5-2023

A STUDY OF VARIOUS DATA SIZES USING MACHINE LEARNING

Sochaeta Koeum

A STUDY OF VARIOUS DATA SIZES USING MACHINE LEARNING

A Project

Presented to the

Faculty of

California State University,

San Bernardino

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

in

Information Systems and Technology

by

Sochaeta Koeum

May 2023

A STUDY OF VARIOUS DATA SIZES USING MACHINE LEARNING

A Project

Presented to the

Faculty of

California State University,

San Bernardino

---

by

Sochaeta Koeum

May 2023

Approved by:

Dr. Conrad Shayo, Committee Chair

Dr. Bailey Benedict, Committee Member

Dr. Conrad Shayo, Department Chair, Information Decision Sciences

ABSTRACT

Social media is a great domain for news consumption; however, it is referred to as a double-edged sword. While it is user-friendly and low-cost, social media is the reason why fake news can spread rapidly, which is detrimental to society, businesses, and many consumers. Therefore, fake news detection is an emerging field. However, some challenges have restricted other researchers from developing a universal machine learning model that is fast, efficient, and reliable to stop the proliferation because of the lack of resources available, such as large-sized datasets. The goal of this culminating experience project is to explore how varying datasets sizes affect the accuracy percentage of a machine learning model. The research questions are: Q1) How do large volumes of fake news datasets affect the accuracy percentage of a machine learning model? Q2) As one increases the volume of data fed into a machine learning model from small to large datasets, what will the cutoff accuracy percentage point be? Various data sizes collected from Kaggle were fed into the machine learning model, Naïve Bayes, to help answer the two questions. Then, all three datasets were combined together to see if the accuracy of the model improves as more data is fed into the model. The results and findings for each question are; 1) Larger dataset sizes do increase the accuracy percentage because there is more data to train and test on. 2) The cutoff accuracy is dependent on the number of unique values within the dataset. Since it is not finite, we can expect that large dataset sizes to have a cutoff accuracy of above 90%, given that the data is

cleaned and pre-processed. Compared to a data size ranging from small to medium, it will achieve an accuracy score of around 70%-90%. An accuracy score below 70% means that the model is highly unreliable and that the dataset size is too small. For instance, Dataset 1 achieved an accuracy score of 66%, Dataset 2 was 83%, and Dataset 3 was 92%. To effectively study and experiment on how to build an optimized model, one must use a large dataset size for analysis. Furthermore, other areas for future studies that appeared from this study are building a new and improved fact-checking website that quickly and accurately processes large databases and how well the model, Naïve Bayes works with different modalities, such as images and videos.

ACKNOWLEDGEMENTS

As a California State University – San Bernardino graduate student, I would like to express deep gratitude towards Dr. Bailey Benedict and Dr. Conrad Shayo. With their help and guidance, I was able to finish this project on time. I truly appreciate their time, efforts, and support during this journey.

DEDICATION

I would like to dedicate this project to my parents, Samouth and Chhorvy Koeum, who have worked hard and made many sacrifices to ensure that I have a better future. They have been nothing but supportive throughout this arduous journey. Also, I would like to give a special shoutout to my sister, Monika Koeum. Without her help and support, I do not think this would be possible. Lastly, I would like to thank my mentors who encouraged me to pursue my Master's degree and my friends who have been by my side.

.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

CHAPTER ONE

INTRODUCTION

Brief Background

Within the last 30 years, the Internet has developed into the world's primary source of information and communication. In 2022, social media had an increase in usage to about 4.89 billion people (Kumar & Singh, 2022). Social media platforms, such as Twitter and Facebook, are the leading source of news. With billions of people acquiring news updates from social media, newspapers and television news editors are compelled to be more present on social media (Kumar & Singh, 2022). However, due to the unmoderated nature of these social platforms, rumors and information can spread at exponential rates, which can lead to dire consequences if not examined cautiously (Hunt et al., 2022). These malicious rumors and misleading information are what society labels as fake news.

Fake News

The spread of fake news is one of the most pressing challenges faced today, as people increasingly access information online. Fake news is fabricated information that is intended to mislead an individual or perpetuate commonly held ideas by a particular political party (Kumar & Singh, 2022). Fake news is detrimental to society and can lead to real-world consequences, such as causing panic, confusion, and violence in politics, academia, the economy, and healthcare (Kumar & Singh, 2022). When the 2016 Presidential Election was approaching, fake news infiltrated all social media platforms, and many argued this was because of Donald Trump (Wright, 2020). Many believe that Donald Trump could not become president if he only persuaded the Republican party (Field, 2021), so he needed to appeal to Independents and Democrats somehow, too. Scholars argue that the only way he could garner support from outside his party was by slandering and making false accusations about Hilary Clinton and the Democratic party, which contributed to his winning of the 2016 presidential election (Field, 2021).

Fake news has also affected the global healthcare. For example, when Covid-19 emerged in the mainstream media in March 2020, rumors of unproven drugs like Ivermectin and dangerous technology lurking in face masks spread rapidly; false information hindered doctors' ability to persuade patients using evidence-based science and online content moderators' ability to persuade the public with accurate health-related information (Madani et al., 2021). The

proliferation of misleading content on Covid-19 confused people and prevented many from getting their vaccines (Madani et al., 2021). These examples demonstrate that social media organizations have historically failed to verify the authenticity of news items; therefore, tweets, videos, and pictures are easily manipulated and can negatively influence the public's opinion (What is Fake News, n.d.).

Stopping the proliferation of fake news has been a primary task for social media organizations, the government, and many other concerned parties. There has been a growing demand to classify and categorize what is real versus fake information. However, the viral spread of false information that has plagued the social media ecosystem has made it harder for people to determine what news to trust online (Collins et al., 2021). A major step in stopping the spread of fake news is finding ways to detect it.

