A COVID-19 Fake News Detection System

Muvazima Mansoor, B.Tech

A Dissertation

Presented to the University of Dublin, Trinity College

in partial fulfilment of the requirements for the degree of

Master of Science in Computer Science (Data Science)

Supervisor: Khurshid Ahmad

August 2022

Declaration

I, the undersigned, declare that this work has not previously been submitted as an exercise for a degree at this, or any other University, and that unless otherwise stated, is my own work.

I have read and I understand the plagiarism provisions in the General Regulations of the University Calendar for the current year, found at `http://www.tcd.ie/calendar`.

I have also completed the Online Tutorial on avoiding plagiarism 'Ready Steady Write', located at `http://tcd-ie.libguides.com/plagiarism/ready-steady-write`.

Muvazima Mansoor

August 19, 2022

Permission to Lend and/or Copy

I, the undersigned, agree that Trinity College Library may lend or copy this thesis upon request.

Muvazima Mansoor

August 19, 2022

A COVID-19 Fake News Detection System

Muvazima Mansoor, Master of Science in Computer Science

University of Dublin, Trinity College, 2022

Supervisor: Khurshid Ahmad

The pandemic bought along with it an infodemic which is the overabundance of information, truthful or not, that is often spread over social media. The spread of misinformation regarding COVID-19 has dire consequences hampering the efforts of health systems worldwide. Social media has made it easy to share personal opinions which are neither entirely truthful nor fake and can be termed as 'imaginative'. This work presents a fake news detection system which can differentiate between informative, imaginative, and fake news. Each type of news is written concerning a target audience. The differences in the style of writing of three types of news are analysed at multiple levels of linguistic description such as – lexical, syntactic, semantic, and pragmatic. Finally, content-based features are leveraged to construct a fake news detection system utilising a Perceptron model that can classify informative, imaginative, and fake news with an F1 score of 0.74, 0.8 and 0.95, respectively. It performs significantly better than the baseline Naïve Bayes model that gave an F1 score of 0.64, 0.58 and 0.35 for informative, imaginative, and fake news, respectively.

Acknowledgments

I sincerely thank Professor Khurshid Ahmad for his constant guidance and support. I am extremely grateful to him for teaching me many concepts patiently and providing invaluable feedback throughout the entire course of my Master's.

I would like to thank my parents, Mansoor Mohammed and Syeda Mudasira, for providing moral support every day. I would also like to thank my brother, Manam Mansoor, for motivating me and looking after me during stressful times.

Finally, I thank all my friends for their moral encouragement throughout the course. Special Mention to Deepayan and Unni for providing valuable inputs and suggestions.

Muvazima Mansoor

University of Dublin, Trinity College
August 2022

Contents

Abstract **iii**

Acknowledgments **iv**

Chapter 1 Introduction **1**
1.1 Problem Definition 3
1.2 Contributions 4
1.3 Structure of the dissertation 4

Chapter 2 Literature Review **5**
2.1 Motivation 5
2.1.1 Infodemic 5
2.1.2 Effects of Infodemic and fake news 6
2.1.3 Conclusion 7
2.2 Fake News Detection Methods 7
2.2.1 Propagation-Based Methods 8
2.2.2 Content-Based Methods 9
2.3 Machine Learning Models 12
2.4 Deep Learning Models 12
2.5 Conclusion 13

Chapter 3 Methods **14**
3.1 Overview 14
3.2 Data Collection 16
3.3 Data Pre-processing 17
3.4 Fake News Detection 19
3.4.1 Lexical Level 20
3.4.2 Syntactic Level 20
3.4.3 Semantic Level 22

3.4.4 Pragmatic Level . . . 24
3.4.5 Machine Learning . . . 26
3.5 Conclusion . . . 29

Chapter 4 Results and Observations **31**
4.1 Dataset . . . 31
4.1.1 Informative . . . 31
4.1.2 Imaginative . . . 32
4.1.3 Fake News . . . 33
4.2 Lexical Level . . . 34
4.3 Syntactic Level . . . 37
4.4 Semantic Level . . . 38
4.5 Pragmatic Level . . . 40
4.6 Machine Learning Models . . . 43
4.7 Discussion of Results . . . 47

Chapter 5 Conclusion and Future Works **49**
5.1 Conclusion . . . 49
5.2 Future Work . . . 50

Bibliography **52**

Appendices **57**
.1 Dataset used for evaluation . . . 58

List of Tables

3.1 Relative frequency distribution of hundred most frequently occurring tokens in Informative corpus and American National Corpus. 20
3.2 Stanford NLTK POS tags 21
3.3 Close Class Words vs Open Class Words 22
3.4 Candidate term selection criteria based on Z-Score Weirdness and Z-Score Frequency 24
3.5 Confusion Matrix 29

4.1 Descriptive Statistics of relative frequency of tokens 34
4.2 The frequency distribution of the top hundred words in the Imaginative corpus and American National Corpus. 34
4.3 Relative frequency distribution of top 100 words in Fake News corpus and American National Corpus. 35
4.4 Distribution of Z-Score of weirdness and Z-Score of relative frequency for informative corpus 39
4.5 Distribution of Z-Score of weirdness and Z-Score of relative frequency for imaginative corpus 39
4.6 Distribution of Z-Score of weirdness and Z-Score of relative frequency for fake news 39
4.7 Ten most frequently occurring candidate terms in informative, imaginative and fake news 40
4.8 Precision, Recall and F1-Score for Multinomial Naïve Bayes 44
4.9 Precision, Recall and F1-Score for Perceptron 44
4.10 Precision, Recall and F1-Score for Binomial Perceptron – Informative vs Fake News 46
4.11 Precision, Recall and F1-Score for Binomial Perceptron – Imaginative vs Fake News 47

List of Figures

1.1 Text cline 2

2.1 Infodemic management ecosystems (1) 6
2.2 Different Fake news detection methods 8
2.3 Life cycle of a news article with the four different fake news detection methods (2) 9

3.1 System Architecture 16
3.2 Data types 17
3.3 Data Pre-processing Steps 18
3.4 Example of Pre-processed Data 19
3.5 Analysis of Different Linguistic Descriptions 19
3.6 Part of speech tagging stanford POS tagger 22
3.7 Multilayer perceptron architecture 28

4.1 Sources used for dataset collection 32
4.2 Relative frequency vs word count for Informative and ANC 35
4.3 Relative frequency vs word count for Imaginative and ANC 36
4.4 Relative frequency vs word count for Fake news and ANC 36
4.5 POS tag distribution in special language corpora 37
4.6 CCW and OCW distribution for special language corpora and the general language corpus 38
4.7 Distribution of emotional words in the special language corpora 41
4.8 OCW and emotion vector for five articles 41
4.9 Euclidean distance between six articles 41
4.10 Distribution of Zmax vs Zmin for Open Class Words 42
4.11 Distribution of Zmax and Zmin for Emotion words 42
4.12 Confusion Matrix for Multinomial Naïve Bayes 43
4.13 Confusion Matrix for the Perceptron 45
4.14 Confusion Matrix for the Binomial Perceptron – Informative vs Fake News 46

4.15 Confusion Matrix for Binomial Perceptron – Imaginative vs Fake News . . 47

Chapter 1

Introduction

With the COVID-19 pandemic came another significant issue known as the "infodemic". Infodemic is the over-abundance of information spread rapidly without checking its authenticity.

In the scenario of COVID-19, the infodemic involved the rapid information spread regarding the source of the coronavirus, so-called "cures", symptoms, opinions regarding the vaccination and death rates, etc. For the general public consuming this vast amount of information so readily available online, it is difficult to determine whether it is truthful. Exaggerated or false information can create a lack of understanding and awareness of the virus, which might hamper the tremendous efforts to stop the spread of the virus. It puts people at risk by advocating a false sense of security. Misinformation also puts people at risk by promoting fake remedies and products. In addition, misinformation instils a suspicion of state mandates and official sources. For instance, fake news such as "Face masks are useless against COVID-19 and are harmful to health" and "Vaccination leads to autism" were widely spread and believed. Information like this led to the formation of anti-vaccination and anti-mask groups worldwide that organised protests against these protective measures.

The spread of misinformation can stem from individuals or criminals looking to profit, government officials or politicians seeking to leverage the situation for their interests and opportunists seeking to discredit truthful sources. This misinformation gains traction when the unaware public shares it on their social network.

The spread of misinformation during the pandemic became so prevalent that social media platforms such as TikTok, Facebook and Twitter started labelling the posts containing misinformation with warning messages to inform the user of a potentially fake piece of information. Early methods to determine fake news involved checking the text's grammar and amount of spelling mistakes. However, with the advancement in technology

and the auto-correct features, it has become tough to distinguish fake news from truthful based on grammar alone. However, some of this news cannot be termed 'fake'. This is because the information stems from sources that are believed to be reliable, and there is no reason to question their authenticity.

Information can be spread from multiple sources. Each source is written for a target audience. Scientific journals and papers assume the reader has some domain knowledge. The articles written for scientific magazines are not as complex as the journals meant for an educated crowd. In comparison, news articles are meant for the general public and are easy to read. These sources, such as journals, scientific papers, books, and news articles, are termed 'informative'. On the flip side, sources such as blogs, opinion editorials, and advertisements are meant for the general public and reflect the view or opinion of the author. These are usually exaggerated and are intended to create a discussion among the readers. Social media posts are short, easy to read and therefore gain more engagement from people. These sources are termed 'imaginative' since they are not backed by scientific research and merely express the view of an individual.

The difference in the interoperability of the different sources, and the ease of access causes misrepresentation of information from an 'informative' source to 'imaginative' and finally to 'fake'. Figure 1.1 represents the text cline where knowledge gets diffused from facts to opinions to rumours.

Figure 1.1: Information gets diffused on a text cline

Misinformation has always been an issue, and the repercussions are worse in an unfamiliar environment like the pandemic. Therefore, the need for accurate information is

paramount to prevent the widespread of the disease and for general wellbeing. This dissertation is an attempt to aid the identification of misinformation regarding the Covid-19 pandemic.

1.1 Problem Definition

As described in the introduction, the need for accurate information regarding the pandemic is paramount. Misinformation in an unfamiliar environment like the pandemic can have serious consequences. There have been various research to determine whether a piece of information is reliable or not. However, due to the diversity and complexity of online data and the dynamic information on social networking platforms, it is difficult to accurately identify 'fake' news or 'imaginative' news. Multiple fact-checkers employ machine learning models to determine the authenticity of a text. However, they exist at a binary level and classify the information as true or false. They do not consider that it could be a combination of both or 'imaginative', which reflects an individual's opinion. In addition, these methods require large training sets and complex machine learning models.

This dissertation aims to distinguish between news that is 'informative', 'imaginative' and 'fake' at different levels of linguistic description. Informative news originates from trusted sources such as journals, peer-reviewed papers, news articles, etc. In comparison, imaginative news comprises opinion editorials which are traditionally located opposite the editorial page and stem from an author not affiliated with the publication. In addition, people's blogs, advertisements, and social media posts also form the 'imaginative' space. For example, advertisements for a 'cure' can be exaggerated since they promote a product.

'Fake' news comprises deliberate misleading information. The intent behind spreading this kind of news is malicious. News sourced from these different sources is analysed for differences at different linguistic description levels, such as the lexical level, syntactic level, semantic level, and pragmatic level. At the lexical level, the words (tokens) are analysed using descriptive statistics and frequency lists of tokens. At the syntactic level, the relationship of words within a sentence is analysed using Part-of-Speech analysis. At the semantic level, the sentence's meaning is analysed using candidate term analysis. At the pragmatic level, the meaning of the entire text is considered to analyse the difference between the three corpora.

Using these different levels of linguistic description, the dissertation hypothesis is:

'A computer program cannot differentiate between informative, imaginative or fake news '.

This dissertation attempts to conduct multiple experiments and tests to accept or reject the above hypothesis using a systematic approach.

1.2 Contributions

This dissertation builds a COVID-19 fake news detection system that identifies whether a given article is informative, imaginative, or fake. It provides more granularity than other fake news detection systems that only classify the article as fake or non-fake. It also offers a systematic approach for analysing the text using linguistic descriptions such as lexical, syntactic, semantic, and pragmatic. In addition, a pre-processed and clean data-set is provided for each informative, imaginative, and fake news class that can be used for other research on the pandemic. Finally, a model is built using a Perceptron that can identify a given article as either informative, imaginative, or fake with 83.5 % accuracy.

1.3 Structure of the dissertation

The structure of the dissertation is as follows –

Chapter 2 describes the motivation behind the dissertation. It also describes the different fake news detection methods that exist and a justification for the method followed in this dissertation.

Chapter 3 describes the method followed to construct a Covid-19 fake news detection model. It describes the data and pre-processing steps. In addition it describes the analysis at multiple linguistic levels such as lexical, syntactic, semantic and pragmatic. Finally, it discusses the machine learning methods used for classification.

Chapter 4 details the data collection and the observation of the analysis at each linguistic level. It also discusses the performance of the two machine learning models - Naive Bayes and Perceptron.

Chapter 5 provides the conclusion, limitations and the scope for future work.

Chapter 2

Literature Review

This chapter describes the background of the project and the related works. The first section describes the concepts of infodemic, misinformation and its repercussions. The second section describes the existing fake news detection methods.

2.1 Motivation

2.1.1 Infodemic

As mentioned in the introduction, the pandemic bought along another significant issue, called the 'infodemic'. Dr Sylvie Briand from WHO defines an infodemic as a "tsunami of information which can be accurate and otherwise that spreads along with a disease outbreak" (1). The infodemic cannot be eradicated due to the power of the internet and the speed of technology. The infodemic is detrimental to the health institutions and systems worldwide. An infodemic management community was formed by WHO (1) to monitor, understand and combat the harm caused due to the infodemic. Figure 2.1 displays the implementation and design model of infodemic management proposed by WHO. The aim was to change the people's behaviour by listening to their concerns, educating about science and risk, developing resilience to untrue information, and finally, empowering and engaging different communities to fight against the infodemic along with the pandemic.

This approach to infodemic management requires a lot of resources and workforce since it involves educating and engaging with the community. However, the dissertation's method involves managing the infodemic at the root level. The information is analysed for linguistic differences to determine whether it is truthful, fake, or merely an opinion. This reduces manual effort since a computer program is built to automate the task of fake news detection.

Figure 2.1: Infodemic management ecosystems (1)

2.1.2 Effects of Infodemic and fake news

The consequences of COVID-19 misinformation are dire, and multiple research papers prove this.

The study outlined in (3) describes people's susceptibility to misinformation worldwide and the role this susceptibility plays in health behaviours. Five countries were considered for the study, including the UK, Mexico, Spain, Ireland, and the USA. The survey was conducted between April and May 2020, and two surveys were conducted in the UK to determine whether the results were consistent over time. The participants were surveyed on gender, age, political ideology, education level, minority status and trust in the government, news, and scientists. The participants were also tested on their numeracy level with the help of different numeracy tests to know whether they could comprehend quantitative data. Concerning COVID-19, the participants were asked if they would get vaccinated, the extent to which they follow health guidelines, whether they trust the steps taken by WHO to combat the virus, and whether they came across information regarding the virus from social media or WHO. In addition, the participants were given nine statements regarding the virus (one ambiguous, two genuine and six false). They were asked to score their reliability on a 1-7 Likert scale. The survey led to many interesting results, the most crucial being that there is a strong link between hesitancy to vaccines and the susceptibility to misinformation, along with a reduced inclination to follow public health guidelines. Another study outlined in (4) explores impact of misinformation on the willingness to comply with the lockdown rules imposed by the UK government. The survey was conducted between 20th to 22nd May 2020, and 2,254 people from the UK were surveyed. The results show a negative correlation between compliance with state-imposed

health measures and the belief in COVID-19 conspiracy theories. Another interesting finding was that younger generation are more susceptible to the conspiracy theories than older people because the younger population uses social media more predominantly than the older generation, which relies more on broadcast media for information. In addition, the exposure to misinformation is much more than what is assumed. For instance, the study in (5) shows that almost half of the population in the UK (46%) have come across COVID-19 misinformation. In addition, among those that have encountered misinformation, 66% (almost two-thirds) experience these misinformation stories daily. This is concerning as prolonged exposure to misinformation increases the susceptibility to fake news (6).

2.1.3 Conclusion

Based on these studies, misinformation threatens society as it hinders the state's efforts to combat the virus. To curb the amount of misinformation on the virus and to help stop the spread of coronavirus, this dissertation aims to build a COVID-19 fake news detection system.

