University of Essex

School of Mathematics, Statistics
and Actuarial Science

MA981 DISSERTATION

Detecting Health Misinformation In Online News Articles Using Naive Bayes And SciBERT

HUU ANN TRAN
2322761

Supervisor: DR LISA VOIGT

April 13, 2025
Colchester

Contents

Abstract 8

1 Introduction 9

1.1 General Introduction 9

1.2 Consequences of health misinformation 10

1.3 Outline 12

2 Literature Review 13

2.1 Overview 13

2.2 Techniques and Models 13

2.2.1 Traditional Machine Learning Approaches 14

2.2.2 Deep Learning Approaches 14

2.3 Detection in Health News Articles 15

3 Research Aims 17

4 Background 18

4.1 (Multinomial) Naive Bayes 18

4.1.1 Overview 18

4.1.2 Hyperparameters 19

4.1.3 Feature Extraction 20

4.2 BERT and SciBERT 20
4.3 Model Evaluation 21
4.3.1 Accuracy 22
4.3.2 Precision 22
4.3.3 Recall 23
4.3.4 F1 score 23
4.3.5 Macro average 23
4.3.6 Weighted average 23
5 Methodology 25
5.1 General Overview of Data 25
5.2 Data Cleansing and Preparation 26
5.3 Exploratory Data Analysis 27
5.3.1 Distribution of words counts 27
5.3.2 News sources 31
5.3.3 Most common topics 32
5.3.4 Top words per label in article texts 32
5.3.5 Comparison words and words of uncertainty 32
5.4 Methods 34
5.4.1 Metrics 34
5.4.2 Multinomial Naive Bayes 36
5.4.3 SciBERT 38
6 Results 39
6.1 Multinomial Naive Bayes 39
6.1.1 Random Classifier 39
6.1.2 Initial model - Bag-of-Words and no hyperparameter tuning . . . 39

6.1.3 TF-IDF versus Bag-of-Words 41

6.1.4 Grid search for best alpha and fit prior 41

6.1.5 Removal of custom stop words 42

6.1.6 Changing threshold for unsatisfactory class 42

6.1.7 Performance on the test set 43

6.2 SciBERT 43

7 Discussion **45**

7.1 Comparison of MNB and SciBERT results 45

7.2 Limitations and Potential Issues 46

7.3 Drivers of health misinformation and its spread 47

8 Conclusion **49**

A Appendix **50**

List of Figures

5.1 The number of articles labeled as ’satisfactory’ is more than twice that of those labeled ’unsatisfactory.’ 28

5.2 The majority of articles had approximately 200 to 900 words 29

5.3 ’Satisfactory’ articles appear longer than ’unsatisfactory’ articles. Satisfactory articles had a mean of 690.48,standard deviation of 384.90, and median of 668 words. Unsatisfactory articles had a mean of 499.54, standard deviation of 311.57, and median of 433. Welch’s t-test of independence reveal that the means are indeed significantly different. 30

5.4 Reviews come from a range of different news sources in US media. Ratings vary considerably per news source 31

5.5 In ’satisfactory’ articles, the words ’health’, ’studies’, ’medical’, ’results’,’medicine’, and ’surgery’ appear more than in the unsatisfactory articles. These words may help indicate whether an article gives a balanced view of treatment. 33

5.6 Bar graphs showing the top 30 words in satisfactory and unsatisfactory titles, respectively, with stopwords removed. The figures show that in the top used words in satisfactory titles the modal words of probability like ’could’, ’might’ and ’more’ are used relatively more than in the unsatisfactory titles. This indicates that satisfactory labelled articles may use more words that show more uncertainty as not to overstate the benefits of a particular treatment. 35

6.1 Summary of results for Multinomial Naive Bayes. Meanings of superscript: **B** means BoW, **1** means alpha=1, **0.5** means alpha=0.5, **F** means fit prior=true, and **S** means custom stop words. Overall, the Multinomial Naive Bayes model performed better than the random classifier, however, the hyperparameter tuning etc. made minimal difference. Bag of Words also performed much better than TF-IDF. 40

6.2 Over the course of each epoch, the model performed very well on the training sets, but did not do as well on the validation sets. On the test set, however, the model performed better than the highest scores for the validation set. This suggests the model is able to performed well on unseen data. 44

A.1 List showing all the news sources evaluated in HealthStory content . . . 51

List of Tables

5.1 Table showing the criteria used to evaluate the news articles. Adapted from *Ginger cannot cure cancer: Battling health news with a comprehensive data repository* [1] 26

5.2 Table showing how ratings of news stories are assigned on the HealthNewsReview.org website based on the 10 criteria shown in 5.1. The corresponding labels used in this paper are also shown. Ratings were only given for articles that had two or fewer criteria deemed ’not applicable’. For articles containing criteria deemed not applicable, the denominator was adjusted accordingly in the percentage calculation. Adapted from HealthNewsReview.org [2] 27

5.3 Table showing the the distributions of the training set, validation set, and test set. The data was split with stratification to reflect the class imbalance in the dataset. 36

Abstract

Despite a large amount of research concerning misinformation detection for COVID-19 on social media, very few studies focus on general health misinformation detection in online news articles. The aim of this dissertation is to investigate whether Multinomial Naive Bayes can be used to classify generic health misinformation in online news articles in alignment with health misinformation criteria, and compare it to the performance of the domain-specific pre-trained model SciBERT. 1180 labelled news articles and reviews from the Health News Review website were used to train the models. The results were as follows: overall, the Multinomial Naive Bayes model achieved a weighted F1 score of 0.66, while the SciBERT model achieved 0.72 for the same metric. Thus, both models demonstrated reasonable ability in health misinformation detection, with SciBERT showing a clear improvement in performance.

CHAPTER 1

Introduction

1.1 General Introduction

The rise of the Internet has led to a wealth of information available instantly, and while much of this information is very helpful, a substantial amount of harmful information also exists online. For the purpose of this dissertation, health misinformation will be defined as incorrect or misleading information related to health, whether intentional or unintentional. Health misinformation in particular can be especially harmful compared to other fake news, as it can have direct consequences on a population's quality of life and even mortality risk, be it on an individual level or on the whole population. The nuanced nature of health misinformation also makes it harder to detect than general fake news. Previous studies have shown Naive Bayes performs well on health misinformation detection on webpages and general fake news, and that BERT-based models tend to perform better than traditional models. However, very few studies have investigated their performance on general health misinformation detection in online news articles.

The detection of health misinformation has been found to be more difficult than that of other fake news [3], as it is often correct facts presented in a way that may be incorrect, and so binary labels such as "fake" and "real" do not capture the nuance of health misinformation [4]. For example, health articles may not grasp the quality of the evidence for a particular intervention or may not adequately explain its potential harms. A meta-analysis of 18 studies assessing the quality of news reports by Oxman et al [5] found that many reports give an unbalanced and overly simplistic portrayal of the potential consequences of health interventions, with around 93% not selectively reporting "statistically significant" results (95% CI 87%-97%), only around 53% quantifying any effects (95% CI 36%-69%), and around 40% not mentioning, discussing or explaining potential harms of an intervention (95% CI 23%-61%).

Health misinformation is also very time-consuming and resource intense to correct, especially with the scarce human resources that are better used for the provision of healthcare. Previous websites have attempted to provide a service for the correction or evaluation of health misinformation online, however, it is impossible for all information posted online to be reviewed given the vast amounts of articles, news releases etc. that are posted daily. These websites also rely on funding and a steady workforce, and if any of these are unavailable, may result in the forced closure of the website, as was the case for HealthNewsReview.org. For this reason, good automation of health misinformation is urgently needed. Below are some of the motivations behind wanting to tackle health misinformation in general.

1.2 Consequences of health misinformation

Health misinformation has a wide range of consequences, all to varying degrees. At its most extreme, it can harm public health, as seen in the case of vaccination misinformation, where they are only effective if the majority of the population are vaccinated. Misinformation about vaccines or diseases can contribute to increased hesitancy or refusal to vaccinate [6], which, in turn, allows pathogens to spread more easily [7]. This not only increases the risk of outbreaks but may also result in higher mortality rates, depending on the pathogen. This has implications for the general population as well as the individual, as some individuals, such as those with weakened immune systems, cannot be vaccinated. Their survival depends on high vaccination rates in the community and when a significant portion of the population refuses vaccination, it jeopardises the protection of these vulnerable groups [7].

Despite evidence that vaccination is one of the most successful public health measures, anti-vaccination movements have been linked to reductions in vaccine acceptance rates and increases in vaccine-preventable disease outbreaks and epidemics [8]. To illustrate, a survey conducted by Barua et al. [9] found that the propagation of misinformation on social media strongly undermined individual responses during the COVID-19 pandemic (2019-2020), particularly religious misinformation beliefs, conspiracy beliefs, and general misinformation beliefs, while assessing the credibility of misinformation influenced individual responses positively. This undermining of responses likely contributed to higher mortality rates, as it led some individuals to dismiss the severity of the virus or rely on ineffective alternative treatments [10]. Additionally, Gangarosa et al. [11] found that incidences of pertussis were 10 to 100 times lower in countries that were not influenced by anti-vaccine movements, which often emphasised adverse events which were unrelated to the vaccination.

Health misinformation can also have consequences on the individual level. For instance, speculation surrounding the effectiveness of hydroxychloroquine (a chemical found in cleaning products) as a cure for COVID-19 resulted in increased hospital admissions in Nigeria from hydroxychloroquine poisoning, and even led to deaths in some countries [12]. Though direct effects of an individual misinformation claim may be insignificant and small in number, these examples show that it can have severe consequences depending on the claim and how widespread a claim is. Furthermore, according to Southwell et al. [13], though a specific claim about a homeopathic product may not directly make a person buy the product, repeated exposure to the idea that folk wisdom is better than contemporary medicine may result in this belief, and this could potentially lead to more mistrust in peer-reviewed medical science [13] and perhaps delays or preventions in getting evidence-backed treatment. Despite an increased mortality rate for people who use alternative medicines like herbal remedies, vitamins, and dieting in place of standard cancer treatments, 39% of people in the United States believe that these alternative medicines can cure cancer alone [14].

Health misinformation can also have broader societal impacts beyond the health domain, such as fuelling discord to create a hostile political environment or increasing violence against ethnic and minority groups [15]. For example, during the pandemic, rumours claiming that Muslims were spreading the virus resulted in violent attacks in

India [10]. During this time health misinformation also incited violence toward people of Asian origin [16] and healthcare workers [17]. Indeed, a study on misinformation and the US Ebola communication crisis found that 45% of tweets with misinformation had a significantly higher amount of discord-inducing statements compared to tweets that did not contain misinformation (26%), and likewise, tweets containing misinformation had a significantly higher proportion of political content (36% versus 24%) [18].

These examples demonstrate the potential for health misinformation to harm public health both on the societal and individual level, exacerbate societal tensions, incite violence, and politicise public health crises if spread widely enough. It is therefore important to tackle health misinformation on all online media that has a wide distribution, and this is especially true for the case of online news articles regarding health, where research is relatively limited.

1.3 Outline

The remainder of this dissertation is structured as follows: in chapter 2 I provide an overview of the literature with particular focus on the studies that guided this paper. The research aims are then summarised in chapter 3. Chapter 4 offers the theoretical background of the algorithms and metrics used. Chapter 5 gives a general overview of the data, data preparation, exploratory data analysis, and methods, followed by the results in chapter 6. These are then discussed further in chapter 7 including the limitations, potential issues, and further research. Finally, I end with the conclusion in chapter 8.

Chapter 2

Literature Review

2.1 Overview

Research on health misinformation detection predominantly focuses on misinformation related to pandemics, especially the COVID-19 pandemic. Vaccine misinformation is another widely explored topic. Beyond these, other health-related topics such as drugs, treatments, and non-communicable diseases are occasionally addressed [19].

In terms of languages most of the research was conducted in English, though there are some research papers in other languages, such as German [20], Dutch [21], and Chinese [22]. A limited number of studies incorporate multilingual datasets [19].

There have also been a handful of studies on multimodal health misinformation detection, however, these too tend to focus on social media rather than on other media such as news stories.