Machine Learning and Algorithms

The abundance of news articles and news-related social media posts has made it difficult for humans to use their analytical abilities to filter out fake news. Fortunately, there have been extensive studies done on machine learning where researchers found that "computational machine learning algorithms have proven useful where data volumes overwhelm human analysis abilities" (Jain & Kashe, 2018, p. 1). Within artificial intelligence (AI), machine learning use data and algorithms (i.e., classifiers) to study the way humans learn to develop software

for computer vision, speech recognition, natural language processing, and other applications (Jordan & Mitchell, 2015). The major goal of machine learning is to automatically train computers by using desired input-output behavior rather than by manually programming the desired response for all inputs (Jordan & Mitchell, 2015).

Machine learning has powered and transformed the digital era in so many ways. For example, “the effect of machine learning has also been felt broadly across computer science and across a range of industries concerned with data-intensive issues, such as consumer services, the diagnosis of faults in complex systems, and the control of logistics chains” (Jordan & Mitchell, 2015, p. 255). For instance, machine learning was used in other disciplines, such as, predicting heart failure disease (Alotaibi, 2019).

Artificial Intelligence (AI) based machine learning algorithms are crucial for machine learning models. They are the backbone of machine learning because they can convert a dataset into a testing and training model, which can then be analyzed and interpreted. However, the lack of an abundance of data on fake news has made it difficult for researchers to assess how well these models perform.

Problem Statement

Fake news detection has gain popularity; however, the shortage of a comprehensive publicly available benchmark dataset makes it harder to study

this topic (Sharma & Garg, 2021). Ahmed et al. (2017) also added fake news detection is a prevalent and challenging topic for researchers because of the lack of resources, such as, datasets and published literature. In this culminating experience project, the outcomes of varying sizes of fake news data will be analyzed to augment for the idea that the availability of large datasets can actuate research in this discipline and produce better prediction models that can help mitigate the issue of fake news.

Research Questions

Q1. How do large volumes of fake news datasets affect the accuracy percentage of a machine learning model?

Q2. As one increases the volume of data fed into a machine learning model from small to large datasets, what will the cutoff accuracy percentage point be?

Objective

The overall objective is to build an optimized model by using the results of different data sizes to determine which is the best to help detect fake news.

Organization of this Project

Chapter 1 provided an overview of the structure of this culminating experience project. In Chapter 2, past literature and related works will be reviewed. The sequential chapter, Chapter 3, covers the research methods that will be implemented to answer the research questions. Chapter 4 consists of the analysis and findings of the research results. Chapter 5 concludes this project with a discussion of the questions, conclusion, and recommendations for further study.

CHAPTER TWO

LITERATURE REVIEW

Many theories and research have been presented to explain the phenomenon of the detection of fake news. This review will focus on three major themes that will help answer my questions. These themes, such as machine learning models, similar experiments and results, and dataset length and diversity, have appeared repeatedly across the literature. Although the literature represents these themes in various contexts, this section will primarily be dependent on the applications of using machine learning to understand how well it works with varying sizes of data.

Machine Learning Models

There are many competing machine learning models that have produced consistent results of the identification of fake news. For instance, Pal et al. (2023) used a variety of machine learning algorithms like Passive Aggressive Classifier, Naïve Bayes, Logistic Regression, Decision Tree, Long short term memory (LSTM) and Bidirectional Encoder Representations from Transformers (BERT) to train their news articles dataset collected from the internet. Their study concluded that Passive Aggressive Classifier performed well on a large dataset, and the Decision Tree algorithm achieved an accuracy of 99.6% and LSTM achieved a recall of 99.8%. Ahmad et al. (2020) used various of ensemble methods by training a combination of different machine learning algorithms, such as Random

Forest, Linear Support Vector Machine (SVM), Decision Trees, Logistic Regression, and LSTM on four different datasets that contained both fake and real news articles from different domains. Their results suggested that the worst performing algorithm was LSTM because the accuracy score was 62%, while the best algorithm was Logistic Regression because the model is fine-tuned with different hyperparameters. Shaikh and Patil (2020) cautioned that resources like datasets are limited. Nevertheless, they presented some solutions for fake news detection by presenting different techniques of classification. With the help of a feature extraction technique, Term Frequency-Inverted Document Frequency (TF-IDF), they used Support Vector Machine (SVM), Naïve Bayes, and Passive Aggressive Classifier to help build their model. They found that the SVM algorithm with the addition of TF-IDF was the most effective combination for fake news detection because the accuracy was 95.05%.

A systematic review by Manzoor and Singla (2019) focused on the methodologies and limitations of machine learning in fake and fabricated news detection. Based on their findings, they concluded that there are various machine learning approaches that were successful to help detect fake news. However, the challenges that continue to arise are the evolution of the characteristics and features of fabricated information. For example, they identified how fake news can exist in different forms such as visual, user, knowledge, style, and stance-based. As technology continues to evolve, it will be harder to detect fake news. Taskin et al. (2022) used Natural Language Processing methods along with SVM

algorithm to detect fake Turkish-language posts. Their study revealed that their F-1 score had a success rate of 0.90 by using SVM.

A study conducted by Pandey et al. (2022) used five different algorithms such as K-Nearest Neighbor, SVM, Decision Tree, Naïve Bayes, and Logistic Regression to detect fake news. In addition to their paper, the data that they were working with had imbalances that were pre-processed to help improve the algorithm's efficiency. In their experiment, KNN had an accuracy score of 89.98%, Logistic Regression was 86.89%, Decision Tree was 73.33%, and SVM was 89.33%

Out of all these competing machine learning models, Naïve Bayes is the best classifier because the procedures are simple and easy to use. Moreover, the proximity and training time of Naïve Bayes are excellent (Albahr & Albahar, 2020). Some of the methods presented above will be used in this culminating experience project, such as using NLP to pre-process the data and the results of others will be used to foreshadow what the Naïve Bayes model should produce in terms of accuracy.