2.2 Fake News Detection Methods

This section describes the different studies on fake news detection to provide a comprehensive view of the current methods.

To combat fake news, many manual fact-checking tools and websites exist. However, with the infodemic and the amount of information present, particularly on social media, the manual fact-checkers are unable to scale well (7). Therefore, automatic fake news detection methods have been developed to overcome the forthcomings of manual fact-checkers.

Methods of automatic fake news detection methods are of two types - content-based and propagation-based methods (8). Content-based strategies leverage the content of the news to identify whether it is fake or not. Whereas Propagation- based methods leverage the social context information to identify whether a piece of information is fake or not. However, content and propagation-based methods are not independent of each other. Combining the strengths of both these methods is proven to be beneficial. For instance, (9) describes an approach to combine the three features of fake news – source, response, and text. A Recurrent Neural Network (RNN) is used for modelling the activity of a user on a news piece. This approach gave a better accuracy than just using either one of the characteristics individually to predict fake news. Similarly, (10) used social

and content-based features to train a probabilistic classifier with an accuracy of 90.62 per cent.

Figure 2.2 displays the different fake news detection methods reviewed in this chapter. The green boxes represent the method followed in this dissertation.

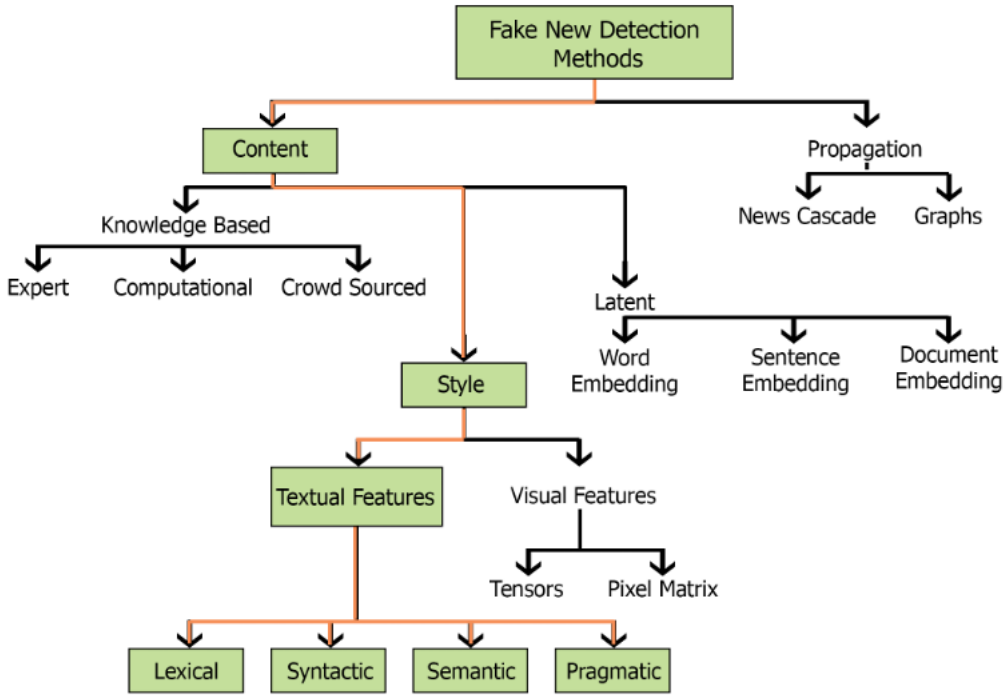

Figure 2.2: Different Fake news detection methods

2.2.1 Propagation-Based Methods

The social context dissemination of news on social media involves a tri-relationship among the users, news pieces and the publishers. Kai Shu et al. (11) leverage this tri-relationship to determine fake news. The user-news and publisher-news interactions are modelled using embeddings to create a fake-news detector framework. A similar approach was followed by (12), where the patterns in a social network are represented and leveraged at different network levels, such as the ego level, node level, triad level, network level and community level. Zhiwei Jin et al. (13) outline a method to automatically identify fake information in microblogs by leveraging social viewpoints that are conflicting in a 'credibility propagation network'. The conflicting views are identified using an unsupervised topic model. The network is constructed by using both opposing and supporting viewpoints. Propagation

of credibility in this network is formulated as a graph optimisation problem. As shown in Figure 2.3, a news article goes through three stages in its life cycle - creation, publication and spread in the media (2).

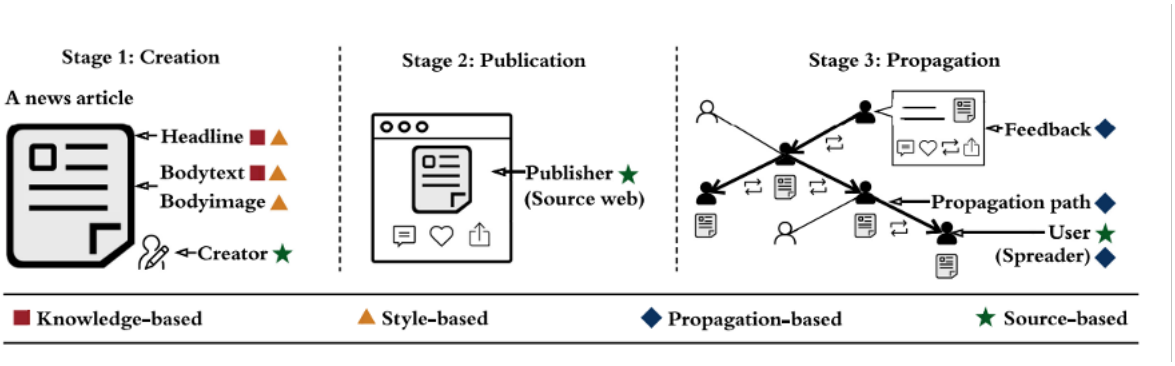

Figure 2.3: Life cycle of a news article with the four different fake news detection methods (2)

A challenge faced by propagation-based methods is that it is tough to determine whether the information is fake before the third stage (before the information is propagated). This limits the detection of fake information at the early stages. Detection of fake information at the early stages is essential because the longer a person is exposed to fake information, the more likely they are to believe it (14).

2.2.2 Content-Based Methods

In content-based methods, the news content can be represented from multiple perspectives such as – Knowledge, Style, and Latent representation (8).

Knowledge-based

Knowledge-based methods employ fact-checking to detect fake news. Fact-checking involves determining the authenticity of unknown information with known and verified facts. Knowledge-based methods are of three types: crowd sourcing, expert, and computational (15).

Crowd sourcing-oriented methods involve a large population of fact-checkers. Collective intelligence is used to identify whether a piece of information is fake. Fact-check websites that are crowd-sourced like Fiskkit (16) allow users to rate different sentences in an article and provide a tag that best describes an article.

However, crowd sourcing-oriented methods are not very credible due to potential bias and are difficult to manage. Therefore, filtering needs to be done for users that are not

credible, and inconsistent results need to be rectified, which becomes progressively more challenging as the number of users grows (2).

On the other hand, Expert-oriented methods involve a selective group of domain experts as fact-checkers. Unlike crowd sourcing methods, they are easy to manage and credible. Fact-checking websites that experts verify, such as PolitiFact (17), provide a 'truth-o-meter' that ranges from different intensities of False-Truth against popular tweets, news articles and social media posts. However, expert-oriented methods are hard to scale with a large amount of information.

Both crowd sourcing and expert-oriented methods involve manual fact-checking at some level.

To overcome the scalability issues associated with these methods, automatic fact-checkers are used. These automatic fact-checkers involve using Natural Language Processing (NLP), Information Retrieval (IR) and Machine Learning along with graph theory (18). 'Knowledge' is defined as tuples of Subject, Predicate, Object (SPO) that are derived from the text (8). For example, an SPO tuple for the sentence 'Narendra Modi is the Prime Minister of India' would be ('Narendra Modi', 'Profession', 'Prime Minister'). Link Prediction algorithms (19) (20) are used in Knowledge-based methods to determine the authenticity of news by extracting knowledge (SPO tuples) from the information and comparing it with ground truth data called Knowledge Vault (21) within a Knowledge Graph (KG). However, post-processing is required to infer the knowledge as KG's are incomplete (22). In addition, fact-checking recent information requires the knowledge in the KG to be timely. Finally, Knowledge-based approaches cannot distinguish between news that is false and fake (intentionally false) (2).

Style-based

Style-based methods can determine the intention behind a news article (whether the intent is to deceive the people or not). In contrast, knowledge-based methods can only determine the authenticity of the information. The Undeutsch Hypothesis (23) states that the style of writing for a true statement is different from a fake one. Style-based methods leverage the above hypothesis and identify the quantifiable features that make fake news different from real news. This facilitates building an ML model that can automatically detect phoney information using these features. These quantifiable features (ML) can be classified as textual and visual.

In a traditional Machine Learning framework, textual features are usually used to identify whether a piece of information is fake. These textual features can be differentiated at four linguistic descriptions such as – lexical, semantic, syntactic and discourse (24).

At the lexical level, the frequency lists of tokens in an article are analysed using the Bag of Words (BOW) model (12).

At the syntactic level, features are divides into deep-syntactic and shallow-syntactic features (25). The frequency of Part-of-speech tags (Nouns, adjectives, determiners, etc.) are used for evaluating the shallow-syntactic features. In comparison, deep-syntactic features involve assessing the frequency of rewrite rules. These re-write rules can be obtained from 'Probabilistic Context-Free Grammar' (PCFG) trees (8) (25).

At the semantic level, the sentiments of the news article are analysed by investigating psycho-linguistic attributes. Zhou et al. (8) separate the news content into the body and headline at the semantic level. This is because disinformation-related articles focus on the body, whereas clickbait articles focus on the headline.

At the discourse level, the relationship between the sentences of an article is investigated. The relationship between the sentences can be found using an RST parser (26). For a news article, the association is modelled using a tree in which the leaf nodes are the sentences and the non-leaf nodes are the relationship between the sentences.

To capture the sequence of tokens (rewrite rules, POS tags) at the different linguistic levels, Term Frequency Inverse Document Frequency (TFIDF) can be utilized (27).

Visual features comprise the images present in the news article. However, little research has been done to identify whether the information is fake by leveraging the photos (28). Non-latent features can be used to represent visual features as described in (29), where different features like a clarity score, clustering score, diversity score, coherence score, similarity distribution, etc., are derived from the images. However, if these images need to be processed by a deep learning model, a latent representation of the pictures is in the form of a tensor or a pixel matrix (30).

Latent Representations

Latent representation of a piece of information is the automatic generation of features using deep learning techniques such as Text-CNN (31) (32) (33) or tensor/matrix factorisation.

Latent representation could be derived at the word level (For example, Word2Vec (34)), sentence level or document level (For example, Doc2Vec) (35). The results are embeddings which can be input directly into Machine Learning or Deep Learning classifiers.

Although these features help detect fake news, they are difficult to understand. This limits the public's comprehension of fake news.

Using these textual (non-latent) and latent features of a piece of information, traditional Machine Learning models or Deep learning models can be built to identify fake

news.

2.3 Machine Learning Models

As discussed earlier, information content can be represented using non-latent or latent representations in a Machine Learning framework. These features are derived from the text at various levels (lexical, semantic, syntactic and discourse). They can also be derived from images in the news. In general, Machine Learning models can be supervised, unsupervised or semi-supervised. Supervised models (classifiers) are commonly used for identification of fake news using style-based methods. For instance, (36) classified fake news using multiple ML models such as Naïve Bayes, K Nearest Neighbour, Decision trees, and Random Forest. The highest accuracy was obtained using the Naïve Bayes approach.

Zhou et al. (8) show that non-latent features (textual and visual) perform much better than latent features. In addition, leveraging features from multiple linguistic levels (lexical, semantic, syntactic and discourse) perform much better than using features from any single level. Finally, rewrite rules and frequency of tokens perform better at fake news detection than other features.

2.4 Deep Learning Models

In a deep learning framework, latent representation of the information content is more widely used than non-latent representations. As discussed earlier, the information content is represented using embeddings at the word level (Word2Vec (34)) or the document level (Doc2Vec (35)). The images are usually represented using tensors or pixel matrices.

These embeddings are fed into a deep learning network such as Convolutional Neural Networks (CNNs) (31) (32) (33), Long Short-Term Memory (LSTMs) (36), Transformers (37) etc., to get the latent representation of the data. These features are then provided to a classifier like softmax to classify whether the news is fake.

For example, (37) uses a combination of BERT transformer and CNN to give an accuracy of 98.90 per cent in fake news detection. BERT creates an embedding for the text, and CNN is used to classify these embeddings into fake or not fake.

2.5 Conclusion

As visualised Figure 2.2, the green boxes with the orange path are chosen as the method followed in this dissertation based on the extensive research on current techniques to identify fake news. The following approach is summarised as follows:

1. Content-Based Fake News Detection.
2. Style-Based textual features are extracted at different linguistic levels.
    2.1 Lexical level – frequency lists using the BOW model
    2.2 Syntactic level – Part-of-speech analysis
    2.3 Semantic level- Candidate term analysis
    2.4 Pragmatic level – emotion analysis
3. TF-IDF is used to capture the sequence of tokens at different linguistic levels.
4. Non-latent features extracted in the above step are fed to ML models such as Naïve Bayes and Perceptron.

Chapter 3

Methods

3.1 Overview

As mentioned in the literature review, fake news can be detected using Knowledge or Content-Based methods. The focus of this dissertation is the Content-Based method, particularly the Style-based using non-latent textual features. Fake news detection using the above approach falls under the Text Analytics domain of Natural Language Processing (NLP). NLP itself branches from a more extensive field called Artificial Intelligence.

Text Analytics uses a combination of ML, linguistic techniques, and statistics to derive insights from significant amounts of unformatted or unstructured data.

Almost fifty per cent of the world's population is active on social media. This causes an enormous amount of daily data in the form of tweets, posts, blogs, reviews, comments, discussions, etc. Most of this data present online is unformatted or unstructured. Deriving insights from these large amounts of data can benefit multiple organisations. These insights can help improve customer satisfaction, improve profitability and, in our case, detect fake news.

In text analytics, insights are derived from a text by leveraging the grammatical structure of the language used. These grammatical structures are identified and understood using Part-of-Speech tagging in multiple text analytics methods. Part-of-Speech tagger developed by Stanford (38) is commonly used in multiple Natural Language Processing tasks. However, not much attention is paid to leveraging keywords in specialist texts. Specialist-texts or domain-specific texts contain specific keywords that are repeated often and can represent the core ideas of the text (39). The language used in these specialist texts concerns a target audience.

This dissertation considers three specialist texts– informative, imaginative, and fake news. All these three corpora are COVID-19 specific, but the style of writing differs due

to their different target audiences.

These texts contain a large number of domain-specific terms. The author could introduce these terms, which eventually become a part of the language. The frequency of usage of these terms varies greatly in specialist texts compared to general language texts. For example, the word 'coronavirus' has a much higher frequency of occurrence in specialist texts about COVID-19 compared to a general language corpus.

Therefore, this idea will be used to determine the terms used in the three specialist corpora and their frequency of occurrence. In addition, this will help to determine the areas of focus of these corpora and how it differs from the other specialist texts and the general language corpus.

Thus, the three corpora are analysed at different linguistic levels using frequency analysis, Part-of-Speech tagging, domain-specific terms, etc. These concepts are explained in further detail in this section.

An important branch of Natural Language Processing is Sentiment Analysis which helps us determine the sentiment associated with a piece of text. Therefore, to analyse these corpora at the pragmatic level, sentiment analysis using the Bag of Words model was performed to determine the sentiment associated with each corpus.

Figure 3.1 displays the system architecture as well as the steps involved in 'COVID-19 fake news detection' that are followed in this dissertation, and this section describes each of those steps in detail.

Key Terms Used

1. Open Class Words – words specific to the domain. Ex – Coronavirus, vaccine
2. Closed Class Words – common words in the language. Ex – the, and, or, but

These two classes are mutually exclusive from each other.