2.2 Techniques and Models

Various methodological approaches are employed in the literature: deep learning, traditional machine learning, and a combination of both. Generally, for the literature containing deep learning, pre-trained transformer models performed the best, and ensemble methods such as random forests performed better than most other traditional machine learning methods [19]. The literature on general fake news detection, which also encompasses health misinformation, also gave similar results: pre-trained transformers (more precisely BERT-based models) outperformed other models, however, out of the traditional machine learning methods, Naive Bayes was the most effective [3].

2.2.1 Traditional Machine Learning Approaches

As mentioned above, Naive Bayes consistently outperforms other traditional machine learning models for detecting general fake news, achieving an accuracy of 93% [3]. Its simplicity and low computational requirements make it particularly suitable in the case of hardware constraints [3].

It has also been shown to perform well in health-specific studies, for example, in a 2020 study by Meppelink et al. [21]. In this study, the authors investigated whether supervised machine learning models (namely Multinomial Naive Bayes and Logistic Regression) can classify webpages about early childhood vaccination as 'reliable' or 'unreliable' in accordance with evidence-based vaccination guidelines. The authors collected 648 different Dutch webpages and compared the use of a count vectorizer approach versus a TF-IDF approach for each model. The study revealed that the Naive Bayes classifier with count vectorizer exhibited the most consistent overall performance, achieving a macro F1 value of 0.88. Furthermore, this model demonstrated a strong predictive performance on an out-of-sample dataset about HPV vaccination, achieving an average F1 score of 0.87 for evaluating reliable information. Despite its focus on Dutch webpages and childhood vaccination, the study suggests potential applications for these methods in other health-related text classification tasks. Further research could examine whether similar performance metrics could be seen in a more general health dataset and if it also applies to the English language.

2.2.2 Deep Learning Approaches

More recent papers evaluated the use of transformers in both social media content and news articles, and various papers show that BERT-based models and other pre-trained models performed the best for detecting fake news compared to other models, especially when labelled data was limited. Performance was then followed by deep learning models, then traditional models [3].

In health misinformation detection, some studies show that more domain-specific BERT-based models like SciBERT perform better than the general BERT model. Kumari et al [23] also found that a self-ensemble SciBERT was more effective in detecting health misinformation in news articles than the traditional BERT model, gaining a weighted F1 score of 0.715. Kontonya and Toni [24] found similar results; in their study on explainable automated fact-checking for public health claims, they found that SciBERT performed better than the general purpose model BERT in both summarisation evaluation and explanation quality assessments, and even outperformed other models trained on domain-specific data (BIOBERT v1.1 and BIOBERT v1.0). As this study only uses specific health claims, however, it is uncertain that these models would generalise well on unseen data given that some claims are very specific, for instance claims about Obamacare. These studies do suggest, however, that SciBERT is more effective than BERT in detecting health misinformation.

2.3 Detection in Health News Articles

Research is particularly focussed on social media with X (formerly known as Twitter) being the most popular, though there are some papers which also use news websites and search engines. Inputs in the research include social media posts, articles and/or health claims as well as user information [19].

One paper that does use news articles, however, it the 2021 paper by Zuo et al [4], which uses reviews of news stories and press releases taken from the health news evaluation website HealthNewsReview.org. In this study, they compare feature-based models (Support Vector Machine and Gradient Booster) with transformer models (BERT ALBERT, XLNet, RoBERTa, DistillBERT, Longformer) in a series of supervised classification tasks regarding the qualitative aspects of misinformation in health news. Each model is evaluated over six different accuracy criteria, with "satisfactory" and "not satisfactory" as the target labels, capturing the nuance of health misinformation. In their study, they found that for most of the criteria, less than half of the news articles are satisfactory, and even reputable news sources also fail to meet the these criteria established by domain experts. This implies that even reputable sources require an analytical view of their content, and that judging the reliability of news information based on the news source is often not sufficient when it comes to the accuracy of health information.

Results from their study also showed that feature-based models seemed to perform better than transformer-based models for criteria that were more specific, such as the cost of an intervention, however, they also note that this could be due to the sample size of the dataset after undersampling [4, p.79]. The macro-average F1 scores for these ranged from 64.4 to 68.1, with the best feature-based model being Gradient Booster. For the transformer-based models, the macro-average F1 score ranged from 58.8 to 62.3, with the best transformer-based model being RoBERTa.

However, it is important to note that the "satisfactory" labels and "not satisfactory" labels were not balanced for each criterion aside from one, and so macro-average of the F1 scores may have been skewed by the majority classes. They had also chosen to remove 6 of the 12 criteria from which the news pieces were evaluated, as they either "requir[ed] highly topic-specific medical knowledge" [4, p77] (which was done "to reflect the extent of medical knowledge available to the lay reader" [4, p77]) or were specific to either news stories or PR news releases, i.e. not both, which is what they had used in their study. This removes important evaluation criteria and raises the question of whether it would have been better to focus on one type of news so that all criteria for a specific news type could have been included. Nevertheless, the study suggests potential for using AI for evaluating general health misinformation in news articles, which is relatively scarce in the literature.

CHAPTER 3

Research Aims

The above literature review highlights some of the potential areas for further research. As such, this dissertation seeks to address the following research objectives:

1. Can Naive Bayes be used to classify generic health misinformation in online news articles in alignment with health misinformation criteria? Which parameters give the best performance?

2. How does performance of Naive Bayes compare with the domain-specific pre-trained transformer model SciBERT?

Before addressing these questions, however, background information of the relevant models and performance metrics is given below to facilitate understanding.

CHAPTER 4

Background

4.1 (Multinomial) Naive Bayes

4.1.1 Overview

Naive Bayes classifiers are a family of probabilistic classifiers based on Bayes' Theorem, which is used to calculate the probability of an event given that another event has occurred. In Naive Bayes classifiers, they obtain the probability that a particular observation in a dataset belongs to a certain class, given a set of features. They are often used for text classification problems such as spam filtering and fraud detection. It assumes that the features that go into the model are independent of each other. For text classification the features may be the words in a text, with each word representing one feature. The formula for Naive Bayes is follows:

$$P(C_k|\mathbf{x}) = \frac{P(\mathbf{x}|C_k) \cdot P(C_k)}{P(\mathbf{x})} \tag{4.1}$$

where the posterior probability $P(C_k|\mathbf{x})$ is computed for each class C_k (e.g. "misinformation" or "no misinformation") to classify the data input x into the class with the highest posterior probability.

The different types of Naive Bayes classifiers are Gaussian Naive Bayes (which uses continuous data), Multinomial Naive Bayes (which uses discrete data), and Bernoulli Naive Bayes. The one that will be used in this paper is Multinomial Naive Bayes.

4.1.2 Hyperparameters

Alpha

There are a few hyperparameters that can be tuned, one of which is the smoothing parameter, often denoted as alpha a. This parameter helps to prevent zero probabilities for unseen features by adjusting the likelihood estimate $P(\mathbf{x}|C_k)$. Without smoothing, the likelihood for a particular feature $P(x_i|C_k)$ is calculated as the count of a particular feature x_i in a class C_k divided by the total features in the class C_k (see equation 4.2).

The smoothing parameter adds a small constant a to the numerator and scales the denominator of the likelihood of a particular feature (see equation 4.3). For instance, if the word "nutrients" was not observed in the training data for a particular class, this would make the entire likelihood for that word equal to zero. In a Naive Bayes model, the likelihood $P(\mathbf{x}|C_k)$ is a product of the likelihoods of individual features, $P(x_i|C_k)$. Thus, if even one feature has a likelihood of zero, the entire product $P(\mathbf{x}|C_k)$ becomes zero, which prevents the model from correctly classifying a data point in the test set as it assumes that the feature (in this case "nutrients") is entirely absent in the given class, based soley on the training data.

$$P(x_i|C_k) = \frac{\text{Count of } x_i \text{ in } C_k}{\text{Total features in } C_k} \tag{4.2}$$

$$P(x_i|C_k) = \frac{\text{Count of } x_i \text{ in } C_k + a}{\text{Total features in } C_k + a \cdot V} \tag{4.3}$$

Class prior

Another hyper-parameter in Naive Bayes is the class prior parameter. In the Naive Bayes model, the prior probabilities $P(C_k)$ quantify how likely a given class is based solely on the distribution of the data (i.e. before new information is introduced to the model, hence the name "prior probability") and is calculated by the number of instances in a class divided by the total number of instances in the dataset. For example, if there are 60 out of 100 news articles that contain misinformation in the training set, the prior probability of the class ’contains misinformation’ would be 0.6. They serve as a baseline prediction when the feature data x provides no additional information, and give weight to the majority class if a the dataset is imbalanced. The prior parameter gives the option to adjust the priors instead of having the model derive them from the class frequencies in the data. This provides more control over the model’s probability calculations, for example if the class distribution in the training data does not reflect that of the real-world, or if the dataset is imbalanced and one wishes to make the model more sensitive to the minority class.

4.1.3 Feature Extraction

Two different methods that can be used to transform text into numerical vectors for text classification with Multinomial Naive Bayes are Bag-of-Words (BoW) and Term Frequency-Inverse Document Frequency (TF-IDF). Bag-of-Words essentially uses the frequency of each word to represent text, with each word being a feature in a vector. It treats all words equally, and does not take into account word order, meaning, or syntax. Given the nature of BoW, longer documents might skew the representation. On the other hand, TF-IDF still considers the term frequency but also adjusts the weight of each word so that rarer words are given more weight and higher frequency words are given less weight. Unlike BoW, it is not as sensitive to the document length.

4.2 BERT and SciBERT

SciBERT [25] is a pre-trained model based on the transformer model BERT [26], but differs in that it had been trained on 1.14 million scientific texts taken from the Semantic Scholar corpus (a research tool for scientific literature) rather than on the BooksCorpus and English Wikipedia. Since this paper deals with medical information, SciBERT was chosen as opposed to the generic BERT model because its training in scientific text may give a better performance, and has also been shown to perform better on health misinformation detection, as mentioned previously in section 2.2.2.

SciBERT stands for Scientific Bidirectional Encoder Representations from Transformers and leverages a transformer-based neural network to understand language. Unlike traditional language representation models, which process text sequentially and in one direction (either from left to right or right to left), SciBERT uses a bidirectional approach, meaning that it processes the full context of a word by considering the entire sentence simultaneously. To illustrate, in the sentence "diets high in... are linked to an increased risk of heart disease", a unidirectional approach would only predict the word based on the words "diets high in". Possible predictions might therefore be relatively generic, such as "proteins" or "calories", since the context of heart disease is not known. In the bidirectional approach, the model is able to utilise both the preceding and succeeding words to predict the unknown word, and would therefore understand the context of heart disease in this example. Thus, it would be able to give more accurate predictions such as "cholesterol" or "saturated fat", which have been scientifically linked to heart disease.

It was pre-trained on a large amount of unlabelled scientific text data with two different Natural Language Processing tasks: masked language modelling (MLM) and next sentence prediction (NSP). In MLM training words are masked in a sentence and the model has to predict the masked words based on their context. In NSP training the model has to predict whether a second sentence follows the first given a pair of sentences, or if the second sentence is simply random. The model can then be fine-tuned depending on the desired task.

4.3 Model Evaluation

The following section will give short explanations of how binary classification tasks are evaluated.

In a binary classification task, data instances are typically classified as either positive or negative. A positive label indicates the presence of an outcome to be observed, for example, the presence of misinformation. Hence a negative outcome is the absence of the particular outcome, in this example, no presence of the misinformation. Each predicted label can therefore fall into one of four categories:

- **True positive (TP)** - This occurs when a positive outcome is correctly predicted (e.g. the model correctly predicted the presence of misinformation)
- **True negative (TN)** - When a negative outcome is correctly predicted (e.g. the model correctly predicted that there is no misinformation present)
- **False positive (FP)** - When a negative outcome is incorrectly predicted as positive (e.g. the model predicted the presence of misinformation when in fact there was none)
- **False negative (FN)** - When a positive outcome is incorrectly predicted as negative (e.g. the model predicted no presence of misinformation when in fact there was)

There are several performance metrics used for the evaluation of machine learning models. For classification tasks, some of the most common are accuracy, precision, recall, and F1 score. More detailed explanations of each are found below.