Limitations of Dataset Length and Diversity

Several studies have used Naïve Bayes models to detect fake news but have only used small sample sizes that lacked diversity and length. For example, Qalaja et al. (2022) proposed a study to detect fake news in Covid-19 related posts. They only used one dataset that comprised of 675 tweets collected from

Twitter, of which 540 were used for training and 135 were used for testing. Their model only achieved an accuracy of 72%.

In another study, Granik and Mesyura (2017) only used Facebook as their source of data collection to detect fake news. They collected around 2,000 posts that were trained and tested. Because they collected a slightly larger sample, they achieved an accuracy score of 74%. Like Granik and Mesyura (2017) in terms of data frequency and source, Rachana et al. (2021) only collected Facebook news feeds because they wanted to understand how to differentiate real and fake news. They collected around 2,000 that were used for training and 927 articles for testing: in total 2,927 news posts. Their model achieved an accuracy score of 74%, indicating prediction accuracy remains a challenging task. Rachana et al. (2021) suggested that increasing the data size may increase the accuracy of the prediction model. In addition, Jain et al. (2018) tested the efficiency of the model by using 6,335 Facebook posts that contained 4 columns of information: index, title, text, and label. In this study, they only used the title and text to test their prediction model. For both attributes, they achieved an accuracy score of 80%. These researchers only built their model to detect fake news using one dataset sourced from a single domain. In addition, the length of the dataset was not sufficient, and some models were found to suffer from overfitting (Rachana et al., 2021).

Due to the limited number of resources and small sample sizes, these researchers were unable to produce a model that yielded a high accuracy.

Therefore, three different data sizes and diverse datasets will be used to see which sample size produces the best results for fake news detection.

Limitations of Experiments and Results

Many have focused on exploring different methods used to detect fake news and the problems that occurred but did not produce a definitive result. Bondielli and Marcelloni (2019) investigated different ways to detect fake news automatically. In addition, they advised that obtaining relevant data is problematic for fake news detection research. A literature survey by Dwivedi et al. (2020) explored various fake news detection methods. In addition, Zhang et al. (2020) provided an overview of the existing datasets and similar approaches to detect fake news. However, these researchers did not run any experiments or provide any results. In this experiment, the results will be extracted from the model and analyzed.

Accuracy Benchmark

Machine learning models require a copious amount of data to perform well. Unfortunately, it is difficult to determine how much is too much or too little. It is subjective to define the exact number of samples a small, medium, and large dataset should consist of because the performance is dependent on the machine learning model. For instance, Decision Trees are unfavorable in settings with a lot of data because they start to produce uncertain outcomes that take longer to

train, which is not ideal in most cases where detection needs to be done in real-time (Khanam et al., 2020). Many researchers do not have the exact sample size of what they deem as small, medium, and large. Fortunately, other metrics, such as precision, recall, AUC curve, and F-1 score can be used to evaluate the model's performance (Rathakrishan & Sathivanarayanan, 2023). However, an accuracy greater than 70% is considered a good performing model (Rathakrishan & Sathivanarayanan, 2023). In order to build an optimized model for detecting fake news, the accuracy should be above 90%.

CHAPTER THREE

RESEARCH METHODS

This chapter provides the research methods used to answer the research questions for this culminating experience project. We will start with the methods used for answering both questions.

Q1. How do large volumes of fake news datasets affect the accuracy percentage of a machine learning model?

Q2. As one increases the volume of data fed into a machine learning model from small to large datasets, what will the cutoff accuracy percentage point be?

Past researchers have only produced research on Naïve Bayes using small sample sizes ranging from 600 to 6,000 news posts. For instance, Jain and Kasbe (2018) only used 6,335 Facebook news posts. Unfortunately, the algorithm for detecting fake news did not perform well on small data sizes because it only achieved a testing accuracy score of 80%. Researchers suggested that working with larger sample sizes may increase the accuracy of detection (Jain & Kasbe, 2018, Granik & Mesyura, 2017). However, they did not provide a threshold for how much data should be considered. In this culminating

experience project, I decided to test the limitations of the Naïve Bayes model for detecting fake news by using varying sample sizes to answer the research questions listed above.

Datasets

To answer Q2: *As one increases the volume of data fed into a machine learning model from small to large datasets, what will the cutoff accuracy percentage point be?,* here are the methods presented below:

The datasets were collected from Kaggle, a public data-storing website for research purposes, and the columns were used to filter out files that were not for commercial use and in csv form (Figure 1). After doing so, the list of available datasets for small was 28, medium was 24, and large was 2. Based on Figure 1, there were not that many available fake and real news datasets available to work with. In addition, datasets that were not prelabeled with real or fake were ruled out because the model is supervised learning so the data must be labeled. The data sizes ranged from 30 to 2,055,271 (Figure 2) and the averages were calculated for each set; n(small) = 6,864, n(medium) = 26,667, and n(large) = 1,539,746. A small dataset size used in existing works consisted of ranges from 18 to 1030 (Athnian, et al., 2021). Since the dataset sizes contained a variety of ranges listed on Kaggle, different sizes were randomly selected between 30 and 2,055,271. For the small dataset, we chose to replicate similar studies that use less than 2,000 samples to see if the model can achieve an accuracy score of

above 70% because previous studies failed to do so. For the medium dataset, a dataset close to the average was chosen to provide a suitable benchmark for future reference. Lastly, since it is commonly known that the more data you train, the higher the accuracy, a random dataset between the average of the middle range and the average of the large range, which was around 72,000 was selected to see if there is any significance in testing on a dataset containing less than two million samples.