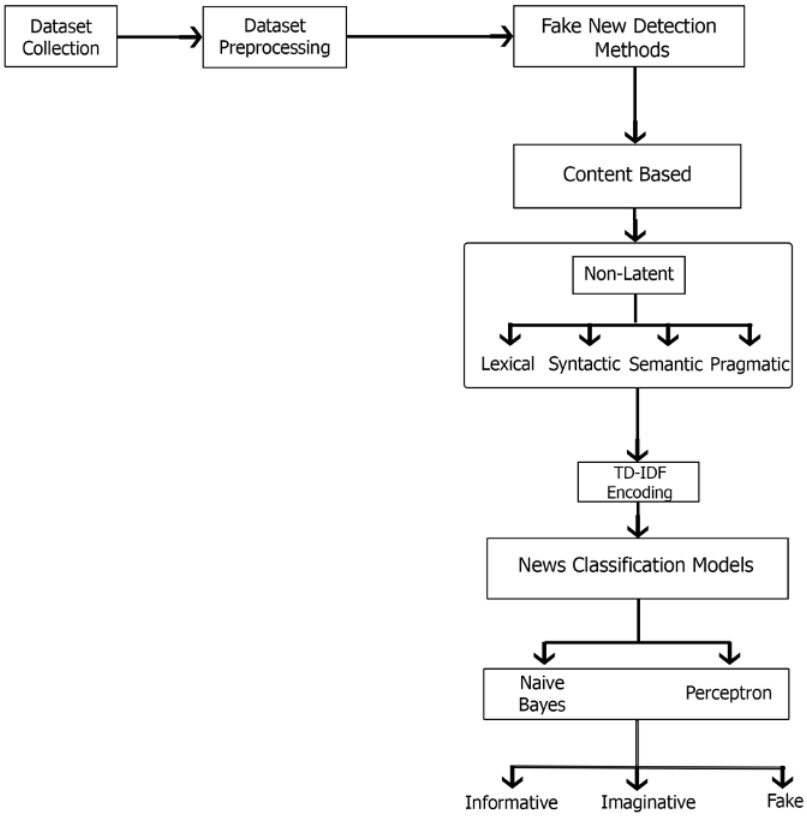

Figure 3.1: System Architecture

3.2 Data Collection

The first step of this research was to sample data randomly and generate a dataset of informative, imaginative, and fake news.

These three sources of informative, imaginative, and fake news form a special language corpus since they are domain-specific and are written concerning a target audience. Figure 3.2 displays the three domains of data. The informative domain consists of scientific research, proven facts, peer-reviewed papers, books, and expert opinions on the COVID-19 pandemic. In comparison, the imaginative domain consists of views by individuals in the form of posts or blogs on social media, advertisements for covid relief products and blogs on the pandemic. Finally, fake news consists of news or tweets written to mislead the public and cause panic.

The rationale behind choosing these three distinct genres of information was to analyse the difference between scientific research or fact, an opinion by an individual or a fake piece of information. The informative source is meant to educate the public and provide ground truth. Whereas the imaginative source can be truthful or not, it merely voices an

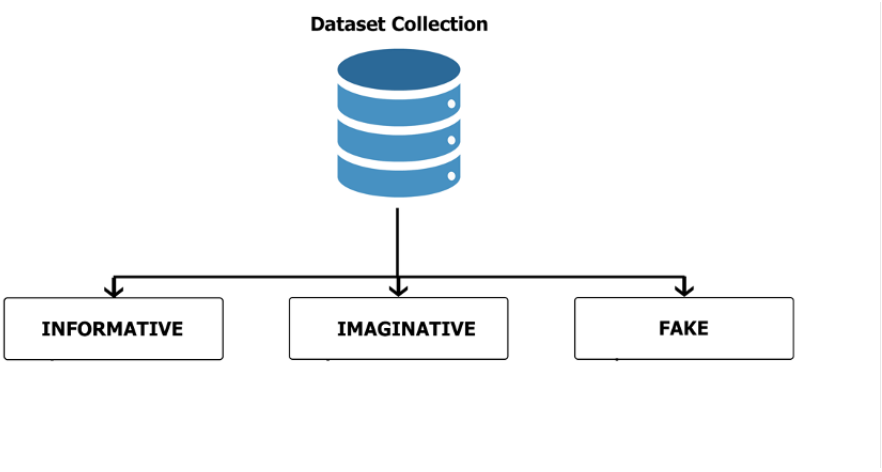

Figure 3.2: Data types

individual's thoughts and is not backed by research. However, fake news is entirely false and causes more harm than good. It can interfere with the government's efforts to curb the spread of the virus.

Differences are analysed concerning the writing style and the terminology used in these corpora to determine whether it could be a differentiating factor in building a fake-news detection model.

3.3 Data Pre-processing

As mentioned in the overview, most of the existing data is unstructured and unformatted. Therefore, an essential step before any analysis is pre-processing the data. Since the primary objective of this dissertation is to create a computer program that can detect fake news, the data that is input into the model needs to be machine-readable. It is also necessary for the data to be compatible with all the components in the system to ensure the algorithm works as intended. Figure 3.3 displays the different pre-processing steps taken to clean the data.

The first step in the pre-processing is ensuring the encoding of all the texts is uniform. This is done to ensure that the data is not misrepresented. This is essential since the text could be written in machines with different operating systems and encoding formats. Therefore, to standardise all texts, the encoding should be made uniform.

Based on the different experiments, 'UTF-8' (40) was the encoding format chosen for all the input text. UTF-8 was selected as the encoding format as it is the most widely used encoding for the World Wide Web. It makes up 97.7 per cent of the encoding of all web pages (41).

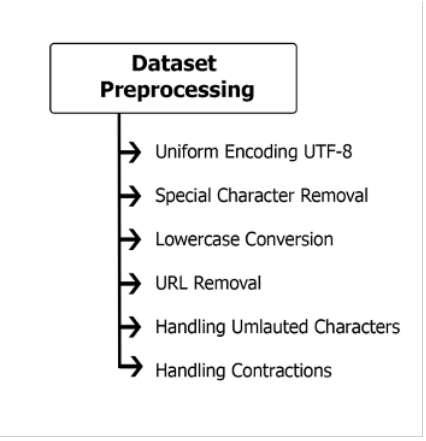

Figure 3.3: Data Pre-processing Steps

The second step in pre-processing the data was removing special characters from the text. Although these special characters have a meaning and are used for a specific purpose, they do not contribute much to the textual analysis and can be removed. In addition, since the specialist texts are compared with a general language corpus, it is essential to standardise the texts across all three corpora so that the words or tokens can be compared.

The third step in the pre-processing was to convert all the text to lowercase and remove any URL's from the data. Since most data is sourced from websites, removing any hyperlinks that might be found in the information is essential. In addition, converting all text to lowercase helps us easily compare the texts in the different corpora.

Removal of special characters, URLs and conversion to lowercase is done with the help of the RegEx (42) library in Python. It stands for 'Regular Expression' and can be used to find patterns in a string and replace them with another or altogether remove them.

The fourth step in the pre-processing was to standardise the umlauted characters in the string. Umlauted characters are the accent characters such as - ä, ë,ï,ö,ü. This was done with the help of the unidecode (43) library in Python.

The final step in the pre-processing was to split the contractions present in the text. For example, 'haven't' would be divided into 'have not'. This was done with the help of the contractions (44) library in Python.

Figure 3.4 summarises the results of the pre-processing.

As shown in Figure 3.4, special characters like '#' were removed. In addition, all characters were converted to lowercase. The downside is that abbreviations like the 'US' were converted to 'us', thereby changing the meaning. It is also observed that 'can't' has changed to 'can not', and the hyperlink has been removed.

After the data has been collected and pre-processed, it is fed into the fake-news detection model.

Figure 3.4: Example of Pre-processed Data

3.4 Fake News Detection

As described in the literature review (Chapter 2), fake-news detection models are content-based or propagation-based. Due to the success of content-based fake news detection models in the past, different content-based approaches were implemented in this dissertation. To convert the text into a numerical representation so that it can be input to a Machine Learning model, a TF-IDF (45) encoding is used which captures the style-based textual features at different linguistic levels. These models are then trained to classify the news as informative, imaginative, or fake.

To determine the intention behind a piece of information, the style of the text is leveraged. The style of the text can be grasped using multiple linguistic description levels such as – lexical, syntactic, semantic, and pragmatic.

The different linguistic levels for the pre-processed text are analysed using a Bag-of-Words model. The Bag-of-Words (BOW) is a Natural Language Processing (NLP) method for data representation. It uses the occurrence of tokens in the text and disregards the order of the tokens. Figure 3.5 displays the different linguistic levels considered in this dissertation and the analysis followed in each level.

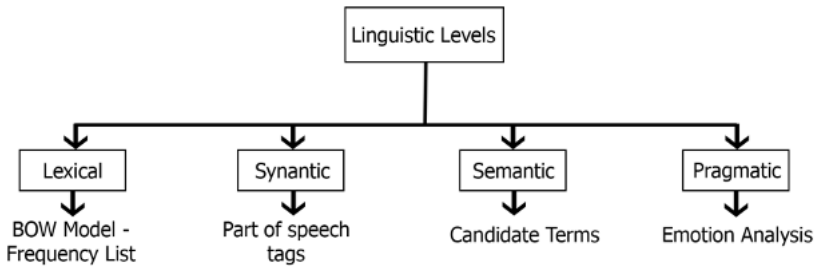

Figure 3.5: Analysis of Different Linguistic Descriptions

3.4.1 Lexical Level

At the lexical level, the pre-processed text is analysed at the lowest granularity by considering the unique words or tokens and their relative frequencies in the corpus. This is based on the concept of Lexical Cohesion. It concerns how similar words are used to connect the different elements of a text. Collocation and repetition are different forms of lexical cohesion. In this dissertation, the repetition form of lexical cohesion is leveraged to analyse the text at this level.

The relative frequencies of words are found in the Special Language Corpus (Informative, Imaginative, and fake) and ordered based on their frequency of occurrence (highest to lowest). The relative frequency of these tokens in a General Language Corpus (American National Corpus) is computed and compared, as shown in Table 3.1. This is similar to the approach followed by Ahmad et al. in (46). This is done to see if the relative frequency distribution of the hundred most frequently occurring tokens in the special language corpus varies from the general language corpus.

Table 3.1: Relative frequency distribution of hundred most frequently occurring tokens in Informative corpus and American National Corpus.

Word count	**Tokens**	**Cumulative Freq (SLC)**	**Cumulative Freq (ANC)**
1-10	['the', 'of', 'and', 'to', 'a', 'in', 'for', 'on', 'is', 'with']	20.58077175	17.87232273
11-20	['that', "'s", 'it', 'was', 'as', 'by', 'from', 'he', 'at', 'where']	5.280750971	4.358590909
21-30	['are', 'be', 'has', 'not', 'an', 'those', 'this', 'his', 'have', 'or']	3.685009835	3.134436364
31-40	['said', 'who', 'but', 'were', 'we', 'trump', 'they', 'people', 'had', 'new']	2.928998546	1.869354545
41-50	['coronavirus', 'been', 'about', 'you', 'their', 'sars', 'virus', 'which', 'covid', 'president']	2.148756125	1.0744
51-60	['can', "'s", 'health', 'would', 'than', 'one', 'patients', 'all', 'other', 'i']	1.788015161	1.138263636
61-70	['also', "n't", 'when', 'will', 'after', 'times', 'if', 'some', 'mr', 'what']	1.536980644	1.319531818
71-80	['its', 'could', 'time', 'mr', 'there', 'news', 'disease', 'vaccine', 'no', 'more']	1.323443801	0.775409091
81-90	['york', 'these', 'may', 'that', 'data', 'have', 'more', 'two', 'like', 'she']	1.193443783	0.977995455
91-100	['cov', 'many', 'how', 'our', 'cases', 'cough', 'care', 'those', 'into', 'up']	1.082949825	0.519081818

3.4.2 Syntactic Level

At the syntactic level, emphasis is made on how the tokens in a sentence are linked to each other. The relationship of the tokens within a sentence is analysed using Part-of-speech tagging. It is a Natural Language Process which marks up a token in a text and designates a part of speech based on the context and the meaning of the token. One of the fundamental steps in learning a new language is recognising different parts of speech such as verbs, adverbs, nouns, etc. This is the same as part-of-speech tagging implemented using Stanford's NLTK tagger (38). One such example of POS tagging is given in Figure 3.6. It is observed that the token 'arm' is classified as a Noun (NN) in the first sentence and as a Verb (VB) in the next sentence. This shows that the Stanford NLTK tagger considers the context of the token in a sentence. A detailed description of what each Part-Of-Speech tag stands for is provided in Table 3.2.

Table 3.2: Stanford NLTK POS tags

POS	**Description**	**Example**
CC	Coordinating Conjunction	and
CD	Cardinal Digit	1,2
DT	Determiner	the
EX	Existential there	there exists
FW	Foreign word	different language
IN	Preposition/subordinating conjunction	in
JJ	Adjective	large
JJR	Adjective, comparative	larger
JJS	Adjective, superlative	largest
LS	List marker	1)
MD	Modal	could
NN	Noun, singular	chair
NNS	Noun, plural	chairs
NNP	Proper Noun, singular	Jamie
NNPS	Proper Noun, plural	Students
PDT	Predeterminer	all
POS	Possessive ending	Sibling\'s
PRP	Personal pronoun	I, she, he
PRP$	Possessive pronoun	mine, hers
RB	Adverb	bad
RBR	Adverb, comparative	bigger
RBS	Adverb, superlative	worst
RP	Particle	take up
TO	to	come 'to' the shop.
UH	Interjection	umm
VB	Verb, base form	shake
VBD	Verb, past tense	shook
VBG	Verb, gerund/present participle	shaking
VBN	Verb, past participle	shaken
VBP	Verb, sing. present, non-3d	shake
VBZ	Verb, 3rd person sing. present	shakes
WDT	wh-determiner	which
WP	wh-pronoun	who
WP$	Possessive wh-pronoun	whose
WRB	wh-abverb	when

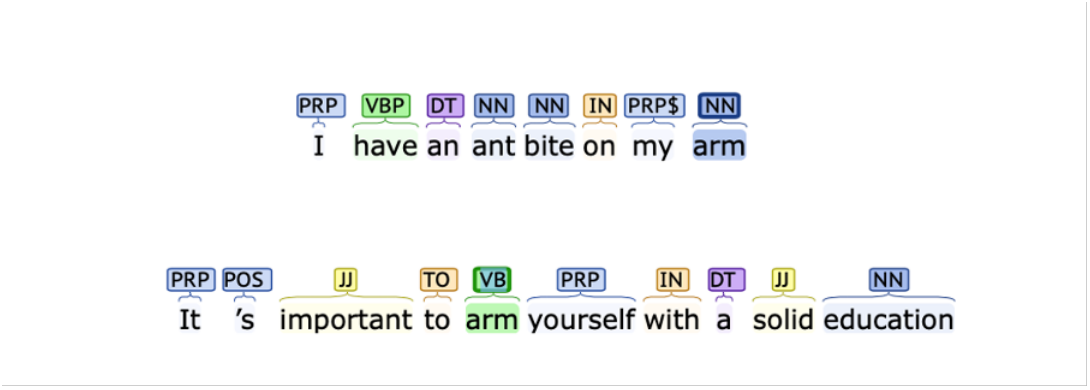

Figure 3.6: Part of speech tagging stanford POS tagger

The words can be classified as either 'Open Class Words ' (OCW) or 'Close Class Words' (CCW) based on the Part-of-Speech tag assigned to them. These classes are mutually exclusive of each other. Open Class Words comprise nouns, adverbs, adjectives and verbs. For example, 'Coronavirus', 'Trump', and 'Facebook' are Open Class Words. These are called Open Class Words as it is possible to add new words to this language class (For instance, the term 'Facebook' did not exist 50 years ago). At the same time, Close Class Words comprise pronouns, determiners, conjunctions, prepositions, and modal verbs. For example, words like 'the', 'and',' but', which act as the glue to the language, are called Close Class Words. New words are usually not added to this class; therefore, they are known as Close Class Words. These properties are summarised in Table 3.3.

Table 3.3: Close Class Words vs Open Class Words

CLOSE CLASS WORDS (CCW)	**OPEN CLASS WORDS (OCW)**
Pronouns, Modal Verbs, Determiners, Prepositions and Conjunctions	Nouns, Verbs, Adjectives and Adverbs
New words are not added to these language classes	New words are added to these language classes

3.4.3 Semantic Level

At the semantic level, emphasis is made on the sentence's meaning. It relies on the syntactic relationship of words in a sentence to interpret meaning. The purpose of a sentence is captured in the Open Class Words present in the sentence. Therefore, the analysis at this level is done on the Open Class Words. The first step in the study is to extract keywords known as Candidate Terms. These are the terms that occur very frequently in the corpus and provide an idea of the domain of the corpus. To determine

the Candidate terms, a 'weirdness' index and the corresponding 'z-score' is computed. This is based on the method followed by Ahmad et al. (39).