4.3.1 Accuracy

This measures the performance of the model. It is calculated by:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.4}$$

Accuracy can be a misleading metric for imbalanced datasets as it evaluates the overall correctness of a model, without considering the distribution of classes. For example, if only 5% of news articles contain misinformation, a model that always predicts "non-misinformation" would achieve 95% accuracy, despite failing to detect misinformation, which is often the critical focus in such cases.

4.3.2 Precision

The precision measures how accurate a model's positive predictions are. This is calculated by:

$$\text{Precision} = \frac{TP}{TP + FP} \tag{4.5}$$

4.3.3 Recall

Recall measures the ability to identify a class and is used when minimising false negatives is essential, for example, in medical diagnoses where having a maximum detection rate is essential. The equation is as follows:

$$\text{Recall} = \frac{TP}{TP + FN} \tag{4.6}$$

4.3.4 F1 score

This is the harmonic mean of precision and recall so evaluates the overall performance of a classification model. It is used to help minimise false positives and false negatives.

$$\text{F1 Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{4.7}$$

4.3.5 Macro average

To calculate the macro average of a metric, the metric for each class is calculated independently and the arithmetic mean of these is taken. Each class contributes equally to the final result, regardless of the size of the class.

$$\text{Macro Average} = \frac{1}{N} \sum_{i=1}^{N} M_i \tag{4.8}$$

where N is the number of classes and M_i is the performance metric for class i.

4.3.6 Weighted average

The weighted average calculates the metric for each class and computes the average, however, unlike the macro average, takes the number of instances in each class into consideration. In an imbalanced dataset, the majority class would therefore have a more proportional influence on the final result.

$$\text{Weighted Average} = \sum_{i=1}^{N} \frac{n_i}{N} \cdot M_i \tag{4.9}$$

where N is the number of observations in the dataset, n_i is the number of instances in class i, and M_i is the performance metric for class i.

CHAPTER 5

Methodology

5.1 General Overview of Data

The dataset used in this paper is taken from the FakeHealth data repository, which was created by Dai et al. along with the paper *Ginger cannot cure cancer: Battling fake health news with a comprehensive data repository* in 2020 [1][27]. It consists of two datasets: HealthStory and HealthRelease, which contain information from news media (e.g. The Guardian, New York Times, Buzzfeed, WebMD - a full list can be seen in Figure A.1) and news released from other institutes (e.g. research centres, universities, and companies) respectively. Each dataset contains information surrounding the news content, news reviews, as well as the social engagements and user networks from X. For this paper only the news content and news reviews from HealthStory will be used. This was chosen as it is generally more common for members of the public to read news articles rather than releases.

The News Content file contains various data including the article title, text, image links, and metadata from 40 different news media. The news review data is a collection of news review data taken from the industry independent, US-based but now defunct [28] website HealthNewsReview.org, which can be viewed using the Way Back Machine [29]. On this website each news article is reviewed by two to three experts from the healthcare domain and evaluated over ten different criteria, such as whether the article adequately explains or quantifies the harms and of the intervention, whether it grasps the quality of the evidence, and whether the story uses independent sources and identifies conflicts of interest. A full list of the criteria can be seen in Table 5.1 (note that only the criteria relevant to the HealthStory data is included in the table).

Each story is then given a rating from 0 to 5 dependent on what percentage of the ten criteria is met [29]. If a criteria is deemed irrelevant for the article then it is classed as ’not applicable’, and only the total scores for articles that have two or fewer ’not applicable’ ratings are posted (the denominator is adjusted to 8 or 9 accordingly if ’not applicable’ ratings are present). Due to the nature of the criteria, generally it was only possible for stories that discuss one specific treatment, test, product or procedure for one specific condition to be reviewed, and so news stories that discussed multiple interventions or multiple uses for an intervention were not eligible for review.

Number	Criteria Questions
C1	Does it compare the new approach with existing alternatives?
C2	Does it adequately explain/quantify the harms of the intervention?
C3	Does it seem to grasp the quality of the evidence?
C4	Does it adequately quantify the benefits of the treatment / test / product / procedure?
C5	Does it establish the true novelty of the approach?
C6	Does it establish the availability of the treatment/test/product/procedure?
C7	Does it commit disease-mongering?
C8	Does it adequately discuss the costs of the intervention?
S9	Does the story use independent sources and identify conflicts of interest?
S10	Does the story appear to rely solely or largely on a news release?

Table 5.1: Table showing the criteria used to evaluate the news articles. Adapted from *Ginger cannot cure cancer: Battling health news with a comprehensive data repository* [1]

5.2 Data Cleansing and Preparation

Altogether there were 1638 rows (i.e. articles) for the HealthStory Content dataframe and 1690 rows (i.e. reviews) in the HealthStory Reviews dataframe. They were merged together based on the original news story title as this was the only variable which was common between the two and were unique enough to ensure that the correct review was matched with the relevant article. Before merging, the titles were converted to

Percentage of criteria judged satisfactory	Rating	Label Assigned
0%	0	Unsatisfactory
1-20%	1	Unsatisfactory
21-40%	2	Unsatisfactory
41-60%	3	Satisfactory
61-80%	4	Satisfactory
81-100%	5	Satisfactory

Table 5.2: Table showing how ratings of news stories are assigned on the HealthNewsReview.org website based on the 10 criteria shown in 5.1. The corresponding labels used in this paper are also shown. Ratings were only given for articles that had two or fewer criteria deemed ′not applicable′. For articles containing criteria deemed not applicable, the denominator was adjusted accordingly in the percentage calculation. Adapted from HealthNewsReview.org [2]

.

lower case and the whitespace and punctuation were removed, to ensure that possible typing errors or discrepancies regarding spaces and punctuation between the two were removed. Rows with duplicated titles and texts were also removed, altogether leaving 1180 rows in the dataframe.

In a similar way to Shu et al. [30] and Dai et al [1], the binary labels ′satisfactory′ and ′unsatisfactory′ were added according to whether rating was less than 3, with articles with 3 or more rated as ′satisfactory′. These will be the target labels for the models. In contrast to the authors of the original dataset, the labels ′satisfactory′ and ′unsatisfactory′ were chosen instead of ′fake′ and ′real′ news to capture the nuance of health misinformation in the articles, which is not always clearly defined as mentioned previously. In total there were a lot more articles with the label ′satisfactory′, with 844 in total, compared to 336 unsatisfactory labels, so care must be taken to ensure that bias toward the majority class is minimized as much as possible in the algorithms.

5.3 Exploratory Data Analysis

5.3.1 Distribution of words counts

Figure 5.2 shows that articles could be up to 4000 words in length, however, there is a positive skew with the mean length of articles being 635.95 words with a standard deviation of 375.38.

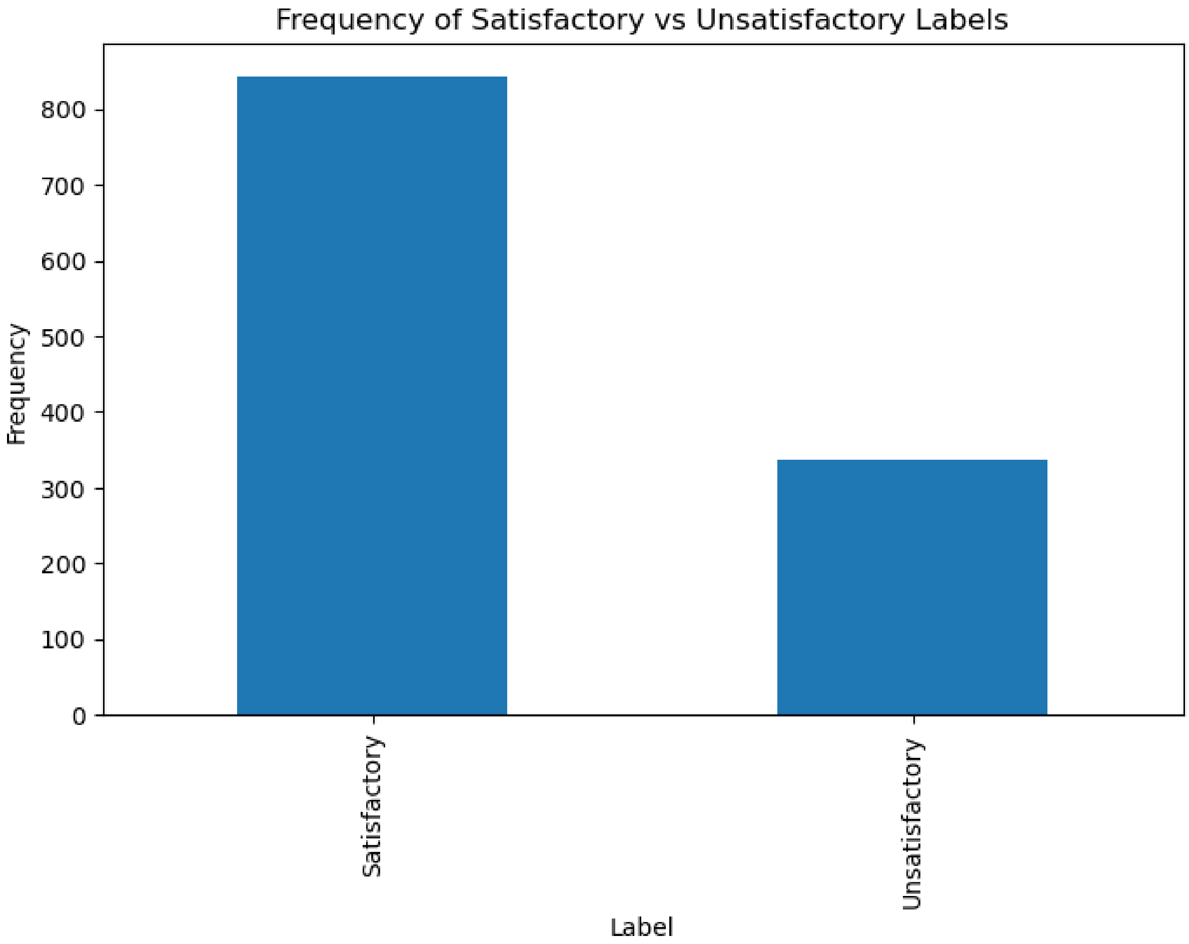


Figure 5.1: The number of articles labeled as ’satisfactory’ is more than twice that of those labeled ’unsatisfactory.’

The histograms in figure 5.3, however, show that there may be a significant difference of the word counts between each label. It was calculated that the mean word count of the satisfactory articles was 690.48 with a standard deviation of 384.90, and that the median was 668 words. In comparison, the mean word count of the unsatisfactory articles was 499.54 with a standard deviation of 311.57, and the median was calculated to be 433. This suggests that the satisfactory labels were generally longer than the unsatisfactory news articles.