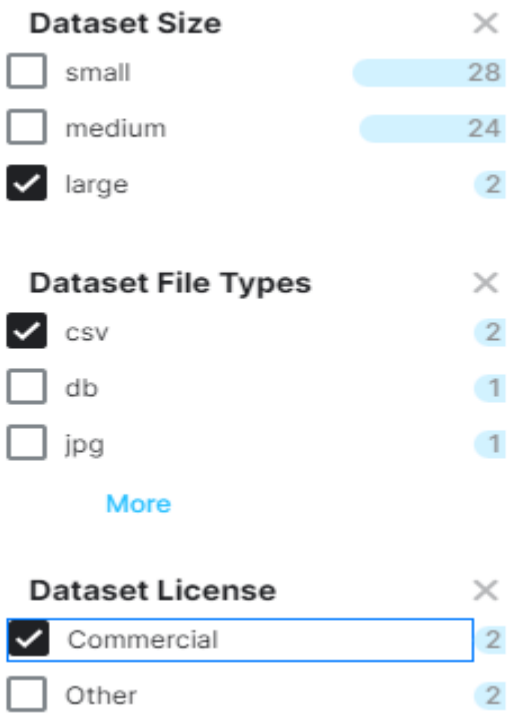

Figure 1: Filtered out file type and license (Kaggle)

	Small		Medium		Large
Fake News.csv	21,724	WELFake_Dataset.csv	72,134	Ciudadanos.csv	660124
Source Based Fake Ne	2006	Getting Real about Fa	12999	Esquerra.csv	649099
Data.csv	3352	Fake or Real News.csv	6256	Podemos.csv	2561356
New Task.csv	9947	Fake_Train.csv	20,800	PSOE.csv	1866559
FakeClaimDataset.csv	10232	Disinformation and Fa	6241	Populares.csv	1446069
Fakeandtruenews.csv	30	BanFake News.csv	50000	2020_vox.csv	2055271
dataset.csv	759	Fake_news_Trrain.csv	18239		
		Fake News Detection	25117		
		Politifact_Real_Fake.(	1056		
		News.csv	7796		
		FakeCovidFactCheck.(	7623		
		FakeNews Classificati	38,729		
		Cleaned_fake_andRe	31,429		
	6,864		26,667		1539746.333

Figure 2: Averages on small, medium, and large datasets available on Kaggle

Dataset 1, *News_Articles*, is the small dataset. It contained news articles from WhatsApp and Twitter. The raw dataset has 2,006 news articles and 10 attributes: *author, published, title, article text, language, site url, image url, article type, real or fake, title without stopwords.* Details about the attributes in Dataset 1 are provided in Table 1.

Table 1: Details of Dataset 1 -- News_Articles (Kaggle)

Sr. No	Parameter/ Attribute	Information
1	*author*	Author of the news article
2	*published*	Date published
3	*title*	Title of news articles
4	*Article_text*	Details about the news content
5	*language*	96% articles were in English and 3% are in German
6	*Site url*	URL of where the article

		is from
7	*Image url*	URL of where the image is from
8	*Article type*	Bias or not bias
9	*Real or fake*	Fake (62%) or real (38%)
10	*Title without stopwords*	Stopwords are removed from title

Dataset 2, *FakeNewsNet*, is the medium dataset that contained a distribution of real and fake news articles from published websites and Twitter posts. The raw dataset has 23,196 news articles and 5 attributes: *title, news_url, source_domain, tweet_num, and real.* Details about the attributes in Dataset 2 are provided in Table 2.

Table 2: Details of Dataset 2 -- FakeNewsNet (Kaggle)

Sr. No	**Parameter/ Attribute**	**Information**
1	*title*	Title of the article
2	*News_url*	URL of the article
3	*Source_domain*	Web domain where articles were posted
4	*Tweet_num*	Number of retweets for an article
5	*Real*	1 = real

		0 = fake

Dataset 3, *WELFake_Dataset*, contained a larger sample size. The dataset consisted of 72,134 news articles collected from BuzzFeed Political, Kaggle, McIntire, and Reuters with 3 attributes: title, text, and label. A description of the attributes in Dataset 3 is provided in Table 3.

Table 3: Details of Dataset 3 -- WELFake_Dataset(Kaggle)

Sr. No	**Parameter/ Attribute**	**Information**
1	*title*	Title of the article
2	*text*	Details about news content
3	*label*	1 = real 0 = fake

The title of the news articles was used to evaluate how the accuracy of the model changes when the data sizes increase. Some of the parameters will be considered a disadvantage because it can skew the accuracy of the prediction model. For example, full-text data are harder to use and train in large quantities than title data (Mai et al., 2018); therefore, a comprehensive training will not be performed.

Python was used to prepare the data. It is a programming language commonly used for scientific computing, such as machine learning and artificial intelligence. Within the computing program, there are powerful and interactive libraries that make programming simpler and more convenient for users. The most common libraries, Scikit-learn, NumPy, Nltk, and Pandas were imported to help answer the questions. Scikit-learn is infamous for working with complex data and works in association with NumPy. In the next section, the process of cleaning the data for modeling will be presented.

