-19 news.

- **Topic Modeling:**
  Is a machine learning technique that automatically analyzes text data to determine cluster words for a set of documents, this is known as 'unsupervised' machine learning because it doesn't require a predefined list of tags or training data that's been previously classified by humans. The main goal of topic modeling is to know more about the hidden topics in the texts, which gives an idea about the inner structure for it.

3.5 Algorithms Used:

Three different classification algorithms were chosen and tested on the dataset, and each one's results will be evaluated, and then compared; to see which algorithm gives the highest accuracy score. The chosen algorithms are:

- **Multinomial Naive Bayes:**
  Naive Bayes is a machine learning algorithm used for classification, it's usually used when we have independent features; Whereas Multinomial NB is specific instance of Naive Bayes, which uses a multinomial distribution for each of the

features, multinomial NB works well for data which can easily be turned into counts, such as word counts in text, which fits our data. [17]
This is a particularly strong hypothesis in the case of text classification because it supposes that words are not related to each other. But it knows to work well given this hypothesis.
Given an element of class y and vector of features X = (x_1,.., x_n). The probability of the class given that vector is defined as

$$P(y|\mathbf{X}) = \frac{P(y) * P(\mathbf{X}|y)}{P(\mathbf{X})}$$

Thanks to the assumption of conditional independence, we have that

$$P(x_i|y, x_1, ..., x_{i-1}, x_{i+1}, ..., x_n) = P(x_i|y)$$

Using Bayes rules, we have that

$$P(y|x_1, ..., x_n) = \frac{P(y)\prod_{i=1}^{n} P(x_i|y)}{P(x_1, ..., x_n)}$$

Because P (x_1, ..., x_n) is constant, we have the classification rule

$$\hat{y} = \underset{y}{argmax} P(y) \prod_{i=1}^{n} P(x_i|y)$$

- **Binary Logistic Regression:**
  Logistic Regression is a special case of linear regression used to classify the input data into binary categories, in this project [fake or real]. The reason we chose Binary logistic regression is that we only want two possible outputs [fake (0) or real (1)]. [18]
  Linear regression which outputs continuous number values, logistic regression changes its yield utilizing the calculated sigmoid capacity to restore likelihood esteem which would then be able to be mapped to at least two discrete classes. The LR model uses gradient descent to converge onto the optimal set of weights (θ) for the training set. For our model, the hypothesis used is the sigmoid function:

$$h_\theta(x) = g(\theta^T x) = \frac{1}{1 + exp(-\theta^T x)}$$

- **Support Vector Machine (SVM):**

  SVM is a classification algorithm which produces significant accuracy with less computation power, Support Vector Machine can be used for both regression and classification tasks, but it is widely used in classification objectives. [20]
  SVM are founded on the idea of finding a hyperplane that best divides a dataset into two classes. Support vectors are the data points nearest to the hyperplane, the points of a data set that, if deleted, would alter the position of the dividing hyperplane. Because of this, they can be considered the critical elements of a data set. The distance between the hyperplane and the nearest data point from either set is known as the margin. The aim is to choose a hyperplane with the greatest possible margin between the hyperplane and any point within the training set, giving a higher chance of new data being classified correctly.

CHAPTER FOUR

RESULTS

Chapter 4

4.1 Introduction:

This chapter discusses the results of data analysis and classification algorithms, It also compares the results together to find out which analysis method, machine learning classifier is the most appropriate among other selected algorithms.

4.2 Results

Data analysis:

Word cloud:

Figure 4-1: fake news word cloud

Figure 4-2: real news word cloud

Word sentiment:

In this project both polarity and subjectivity were applied on corpus data.

Polarity is float which lies in the range of [-1, 1] where -1 means negative, 1 means positive.
Subjectivity is also a float which lies in range of [0, 1] where 0 is very objective (fact), 1 is very subjective (opinion).