Weirdness Index Calculation

The weirdness index is calculated to identify the domain-specific keywords in a special language corpus by comparing the relative frequency of a word in the special language corpus to the frequency of the same word in the general language corpus. The general language corpus considered is the American National Corpus which consists of 22 million terms of spoken and written American English. This is chosen as the reference corpus because it represents the current language usage as it is updated regularly. The American National Corpus is available as a CSV (Comma Separated Values) file which can be quickly loaded into a python program. It contains four fields – sort-order, word, frequency, and word class. Sort-order is the position of the word considering the frequency of occurrence. The word class represents the part-of-speech tag discussed in the previous section.

Using the data from the special language corpus, the relative frequency of each word in American National Corpus is calculated.

The weirdness index is computed by the division of relative frequency of the word in special language corpus with the relative frequency in the American National Corpus.

The formula for Weirdness Index is given in Equation 3.1.

$$Weirdness\ Index = \frac{Relative\ frequency\ of\ a\ word\ in\ SLC}{Relative\ frequency\ of\ a\ word\ in\ GLC} \tag{3.1}$$

A high weirdness score indicates that the word frequency is high in the special language corpus and not so high in the general language corpus. This means that the word is domain specific due to the unusually high occurrence in the corpus.

However, some drawbacks exist to using just the weirdness index to determine the domain-specific words. First, since the language is constantly evolving, some new words might not have been updated in the general language corpus giving a weirdness score of infinity. Similarly, the weirdness index score is infinity for misspelt words and names not in the general language corpus. This leads to a false positive as such a high weirdness score would classify the word as domain-specific.

Z-Score Calculation

Two Z-scores are computed to overcome the drawback of computing just the weirdness index to obtain the domain-specific words. Z-Score determines the number of standard deviations below or above the mean; the value lies. The first Z-Score is calculated for

the weirdness index of the word, and the second Z-Score is calculated for the relative frequency of the word in the special language corpus. The formula for Z-Score is given in Equation 3.2.

$$Z - Score = \frac{x - \mu}{\sigma} \tag{3.2}$$

Where x is the word's relative frequency or weirdness index, μ is the mean of the relative frequency or weirdness index, and σ is the standard deviation of the relative frequency or weirdness index.

The Z-Score value can be interpreted in the following way -

- Z-Score = 0 indicates that the frequency is equal to the mean.
- Z-Score = 1 indicates that the frequency is one standard deviation above the mean.
- Z-Score = -1 indicates that the frequency is one standard deviation below the mean.

Candidate Terms

Using Equation 3.2, the Z-Score of weirdness and Z-Score of relative frequency are found. Only the words having a Z-Score of weirdness AND Z-Score of relative frequency greater than one are considered as Candidate terms. This avoids the false positives obtained by just using the weirdness index method, where the misspelt words and words not present in the American National Corpus were tagged as domain-specific words due to the weirdness score of infinity. The selection of the candidate terms can be visualised in Table 3.4.

Table 3.4: Candidate term selection criteria based on Z-Score Weirdness and Z-Score Frequency

Z-Score		**Weirdness**	
		>=1	**<1**
Frequency	**>=1**	CANDIDATE	NOT CANDIDATE
	<1	NOT CANDIDATE	NOT CANDIDATE

3.4.4 Pragmatic Level

At the pragmatic level, the overall text and its sentiment are considered to analyse the differences between the three special language corpora – informative, imaginative, and fake. It is essential to know the intent or emotion behind a piece of information to determine the kind of impact it will have on the public. The sentiment analysis of the text will help in doing so.

Sentiment Analysis:

Sentiment Analysis is a NLP tmethod to determine the emotion, opinion, or judgement behind the natural language. The primary step in sentiment analysis is identifying the polarity of a given text. The polarity is measured in terms of Neutral, Positive or Negative. It works using a Bag of Words model, where the polarity of each word is found, and a score is assigned. A +1 score is assigned to a positive word, a -1 score is assigned to a negative word, and a score of 0 is assigned to a neutral word. However, more advanced sentiment analysis would go beyond just analysing polarity and looking at different levels of emotions. In this dissertation, the sentiment analysis uses multiple emotion states such as Active, Passive, Strong, Weak, Positive and Negative. For each word in the corpus, a corresponding emotional state was found with the help of the General Inquirer corpus (47). The general Inquirer system is a tool for identifying the word's context. It consists of 182 attributes, of which 6 (Active, Passive, Strong, Weak, Positive and Negative) were considered for analysis.

The relative frequency of each emotional state was found and compared for each special language corpus.

Since the pragmatic analysis considers the text as a whole, the news articles can be represented as a vector using emotion words and close class words.

Open Class Word and Emotion vector:

Each news article is represented using the relative frequencies of different open class tags such as – NN (noun), NNP (Proper Noun, Singular), NNPS (Proper Noun, Plural), NNS (Noun, Plural), JJ (Adjective), RB (Adverb) and JJS (Adjective, Superlative) as described in Table 3.2. In addition, the relative frequencies of emotional states such as Positive, Negative, Active, Passive, Strong and Weak are used to represent a piece of information. Each news article is now represented using a 1 x 13 size vector. The similarities between the news articles can be found by computing the distance between these vectors. Euclidean distance was used to determine the length. The distance module from SciPy (48) was used to implement the Euclidean distance function in Python.

Euclidean Distance: The Euclidean distance between the two vectors can be found using the Pythagorean theorem. In general, the distance between a vector p and q is given by Equation 3.3.

$$d(p,q) = \sqrt{(p_1 - q_1)^2 + (p_2 - q_2)^2 + \cdots + (p_i - q_i)^2 + \cdots + (p_n - q_n)^2} \qquad (3.3)$$

Where pi and qi represent the relative frequencies of either OCW or emotion class in vector p and vector q, respectively.

3.4.5 Machine Learning

One issue with using Machine Learning for natural language is that the input to these algorithms is numerical. Since natural language is text, it is essential to convert the text into numerical values so they can be fed into different machine learning models for classification. These numerical values must be able to represent the text to a good extent to achieve good classification accuracy. A TF-IDF vector representation is used to represent the text.

TF-IDF

One of the most popular ways to generate a vector from text is using Term Frequency – Inverse Document Frequency (TF-IDF). It measures the important a token is in a corpus. TF-IDF consists of three important concepts – Term Frequency, Document Frequency, and Inverse Document Frequency.

- Term Frequency – It is the frequency of the term T in document D. This can be formulated by –

$$TF(T, D) = \frac{number\ of\ times\ T appears\ in\ D}{Total\ number\ of\ terms\ in\ D}$$

- Document Frequency – The document frequency determines the meaning of the text. It is the number of documents in which the term T is present. This can be formulated by –

$$DF(T) = occurrence\ count\ of\ T\ in\ N\ documents$$

- Inverse Document Frequency – Inverse document frequency determines the relevance of a word in the corpus. The term frequency, TF considers all words equal; however, close class words like 'and', 'the', of', etc., appear multiple times in a corpus but have low significance. Therefore, it is essential to scale down the frequently occurring terms and scale up the rare terms. This can be formulated by –

$$IDF(T) = \log_e \frac{Number\ of\ Documents}{Number\ of\ Documents\ that\ have\ T}$$

  The first step is calculating the Term frequency for all unique words in each document.

If the value of IDF is close to 0, then the term is common, and a higher value of IDF means that the word is rare. The logarithm is taken to avoid having very large values of IDF.

The term frequency and the inverse term frequency can be combined to give the formula for TF-IDF –

$$TF - IDF = TF\ (T, D) * IDF(T)$$

The higher the score of TF-IDF, the higher the significance of the word in that document. After finding the TF-IDF score for each word in the corpus, a vector can be created for each news article. The vector size would be 1 x the number of distinct words in the corpus.

The disadvantage of TF-IDF is that the calculations can become computationally expensive if the corpus is too large. The TfIdfVectorizer module from scikit-learn (49) was used to vectorise the articles and feed them into multiple machine learning models.

Once the text has been represented as a vector in numerical form using TF-IDF, it can be easily fed into different classification models. Two Machine Learning models are used to build fake news detection models – Naïve Bayes and a Perceptron. Naïve Bayes was chosen as it is commonly used in spam detection and gave the best accuracy for fake news detection in (36). The Perceptron was chosen to implement a deep learning technique and distinguish its performance with a simple Machine Learning model like the Naïve Bayes.

Naïve Bayes

Naïve Bayes is a simple machine learning algorithm that is probabilistic and uses the Bayes theorem. It assumes that the features are independent of each other. Due to this assumption of feature independence, it is known as 'naïve'. Equation 3.4 is the formula for Bayes theorem. It is useful for dealing with conditional probabilities.

$$P(A \mid B) = \frac{P(B \mid A) \times P(A)}{P(B)} \tag{3.4}$$

Where, P(A|B) is the posterior probability that hypothesis 'A' is true given the data 'B'.

In the context of fake news detection using news articles, Equation 3.4 can be rewritten as Equation 3.5.

$$P('Fake' \mid Article) = \frac{P(Article \mid' Fake') \times P('Fake')}{P(Article)} \tag{3.5}$$

Where P('Fake'| Article) is found which is the probability that an article is 'fake' given the article and P(Article | 'Fake') is the probability that the article will be found in 'fake' dataset. If this article is not present in the training data, then the probability P(Article | 'Fake') will be zero. However, this is where 'naïve' part of naïve bayes comes into picture. The article is split into different words since it assumes that each word is independent of the other words. Then, the probability that each word lies in the training data is found and combined to give the probability for the entire article.

Perceptron

The perceptron is a type of neuron that is artificially built to do computations to the input, which helps to detect the features present in the input. Multiple perceptrons work together to compute difficult problems, and this is known as multi-layer perceptron. Generally, it consists of input and output layers with several hidden layers in between as shown in Figure 3.7. It can be classified into Deep and Shallow Neural networks based on the number of hidden layers. The perceptron consists of 4 main parts – Inputs, Weights,

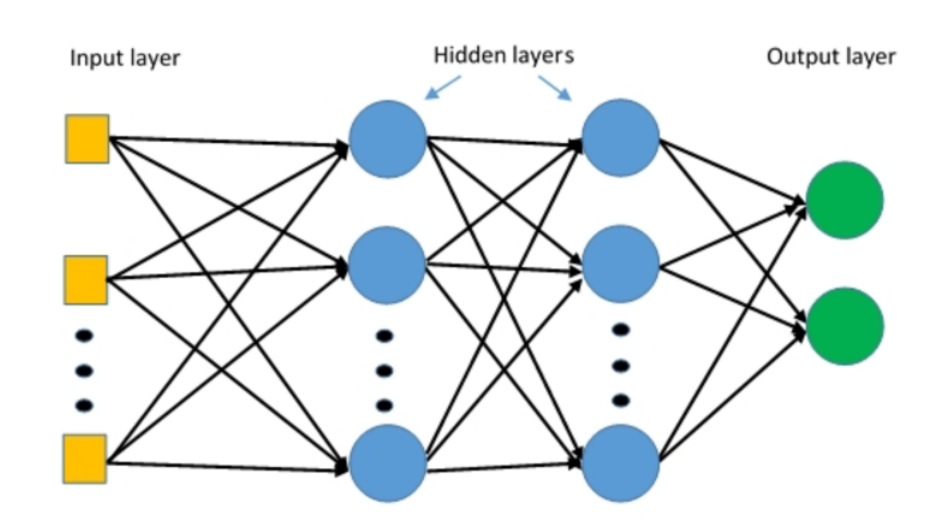

Figure 3.7: Multilayer perceptron architecture

Weighted sum and Activation function. The process can be briefly explained as follows:

1. Random initialisation of values to all perceptrons.
2. Compute the weighted sum.
3. Add bias, and this is fed to the activation function.
4. Output is fed to the following layers if any.

5. Once it reaches the output, the error is computed.

6. If a large error is present, they are propagated back, the weights are updated, and the process is repeated.

7. If the error is small, the model can be said to be trained.

Evaluation Metrics

To evaluate the performance of the machine learning models, a confusion matrix shown in Table 3.5 can be used. Using the values obtained from the confusion matrix, the following

Table 3.5: Confusion Matrix

	Predicted Fake	Predicted Real
Actual Fake	True Negative (TN)	False Positive (FP)
Actual Real	False Negative (FN)	True Positive (TP)

evaluation metrics can be calculated -

1. Precision – It is the proportion of the predicted positives that are actually positive. In this dissertation's use case, it finds how many articles that are predicted to be 'fake' are actually fake. Equation 3.6 provides the formula to calculate the Precision.

$$\text{Precision } = \frac{TP}{TP + FP} \tag{3.6}$$

2. Recall – It is the proportion of true positives that are actually predicted positive. In this dissertation's use case, recall measures how many 'fake' news articles were rightly predicted as 'fake' by the model. Equation 3.7 provides the formula to calculate the Recall.

$$\text{Recall } = \frac{TP}{TP + FN} \tag{3.7}$$

3. F1-Score – F1 score uses both Precision and Recall to give a single metric which is the harmonic mean of Recall and Precision. Equation 3.8 provides the formula to calculate the F1-Score.

$$\text{F1 Score } = 2 * \frac{Precision * Recall}{Precision + Recall} \tag{3.8}$$

3.5 Conclusion

In conclusion, the following method is used to build a fake news detection model-

1. Informative, imaginative, and fake news corpus are collected and pre-processed.
2. The pre-processed corpora are analysed at different levels of linguistic description.
3. At the lexical level, the three corpora are distinguished using frequency analysis and compared with a general language corpus.
4. At the syntactic level, part-of-speech analysis is used and two mutually exclusive classes – Open Class Words and Close Class Words are derived.
5. At the semantic level, candidate terms are found using weirdness index and Z-scores.
6. At the pragmatic level, sentiment analysis is performed and a vector is generated to represent the news article.
7. Two machine learning models – Naïve Bayes and Perceptron are used to construct a fake news detection model that are evaluated using Precision, Recall and F1-Score.

Chapter 4

Results and Observations

This chapter describes the various experiments conducted to test the fake news detection model and the results obtained. It presents the analysis performed at each linguistic level and provides the differentiation between the three corpora – informative, imaginative, and fake if any exists. The dissertation hypothesis that 'A computer program cannot differentiate between an article that is informative, imaginative or fake' is tested at linguistic descriptions such as lexical, semantic, syntactic, and pragmatic. Finally, two machine learning models are trained to differentiate the three corpora, and their performance is presented.

4.1 Dataset

As described in the previous chapter, three different corpora were collected – informative, imaginative, and fake. These special language corpora are compared with each other and a reference general language corpus, the American National Corpus (ANC). The dissertation aims to construct a model that can differentiate between these corpora and analyse the fundamental differences. Figure 4.1 displays the different sources from which the data was collected.

4.1.1 Informative

The 'informative' data was gathered from multiple sources.

The first source was LexisNexis (50). This website allows users to download historical news – local and global archives. LexisNexis provides access to about 83 billion articles obtained from over 10,000 sources. It gives the flexibility to get news by publication name, keywords, date, and publication type. The news articles can be downloaded in

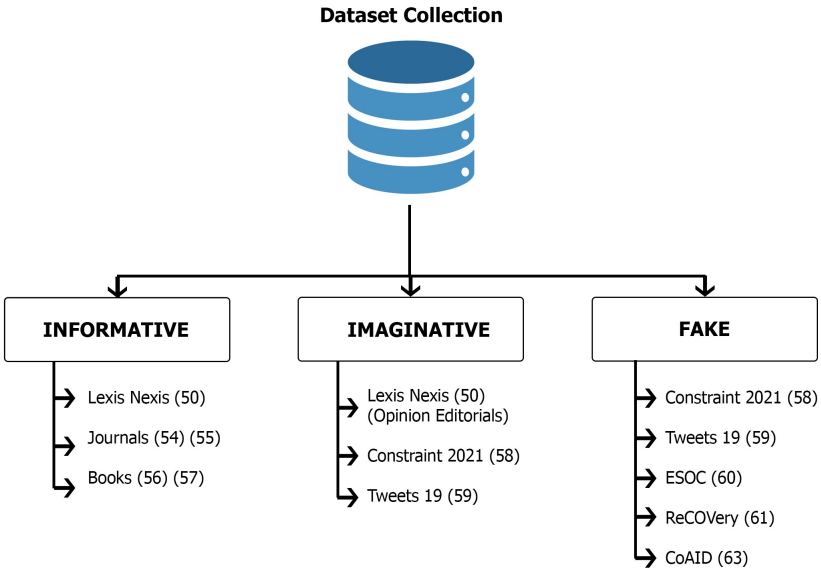

Figure 4.1: Sources used for dataset collection

multiple formats such as PDF, CSV, TXT, RTF, etc. Ahmad et al. (51) used LexisNexis to generate news articles for their research.