Data Cleaning

Clean data is important for producing high-quality machine learning models. In all three datasets, there were many tweets, broken characters, missing values, and symbols (@ and #) that can skew the results and analysis. However, because the datasets are so large, it is impossible to do minimal cleaning by the naked eye and is extremely time-consuming. Therefore, the datasets were preprocessed through Google Collaboration, which is commonly used by Python developers. Before cleaning the datasets, the first five rows were viewed using the head() function to see how the data is presented without having to look at the entire csv file. Then, the data was checked to eliminate any duplicate values, if found. The drop_duplicates() function was then used to remove any values that were similar. Then, the isnull().sum() function was used to check for missing values. The new net data size for Datasets 1, 2 and 3 are shown in Table 4. After the data processing was completed, a Natural Language

Processing (NLP) Application was used to process the text and provide numerical values.

Table 4: NLP Data Processing Results

	Initial Data Shape	Duplicate Entries	Missing Values (Titles)	New Net Total
Dataset 1	2,096	10	0	**2,086**
Dataset 2	23,196	137	0	**22,730**
Dataset 3	72,134	0	558	**71,537**

Natural language processing is an important process for computers to understand human language. For example, it can help computers read text, hear speech, interpret, measure sentiment, and determine importance (Rouse, 2022). Natural Language Toolkit (NLTK) is a Python library that turns text into numerical values (What is Natural Language Processing, n.d.). By using the natural language toolkit package (nltk.dowload()), stop words and punctuations that were present in the title were eliminated. Stopwords are irrelevant words in a stop list that are removed before or after the processing of natural language data. Table 5 presents an example of the types of words filtered out during the stop word removal process and Figure 3 are examples of how it was applied to Dataset 2. The same process was also applied to Datasets 1 and 3 but is not shown.

Table 5: Examples of texts with and without stopwords

Examples Containing StopWords	Without StopWords
I like reading, so I read.	Like, reading, read

I love to cook.	Love, cook

```
0    [Kandi, Burruss, Explodes, Rape, Accusation, R...
1    [Peoples, Choice, Awards, 2018, best, red, car...
2    [Sophia, Bush, Sends, Sweet, Birthday, Message...
3    [Colombian, singer, Maluma, sparks, rumours, i...
4    [Gossip, Girl, 10, Years, Later, Upper, East, ...
Name: combined, dtype: object
```

Figure 3: Stopwords and punctuations removed from Dataset 2

Naïve Bayes Supervised Learning Model

Supervised learning is known to be a valuable solution for eliminating manual classification work and for making future predictions. It uses a labeled dataset that is then trained by a machine learning algorithm to produce a desired output. Naïve Bayes was used since it works well with text classification and has the best converging and training time (Albahr & Albahar, 2020). For a supervised learning algorithm to work, both the feature, which is the attribute of the input data, and target (dependent) variables must be labeled. The goal is to find patterns between feature and target variables in hopes of measuring the model's accuracy on an unlabeled dataset.

Implementation Steps

The implementation steps for Q1. ***How do large volumes of fake news datasets affect the accuracy percentage of a machine learning model? and Q***2. ***As one increases the volume of data fed into a machine learning model***

from small to large datasets, what will the cutoff accuracy percentage point be? are presented below:

1. Prepared Datasets 1, 2, and 3 using pre-processing and processing
2. Picked an algorithm suitable for providing the best accuracy (Naïve Bayes)
3. Tested the model using the three different sample sizes (small, medium, large)
4. Analyzed the model and results
5. Saved the trained model

Google Collaboration, which offers a variety of programming tools was used to build a machine-learning model. First, pandas, a Python library, was imported to help read the dataset and then two variables ‘X’ and ‘Y” were created for the training and testing process. Then, the Naive Bayes classifier was used to help build the machine-learning model. Afterwards, the data was divided into training and testing because data splitting for modeling is very important (Kebonye, 2021). For the Naive Bayes model, 80 percent of the dataset was selected for training and 20 percent for testing because more trained data means that the model is better equipped for when it is tested on the dataset (Nguyen et al., 2021). By using the Scikit-learn model selection method, we can train, test, and fit the dataset by comparing how varied sizes of data affected the accuracy percentage.

CHAPTER FOUR

DATA ANALYSIS AND FINDINGS

To answer the questions, this section provides an overview of the results after applying the methods from Chapter 3 to help answer Q1 and Q2. This project produced the following results from a Naïve Bayes classifier and is presented below in a graph divided into training and testing. To recap, the Naïve Bayes algorithm was utilized to evaluate how well it perform on various sizes of fake news data.

Q1. How do large volumes of fake news datasets affect the accuracy percentage of a machine learning model?

About 80% of the data was trained and 20% of the data was tested. The results showed a linear relationship for the training accuracy (Figure 4). Datasets 1, 2, 3, and combined all produced an accuracy of above 90%. In addition, the training scores for precision, F-1, and recall were all excellent. However, Dataset 2 slightly underperformed because the scores were below 90%. Table 6 displays the training results for Datasets 1, 2, 3, and combined. In the testing phase, we can observe that as the data size increases, so does the accuracy percentage.