This has implications on the models are trained. For example, in the Multinomial Naive Bayes model, this may skew the model towards the class with the longer articles if it uses the Bag-of-Words method for feature extraction. As it has already been shown that there are a lot more of the satisfactory class than the unsatisfactory class, this may exacerbate overfitting and reduce the model’s ability to generalise, if no mitigation

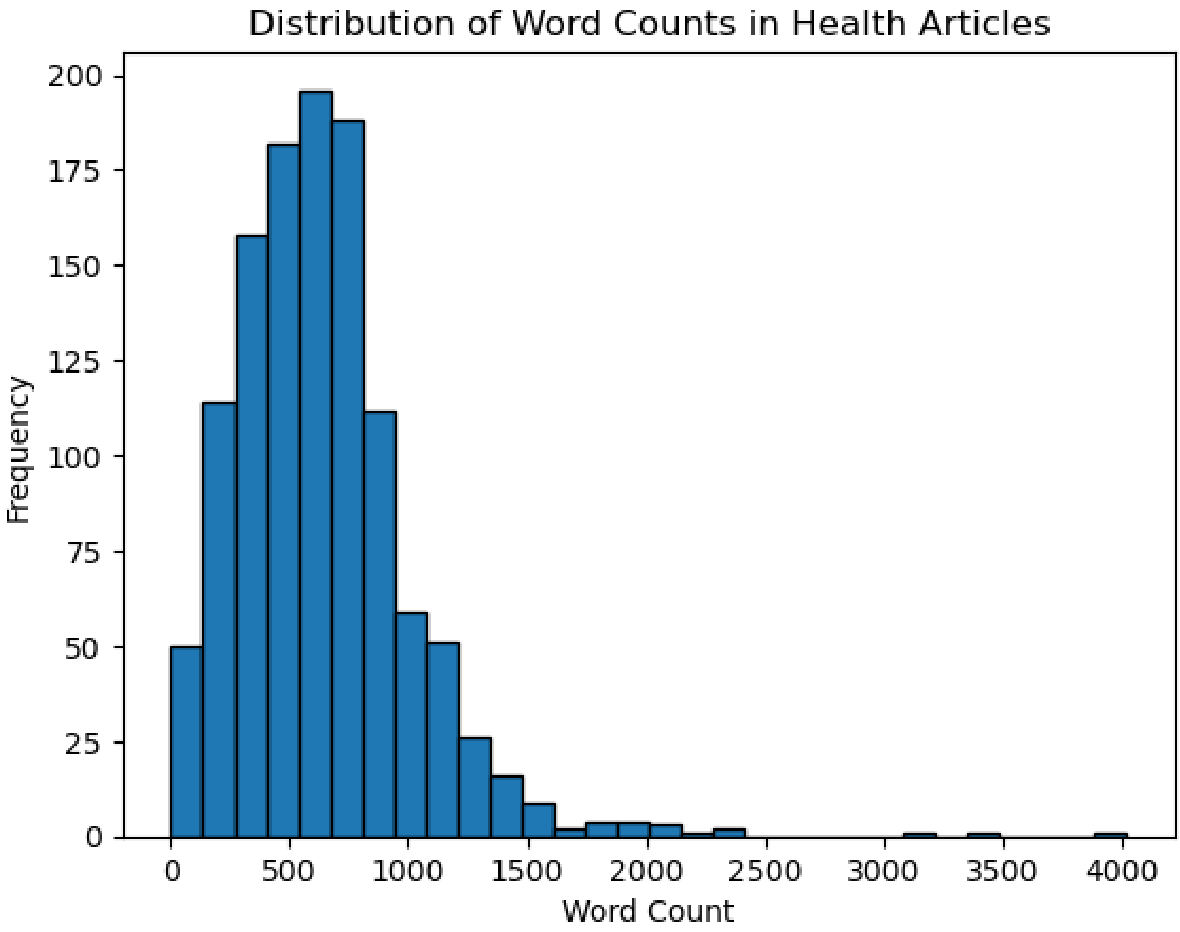


Figure 5.2: The majority of articles had approximately 200 to 900 words

strategies are utilised. Similarly, in the case of SciBERT, if the classes have significantly different article lengths, it may exhibit bias toward the satisfactory class, simply because it has more tokens (i.e. information) to process.

To investigate whether there was indeed a significant difference in the means of word counts for each of the target labels, an independent samples Welch's t-test was used. Normally, this test requires a normal distribution, but due to the Central Limit Theorem, it can still be used as the sample sizes (i.e. the number of news articles) in each class is larger than 30. This specific version of the t-test was also chosen due to the significant differences in the variances between the satisfactory articles and unsatisfactory articles (p-value: 0.0009). Results of the Welch's t-test gave a p-value of 5.3587e-18. At a 5% significance level, this means that there is a significant difference in the means of word counts between satisfactory articles and unsatisfactory articles.

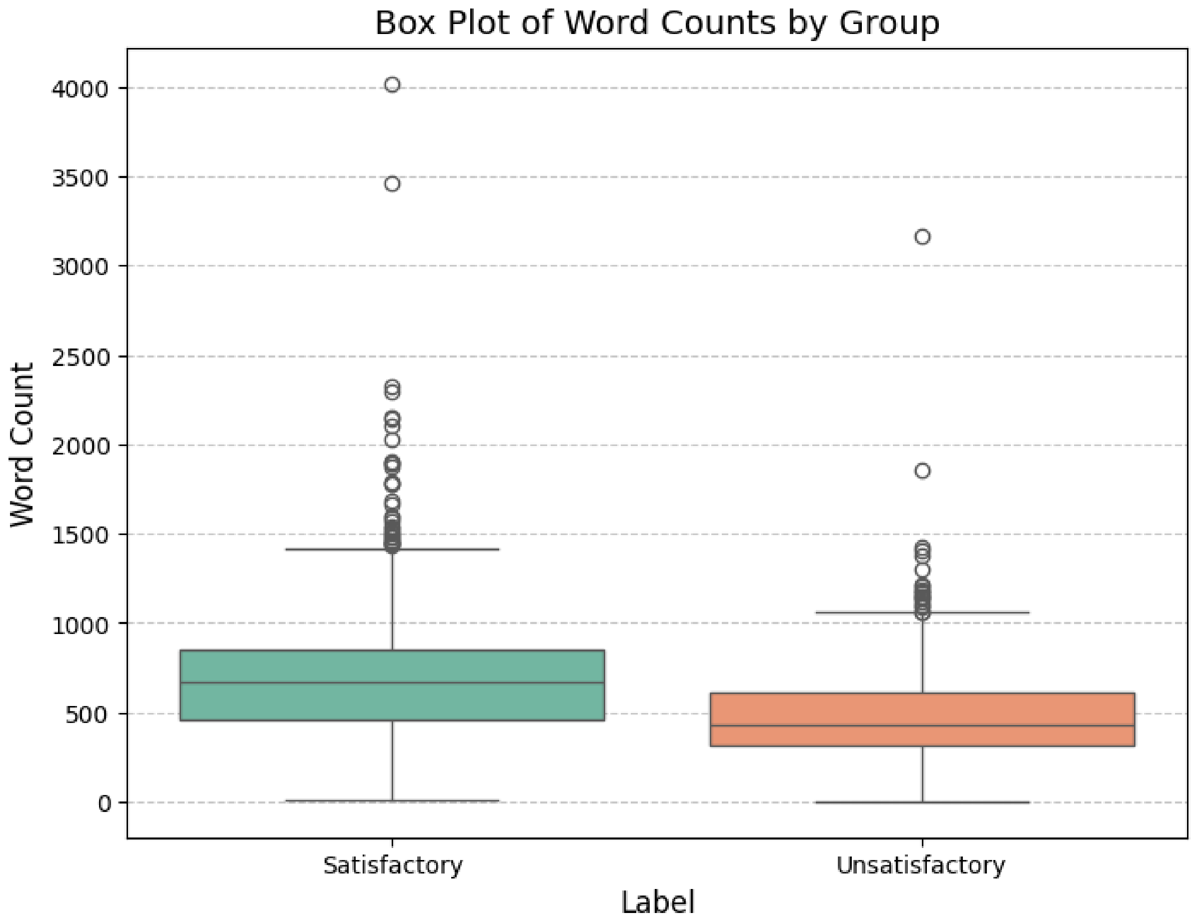


Figure 5.3: ’Satisfactory’ articles appear longer than ’unsatisfactory’ articles. Satisfactory articles had a mean of 690.48,standard deviation of 384.90, and median of 668 words. Unsatisfactory articles had a mean of 499.54, standard deviation of 311.57, and median of 433. Welch’s t-test of independence reveal that the means are indeed significantly different.

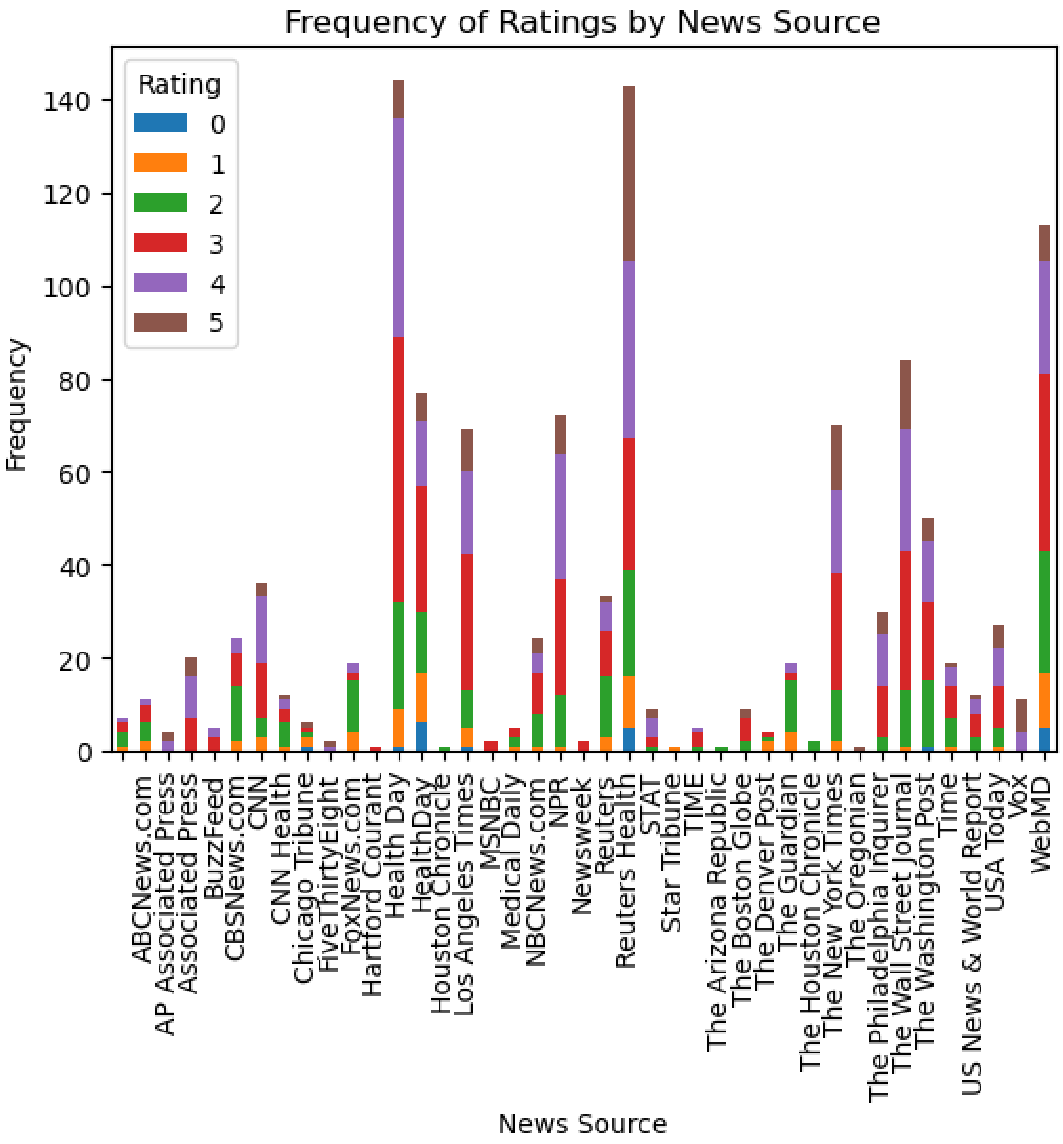


Figure 5.4: Reviews come from a range of different news sources in US media. Ratings vary considerably per news source

5.3.2 News sources

Figure 5.4 shows the articles came from a range of different news sources in US media, with a varying number of reviews and ratings per news organization. It is important to note, however, that this data is not representative of the news organisations for a few reasons: not all articles from an organisation could be reviewed due to eligibility of articles, funding and staffing restraints; and the news organisations chosen for review had changed several times throughout the course of the HealthNewsReview history.

5.3.3 Most common topics

In terms of the most common topics, an overview of the most common tags reveal that articles relating to various forms of cancer, as well as general health (such as weight loss and exercise), surgery, and depression were the most common articles to be reviewed. A list of the most common tags can be found in the appendix for more information.