To get informative articles related to COVID-19, news articles from trusted publications such as New York Times (52) and Financial Times (53) were queried using the keywords 'COVID-19' and 'coronavirus'. Five hundred news articles were downloaded in bulk from December 2019 to March 2022 in CSV files along with author name, date of publication, article length, publication type, etc. The range of dates was chosen from December 2019 to March 2022 to capture information from the start of the pandemic till the current stage of the pandemic.

The other source for 'informative' data was a dataset collected from journals and books such as "Rapid expert consultations on the COVID-19 pandemic" (54) by the National Academy of Sciences, Science Immunology (55), "Imperial College COVID-19 response team" (56), and book on "Severe Acute Respiratory Syndrome (SARS)" (57).

4.1.2 Imaginative

The 'imaginative' corpus was compiled using multiple sources as well. The first source was LexisNexis (50), where the articles were queried with publication type as 'Opinion Editorials'. Opinion Editorials are editorials that typically reflect the view or opinion of the author. These articles are unsigned and contain the heading 'opinion' to warn the readers that the piece of information entailed reflects an opinion and not a fact. Five hundred Opinion-Editorials were sourced from December 2019 to March 2022 using the

keywords ‘COVID-19’ and ‘coronavirus’.

In contrast to the procedure for getting ‘informative’ data explained in section 4.1.1, publication-type was not provided in the ‘imaginative’ case. This is done to get the opinions of authors across multiple publication types to remove any bias towards a particular view. Whereas ‘informative’ data required a trusted source, and therefore, selected publications were chosen as the source. The other source for imaginative data was social media posts by people regarding the pandemic. Social media has emerged as the go-to platform for people to voice their opinion and has also become instrumental in influencing the general public. Therefore, it was essential to include this genre of information which is neither a fact nor fake but constitutes most of the content regarding the pandemic.

The social media posts were obtained from a dataset created by ‘CONSTRAINT 2021’ (58) to combat COVID-19-related fake news. It contains 10,700 posts from social media platforms such as Instagram, Facebook, and Twitter. Each post is manually labelled either ‘fake’ or ‘real’. In addition, tweets were taken from the ‘Tweets-19’ dataset from (59), which also aims to combat COVID-19-related fake news.

The social media posts corresponding to the ‘real’ label were chosen as the ‘imaginative’ data as real people write them. In contrast, social media posts annotated with ‘fake’ were selected as ‘fake news’ data as described in section 4.1.3.

4.1.3 Fake News

The ‘fake news’ corpus was built using multiple sources. One source was the ‘Empirical Study of Conflict (ESOC) COVID-19 Misinformation’ Dataset (60). It consists of 5,613 misinformation stories and information such as the country of origin, type of misinformation, dissemination channels, the motive behind the story and the source. The misinformation stories are collected from December 2019 to December 2020. The research in (60) aimed to examine the trend in misinformation locally and globally.

The second source for ‘fake’ news was the ‘ReCOVery’ Dataset (61) by Zhou et al., which consists of 2,029 fake news pieces collected between January 2020 to May 2020. The authors allocate each piece a ‘credibility’ score using the news website’s NewsGuard (62) score. NewsGuard allows the construction of a ‘fake’ news dataset by providing a ground truth.

The third source for ‘fake’ news was the ‘CoAID’ Dataset (63) which is short for ‘Covid-19 heAlthcare mIsinformation Dataset’. It consists of 516 social media posts and 1,896 news articles. Fact-checking websites and trusted websites were used to create this dataset.

Finally, the social media posts labelled ‘fake’ from the ‘Constraint 2021 (58)’ and the

'Tweets-19' Dataset described in section 4.1.2 were added to the 'fake news' corpus as well.

Table 4.1 displays the descriptive statistics of the relative frequency of tokens in the informative, imaginative, fake and American National Corpus datasets. The informative corpus consists of 69,106 unique tokens, the imaginative corpus consists of 27,175 unique tokens, the fake news corpus consists of 18,555 unique tokens, and American National Corpus consists of 293,867 unique tokens. Different statistical measures such as the mean, standard deviation, skewness, kurtosis, and the Z-score (minimum and maximum) are tabulated.

Table 4.1: Descriptive Statistics of relative frequency of tokens

	Informative ($N_{inf} = 69,106$)	**Imaginative** ($N_imag = 27,175$)	**Fake News** ($N_{fake} = 18,555$)	**ANC** ($N_{ANC} = 293,867$)
$Mean \times 10^4$	0.1447	0.368	0.539	0.031
$StdDev \times 10^2$	0.0315	0.0479	0.0592	0.0029
$Skewness$	108.39	54.82	48.637	232.88
$Kurtosis$	15114.04	3949.5	3143.83	70716.34
Z^f_{min}	-0.044	-0.075	-0.0798	-0.0207
Z^f_{max}	170.82	93.41	82.309	365.627

4.2 Lexical Level

At the lexical level, the analysis is done at the word level. The distribution of the relative frequencies of tokens are found for the different corpora and are arranged in decreasing order of relative frequency. Table 3.1 displays the relative frequencies of the hundred most frequently occurring words in the informative corpus. Table 4.2 displays the relative frequency of the top hundred words for the imaginative corpus, and Table 4.3 shows the top hundred words for the fake news corpus. In all three cases, it is observed that the distribution of the top hundred words for the three corpora is very similar to the American National Corpus.

Table 4.2: The frequency distribution of the top hundred words in the Imaginative corpus and American National Corpus.

	Tokens	**Cumulative Freq (SLC)**	**Cumulative Freq (ANC)**
0	['the', 'of', 'to', 'in', 'and', 'a', 'but', 'is', 'for', 'are']	20.426	17.120
1	['cases', 'on', 'we', 'coronavirus', 'with', 'from', 'that', 'covid', 'new', 'it']	6.353	3.172
2	['has', 'as', 'have', 'this', 'be', 'at', 'people', "'s", 'not', 'by']	4.634	3.447
3	['our', 'can', 'you', 'tests', 'been', 'who', 'will', 'there', 'deaths', 'health']	3.199	1.253
4	['states', 'or', 'was', 'more', 'number', 'an', 'they', 'that', 'than', 'but']	2.617	2.393
5	['india', 'all', 'vaccine', "'s", 'if', 'now', 'virus', 'about', 'today', 'which']	2.186	1.135
6	['your', 'total', 'their', 'state', 'one', 'day', "n't", 'no', 'were', 'testing']	1.861	1.042
7	['may', 'says', 'more', 'patients', 'case', 'positive', 'over', 'us', 'reported', 'other']	1.444	0.567
8	['hospital', 'also', 'when', 'he', 'during', 'after', 'days', 'data', 'do', 'pandemic']	1.266	1.054
9	['only', 'these', 'last', 'world', 'some', 'how', 'time', 'rate', 'have', 'up']	1.172	0.876

Table 4.3: Relative frequency distribution of top 100 words in Fake News corpus and American National Corpus.

	Tokens	Cumulative Freq (SLC)	Cumulative Freq (ANC)
0	['the', 'to', 'of', 'a', 'in', 'and', 'but, 'is', 'for', 'covid']	20.982	16.714
1	['that', 'on', 'from', 'it', 'with', 'has', 'are', 'covid', 'by', 'this']	6.258	4.258
2	['not', 'as', "'s", 'be', 'was', 'people', 'you', 'who', 'virus', 'will']	4.233	3.050
3	['have', 'at', 'can', 'new', 'an', 'been', 'we', 'they', 'that', 'or']	3.071	2.412
4	['cases', 'vaccine', "n't", 'he', 'trump', 'but', "'s", 'about', 'says', 'if']	2.381	1.726
5	['all', 'health', 'video', 'no', 'president', 'there', 'after', 'said', 'which', 'china']	1.916	0.958
6	['government', 'were', 'hospital', 'their', 'world', 'news', 'your', 'india', 'chinese', 'one']	1.687	0.645
7	['than', 'against', 'during', 'patients', 'pandemic', 'us', 'had', 'i', 'do', 'deaths']	1.500	0.634
8	['his', 'being', 'positive', 'man', 'shows', 'now', 'because', 'up', 'more', 'also']	1.330	0.888
9	['just', 'so', 'coronavirus', 'due', 'have', 'what', 'other', 'when', 'only', 'claim']	1.160	0.968

These tables can be visualised in terms of graphs for a better understanding. For all three corpora, the relative frequency of the top hundred words is plotted along with the corresponding frequency distribution in the ANC corpus, as shown in Figure 4.2, 4.3, and 4.4. In addition, the most frequently occurring word in each interval is plotted to show how the terminology varies in the three special language corpora.

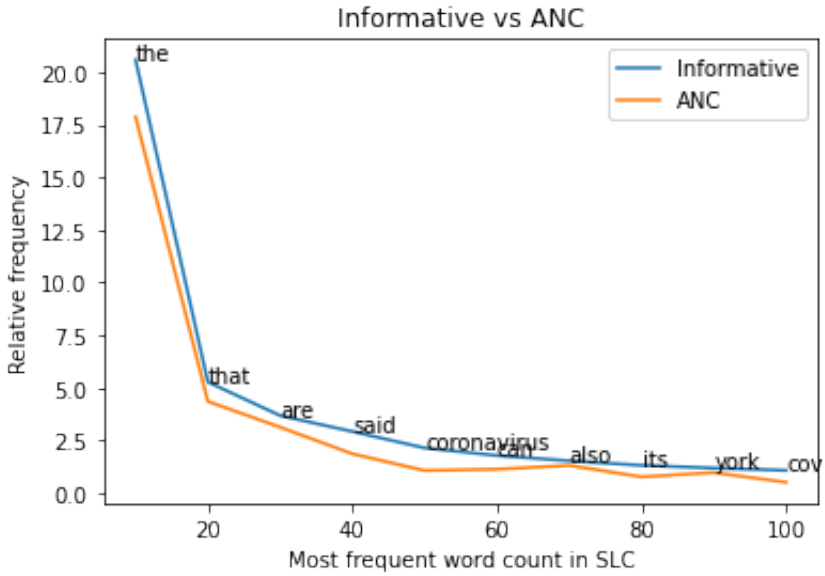

Figure 4.2: Relative frequency vs word count for Informative and ANC

According to 4.2, 4.3, 4.4, the distribution of relative frequency for the special language corpora in all three cases follows a very similar graph to the ANC. This is known as the 'Zipf' curve. Zipf's law states that the "frequency of any word is inversely proportional to its rank in the frequency table" (64). This means that the word that occurs most frequently ('the' in the English language) will occur twice as often compared to the 2nd most frequent word ('of' in the English language) and thrice as often compared to the 3rd most frequent word and so on. Therefore, regardless of the domain of the corpus, the Zipf law will hold. It is also observed that the informative corpus has frequently occurring terms like 'coronavirus' and 'cov'. In contrast, the imaginative corpus has frequently

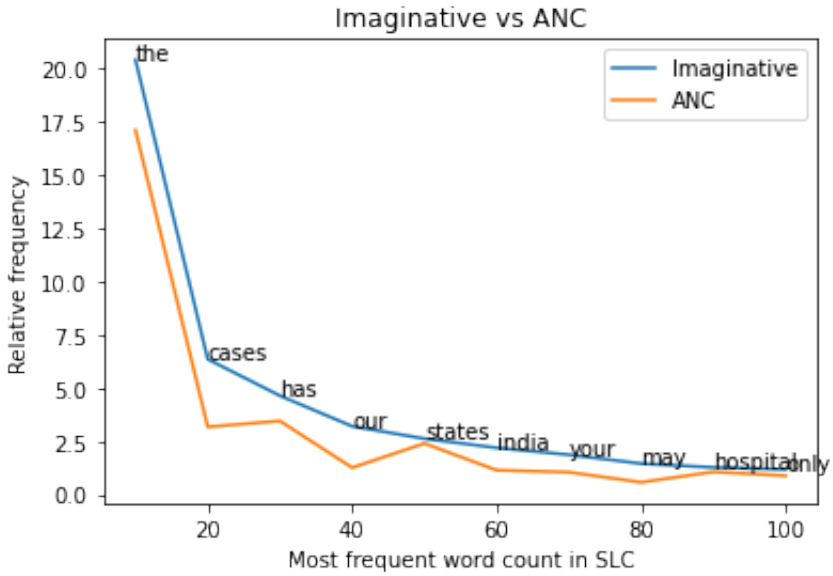

Figure 4.3: Relative frequency vs word count for Imaginative and ANC

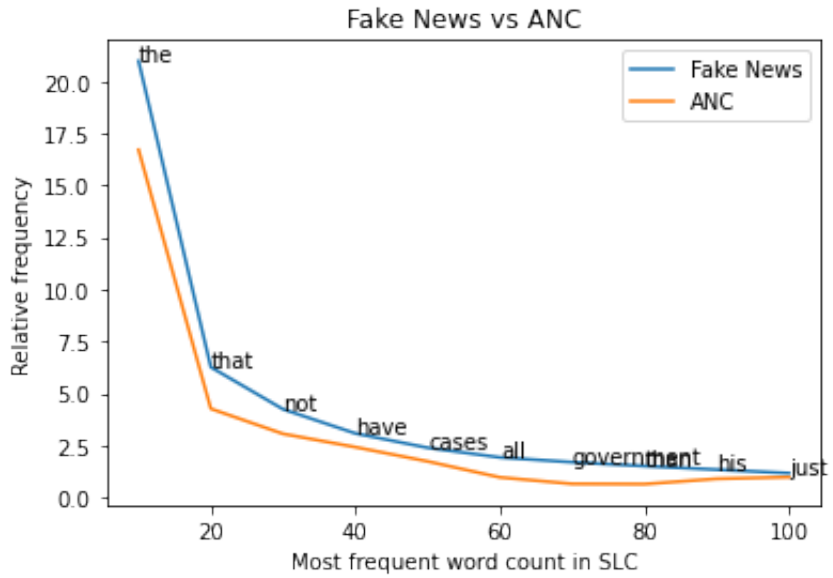

Figure 4.4: Relative frequency vs word count for Fake news and ANC

occurring terms like ‘cases’ and ‘states’, and the fake news corpus has frequently occurring terms like ‘cases’ and ‘government’. This gives an insight into the content of the three corpora.

4.3 Syntactic Level

At the syntactic level, the relationship between the words of a sentence is analysed. As mentioned in the previous chapter, this is done with the help of part-of-speech tagging. The percentage of different part-of-speech tags in the special language corpora are analysed and compared. Figure 4.5 displays the distribution of all the part-of-speech tags in the informative, imaginative, and fake news corpus. The part-of-speech tags are arranged in descending order of their occurrence. It is observed that Nouns make up most of the part-of-speech class, and the fake news corpus has the highest proportion of Nouns compared to the informative and the imaginative corpus. This is followed by the prepositions, and the adverbs, where the informative corpus is observed to have more adverbs than the informative and fake news corpus.