Table 6: Training Results from using Naive Bayes Model

	Train	Precision	F-1 Score	Recall	Accuracy	Size
Dataset 1	80%	0.93	0.93	0.93	93%	1,628
Dataset 2	80%	0.88	0.87	0.87	91%	18,184

Dataset 3	80%	0.94	0.94	0.94	94%	57,229
Combined	80%	0.91	0.91	0.91	91%	67,256

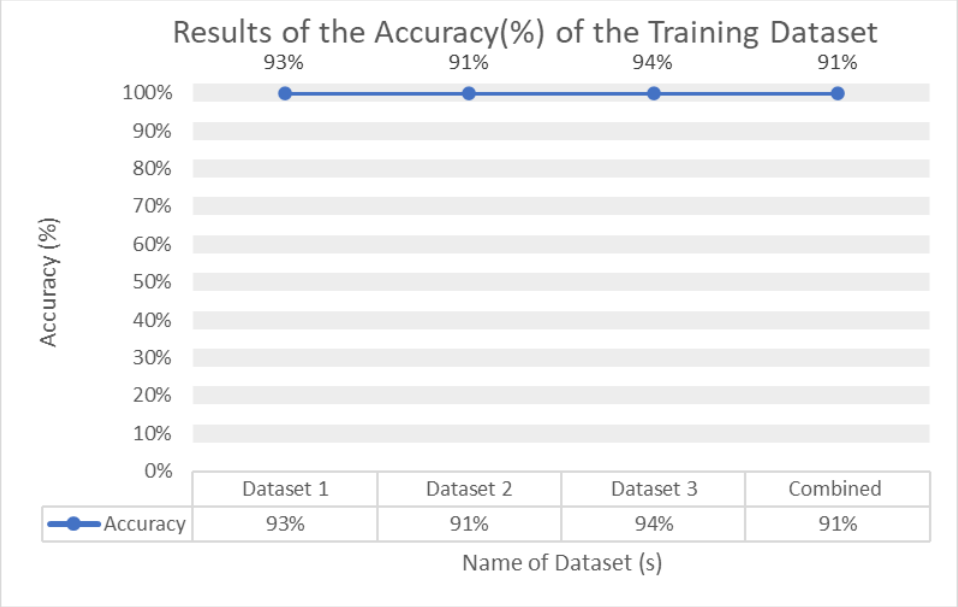

Figure 4: Accuracy (%) of the Training Dataset(s)

Q2. As one increases the volume of data fed into a machine learning model from small to large datasets, what will the cutoff accuracy percentage point be?

In Table 7, Dataset 1 had an accuracy of 66%, Dataset 2 had an accuracy of 83%, and Dataset 3 has an accuracy of 92%. But when all the datasets were combined for testing, the testing accuracy of Naïve Bayes dropped to 88%. The

testing results of precision (0.67), F-1 (0.66), and recall (0.66) were extremely low for Dataset 1. Table 7 contains all the results for the testing data.

Table 7: Testing Results from using Naive Bayes Model

	Test	Precision	F-1 Score	Recall	Accuracy	Size
Dataset 1	20%	0.67	0.66	0.66	66%	407
Dataset 2	20%	0.83	0.83	0.83	83%	4,546
Dataset 3	20%	0.92	0.92	0.92	92%	14,308
Combined	20%	0.88	0.88	0.88	88%	16,814

CHAPTER FIVE

DISCUSSION, CONCLUSION,

AND AREAS FOR FURTHER STUDY

Discussion

Chapter 5 discusses the research findings, provides a conclusion and areas for further study.

Q1. How do large volumes of fake news datasets affect the accuracy percentage of a machine learning model?

In this study, large volumes of data did affect the accuracy percentage of the Naïve Bayes model. For instance, the largest dataset, Dataset 3, had more data that was fed into the model for training; therefore, based on Table 7, the model performed extremely well compared to the other competing datasets. The predictive model of Dataset 3 achieved an accuracy score of 92%. However, when all the datasets were combined for testing, the accuracy dropped to 88%. Although it is not a significant decrease in accuracy, it does present some open questions about the model such as whether issues with overfitting were present and whether the Naïve Bayes model stops performing well at a certain threshold. The conclusion is that large data sizes containing many unique features increase the accuracy percentage. An area for further study includes using other attributes other than the title to test for accuracy.

Q2. As one increases the volume fed into a machine learning model from small to large datasets, what will the cutoff accuracy percentage point be?

To determine the accuracy percentage, three different data sizes were randomly selected. The cutoff accuracy percentage depends on how many values you use because data is not stagnant. It continues to increase over time; therefore, we cannot definitively produce a finite number of how much should be fed into a machine to produce an outcome of 70%, 80%, 90%, or 100%. However, we can conclude that certain types of data ranges will always produce a certain percentage. For instance, a small data size will produce an accuracy of 70%, a medium data size will produce an accuracy of 70-80%, and a large data size will produce an accuracy of 90-100%. An area for further study is choosing the largest dataset and dividing it into parameters in 100 folds so that an accurate benchmark will be given.

Conclusion

Fake news is ubiquitous. It can change viewpoints, opinions, attitudes, and behavior about a particular topic. Therefore, we must be vigilant before consuming information. Automatic detection of fake news using machine learning models makes it possible to slow or mitigate the rapid spread of fake news. Machine learning algorithms can process large datasets created by the plethora of news articles quickly and effectively.

Areas for Further Study

This culminating experience project only touches the surface level of this challenging topic. There is still much more work that needs to be done. For example, someone can use the model and build a novel fact-checking website that can process a bunch of data in a few seconds. Building a website will give a better impression and understanding of how well the model performs in real-time. In addition, one may consider how the machine learning model, Naïve Bayes, affects different modalities, such as images and videos since it only works well with text classification.