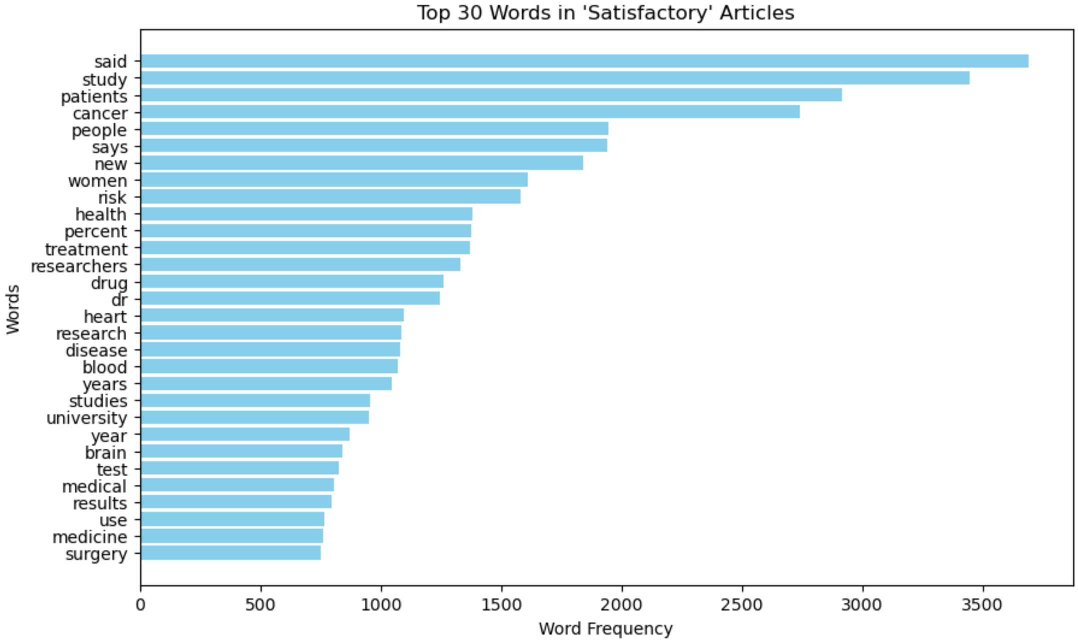
5.3.4 Top words per label in article texts

Figure 5.5 shows bar graphs of the top words in both of the satisfactory and unsatisfactory article texts respectively (with common stopwords removed - using the list from Sklearn's Count Vectorizer). The top five words were the same for both classes, with the top words being 'said', 'study', 'patients', 'cancer', and 'people'. However, they also show that in the text that was labelled 'satisfactory' the word 'health' is used relatively more, and there is the presence of some words in the top used words that are not present in the unsatisfactory texts, such as 'studies', 'medical', 'results', 'medicine', and 'surgery'. This suggests that the relative frequency of particular words could give an indication as to whether the articles also compare a particular treatment to other treatments that are already available, thus indicating a balanced view of a treatment. The presence of the word 'studies' in particular could be an indication of this, as the word 'study' is present in both the satisfactory and unsatisfactory top used words.

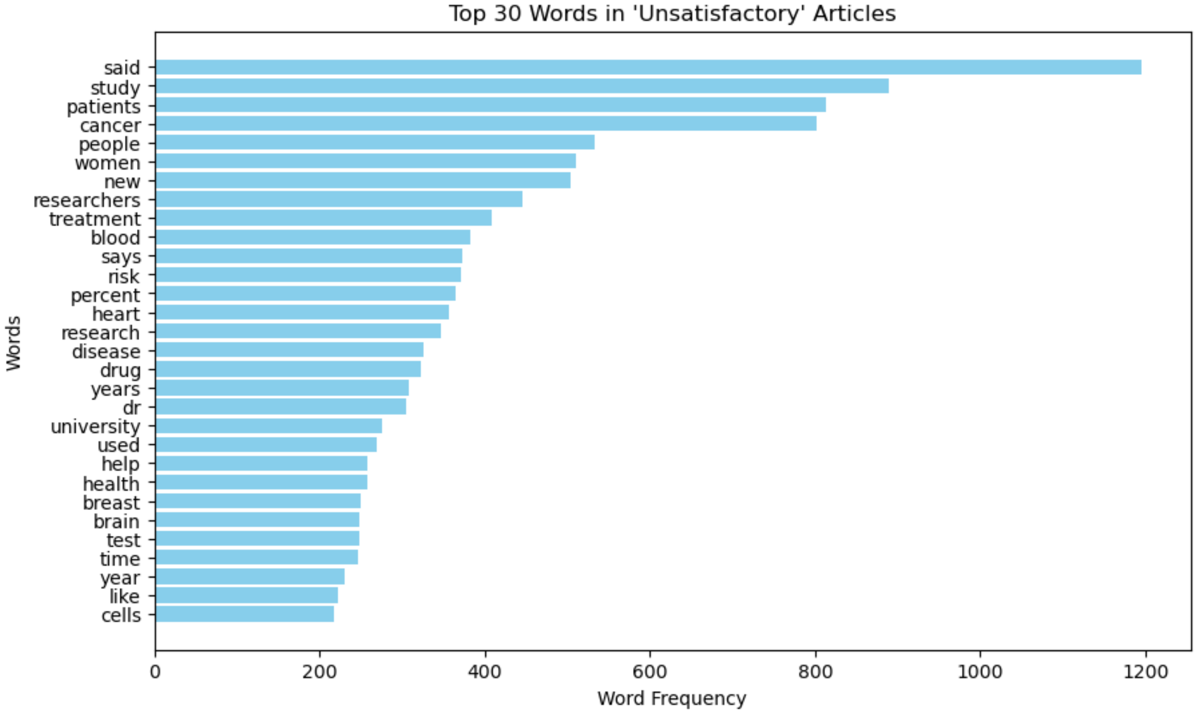
5.3.5 Comparison words and words of uncertainty

A visual inspection of the dataframe also appeared to suggested that the titles of the articles labelled 'satisfactory' had more modal words of probability, such as 'may' and 'could'. To investigate this, bar graphs of the top words in the titles of satisfactory and unsatisfactory were also plotted, this time with no removal of stopwords, and these can be seen in figure 5.6.

Indeed, the figures show that in the top used words in satisfactory titles the words 'could', 'might' and 'more' are used relatively more than in the unsatisfactory titles. This indicates that satisfactory labelled articles may use more comparison words and words that show more uncertainty as not to overstate the benefits of a particular treatment and



(a)



(b)

Figure 5.5: In 'satisfactory' articles, the words 'health', 'studies', 'medical', 'results','medicine', and 'surgery' appear more than in the unsatisfactory articles. These words may help indicate whether an article gives a balanced view of treatment.

could potentially be used in the detection of health information. However, using singular words alone may not be enough as the meaning of the text may change depending on the words surrounding it, for example ′may′ and ′shows′ imply uncertainty and certainty respectively, however, when put together, ′may show′ implies uncertainty.

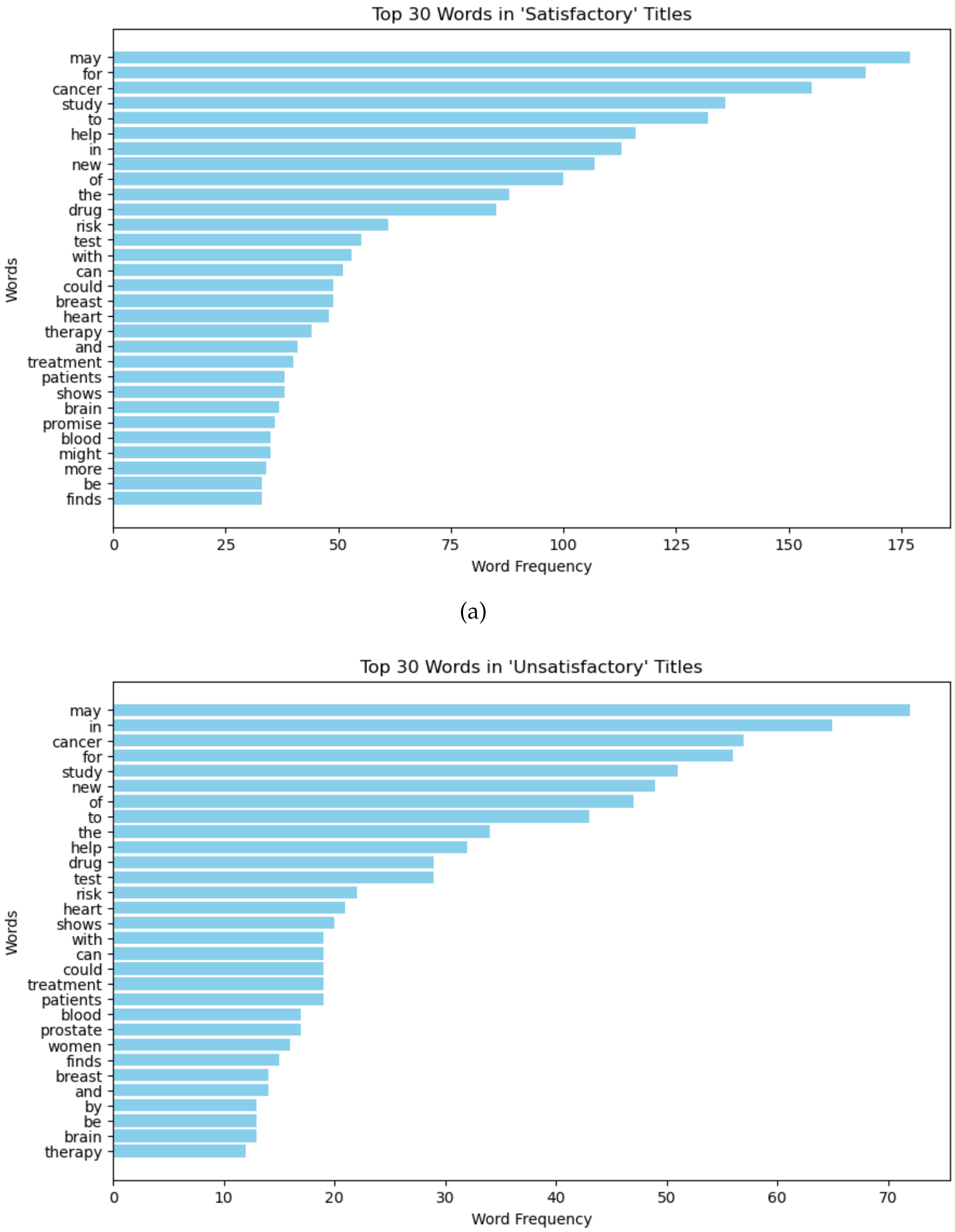
Similar to the general word counts between the two classes, a Welch's t-test was carried out to see if there was a significant difference in the mean frequency of comparison words and words of uncertainty in the article texts. The words chosen for the test were ′could′, ′may′, ′might′, ′should′, ′would′, ′more′, and ′less′, as these words also convey uncertainty or, in the case of ′more′ and ′less′, are commonly used for comparisons. These, together with the article texts, were then stemmed (i.e. had prefixes and suffixes removed, thereby transforming them into their root form) and the frequency of the above words were calculated for each observation in the dataframe. The test revealed that there was indeed a significant difference in the mean frequencies of comparison words and words of uncertainty, giving a p-value of 4.5305e-17.

5.4 Methods

As mentioned previously, this paper aims to see whether using Multinomial Naive Bayes or the transformer model SciBERT can be used to detect whether an article gives satisfactory versus unsatisfactory information regarding health interventions based on the text. Naive Bayes was chosen as it has been shown to be one of the most effective traditional methods when it comes to fake news detection when there are hardware constraints [3]. Its performance will then be compared to using the transformer model SciBERT, as transformer models have also been shown to perform the best at classification tasks (outperforming traditional models), especially when there is limited labelled data [3], which is the case for this particular dataset.