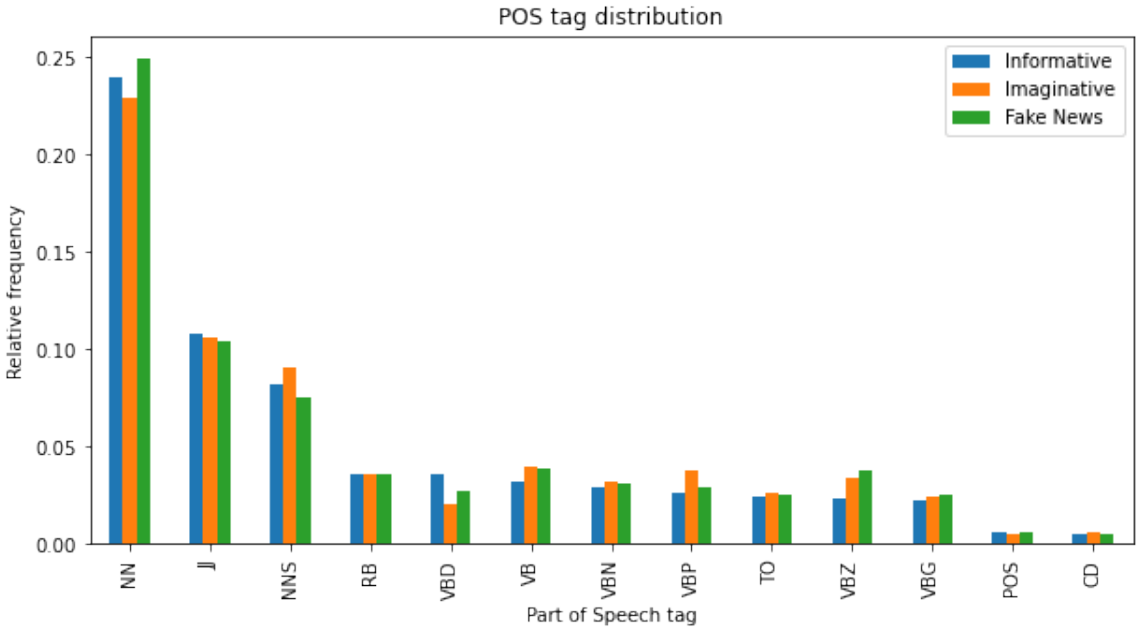

Figure 4.5: POS tag distribution in special language corpora

These part-of-speech tags can be split into Close Class Words (CCW) and Open Class Words (OCW). These word classes are mutually exclusive of each other. The Close Class Words comprise conjunctions, pronouns, determiners, modal verbs, conjunctions, and prepositions. Open Class Words contain verbs, adverbs, adjectives, and nouns. Figure 4.6 displays the distributions of Close Class Words and Open Class Words for the special language corpora (informative, imaginative, fake news) and the general language corpus (ANC). The following observations can be made –

1. The proportion of Open Class Words is greater than the Close Class Words for all corpora.

2. Special language corpora have a greater proportion of Open Class Words than ANC.

3. ANC has a higher proportion of Close Class Words than the special language corpora.

4. Among the special language corpora, the imaginative corpus has the highest proportion of Open Class Words, closely followed by fake news and informative corpus.

5. The informative corpus has the highest proportion of Close Class Words among the special language corpora.

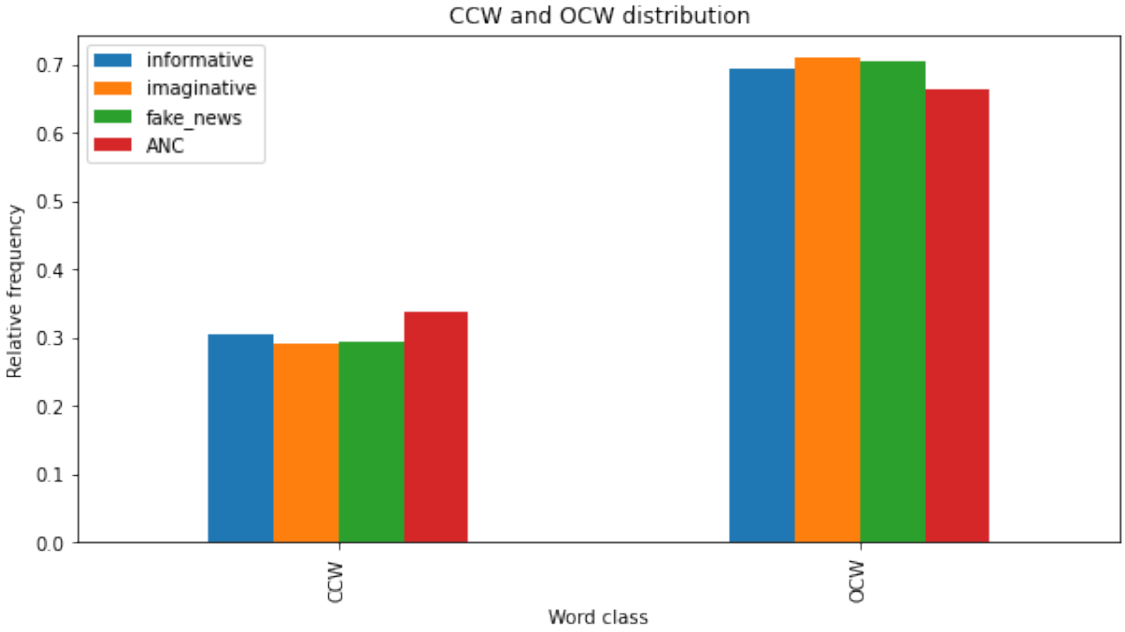

Figure 4.6: CCW and OCW distribution for special language corpora and the general language corpus

4.4 Semantic Level

At the semantic level, the meaning behind the sentence is analysed. This is done with the help of keyword extraction and candidate term analysis. As described in the previous chapter, a term is labelled as a 'Candidate' term when the Z-Score of relative frequency and the Z-score of weirdness are greater than 1. Table 4.4 displays the distribution of Z-Score of Weirdness and Z-Score of relative frequency in the informative corpus.

It is observed that the informative corpus contains 1.529% of candidate terms. Table 4.5 displays the distribution of the Z-Score of Weirdness and Z-Score of relative frequency

in the imaginative corpus. It is observed that 1.3445% of the corpus are candidate terms.

Table 4.4: Distribution of Z-Score of weirdness and Z-Score of relative frequency for informative corpus

Informative			
Z-Score		**Weirdness**	
		>=1	**<1**
Frequency	**>=1**	1.529%	9.748%
	<1	10.209%	78.512%

Table 4.5: Distribution of Z-Score of weirdness and Z-Score of relative frequency for imaginative corpus

Imaginative			
Z-Score		**Weirdness**	
		>=1	**<1**
Frequency	**>=1**	1.344%	10.469%
	<1	7.735%	80.45%

Table 4.6: Distribution of Z-Score of weirdness and Z-Score of relative frequency for fake news

Fake News			
Z-Score		**Weirdness**	
		>=1	**<1**
Frequency	**>=1**	1.187%	11.256%
	<1	7.429%	80.126%

Similarly, Table 4.6 displays the distribution of Z-Score of Weirdness and Z-Score of relative frequency in the fake news corpus. It is observed that 1.187% of the fake news corpus are candidate terms. In all three corpora, it is observed that most of the words (around 80% of the words in the corpus) have a Z-Score of relative frequency <1 and a Z-Score of weirdness <1. The informative corpus has the highest percentage of candidate terms compared to the imaginative and fake news corpus.

Table 4.7 displays the ten most frequently occurring candidate terms in each corpus. It is observed that 'Coronavirus' is the most frequently occurring candidate term in imaginative and fake news. At the same time, 'Trump' is the most frequently occurring term in imaginative news, followed by the word 'coronavirus'. Words common to the different corpora are coloured similarly to compare their ranks. This analysis shows the difference in terminology between the three corpora and gives an insight into the content.

Table 4.7: Ten most frequently occurring candidate terms in informative, imaginative and fake news

Informative	Imaginative	Fake News
Trump	Coronavirus	Coronavirus
Coronavirus	Pandemic	India
SARS	RT	Corona
China	Corona	Lockdown
Pandemic	Quarantine	Outbreak
Epidemiology	Epidemic	Quarantine
Closures	Bioweapon	Virology
Microbiology	Lockdown	Bioweapon
Platelets	Masks	Ebola
Pneumonia	Vaccination	China

It is observed that the informative corpus contains more scientific candidate terms such as 'SARS', 'Epidemiology', 'Microbiology', 'Pneumonia'. In contrast, the imaginative and fake news corpus contains more negative and general candidate terms like 'Corona', 'Quarantine', 'Bioweapon', 'Lockdown', etc.

4.5 Pragmatic Level

At the pragmatic level, the entire text and its sentiment are analysed, and a vector is generated to represent an article.

Sentiment Analysis

The sentiment analysis is done using a Bag of Words model where the relative frequency of Active, Passive, Strong, Weak, Positive and Negative words are found and compared. Figure 4.7 shows the distribution of the emotion words in the three special language corpora – informative, imaginative, and fake news. It is observed that 'Strong' is the most commonly occurring emotion among all three corpora, followed by Active, Positive, Negative, Passive and Weak. It is seen that the imaginative corpus has more 'Strong' emotion than the informative and fake news corpus. In addition, fake news has more 'Negative' emotions than the others.

Vector Generation

An article is represented as a vector using the relative frequencies of Open Class Words and the emotion words. An example of the vector generated for five articles is shown in Figure 4.8.

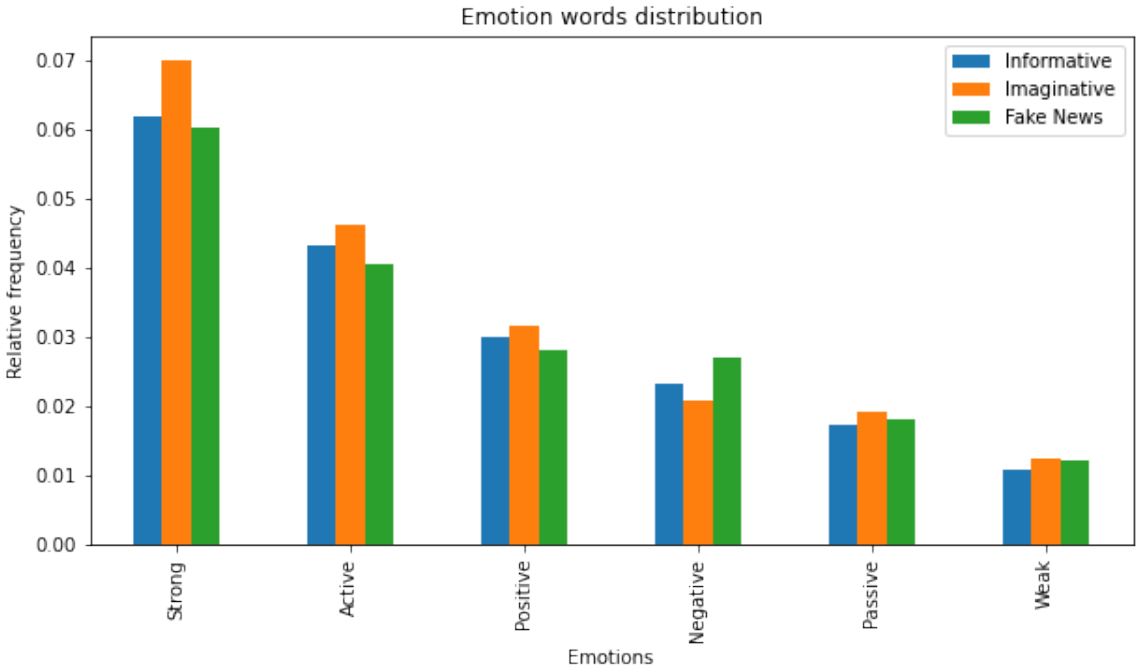

Figure 4.7: Distribution of emotional words in the special language corpora

Title	Publication Name	Date	Length	Author	Body	NN	NNP	NNPS	NNS	JJ	RB	JJS	Positive	Negative	Active	Passive	Strong	Weak	Text type
W.H.O. Also Is Confront	The New York Times	2020-02-07 00:00:00	1709 word	Matt Richt	Working w	0.150532	0.124231	0.002238	0.073307	0.06883	0.033576	0.00056	0.031337	0.025182	0.049245	0.012871	0.066032	0.002798	informativ
What Happens When Q	The New York Times	2020-02-09 19:02:00	2429 word	Mike McInt	A city coun	0.156789	0.143338	0.003363	0.071459	0.069777	0.022278	0.001681	0.026902	0.028163	0.040353	0.015973	0.060109	0.005885	informativ
Internet Delusion Ooze	The New York Times	2020-02-10 00:00:00	2398 word	Mike McInt	A city coun	0.14872	0.13206	0.003251	0.06989	0.065421	0.022349	0.001625	0.025193	0.027225	0.038602	0.015441	0.0577	0.005689	informativ
The Open Questions Ab	The New York Times	2020-02-11 00:00:00	1443 word	Gabriel Leu	Let's start w	0.146325	0.056865	0	0.080444	0.086685	0.042996	0.004161	0.029126	0.039528	0.049931	0.02982	0.074202	0.014563	informativ
The Urgent Questions Si	The New York Times	2020-02-11 19:16:00	1449 word	Gabriel Leu	Let's start w	0.150175	0.058947	0	0.080702	0.084211	0.042105	0.004211	0.029474	0.04	0.047719	0.030175	0.073684	0.014737	informativ
Maths behind a virus Fi	Financial Times (Lor	2020-02-15 00:00:00	1051 word	Clive Cooks	As coronav	0.162319	0.097585	0.000966	0.057971	0.09372	0.039614	0.001932	0.028019	0.037681	0.039614	0.014493	0.068599	0.01256	informativ

Figure 4.8: OCW and emotion vector for five articles

These vectors generated for each article are used to find the distance between the articles, which can be used as a similarity measure. These distances are stored in the form of a contingency matrix. Figure 4.9 shows an example of the Euclidean distances between six articles. The diagonal elements are zero as the distance between the same article is always zero.

	0	1	2	3	4	5
0	0	0.099287	0.084683	0.056152	0.125383	0.043398
1		0	0.059616	0.120303	0.092997	0.122589
2			0	0.087496	0.06624	0.103105
3				0	0.118831	0.051856
4					0	0.139242
5						0

Figure 4.9: Euclidean distance between six articles

Using the obtained vector representation of the articles, a Zmax (maximum of Z-Score) and Zmin (minimum of Z-Score) plot is drawn for the Open Class Words and the Emotion words. Figure 4.10 shows the distribution of Zmax vs Zmin for Open Class Words, and

Figure 4.11 shows the distribution of Zmax vs Zmin for Emotion words. The mean of OCW and emotion words for all articles in a corpus are plotted and compared. In both plots, it is observed that the fake news and imaginative corpus's OCW and emotion words are clustered closer than the imaginative corpus. A linearly separable plane can be drawn to separate the informative corpus from the creative and fake news corpus.

This leads to an understanding that a machine learning model can be trained to distinguish the imaginative, informative, and fake news

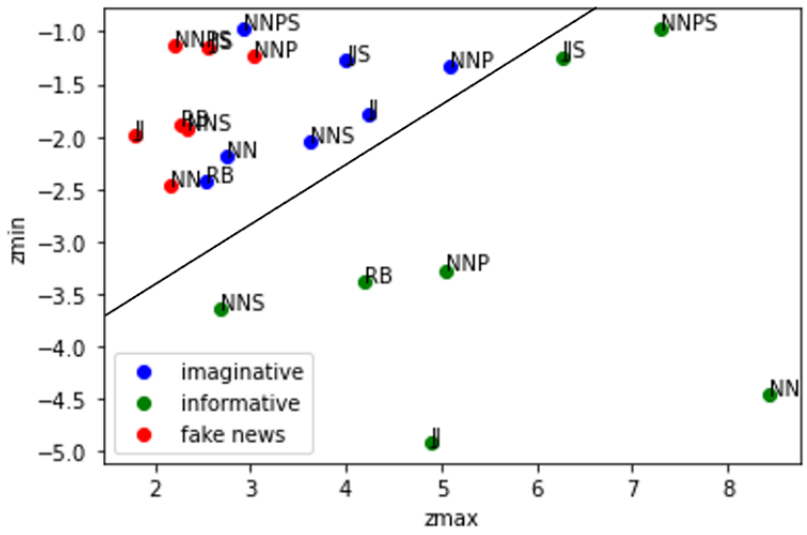

Figure 4.10: Distribution of Zmax vs Zmin for Open Class Words

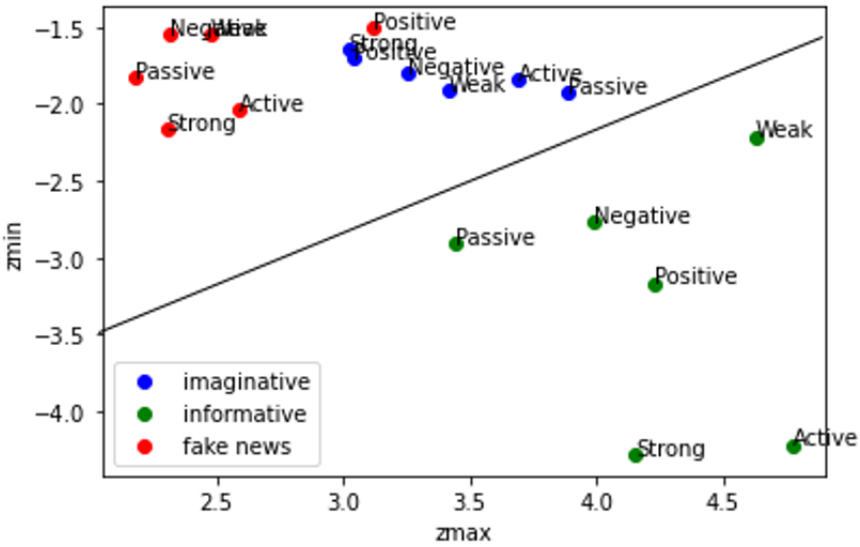

Figure 4.11: Distribution of Zmax and Zmin for Emotion words

4.6 Machine Learning Models

After analysing the informative, imaginative, and fake news corpora at different levels of linguistic description, it is essential to construct a model that can learn these differences to predict whether a given article is informative, imaginative, or fake. This is done using a probabilistic Machine learning model, Naïve Bayes, and a Perceptron model. As the input to the machine learning models needs to be numeric, a TF-IDF vector representation of the articles is used to train and test the models. Finally, the models are evaluated using Precision, Recall and F1-Score.