APPENDIX

CODES

File Edit View Insert Runtime Tools Help All changes saved

+ Code + Text

```
#Description: This program detects real (0) and fake (1) news
```

```
#import the libraries
import numpy as np
import pandas as pd
import nltk
from nltk.corpus import stopwords
import string
```

```
df = pd.read_csv('FakeNewsNet.csv')
```

```
df.head()
```

	title	news_url	source_domain	tweet_num	real
0	Kandi Burruss Explodes Over Rape Accusation on...	http://toofab.com/2017/05/08/real-housewives-a...	toofab.com	42	1
1	People's Choice Awards 2018: The best red carp...	https://www.today.com/style/see-people-s-choic...	www.today.com	0	1
2	Sophia Bush Sends Sweet Birthday Message to 'O...	https://www.etonline.com/news/220806_sophia_bu...	www.etonline.com	63	1
3	Colombian singer Maluma sparks rumours of inap...	https://www.dailymail.co.uk/news/article-33655...	www.dailymail.co.uk	20	1
4	Gossip Girl 10 Years Later: How Upper East Sid...	https://www.zerchoo.com/entertainment/gossip-g...	www.zerchoo.com	38	1

```
df.shape
#tells you the number of rows
```

```
(23196, 5)
```

```
df.drop_duplicates(inplace=True)
df.shape
#Trying to see if there are duplicates in the data and if it got deleted
```

```
(23059, 5)
```

```
[ ]  def process_text(text):
       nopunc = [char for char in text if char not in string.punctuation]
       nopunc = ''.join(nopunc)

       clean_words = [word for word in nopunc.split() if word.lower() not in stopwords.words('english')]


       return clean_words
```

```
[ ]  df['title'].head().apply(process_text)
     #list doesnt contain any stopwords or punc
```

```
0    [Kandi, Burruss, Explodes, Rape, Accusation, R...
1    [Peoples, Choice, Awards, 2018, best, red, car...
2    [Sophia, Bush, Sends, Sweet, Birthday, Message...
3    [Colombian, singer, Maluma, sparks, rumours, i...
4    [Gossip, Girl, 10, Years, Later, Upper, East, ...
Name: title, dtype: object
```

```
     df['title']
```

```
0        Kandi Burruss Explodes Over Rape Accusation on...
1        People's Choice Awards 2018: The best red carp...
2        Sophia Bush Sends Sweet Birthday Message to 'O...
3        Colombian singer Maluma sparks rumours of inap...
4        Gossip Girl 10 Years Later: How Upper East Sid...
                              ...
23191    Pippa Middleton wedding: In case you missed it...
23192    Zayn Malik & Gigi Hadid’s Shocking Split: Why ...
23193    Jessica Chastain Recalls the Moment Her Mother...
23194    Tristan Thompson Feels "Dumped" After Khloé Ka...
23195    Kelly Clarkson Performs a Medley of Kendrick L...
Name: title, Length: 22730, dtype: object
```

```
[ ]  #convert text into matrix token counts
     from sklearn.feature_extraction.text import CountVectorizer
     message_bow = CountVectorizer(analyzer=process_text).fit_transform(df['title'])
```

```
[ ] #split data into 80% training and 20% testing
    from sklearn.model_selection import train_test_split
    X_train, X_test, y_train, y_test = train_test_split(message_bow, df['real'], test_size = 0.20, random_state=0)
```

```
[ ] message_bow.shape
    #there are more differnt rows
```

```
(22730, 27817)
```

```
[ ] from sklearn.naive_bayes import MultinomialNB
    classifier = MultinomialNB()
    classifier.fit(X_train, y_train)
```

```
▾ MultinomialNB
MultinomialNB()
```

```
print(classifier.predict(X_train))
print(y_train.values)
#the model classified the first
```

```
[1 1 1 ... 0 1 0]
[1 1 1 ... 0 1 0]
```

```
[ ] from sklearn.metrics import classification_report
    pred = classifier.predict(X_train)
    print(classification_report(y_train, pred))
```

```
                precision    recall  f1-score   support

             0       0.84      0.78      0.81      4333
             1       0.93      0.95      0.94     13851

      accuracy                           0.91     18184
     macro avg       0.88      0.87      0.87     18184
  weighted avg       0.91      0.91      0.91     18184
```

```
[ ] from sklearn.metrics import classification_report
    pred = classifier.predict(X_test)
    print(classification_report(y_test, pred))
```

```
                precision    recall  f1-score   support

             0       0.63      0.64      0.63      1070
             1       0.89      0.88      0.89      3476

      accuracy                           0.83      4546
     macro avg       0.76      0.76      0.76      4546
  weighted avg       0.83      0.83      0.83      4546
```

REFERENCES

Ahmed, H., Traore, I., & Saad, S. (2017). Detection of online fake news using n-gram analysis and machine learning techniques. In *Intelligent, Secure, and Dependable Systems in Distributed and Cloud Environments: First International Conference, ISDDC 2017, Vancouver, BC, Canada, October 26-28, 2017, Proceedings 1* (pp. 127-138). Springer International Publishing.

Ahmad, I., Yousaf, M., Yousaf, S., & Ahmad, M. O. (2020). Fake news detection using machine learning ensemble methods. *Complexity*, *2020*, 1-11.

Albahr, A., & Albahar, M. (2020). An empirical comparison of fake news detection using different machine learning algorithms*. International Journal of Advanced Computer Science and Applications, 11(9).*

Alotaibi, F. S. (2019). Implementation of machine learning model to predict heart failure disease. *International Journal of Advanced Computer Science and Applications, 10(6).*

Althnian, A., AlSaeed, D. H., Al-Baity, H. H., Samha, A. K., Dris, A. B., Alzakari, N., Elwafa, A. A., & Kurdi, H. (2021). Impact of Dataset Size on Classification Performance: An Empirical Evaluation in the Medical Domain. *Applied Sciences*, *11*(2), 796. https://doi.org/10.3390/app11020796

Bondielli, A., & Marcelloni, F. (2019). A survey on fake news and rumour detection techniques. *Information Sciences*, *497*, 38-55.

Collins, B., Hoang, D. T., Nguyen, N. T., & Hwang, D. (2021). Trends in combating fake news on social media–a survey. *Journal of Information and Telecommunication*, *5*(2), 247-266.