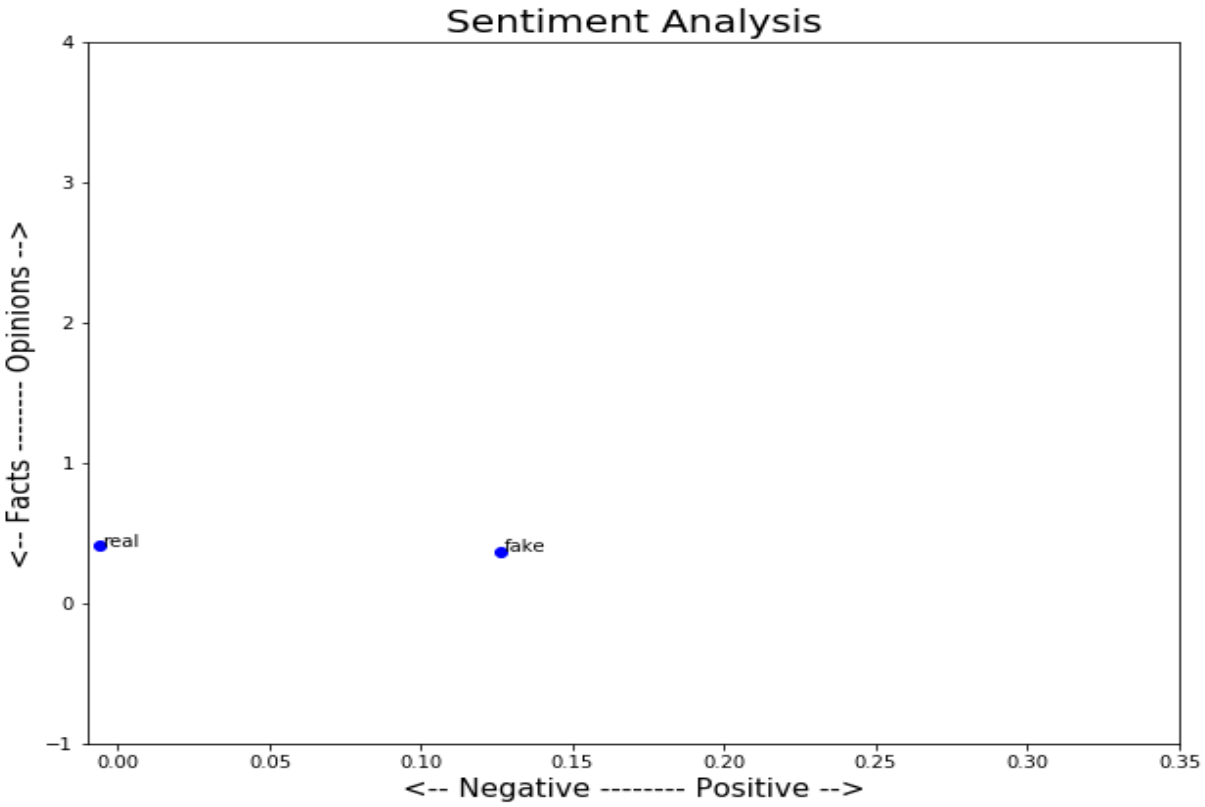

Figure 4-2: word sentiment for both fake and real news

From the plot it is obvious that real news tends to be very negative about covid-19 virus, while fake news is less negative about it.

Topic modeling:

The best analysis for topic modeling in this project when nouns and adjectives were involved only, with number of topics is two, after analysis the results were as the following:

Table 4-1: The two topics extracted from corpus data using topic modeling

	Most repeated words	Topic overview
Topic 0	Infect- kit symptom- person- hospit-bank- posit-south-case	Infections, hospitals and symptoms
Topic 1	Monitor - care- vaccine - corona-record-fall-immune-unnecessary	Vaccine ,care and immuney system

The final results were as following:

Real news were included in topic 0, while fake news were in topic 1.

4.3 Comparison

Data analysis:

Table 4-2: comparison between the three data analysis methods

Data analysis method	Feedback
Word cloud	Promising results, showed most frequented words mentioned in each label.
Data sentiment (Polarity)	Acceptable results, but not enough alone.
Data sentiment (subjectivity)	Not useful at all, since it showed no differences between fake and real news (both were classified as facts).
Topic modeling	Promising and helpful results for exploring the inner structure for both fake and real news.

Machine learning algorithms:

Table 4-3: Comparison between the three algorithms

Classification algorithm	Accuracy score
Logistic Regression	66.6%
Multinomial NB	73.3%
Support vector machine	73.3%

The difference between logistic regression and multinomial NB is that logistic regression for dependent variables, one the other hand multinomial NB assumes that variables are independent.
Support vector machine tries to finds the “best” margin (distance between the line and the support vectors) that separates the classes and this reduces the risk of error on the data, while logistic regression does not, instead it can have different decision boundaries with different weights that are near the optimal point.