Multinomial Naïve Bayes

Naïve Bayes is a probabilistic machine learning algorithm that uses the Bayes theorem to determine whether the article is informative, imaginative, or fake. A multiclassification model is built using the sci-kit-learn (49) module in Python. To avoid class imbalance, 800 articles are chosen randomly for each informative, imaginative, and fake news. These articles are then split into train and tested with random shuffle using an 80-20 split ratio. Figure 4.12 shows the confusion matrix obtained for the multinomial naïve Bayes model. The model gave an accuracy of 0.57. The Precision, Recall and F1 scores for the different classes are shown in Table 4.8.

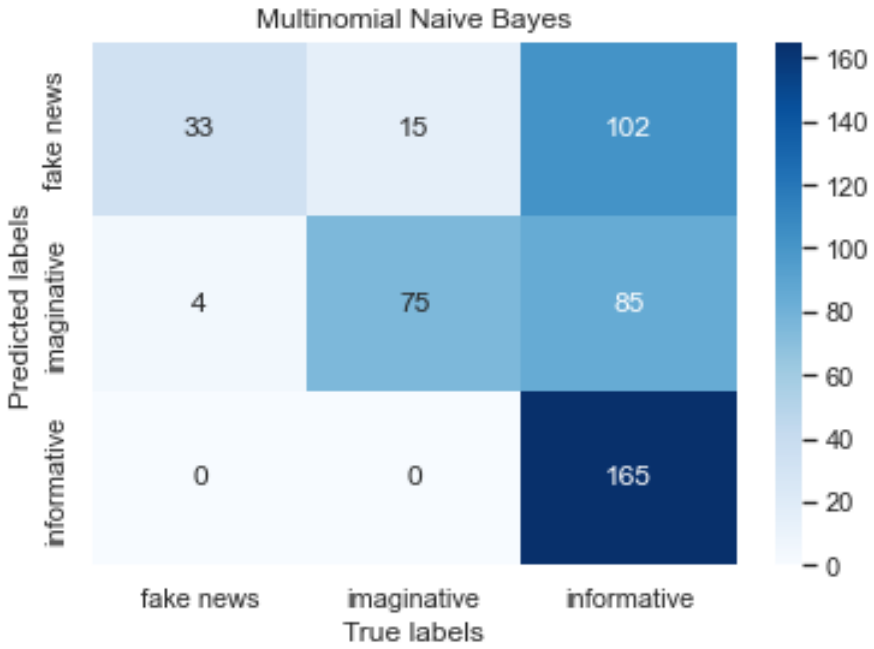

Figure 4.12: Confusion Matrix for Multinomial Naïve Bayes

It is observed that the naïve Bayes model gave a precision of 0.89 for fake news and

Table 4.8: Precision, Recall and F1-Score for Multinomial Naïve Bayes

	Precision	**Recall**	**F1-score**	**Support**
Fake News	0.89	0.22	0.35	150
Imaginative	0.83	0.46	0.59	164
Informative	0.47	1	0.64	165

a recall of 0.22. This means that the model is good at picking fake news as 89% of the predicted fake news is fake. However, out of all the fake news, only 22% of them were classified as fake. Similarly, for 'imaginative', the model performs slightly better with a precision score of 0.83 and a recall of 0.46. This means that 83% of all news predicted as 'imaginative' is genuinely imaginative. However, only 47% of all imaginative news was predicted as 'imaginative'. Finally, it is seen that the Naïve Bayes works best for the informative class with an F1 score of 0.64. Contrary to the fake and imaginative class, the recall is higher than precision. A recall value of 1 means that all the informative articles have been predicted correctly, and a precision score of 0.47 signifies that 47% of articles predicted as 'informative' are genuinely informative.

Overall, the Naïve Bayes classifier predicts most (73%) of the articles as 'informative' even though the training data was not imbalanced. However, when the model predicts an article to be 'fake' or 'imaginative', it is right 89% and 83% of the time, respectively.

Perceptron

Since a Naïve Bayes classification model is a simple probabilistic model, a more complex Machine Learning algorithm, Perceptron, is implemented to predict whether a given article is informative, imaginative, or fake. A multi-classification Perceptron model is built using the Perceptron module from sci-kit-learn. Similar to the procedure for training the Naïve Bayes model, 800 articles each from informative, imaginative, and fake news corpus were randomly split using an 80-20 train-test ratio. Figure 4.13 displays the confusion matrix obtained for the Perceptron model. The model gave an accuracy of 83.5%. This is significantly better than the performance of the Naïve Bayes model. The Precision, Recall and F1 score obtained for different classes are shown in Table 4.9.

Table 4.9: Precision, Recall and F1-Score for Perceptron

	Precision	**Recall**	**F1-score**	**Support**
Fake News	0.81	0.69	0.74	150
Imaginative	0.78	0.81	0.8	164
Informative	0.9	0.99	0.95	165

It is observed that the Perceptron performs much better than the Naïve Bayes in the

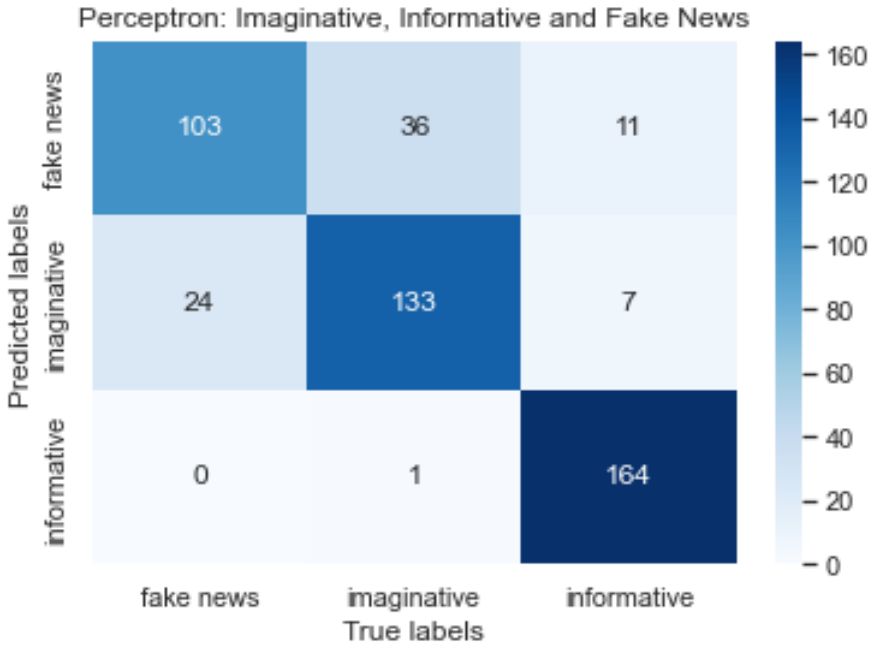

Figure 4.13: Confusion Matrix for the Perceptron

classification of fake news, with an F1 score of 0.74. The precision for fake news was 0.81, which means that 81% of the articles predicted as 'fake' were genuinely fake, and the recall score of 0.69 suggests that 69% of the fake news articles were indeed predicted as 'fake'.

For the imaginative corpus, the Perceptron achieves an F1-score of 0.8 with a precision of 0.78 and a recall of 0.81. This means that 78% of the articles predicted as 'imaginative' are truly imaginative, and 81% of all the imaginative articles were classified as 'imaginative.'

Like the Naïve Bayes, the Perceptron performs the best for the informative corpus giving an F1 score of 0.95 with a precision of 0.9 and a recall of 0.99. This means that 90% of the articles predicted as 'informative' were genuinely informative, and 99% of all informative articles were classified as 'informative'.

Overall, the Perceptron model performs much better than the Naïve Bayes classifier and can distinguish between the three corpora to a reasonable extent.

To further improve the model's accuracy, the data's complexity was reduced, and two binary classification Perceptron models were trained to differentiate fake news from informative and imaginative, respectively. Similar to the multi-class classification, 800 articles were chosen for each class and split into train-test using an 80-20 split.

Informative vs Fake News A perceptron model trained on the informative and fake news corpus alone gave an accuracy of 95%. Figure 4.14 displays the confusion

matrix for the binomial perceptron. Table 4.10 shows the Precision, Recall and F1 score for the binomial perceptron trained to classify informative and fake news articles.

It is seen that the model achieves a high F1 score of 0.95 for fake news and a high F1 score of 0.96 for informative articles.

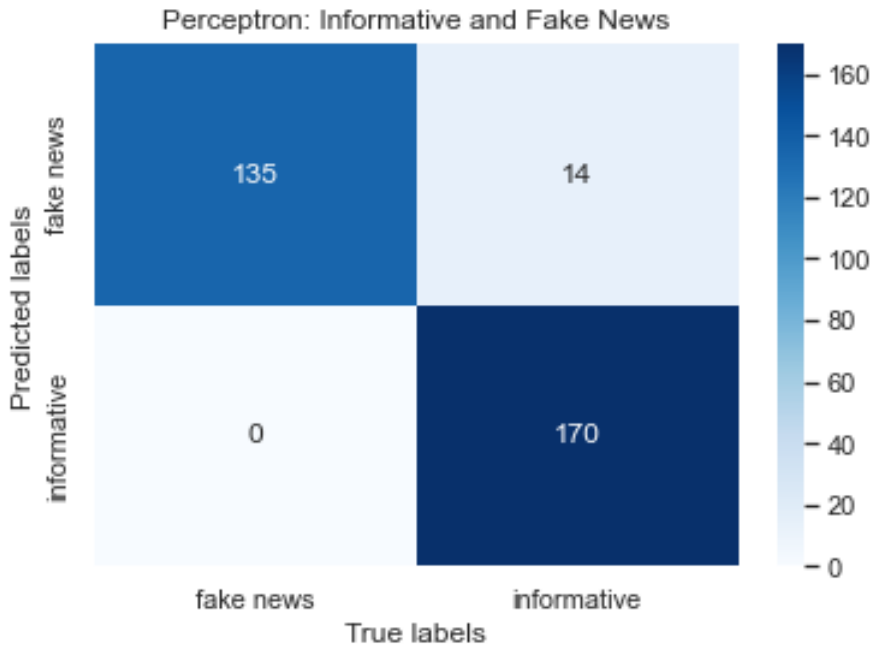

Figure 4.14: Confusion Matrix for the Binomial Perceptron – Informative vs Fake News

Table 4.10: Precision, Recall and F1-Score for Binomial Perceptron – Informative vs Fake News

	Precision	**Recall**	**F1-score**	**Support**
Fake News	1	0.91	0.95	149
Informative	0.92	1	0.96	170

Imaginative vs Fake News A binomial perceptron trained to distinguish imaginative articles from fake news gave an accuracy of 80.3%. Figure 4.15 displays the confusion matrix for the binomial perceptron. Table 4.11 shows the Precision, Recall and F1 score for the binomial perceptron trained to classify imaginative and fake news articles.

It is observed that the model gives an F1 score of 0.78 with fake news and an F1 score of 0.82 with imaginative articles.

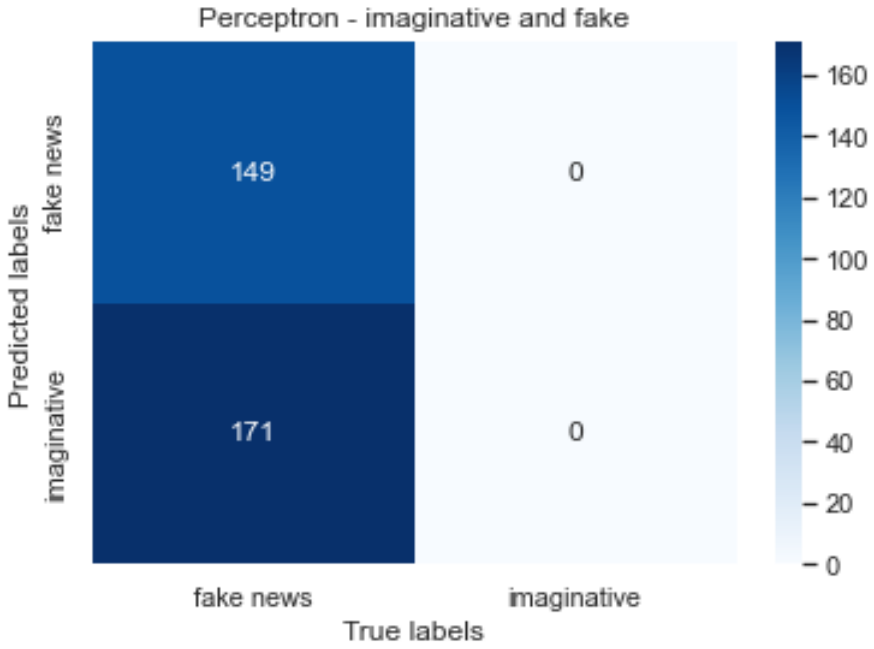

Figure 4.15: Confusion Matrix for Binomial Perceptron – Imaginative vs Fake News

Table 4.11: Precision, Recall and F1-Score for Binomial Perceptron – Imaginative vs Fake News

	Precision	**Recall**	**F1-score**	**Support**
Fake News	0.80	0.77	0.78	149
Informative	0.80	0.84	0.82	171

4.7 Discussion of Results

A fake news detection model was built to differentiate informative, imaginative, and fake news. At the lexical level, the distribution of relative frequency for the special language corpora and the general language corpus follows the Zipf curve. At the syntactic level, it is observed that the special language corpora have a higher proportion of Open Class Words than the general language corpus. In addition, fake news has the highest proportion of nouns, and informative corpus has the highest Open Class Word proportion. At the semantic level, the difference in terminology among the three corpora are observed, and it is found that the informative corpus has more scientific terms compared to the imaginative and fake news, which have more general and negative words. At the pragmatic level, sentiment analysis is performed to determine the emotion behind the text and a vector representation is created to find the distances between the articles. Finally, two machine learning models are trained to classify an article as imaginative, informative, or fake.

The multi-class perceptron model gave better accuracy (83.5%) than the simple naïve Bayes model (57%). In both the models, it was observed that the models were able to identify the informative articles better than the imaginative and fake ones. This could be because informative articles are generally longer as they are taken from journals and research papers compared to the imaginative and fake news, which were taken from blogs, advertisements, and social media posts. In addition, the performance of the binomial perceptron trained with imaginative and fake news corpus (F1 score of 0.78) is worse than that of the binomial perceptron trained with the informative and fake news corpus (F1 score of 0.95). This means that the model can better differentiate informative news from fake news than determine imaginative from fake news. This suggests that imaginative and fake news articles are more similar than informative articles.