5.4.1 Metrics

As such, the main metrics that will be used to evaluate the models will be precision, recall, and F1 score (the harmonic mean of precision and recall). Due to the imbalanced dataset, the weighted averages of each metric was used to avoid skewness in the metrics



(a)

(b)

Figure 5.6: Bar graphs showing the top 30 words in satisfactory and unsatisfactory titles, respectively, with stopwords removed. The figures show that in the top used words in satisfactory titles the modal words of probability like ’could’, ’might’ and ’more’ are used relatively more than in the unsatisfactory titles. This indicates that satisfactory labelled articles may use more words that show more uncertainty as not to overstate the benefits of a particular treatment.

caused by the minority ’unsatisfactory’ class. The accuracy metric, which measures how often the data is correctly classified across all classes, will also be used to compare performance across the models.

For labelling health misinformation articles, it is important to minimise the number of false positive and false negative outcomes, as the truth of the label is the most important thing to consider.

5.4.2 Multinomial Naive Bayes

Before training the model, the words of the article texts were first converted into lowercase, tokenised, and converted into a vector using the CountVectorizer tool in the Scikit-learn library, which utilizes the Bag-of-Words method for Naive Bayes. The data was then split with stratification into a training set, validation set, and test set with the ratio 60:20:20, with parameters optimised on the validation set and the final model tested on the test set. The stratification ensured that the class distribution was preserved across all the splits, which is especially important for imbalanced datasets. The distribution of each set can be seen in Table 5.3.

	Training Set	Validation Set	Test Set
Satisfactory	506	168	169
Unsatisfactory	202	68	67
Total	708	236	236

Table 5.3: Table showing the the distributions of the training set, validation set, and test set. The data was split with stratification to reflect the class imbalance in the dataset.

First, a random classifier was trained on the training set and evaluated on the validation set. This was done to provide a basis of comparison for the Naive Bayes models in general.

To provide a baseline for comparison for hyperparameter tuning, the first model was trained using the Bag-of-Words method for feature extraction, with no text preprocessing steps, and the default parameter settings for the Scikit-learn MultinomialNB class were used for the model. The default parameters are as follows: the smoothing parameter alpha is set to 1, and the class priors are calculated based on the relative frequencies of the satisfactory and unsatisfactory labels in the training data (fit prior=True).

The TF-IDF method for feature extraction was then compared to Bag-of-Words. This was done to see what was the best method of feature extraction for the following reasons: on the one hand, Bag-of-Words might be better at classification based on words of comparison and uncertainty but be sensitive to document length, whereas TF-IDF would be less sensitive to document length but would give more weight to rarer words that can help determine the class. This was important considering the significant differences in the word counts between the satisfactory class and the unsatisfactory class found in section 5.3.1.

Grid search was then carried out to find the best combination of hyperparameters. This was done with 5-fold cross-validation. Here, performance of each model is evaluated by dividing the data into five sections (folds), with four folds used for training and the remaining fold used for testing. This process is repeated 5 times, with each fold used as the test set once, and the results are averaged for evaluation. The 5-fold cross-evaluation was done to get reliable performance metrics for the best hyperparameters.

Then, a custom stop words list was made to see if this would enhance performance. Stop words are common words that are removed from the text as they do not carry useful information. This was based on the default stop words list from the nltk package, however, the words from the statistical test in section 5.3.5 conveying uncertainty or comparison ('could', 'may', 'might', 'should', 'would', 'more' and 'less') were removed from the list to ensure that these were picked up from the model. The best parameters from the grid search were also kept.

Following this, performance of the model was evaluated with a threshold of 0.6 rather than 0.5 for the satisfactory class. This was to see if performance on the minority unsatisfactory class would be improved by classifying articles that were only just classed satisfactory (i.e. had a probability of being satisfactory between 0.5 and 0.6) as unsatisfactory, as it might be better for the model to be slightly critical of an article so that readers can be encouraged to find out more about an intervention.

The best parameters found from the validation set were then performed on the test set and compared with the results of the initial model without tuning and the random classifier.

The results of the above can be seen in section 6.1

5.4.3 SciBERT

For the SciBERT model, the text of the articles were pre-processed in preparation for the model. More specifically, the text was converted into lowercase and URLs, special characters, and extra whitespaces were removed so that the text could be in the correct format for the model.

As with the Multinomial naive Bayes model, the data was split into a 60:20:20 ratio for the training set, validation set, and test set. This was to ensure that there was a fair comparison performance of the model compared to the Multinomial Naive Bayes model.

The article texts were then tokenized using the AutoTokenizer tool from the Pytorch transformers package and encoded. The SciBERT model was then loaded, using cross-entropy loss and the loss function, and the class weights were also computed dynamically, meaning that misclassification of the minority class is penalized more heavily, ensuring that the class imbalance is accounted for during the training process. The model was trained over three epochs.

The results for the SciBERT model can be seen in section 6.2

CHAPTER 6

Results

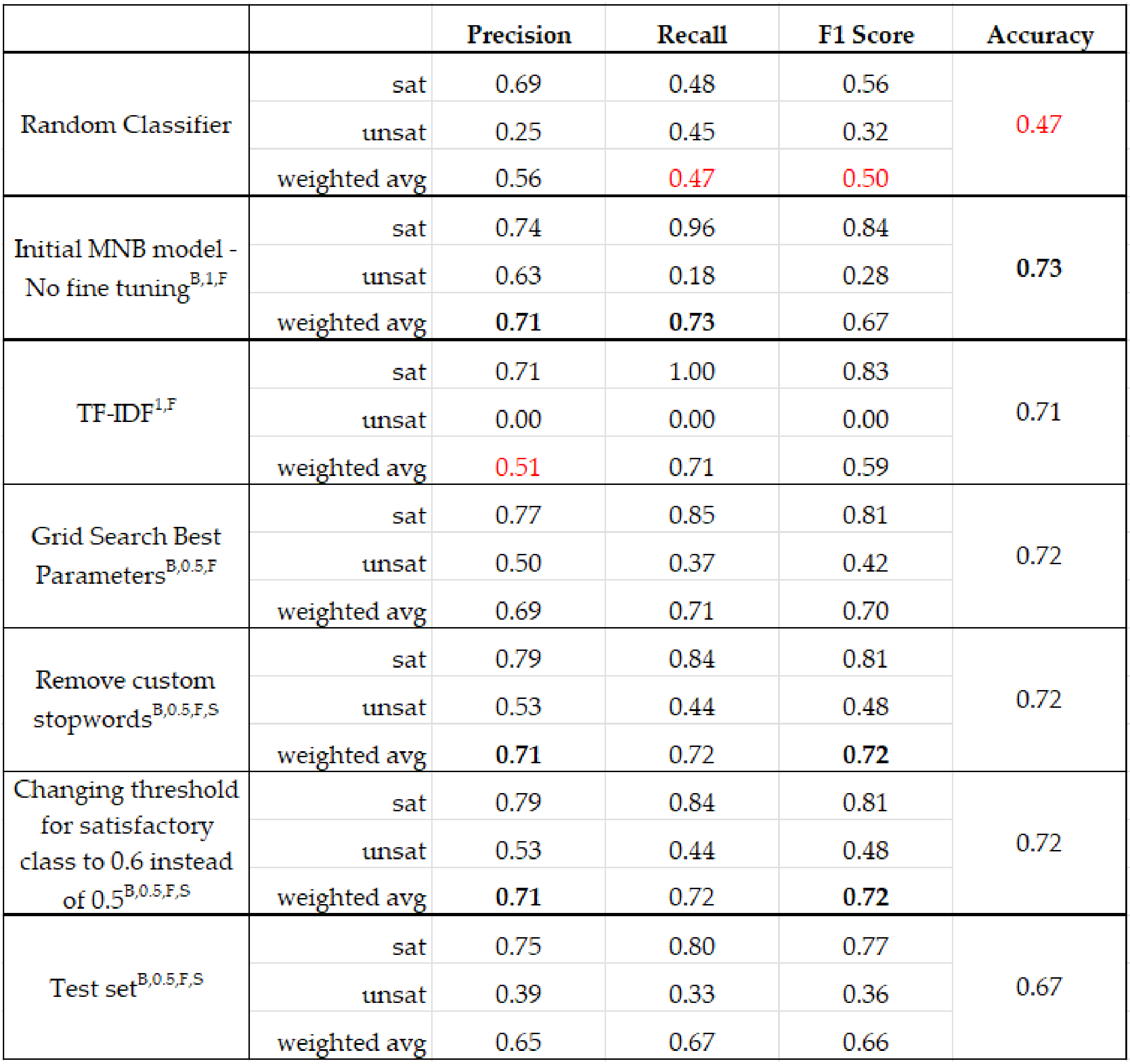
6.1 Multinomial Naive Bayes

6.1.1 Random Classifier

As a baseline for comparison for all of the models, a random classifier was used. This gave an overall accuracy of 0.47, weighted precision of 0.56, weighted recall of 0.47, and weighted F1 score of 0.50. See Figure 6.1 for a summary of the results.

6.1.2 Initial model - Bag-of-Words and no hyperparameter tuning

With the Bag-of-Words method and default hyperparameters (alpha=1, fit prior=True), the initial model achieved an overall score of 0.73 for the accuracy, which is much higher than the accuracy of the random classier. Nevertheless, the model performs roughly the same as if it were to simply label everything as 'satisfactory', which makes up for 72% of the validation set. For the weighted average of the F1 score, it achieved 0.67, however, the report shows that the recall for the satisfactory class was very high (0.96) and was very low for the unsatisfactory class (0.18), meaning that whilst it is highly effective at identifying most of the true positive cases for the satisfactory class, the model often classes 'unsatisfactory' classes as 'satisfactory'. Indeed, performance for the satisfactory articles was high in general, with an F1 score of 0.84.



		Precision	**Recall**	**F1 Score**	**Accuracy**
Random Classifier	sat	0.69	0.48	0.56	0.47
	unsat	0.25	0.45	0.32	
	weighted avg	0.56	0.47	0.50	
Initial MNB model - No fine tuning^{B,1,F}	sat	0.74	0.96	0.84	**0.73**
	unsat	0.63	0.18	0.28	
	weighted avg	**0.71**	**0.73**	0.67	
TF-IDF^{1,F}	sat	0.71	1.00	0.83	0.71
	unsat	0.00	0.00	0.00	
	weighted avg	0.51	0.71	0.59	
Grid Search Best Parameters$^{B,0.5,F}$	sat	0.77	0.85	0.81	0.72
	unsat	0.50	0.37	0.42	
	weighted avg	0.69	0.71	0.70	
Remove custom stopwords$^{B,0.5,F,S}$	sat	0.79	0.84	0.81	0.72
	unsat	0.53	0.44	0.48	
	weighted avg	**0.71**	0.72	**0.72**	
Changing threshold for satisfactory class to 0.6 instead of 0.5$^{B,0.5,F,S}$	sat	0.79	0.84	0.81	0.72
	unsat	0.53	0.44	0.48	
	weighted avg	**0.71**	0.72	**0.72**	
Test set$^{B,0.5,F,S}$	sat	0.75	0.80	0.77	0.67
	unsat	0.39	0.33	0.36	
	weighted avg	0.65	0.67	0.66	

Figure 6.1: Summary of results for Multinomial Naive Bayes. Meanings of superscript: **B** means BoW, **1** means alpha=1, **0.5** means alpha=0.5, **F** means fit prior=true, and **S** means custom stop words. Overall, the Multinomial Naive Bayes model performed better than the random classifier, however, the hyperparameter tuning etc. made minimal difference. Bag of Words also performed much better than TF-IDF.

6.1.3 TF-IDF versus Bag-of-Words

For the next validation set, it was investigated whether using TF-IDF instead of Bag-of-Words would perform better for feature extraction. Again, the default parameters of 1 for alpha and fit prior set to 'True' were used. This performed considerably worse than the previous Bag-of-Words model, with scores of 0.00 for both the precision and recall (and hence F1 score) for the unsatisfactory class, meaning that it was unable to predict the the unsatisfactory class at all, simply labelling everything as satisfactory. This is also confirmed by the score of 1.00 for recall of the satisfactory class. The reason for the poor performance could be that it is precisely the common words which TF-IDF gives less weight to that can help determine whether an article gives satisfactory versus unsatisfactory information regarding health interventions, such as the comparison words and words of uncertainty used in the statistical test in section 5.3.5. For this reason, Bag-of-Words was used for the rest of the parameter tuning and test set.