Dwivedi, R. K., Rai, A. K., & Kumar, R. (2020, February). Outlier detection in wireless sensor networks using machine learning techniques: a survey. In *2020 International Conference on Electrical and Electronics Engineering (ICE3)* (pp. 316-321). IEEE.

Field, P. (2021, February). Fake news was a thing long before Donald Trump – just ask the ancient Greeks*. TheConversation*. Retrieved March 9, 2023, from Fake news was a thing long before Donald Trump — just ask the ancient Greeks (theconversation.com)

Granik, M., & Mesyura, V. (2017). Fake news detection using naive Bayes classifier. In *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)* (pp. 900-903). IEEE.

Hunt, K., Agarwal, P., & Zhuang, J. (2022). Monitoring misinformation on Twitter during crisis events: a machine learning approach. *Risk analysis, 42(8), 1728-1748.*

Jain, A., & Kasbe, A. (2018). Fake news detection. In *2018 IEEE International Students' Conference on Electrical, Electronics and Computer Science (SCEECS)* (pp. 1-5). IEEE.

Jordan, M. I., & Mitchell, T. M. (2015). Machine learning: Trends, perspectives, and prospects. *Science,* 349(6245), 255-*260.*

Kebonye, N. M. (2021). Exploring the novel support points-based split method on a soil dataset. *Measurement, 186*, 110131.

Khanam, Z., Alwasel, B. N., Sirafi, H., & Rashid, M. (2021, March). Fake news detection using machine learning approaches. In *IOP conference series: materials science and engineering* (Vol. 1099, No. 1, p. 012040). IOP Publishing.

Kumar, S., & Singh, T. D. (2022). Fake news detection on Hindi news dataset. Global Transitions Proceedings, 3(1), 289-297.

Madini, Y, Erritali, M, Bouikhalene, B. (2021, June)"Using Artificial Intelligence Techniques for detecting Covid-19 epidemic fake news in Moroccan Tweets," *ScienceDirect.* Retrieved March 21, 2023, from https://www.sciencedirect.com/science/article/pii/S2211379721004034

Mai, F., Galke, L., & Scherp, A. (2018). Using deep learning for title-based semantic subject indexing to reach competitive performance to full-text. *In Proceedings of the 18th ACM/IEEE on Joint Conference on Digital Libraries* (pp. 169-178).

Manzoor, S. I., & Singla, J. (2019, April). Fake news detection using machine learning approaches: A systematic review. In *2019 3rd international conference on trends in electronics and informatics (ICOEI)* (pp. 230-234). IEEE.

Nguyen, Q. H., Ly, H. B., Ho, L. S., Al-Ansari, N., Le, H. V., Tran, V. Q., ... & Pham, B. T. (2021). Influence of data splitting on performance of machine learning models in prediction of shear strength of soil. *Mathematical Problems in Engineering*, 2021, 1-15.

Pal, A., & Pradhan, M. (2023). Survey of fake news detection using machine intelligence approach. *Data & Knowledge Engineering*, *144*, 102118.

Pandey, S., Prabhakaran, S., Reddy, N. S., & Acharya, D. (2022). Fake News Detection from Online media using Machine learning Classifiers. In *Journal of Physics: Conference Series* (Vol. 2161, No. 1, p. 012027). IOP Publishing.

Qalaja, E. K., Al-Haija, Q. A., Tareef, A., & Al-Nabhan, M. M. (2022). Inclusive Study of Fake News Detection for COVID-19 with New Dataset using Supervised Learning Algorithms. *International Journal of Advanced Computer Science and Applications*, 13(8).

Rachana, B., Priyanka, T., Sahana, K. N., Supritha, T. R., Parameshachari, B. D., & Sunitha, R. (2021). Detection of polycystic ovarian syndrome using follicle recognition technique. *Global Transitions Proceedings*, *2*(2), 304-308.

Rathakrishnan, A., & Sathiyanarayanan, R. (2023). Rumor detection on social media using deep learning algorithms with fuzzy inference system for healthcare analytics system using COVID-19 dataset. *International Journal of Computational Intelligence and Applications*, 2341008.

Shaikh, J., & Patil, R. (2020, December). Fake news detection using machine learning. In *2020 IEEE International Symposium on Sustainable Energy, Signal Processing and Cyber Security (iSSSC)* (pp. 1-5). IEEE.

Sharma, D. K., & Garg, S. (2021). IFND: a benchmark dataset for fake news detection. *Complex & Intelligent Systems*, 1-21.

Taskin, S. G., Kucuksille, E. U., & Topal, K. (2022). Detection of Turkish fake news in Twitter with machine learning algorithms. *Arabian Journal for Science and Engineering*, *47*(2), 2359-2379.

What is Fake News? (n.d.). *Center for Information Technology and Society - UC Santa Barbara*. Retrieved March 9, 2023, from https://www.cits.ucsb.edu/fake-news/what-is-fake-news

What is Natural Language Processing? (n.d). *IBM*. Retrieved March 10, 2023 from https://www.ibm.com/topics/natural-language-processing

Wright, C. L. (2020, November 08). Political Fake News and the 2020 Election. *PsychologyToday*. Retrieved March 9, 2023, from https://www.psychologytoday.com/us/blog/everydaymedia/202011/political-fake-news-and-the-2020-election

Zhang, X., & Ghorbani, A. A. (2020). An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, *57*(2), 102025.

[NewsArticles] Retrieved [04 April 2023] from https://www.kaggle.com/datasets/asad1m9a9h6mood/news-articles

[WELFake] Retrieved [23 March 2023] from https://www.kaggle.com/datasets/saurabhshahane/fake-news-classification

[FakeNewsNet] Retrieved [06 April 2023] from https://www.kaggle.com/datasets/algord/fake-news