Chapter 5

Conclusion and Future Works

5.1 Conclusion

The pandemic came along with an infodemic that has dire consequences and is detrimental to the health systems worldwide. This led to the motivation behind this dissertation to build a fake news detection system to help curb the spread of the infodemic that is so prevalent due to the power of social media. Along with fake news, social media has made it easy to share opinions on the virus, which may not be fake but are not entirely factual. Therefore, an extra level of granularity was added to build a model that can differentiate between informative, imaginative, and fake news. The three specialist corpora were analysed for differences at different levels of linguistic description, such as lexical, syntactic, semantic, and pragmatic. This is done to analyse the text at all levels, from the word level to the entire text. The lexical analysis at the word level showed that the distribution of the hundred most frequently occurring words in the specialist texts, as well as the general language corpus, follows a Zipf curve. The syntactic analysis using Part-of-speech tags to determine the relationship of words within a sentence showed that the specialist texts had a more significant proportion of Open Class Words than the general language corpus. The semantic analysis to understand the meaning behind the sentences using candidate terms showed the difference in terminology between the three corpora, which gave an insight into how the readability of the text differs and the kind of content in each corpus. Finally, the pragmatic analysis at the text level uses emotion analysis to understand the sentiment behind the different corpora. In addition, the vector representation provides a measure to represent the text numerically and find the distances between the articles, which can be used to cluster similar articles together. In addition, the Zmax and Zmin distribution plots for the open class words and the emotion words showed a clear separation of informative articles from the fake and imaginative. This

led to an understanding that a machine learning model could be trained to differentiate these classes. A TF-IDF representation of the news articles is used to train two machine learning models – Naïve Bayes and a Perceptron. The multinomial naïve Bayes model trained using equal samples of informative, imaginative, and fake news articles gave an accuracy of 57%. It issued a very high recall for the informative corpus and a low recall for the fake news corpus. This suggests that most of the articles were getting classified as informative. The model complexity was increased, and a Perceptron was trained to learn the class differences better. The multi-class Perceptron performed significantly better with an accuracy of 83.5%. It was able to differentiate between the three types to a reasonable extent giving an F1 score of 0.74 for fake news, 0.8 for imaginative articles, and 0.95 for informative articles. When a model was trained using only two classes to differentiate fake news from informative or differentiate fake news from imaginative, exciting results were obtained. It was found that the binomial perceptron gave an F1 score of 0.95 for fake news and an F1 score of 0.96 for informative. The binomial perceptron trained on imaginative and fake news gave an F1 score of 0.78 with fake news and an F1 score of 0.82 with imaginative. Therefore, the model can differentiate better between fake news and informative compared to fake news and imaginative. This means that fake news is more similar to an imaginative corpus. Based on these discussions and analyses, the hypothesis that 'A computer program cannot differentiate between informative, imaginative, and fake news' is rejected

5.2 Future Work

This section details the limitations faced during this dissertation. Future work based on this method must consider the limitations of this dissertation to expand and build a robust fake news detection model.

1. The informative, imaginative, and fake news articles are of different lengths due to different sources and target audiences. Training the model using articles of similar sizes would help improve the model's performance.

2. The analysis of the different corpora has been done using articles in the English language. Although the fake news detection model can work with any language, it has not been tested. In addition, an analysis of candidate terms and part-of-speech tagging were done using the American National Corpus. Therefore, a reference corpus must be available for the language of interest. A potential future work would be to build a fake news detection system for articles in a different language.

3. This dissertation only leverages the style-based textual features to build a fake news detection model. However, other features, such as the visual features, can also be incorporated. In addition, knowledge-based methods such as leveraging the social context would improve the model's accuracy.

4. This dissertation uses a simple bag of words model that does not consider the order of words in the text. Therefore, latent representation of the text, such as a word embedding (Word2Vec) or document embedding (Doc2Vec), can be used instead of a non-latent representation like TF-IDF.

5. A simple perceptron has been used for fake news detection in this dissertation. However, more complex deep learning methods such as the LSTM could be implemented with more training data.

This project could be extended to other applications as well. One such application is determining the impact of COVID-19 fake news on the stock market. This is currently a work in progress and soon to be published (65).

Bibliography

[1] W. H. Organization *et al.*, "Infodemic management: an overview of infodemic management during covid-19, january 2020–may 2021," 2021.

[2] X. Zhou and R. Zafarani, "A survey of fake news: Fundamental theories, detection methods, and opportunities," *ACM Computing Surveys (CSUR)*, vol. 53, no. 5, pp. 1–40, 2020.

[3] J. Roozenbeek, C. R. Schneider, S. Dryhurst, J. Kerr, A. L. Freeman, G. Recchia, A. M. Van Der Bles, and S. Van Der Linden, "Susceptibility to misinformation about covid-19 around the world," *Royal Society open science*, vol. 7, no. 10, p. 201199, 2020.

[4] B. Duffy and D. Allington, "Covid conspiracies and confusions: the impact on compliance with the uk's lockdown rules and the link with social media use," *The Policy Institute, King's College*, 2020.

[5] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay, "Scikit-learn: Machine learning in Python," *Journal of Machine Learning Research*, vol. 12, pp. 2825–2830, 2011.

[6] https://www.ofcom.org.uk/about-ofcom/latest/features-and-news/half-of-uk-adults-exposed-to-false-claims-about coronavirus, 2022.

[7] R. Zafarani, M. A. Abbasi, and H. Liu, *Social media mining: an introduction*. Cambridge University Press, 2014.

[8] X. Zhou, A. Jain, V. V. Phoha, and R. Zafarani, "Fake news early detection: A theory-driven model," *Digital Threats: Research and Practice*, vol. 1, no. 2, pp. 1–25, 2020.

[9] N. Ruchansky, S. Seo, and Y. Liu, "Csi: A hybrid deep model for fake news detection," in *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management*, pp. 797–806, 2017.

[10] S. Bauskar, V. Badole, P. Jain, and M. Chawla, "Natural language processing based hybrid model for detecting fake news using content-based features and social features," *International Journal of Information Engineering and Electronic Business*, vol. 11, no. 4, pp. 1–10, 2019.

[11] K. Shu, S. Wang, and H. Liu, "Beyond news contents: The role of social context for fake news detection," in *Proceedings of the twelfth ACM international conference on web search and data mining*, pp. 312–320, 2019.

[12] X. Zhou and R. Zafarani, "Network-based fake news detection: A pattern-driven approach," *ACM SIGKDD explorations newsletter*, vol. 21, no. 2, pp. 48–60, 2019.

[13] Z. Jin, J. Cao, Y. Zhang, and J. Luo, "News verification by exploiting conflicting social viewpoints in microblogs," in *Proceedings of the AAAI conference on artificial intelligence*, vol. 30, 2016.

[14] L. E. Boehm, "The validity effect: A search for mediating variables," *Personality and Social Psychology Bulletin*, vol. 20, no. 3, pp. 285–293, 1994.

[15] P. B. Petkar and S. Sonawane, "Fake news detection: a survey of techniques," *International Journal of Innovative Technology and Exploring Engineering*, vol. 9, no. 9, pp. 383–386, 2020.

[16] http://fiskkit.com/, 2022.

[17] https://www.politifact.com, 2022.

[18] S. Cohen, J. T. Hamilton, and F. Turner, "Computational journalism," *Communications of the ACM*, vol. 54, no. 10, pp. 66–71, 2011.

[19] G. L. Ciampaglia, P. Shiralkar, L. M. Rocha, J. Bollen, F. Menczer, and A. Flammini, "Computational fact checking from knowledge networks," *PloS one*, vol. 10, no. 6, p. e0128193, 2015.

[20] B. Shi and T. Weninger, "Discriminative predicate path mining for fact checking in knowledge graphs," *Knowledge-based systems*, vol. 104, pp. 123–133, 2016.

[21] X. Dong, E. Gabrilovich, G. Heitz, W. Horn, N. Lao, K. Murphy, T. Strohmann, S. Sun, and W. Zhang, "Knowledge vault: A web-scale approach to probabilistic knowledge fusion," in *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, pp. 601–610, 2014.

[22] M. Nickel, K. Murphy, V. Tresp, and E. Gabrilovich, "A review of relational machine learning for knowledge graphs," *Proceedings of the IEEE*, vol. 104, no. 1, pp. 11–33, 2015.

[23] U. Undeutsch, "Beurteilung der glaubhaftigkeit von aussagen refveracity assessment of statements. undeutsch," *Handbuch der Psychologie*, vol. 11, pp. 26–181, 1967.

[24] N. K. Conroy, V. L. Rubin, and Y. Chen, "Automatic deception detection: Methods for finding fake news," *Proceedings of the association for information science and technology*, vol. 52, no. 1, pp. 1–4, 2015.

[25] S. Feng, R. Banerjee, and Y. Choi, "Syntactic stylometry for deception detection," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 171–175, 2012.

[26] Y. Ji and J. Eisenstein, "Representation learning for text-level discourse parsing," in *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)*, pp. 13–24, 2014.

[27] V. Pérez-Rosas, B. Kleinberg, A. Lefevre, and R. Mihalcea, "Automatic detection of fake news," *arXiv preprint arXiv:1708.07104*, 2017.

[28] K. Shu, A. Sliva, S. Wang, J. Tang, and H. Liu, "Fake news detection on social media: A data mining perspective," *ACM SIGKDD explorations newsletter*, vol. 19, no. 1, pp. 22–36, 2017.

[29] Z. Jin, J. Cao, Y. Zhang, J. Zhou, and Q. Tian, "Novel visual and statistical image features for microblogs news verification," *IEEE transactions on multimedia*, vol. 19, no. 3, pp. 598–608, 2016.

[30] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *arXiv preprint arXiv:1409.1556*, 2014.

[31] J. Y. Khan, M. T. I. Khondaker, S. Afroz, G. Uddin, and A. Iqbal, "A benchmark study of machine learning models for online fake news detection," *Machine Learning with Applications*, vol. 4, p. 100032, 2021.

[32] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "Eann: Event adversarial neural networks for multi-modal fake news detection," in *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pp. 849–857, 2018.

[33] Y. Wang, F. Ma, Z. Jin, Y. Yuan, G. Xun, K. Jha, L. Su, and J. Gao, "Eann: Event adversarial neural networks for multi-modal fake news detection," in *Proceedings of the 24th acm sigkdd international conference on knowledge discovery & data mining*, pp. 849–857, 2018.

[34] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[35] Q. Le and T. Mikolov, "Distributed representations of sentences and documents," in *International conference on machine learning*, pp. 1188–1196, PMLR, 2014.

[36] R. K. Kaliyar, "Fake news detection using a deep neural network," in *2018* 4^{th} *International Conference on Computing Communication and Automation (ICCCA)*, pp. 1–7, IEEE, 2018.

[37] R. K. Kaliyar, A. Goswami, and P. Narang, "Fakebert: Fake news detection in social media with a bert-based deep learning approach," *Multimedia tools and applications*, vol. 80, no. 8, pp. 11765–11788, 2021.

[38] C. D. Manning, M. Surdeanu, J. Bauer, J. R. Finkel, S. Bethard, and D. McClosky, "The stanford corenlp natural language processing toolkit," in *Proceedings of* $52^{n}d$ *annual meeting of the association for computational linguistics: system demonstrations*, pp. 55–60, 2014.

[39] K. Ahmad, L. Gillam, L. Tostevin, *et al.*, "University of surrey participation in trec8: Weirdness indexing for logical document extrapolation and retrieval (wilder).," in *TREC*, pp. 1–8, 1999.

[40] F. Yergeau, "Utf-8, a transformation format of iso 10646," tech. rep., Yergeau, F, 2003.

[41] https://w3techs.com/technologies/cross/character_encoding/ranking, 2022.

[42] J. Hunt, "Regular expressions in python," in *Advanced Guide to Python 3 Programming*, pp. 257–271, Springer, 2019.

[43] https://pypi.org/project/Unidecode/, 2022.

[44] https://pypi.org/project/contractions/, 2022.

[45] L. Cui and D. Lee, "Coaid: Covid-19 healthcare misinformation dataset," *arXiv preprint arXiv:2006.00885*, 2020.

[46] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *arXiv preprint arXiv:1810.04805*, 2018.

[47] K. Ahmad, A. Davies, H. Fulford, and M. Rogers, "What is a term? the semi-automatic extraction of terms from text," *Translation studies: an interdiscipline*, vol. 267, p. 278, 1994.

[48] P. J. Stone, D. C. Dunphy, and M. S. Smith, "The general inquirer: A computer approach to content analysis.," *MIT press*, 1966.

[49] https://docs.scipy.org/doc/scipy/reference/spatial.distance.html, 2022.

[50] https://www.lexisnexis.com/, 2022.

[51] K. Ahmad, N. Daly, and V. Liston, "What is new? news media, general elections, sentiment, and named entities," in *Proceedings of the Workshop on Sentiment Analysis where AI meets Psychology (SAAIP 2011)*, pp. 80–88, 2011.

[52] https://www.nytimes.com/, 2022.

[53] https://www.ft.com/, 2022.

[54] E. National Academies of Sciences, Medicine, *et al.*, "Rapid expert consultations on the covid-19 pandemic: March 14, 2020-april 8, 2020," *National Academies Press*, 2020.

[55] K. Roltgen, A. E. Powell, O. F. Wirz, B. A. Stevens, C. A. Hogan, J. Najeeb, M. Hunter, H. Wang, M. K. Sahoo, C. Huang, *et al.*, "Defining the features and duration of antibody responses to sars-cov-2 infection associated with disease severity and outcome," *Science immunology*, vol. 5, no. 54, p. eabe0240, 2020.

[56] N. M. Ferguson, D. Laydon, G. Nedjati-Gilani, N. Imai, K. Ainslie, M. Baguelin, S. Bhatia, A. Boonyasiri, Z. Cucunubá, G. Cuomo-Dannenburg, *et al.*, "Impact of non-pharmaceutical interventions (npis) to reduce covid-19 mortality and healthcare demand," *Imperial College COVID-19 Response Team London*, 2020.

[57] J. S. Peiris, Y. Guan, and K. Yuen, "Severe acute respiratory syndrome," *Nature medicine*, vol. 10, no. 12, pp. S88–S97, 2004.

[58] P. Patwa, M. Bhardwaj, V. Guptha, G. Kumari, S. Sharma, S. Pykl, A. Das, A. Ekbal, M. S. Akhtar, and T. Chakraborty, "Overview of constraint 2021 shared tasks: Detecting english covid-19 fake news and hindi hostile posts," in *International Workshop on Combating On line Hostile Posts in Regional Languages during Emergency Situation*, pp. 42–53, Springer, 2021.

[59] T. Chakraborty, K. Shu, H. R. Bernard, H. Liu, and M. S. Akhtar, *Combating Online Hostile Posts in Regional Languages during Emergency Situation: First International Workshop, CONSTRAINT 2021, Collocated with AAAI 2021, Virtual Event, February 8, 2021, Revised Selected Papers*, vol. 1402. Springer Nature, 2021.

[60] F. Alam, F. Dalvi, S. Shaar, N. Durrani, H. Mubarak, A. Nikolov, G. Da San Martino, A. Abdelali, H. Sajjad, K. Darwish, *et al.*, "Fighting the covid-19 infodemic in social media: A holistic perspective and a call to arms.," in *ICWSM*, pp. 913–922, 2021.

[61] S. Siwakoti, K. Yadav, I. Thange, N. Bariletto, L. Zanotti, A. Ghoneim, and J. Shapiro, "Localized misinformation in a global pandemic: Report on covid-19 narratives around the world," *Princeton, NJ: Princeton University*, pp. 1–68, 2021.

[62] X. Zhou, A. Mulay, E. Ferrara, and R. Zafarani, "Recovery: A multimodal repository for covid-19 news credibility research," in *Proceedings of the 29th ACM international conference on information & knowledge management*, pp. 3205–3212, 2020.

[63] https://www.newsguardtech.com/, 2022.

[64] G. Pennycook, T. D. Cannon, and D. G. Rand, "Prior exposure increases perceived accuracy of fake news.," *Journal of experimental psychology: general*, vol. 147, no. 12, p. 1865, 2018.

[65] M. Mansoor, A. Goyal, and K. Ahmad, "Impact of covid-19 fake news on stock markets [to be published]," 2022.

[66] Y. Bang, E. Ishii, S. Cahyawijaya, Z. Ji, and P. Fung, "Model generalization on covid-19 fake news detection," in *International Workshop on Combating Online Ho stile Posts in Regional Languages during Emergency Situation*, pp. 128–140, Springer, 2021.

Appendix

.1 Dataset used for evaluation

- ESOC Covid-19 Misinformation Dataset

  (https://drive.google.com/file/d/15mN4MYKH46kIN1j_OU3QC-wdx5KvCRh5/view)

- CMU-MisCov19

  (https://zenodo.org/record/4024154#.YoNuYi9U1QI)

- ReCOVery

  (https://github.com/apurvamulay/ReCOVery/tree/master/dataset)

- MM-COVID

  (https://drive.google.com/drive/folders/1gd4AvT6BxPRtymmNd9Z7ukyaVhae5s7U)

- CoAID

  (https://github.com/cuilimeng/CoAID)

- FakeCOVID

  (https://gautamshahi.github.io/FakeCovid/)

- Constraint 2021

  Mailed authors of (66) for dataset and stored it in the Google Drive link.

  (https://drive.google.com/file/d/1R6UomPB2nVJXyA1062GExWpot4Q5rCk9/view?usp=sharing)

...