6.1.4 Grid search for best alpha and fit prior

To find the best configuration of the hyperparameters to be used, a grid search was applied, which evaluates all possible combinations of specified alpha values and fit prior options. The alpha values tested were 0.01, 0.1, 0.5, 1.0, 1.5, and 2.0, and the options for the fit prior parameter were 'True' (learn class prior probabilities from the data) and 'False' (assume uniform class priors, i.e. that all classes are equally likely, regardless of their distribution in the training data).

The grid search for the best configuration of the hyperparameters found that the best combination was 0.5 for the smoothing parameter alpha, and 'True' for the fit prior parameter. With these parameters, the model achieved significantly higher results for the recall for unsatisfactory articles, achieving a recall of 0.37, compared to 0.18 for the initial model. This meant that having these parameters resulted in a better prediction of the unsatisfactory class, and ultimately a higher weighted F1 score of 0.70 compared to 0.67. The performance for the satisfactory class remained high, with an F1 score of 0.81.

The reason these parameters resulted in improved performance for the unsatisfactory class compared to the initial model could be because less smoothing allowed the model to rely more heavily on the actual observed word frequencies in the data, giving rare words that are highly distinctive in a class more weight, whilst keeping the common terms in the model as well. The superior performance of the fit prior being set to 'True' is likely due to the class imbalance, as learning the class priors based on the relative frequencies of the satisfactory and unsatisfactory labels in the training data gives more weight to the majority class. It is important to note, however, that these increases were very slight and so may not be a significant difference.

6.1.5 Removal of custom stop words

Next, a custom list of stop words based on the statistical test in section 5.3.5 conveying uncertainty or comparison was made and removed from the model.

This resulted in slightly better performances in the prediction of the unsatisfactory class compared to the best results from the grid search, with the precision increasing from 0.50 to 0.53, recall increasing from 0.37 to 0.44, and F1 score increasing from 0.42 to 0.48. This subsequently led to a slight increase in the weighted F1 score, with 0.72. Again, performance for the satisfactory class was high (F1 score = 0.81). This suggests that the removal of stopwords and the inclusion of the words of comparison and uncertainty could potentially result in both less false positives and false negatives of the unsatisfactory class. Again, the increases were relatively small so these increases may also be due to overfitting on the validation set.

6.1.6 Changing threshold for unsatisfactory class

It was then investigated whether changing the threshold for the unsatisfactory class to 0.6 instead of 0.5 would result in a better prediction for the unsatisfactory class. This, however, gave the same results as the previous model, where custom stopwords were removed. This suggests that having a custom threshold does not make the model any more sensitive to predicting the unsatisfactory class.

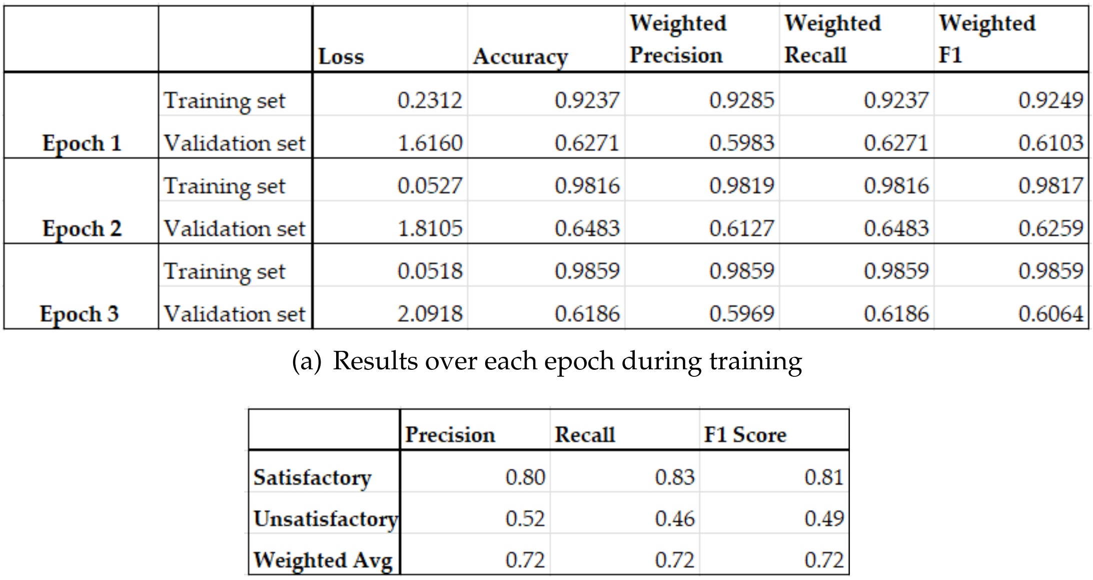
6.1.7 Performance on the test set

Finally, the model was evaluated on the test set to see how the best feature extraction method, data pre-proccesing steps, and hyperparameters performed on unseen data, and to see if the tuned parameters was simply resulting from overfitting on the validation set. A summary of these are as follows: use of the Bag-of-Words for feature extraction, alpha set to 0.5, class priors based on the relative frequencies of the labels (fit prior='True'), and the use of custom stopwords.

Overall, the model had a relatively similar performance to the initial model without fine tuning, with an overall weighted F1 score of 0.66, compared to 0.67 for the initial model. It did, however, achieve higher results for the recall of the unsatisfactory class, with 0.33 compared to 0.18. Although the F1 score for satisfactory articles was lower (0.77), it was still relatively high. Nevertheless, the weighted metrics performance on the test set was significantly better than that of the random classifier, which had achieved 0.56 for the weighted precision, 0.47 for the weighted recall, and 0.50 for the weighted F1 score. The weighted metric scores for the test set were 0.65 for the precision, 0.67 for the recall, and 0.66 for the F1 score. This suggests that the Multinomial Naive Bayes model was able to learn patterns from the data effectively and predict the classes better than random chance, demonstrating its ability to generalise and make meaningful predictions on unseen data.

6.2 SciBERT

Over the course of each epoch, the model performed very well on the training sets, but did not do as well on the validation sets. During the first epoch for the training set, the model received high scores for the metrics, with a weighted average precision score of 0.9285, recall of 0.9237, and F1 of 0.9249, however, received scores of 0.5983 for the weighted average for the precision, 0.6271 for the recall, and 0.6103 for the F1 score. During the second epoch, despite a relatively significant difference in the metrics for the training set (precision: 0.9819; recall: 0.9816, F1:0.9817), this only resulted in a slight increase in the metrics for the validation set, with weighted averages of 0.6127 for the precision, 0.6483 for the recall, 0.6259 for the F1 score. Despite having the highest



		Loss	Accuracy	Weighted Precision	Weighted Recall	Weighted F1
Epoch 1	Training set	0.2312	0.9237	0.9285	0.9237	0.9249
	Validation set	1.6160	0.6271	0.5983	0.6271	0.6103
Epoch 2	Training set	0.0527	0.9816	0.9819	0.9816	0.9817
	Validation set	1.8105	0.6483	0.6127	0.6483	0.6259
Epoch 3	Training set	0.0518	0.9859	0.9859	0.9859	0.9859
	Validation set	2.0918	0.6186	0.5969	0.6186	0.6064

(a) Results over each epoch during training

	Precision	Recall	F1 Score
Satisfactory	0.80	0.83	0.81
Unsatisfactory	0.52	0.46	0.49
Weighted Avg	0.72	0.72	0.72

(b) Test set

Figure 6.2: Over the course of each epoch, the model performed very well on the training sets, but did not do as well on the validation sets. On the test set, however, the model performed better than the highest scores for the validation set. This suggests the model is able to performed well on unseen data.

metrics for the training data in the third epoch (precision: 0.9859; recall: 0.9859; F1: 0.9859), this model had the lowest performance on the validation set, with scores of 0.5969, 0.6186, and 0.6064 for the weighted averages of the precision, recall and F1 score respectively, indicating significant overfitting to the training data.

On the test set, however, the model performed very well, even doing better than the highest scores for the validation set, with 0.72 for all the weighted averages of the precision, recall, and F1 score. This shows the model performed well on unseen data and that the model was not prone to overfitting.

CHAPTER 7

Discussion

7.1 Comparison of MNB and SciBERT results

The scores on the test sets show that the SciBERT model performed better than the Multinomial Naive Bayes model, with a weighted F1 score of 0.72 on the test set, compared to 0.66 for the Multinomial Naive Bayes Model. This could be because in health-related contexts, nuanced language and precise terminology are critical, and so having a model that can consider the context may play a significant role in the performance.

That being said, the difference is not as significant as expected. This suggests that Multinomial Naive Bayes can be a good alternative to the pre-trained model in the context of the detection of satisfactory and unsatisfactory health information, as although it is not as efficient in detection, it is a much less costly and computationally intense method than SciBERT, and can still achieve adequate results.

Furthermore, although both models achieved relatively high weighted scores, both models performed much better in the classification of the majority satisfactory class: the Multinomial Naive Bayes model achieved an F1 score of 0.77 for the satisfactory class and 0.36 for the unsatisfactory class on the test set, and similarly, the SciBERT model achieved an F1 score of 0.81 for the satisfactory class, and 0.49 on the unsatisfactory class. This could be because of the significant difference in the word counts between the two classes. As discussed previously in section 5.3.1, the longer articles in the satisfactory class could have skewed the models towards it, simply because they had more information to process in the satisfactory class.

7.2 Limitations and Potential Issues

The models suggest a potential in using machine learning for the automation of misinformation detection in online health news articles, which could save a lot of time and resources than if they were to be evaluated normally. That said, there are some limitations to this study that need to be put into consideration.

Due to the nature of the dataset, these particular models are only suitable for detecting whether a health intervention is adequately discussed in a health article, and so would not be suitable to be used for articles that discuss more than one intervention or health issue. It is also not clear whether the performance of the algorithms is specific only to this dataset, as it does not give a representative sample of news topics. Nevertheless, the performance of the models show that there is potential for them to be used in the detection of satisfactory and unsatisfactory health information in news articles related to health.

Another issue with these models is that though they can give an indication of whether a particular news article gives a balanced perspective, they do not evaluate the accuracy of the scientific content themselves. Nevertheless, they could encourage readers to read online health news articles more critically, and seek out more information regarding an intervention should the need arise.

Naturally, because the dataset used news articles from US media, more research would have to be done for other countries and languages as well. Other possible research areas could utilise images of the articles as well as other multimodal methods, as it was shown that these perform better than methods that only use the text or only use the images. Additionally, these models could be used together with explainable health misinformation detection, to enable readers to find out more about a particular health intervention, and not just take what they read for granted. This in turn could also improve trust between individuals and medical institutions.

Regarding SciBERT, although having the model pre-trained on scientific literature may give it an understanding of health issues and interventions, in news articles the language may not be as technical to make it easier for the general public to understand. This might be partially accountable for the lower metric results of the validation and test sets compared to the training data. Further research could therefore investigate whether the general model BERT would give a better performance.

7.3 Drivers of health misinformation and its spread

To tackle the problem of misinformation, the drivers must also be understood.

The spread of health-related misinformation has been intensified by the role of the internet and social media. Easy access to the internet and the ability to share information with thousands or even millions of individuals at a time mean that today's society makes it easier than ever to spread both reliable and unreliable information. This information is then spread even further through the effect of echo chambers, whereby like-minded users reinforce a shared narrative and exposure to diverse perspectives is limited [31]. Though having an automated system of detecting health misinformation can save time and resources, the ever-evolving nature of scientific knowledge and medicine mean that even if algorithms are regularly updated, search tools may not be able to keep up with the new evidence [13] and spread on social media, and this could result in delays in receiving correct information, which is especially important to consider when faced with a quickly changing situation such as a pandemic.

It has also been found that false health misinformation diffuses much faster and more broadly than true information. For instance, fake news surrounding the Zika virus was shared three times more on social media than verified stories between February 2016 and January 2017 [32]. The popularity of misinformation over factual messages is also reflected in other health topics, for example drugs [33], dialysis [34], and chronic diseases [35] [36] [37]. This could be because content that evoke strong emotions such as anger, anxiety or disgust are more likely to be shared [38] [39] and may mean that misinformation that rely on scaremongering are even more susceptible to being spread.