CHAPTER FIVE

CONCLUSION and FUTURE WORK

Chapter 5

5.1 Project Evaluation

This section evaluates the different data analysis and machine learning algorithms that were used in this project.

Data analysis:

Table 5-1: Data analysis overview.

Data analysis method	Overview	Reasons why
Word cloud	Successful	Showed repeated words in each news label, which gave an idea what the news talks about the most.
Data sentiment	Failed	Subjectivity distinguishes between objective and opinion sentiments, since fake and real news are written in formal language, subjectivity fails to know the difference.
Topic modeling	Successful	Classifying the data into two topics enabling to enclose each label in one topic to know in general what each label talks about.

Machine Learning Algorithms:

Table 5-2: Machine learning algorithms overview

Machine learning algorithm	Overview	Reasons why
Logistic regression	poor accuracy with 66.6% accuracy	supervised classifier for dependent variables, can have different decision boundaries, with different weights that are near the optimal point
Multinomial NB	Successful with 73.3% accuracy	Because it's suitable for classification with discrete features (e.g., word counts for text classification)
Support vector machine	Successful with 73.3% accuracy	SVM is specified classifier for two group problems which is suitable in fake/ real news classification problem. it is also try to find the best decision line between the two group.

5.2 Conclusions

The following remarks can be concluded after finishing this project:

- For data analysis, both word cloud and topic modeling are suitable for analyzing and recognizing the structure of news corpus data.
- For machine learning algorithms, both multinomial NB and support vector machine algorithms are acceptable, whereas logistic regression showed lower accuracy results.

5.3 Future work

This project can be further improved by applying the following:

- Add more data related to COVID-19, and feeding it to the three training models.
- Perform more data cleaning steps.
- Explore the performance of other classifiers on the same dataset.

References

[1] Yue Liu, Ke Yu , Xiaofei Wu , Linbo Qing , and Yonghong Peng Analysis and Detection of Health-Related Misinformation on Chinese Social Media
[2] Yonghun Jang,Chang-Hyeon ParkYeong-SeokSeo(2019)Fake News Analysis Modeling Using Quote Retweet
[3] SajjadAhmed , Knut Hinkelmann, Flavio Corradini, Combining Machine Learning with Knowledge Engineering to detect Fake News in Social Networks-a survey
[4]Terry TraylorJeremyStraub ; Gurmeet_; Nicholas Snell-Classifying Fake News Articles Using Natural Language Processing to Identify In-Article Attribution as a Supervised Learning Estimator
[5]https://internetofbusiness.com/facebook-machine-learning-fake-news/
[6]https://www.bbc.com/news/stories-52731624
[7]https://www.internetsociety.org/resources/doc/2017/artificial-intelligence-and-machine-learning-policy-paper/?gclid=EAIaIQobChMIxtrvhuPN6gIV2ON3Ch3HtgSoEAAYAiAAEgItBPD_BwE
[8]https://blog.aureusanalytics.com/blog/5-natural-language-processing-techniques-for-extracting-information
[9]https://scikit-learn.org/stable/
[10]https://pandas.pydata.org/
[11]https://matplotlib.org/
[12]https://numpy.org/
[13]https://pypi.org/project/wordcloud/
[14]https://www.nltk.org/
[15]https://docs.python.org/3/library/csv.html
[16]https://github.com/obresearch19/covid19-Fake-News-Check
[17]https://www.machinelearningplus.com/predictive-modeling/how-naive-bayes-algorithm-works-with-example-and-full-code/#:~:text=Naive%20Bayes%20is%20a%20probabilistic%20machine%20learning%20algorithm%20that%20can,classifying%20documents,%20sentiment%20prediction%20etc.&text=The%20name%20naive%20is%20used,is%20independent%20of%20each%20other.
[18]https://towardsdatascience.com/logistic-regression-detailed-overview-46c4da4303bc
[19]https://pythonhealthcare.org/2018/12/14/101-pre-processing-data-tokenization-stemming-and-removal-of-stop-words/
[20]https://medium.com/@bedigunjit/simple-guide-to-text-classification-nlp-using-svm-and-naive-bayes-with-python-421db3a72d34

PS: **All links were explored in AUGUST 2020**