On the other hand, a paper published in 2021 in the *Journal of Medical Research* found that people with better information and science literacy were less likely to spread health misinformation videos on the internet [40]. Though this is only one paper and only focused on health misinformation in vidoes, it could imply that better information and science literacy are the keys to tackling health misinformation, however, more research needs to be done to investigate this.

There are numerous psychological and sociological reasons for why people are susceptible to health misinformation. In terms of psychological reasons, these may include denialism and conspiracy thinking [41], the illusory truth effect (when repeated claims are more likely to be perceived as true compared to novel claims or claims that are not repeated) [42], inattention account (when people rely on quick, intuitive judgements rather than reflective thinking due to the distracting nature of social media) [42], and motivated reasoning (when individuals have a pre-determined goal at the start of their reasoning process, such as wanting to believe vaccines are unsafe due to familial influence) [42].

One of the sociological drivers for the spread of health misinformation is for personal gain. Perhaps one of the most prominent examples is the "Wakefield controversy" which took place in the UK in the late 20th century and early 21st century [43]. For years, many parents refused to get their children vaccinated with the MMR vaccine after a 1998 paper published in the Lancet by Andrew Wakefield suggested a link between the vaccine, some forms of colitis, and autism [44]. This led to outbreaks of measles, mumps, and rubella in the US and UK which were previously under control. However, it was found that Wakefield had a major conflict of interest as he had been paid 81,800 euros by lawyers planning legal action against vaccine manufacturers to conduct tests on ten children [45]. Numerous studies published evidence refuting the study results (e.g. [46] [47] [48]), and the paper was retracted in 2010 [44]. Other sociological drivers of health misinformation include lack of accountability for spreading false health information [49] and the provision of social support for people during health crises, albeit false [50].

These suggest that having tools that encourage critical thinking may not be enough to tackle health misinformation as other psychological and sociological factors play a role. Therefore more interdisciplinary research could be done to see if psychological and sociological theories and artificial intelligence could be combined to tackle the issue.

CHAPTER 8

Conclusion

In conclusion, both the Multinomial Naive Bayes model and the transformer model SciBERT show potential for evaluating health related news articles in terms of whether they give satisfactory or unsatisfactory explanations based on the 10 criteria from the HealthNewsReview website. For the Multinomial Naive Bayes model, using the Bag-of-Words method for feature extraction, a smaller smoothing parameter, class priors, and removal of custom stop words resulted in an overall weighted F1 score of 0.66 on the test set. In comparison, the SciBERT model achieved 0.72 for the weighted F1 score.

Though the SciBERT model performed better than the Multinomial Naive Bayes model, the Naive Bayes model still performed relatively well and is a much more cost-effective and less computationally intense alternative. This can potentially save lots of time and resources than having articles reviewed manually, which is often dependent on funding and a consistent workforce. More research needs to be done on representative data and data from other countries and languages, and there is the potential for improvement through multimodal detection methods and having explainable misinformation detection. Nevertheless, it is important for individuals to be able to critically analyse any information given to them, especially when it can directly impact their health, and building trust between the general public and medical institutions could help individuals make more informed choices.

APPENDIX A

Appendix

Dataset:

EnyanDai. EnyanDai/FakeHealth; 2024. Original-date: 2020-03-29T01:58:58Z. Available from: GitHub - FakeHealth

EDA code:

https://colab.research.google.com/drive/1VQ7$_{7}EURHory6jPHXYutF_{9}bVcQLW6kZ?usp = sharing$

Naive Bayes code:

https://colab.research.google.com/drive/1$_{w}svazIP8gQk8ZoMFGbAHrUksYb74g6S?usp = sharing$

SciBERT code:

https://colab.research.google.com/drive/1P$_{F}b4lyAe2ViiL-2bWeJLBrRPb25sDFf?usp = sharing$

The Guardian
ABCNews
Health Day
FoxNews
The Philadelphia Inquirer
The Washington Post
NBCNews
BuzzFeed
Vox
Los Angeles Times
CBSNews.com
STAT
FiveThirtyEight
Newsweek
HealthDay
MSNBC
WebMD
The Arizona Republic
Hartford Courant
The Oregonian
Reuters
NPR
USA Today
The New York Times
Reuters Health
US News & World Report
The Wall Street Journal
Associated Press
CNN
The Boston Globe
Time
Medical Daily
Chicago Tribune
Houston Chronicle
AP Associated Press
CNN Health
Star Tribune
The Denver Post
The Houston Chronicle

Figure A.1: List showing all the news sources evaluated in HealthStory content

Bibliography

[1] Dai E, Sun Y, Wang S. Ginger Cannot Cure Cancer: Battling Fake Health News with a Comprehensive Data Repository. Proceedings of the International AAAI Conference on Web and Social Media. 2020 May;14:853-62. Available from: https://ojs.aaai.org/index.php/ICWSM/article/view/7350.

[2] HealthNewsReview. How is the Star Score Determined |;. Available from: https://web.archive.org/web/20220626022623/https:/www.healthnewsreview.org/about-us/star-scores/.

[3] Khan JY, Khondaker MTI, Afroz S, Uddin G, Iqbal A. A benchmark study of machine learning models for online fake news detection. Machine Learning with Applications. 2021 Jun;4:100032. Available from: https://www.sciencedirect.com/science/article/pii/S266682702100013X.

[4] Zuo C, Zhang Q, Banerjee R. An Empirical Assessment of the Qualitative Aspects of Misinformation in Health News. In: Feldman A, Da San Martino G, Leberknight C, Nakov P, editors. Proceedings of the Fourth Workshop on NLP for Internet Freedom: Censorship, Disinformation, and Propaganda. Online: Association for Computational Linguistics; 2021. p. 76-81. Available from: https://aclanthology.org/2021.nlp4if-1.11.

[5] Oxman M, Larun L, PÃ©rez Gaxiola G, Alsaid D, Qasim A, Rose CJ, et al. Quality of information in news media reports about the effects of health interventions: Systematic review and meta-analyses. F1000Research. 2022 Jan;10:433. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC8756300/.

[6] Yaqub O, Castle-Clarke S, Sevdalis N, Chataway J. Attitudes to vaccination: A critical review. Social Science & Medicine. 2014 Jul;112:1-11. Available from: https://www.sciencedirect.com/science/article/pii/S0277953614002421.

[7] World Health Organization. How do vaccines work?;. Available from: https://www.who.int/news-room/feature-stories/detail/how-do-vaccines-work.

[8] DubÃ© E, Vivion M, MacDonald NE. Vaccine hesitancy, vaccine refusal and the anti-vaccine movement: influence, impact and implications. Expert Review of Vaccines. 2015 Jan;14(1):99-117. Publisher: Taylor & Francis _eprint: https://doi.org/10.1586/14760584.2015.964212. Available from: https://doi.org/10.1586/14760584.2015.964212.

[9] Barua Z, Barua S, Aktar S, Kabir N, Li M. Effects of misinformation on COVID-19 individual responses and recommendations for resilience of disastrous consequences of misinformation. Progress in Disaster Science. 2020 Dec;8:100119. Available from: https://www.sciencedirect.com/science/article/pii/S2590061720300569.

[10] Spring M. The human cost of virus misinformation. 2020 May. Available from: https://www.bbc.co.uk/news/stories-52731624.

[11] Gangarosa EJ, Galazka AM, Wolfe CR, Phillips LM, Miller E, Chen RT, et al. Impact of anti-vaccine movements on pertussis control: the untold story. The Lancet. 1998 Jan;351(9099):356-61. Publisher: Elsevier. Available from: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(97)04334-1/fulltext.

[12] Islam MS, Sarkar T, Khan SH, Mostofa Kamal AH, Hasan SMM, Kabir A, et al. COVID-19âRelated Infodemic and Its Impact on Public Health: A Global Social Media Analysis. The American Journal of Tropical Medicine and Hygiene. 2020 Oct;103(4):1621-9. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7543839/.

[13] Southwell BG, Niederdeppe J, Cappella JN, Gaysynsky A, Kelley DE, Oh A, et al. Misinformation as a Misunderstood Challenge to Public Health.

American Journal of Preventive Medicine. 2019 Aug;57(2):282-5. Publisher: Elsevier. Available from: https://www.ajpmonline.org/article/S0749-3797(19)30159-X/abstract.

[14] ASCO. National Survey Reveals Surprising Number of Americans Believe Alternative Therapies Can Cure Cancer; 2018. Available from: https://society.asco.org/about-asco/press-center/news-releases/national-survey-reveals-surprising-number-americans-believe.

[15] Borges do Nascimento IJ, Pizarro AB, Almeida JM, Azzopardi-Muscat N, GonÃ§alves MA, BjÃ¶rklund M, et al. Infodemics and health misinformation: a systematic review of reviews. Bulletin of the World Health Organization. 2022 Sep;100(9):544-61. Available from: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC9421549/.

[16] Kim JY, Kesari A. Misinformation and Hate Speech: The Case of Anti-Asian Hate Speech During the COVID-19 Pandemic. Journal of Online Trust and Safety. 2021 Oct;1(1). Number: 1. Available from: https://tsjournal.org/index.php/jots/article/view/13.

[17] McKay D, Heisler M, Mishori R, Catton H, Kloiber O. Attacks against health-care personnel must stop, especially as the world fights COVID-19. The Lancet. 2020 Jun;395(10239):1743-5. Publisher: Elsevier. Available from: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(20)31191-0/fulltext.

[18] Sell TK, Hosangadi D, Trotochaud M. Misinformation and the US Ebola communication crisis: analyzing the veracity and content of social media messages related to a fear-inducing infectious disease outbreak. BMC Public Health. 2020 May;20(1):550. Available from: https://doi.org/10.1186/s12889-020-08697-3.

[19] Schlicht IB, Fernandez E, Chulvi B, Rosso P. Automatic detection of health misinformation: a systematic review. Journal of Ambient Intelligence and Humanized Computing. 2024 Mar;15(3):2009-21. Available from: https://doi.org/10.1007/s12652-023-04619-4.

[20] Mattern J, Qiao Y, Kerz E, Wiechmann D, Strohmaier M. FANG-COVID: A New Large-Scale Benchmark Dataset for Fake News Detection in German. In: Aly R, Christodoulopoulos C, Cocarascu O, Guo Z, Mittal A, Schlichtkrull M, et al., editors. Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER). Dominican Republic: Association for Computational Linguistics; 2021. p. 78-91. Available from: https://aclanthology.org/2021.fever-1.9.

[21] Meppelink CS, Hendriks H, Trilling D, van Weert JCM, Shao A, Smit ES. Reliable or not? An automated classification of webpages about early childhood vaccination using supervised machine learning. Patient Education and Counseling. 2021 Jun;104(6):1460-6. Available from: https://www.sciencedirect.com/science/article/pii/S0738399120306376.

[22] Zhao Y, Da J, Yan J. Detecting health misinformation in online health communities: Incorporating behavioral features into machine learning based approaches. Information Processing & Management. 2021 Jan;58(1):102390. Available from: https://www.sciencedirect.com/science/article/pii/S0306457320308852.

[23] Kumari S, Reddy HK, Kulkarni CS, Gowthami V. Debunking health fake news with domain specific pre-trained model. Global Transitions Proceedings. 2021 Nov;2(2):267-72. Available from: https://www.sciencedirect.com/science/article/pii/S2666285X21000662.

[24] Kotonya N, Toni F. Explainable Automated Fact-Checking for Public Health Claims. In: Webber B, Cohn T, He Y, Liu Y, editors. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). Online: Association for Computational Linguistics; 2020. p. 7740-54. Available from: https://aclanthology.org/2020.emnlp-main.623.

[25] Beltagy I, Lo K, Cohan A. SciBERT: A Pretrained Language Model for Scientific Text. In: Inui K, Jiang J, Ng V, Wan X, editors. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong,

China: Association for Computational Linguistics; 2019. p. 3615-20. Available from: https://aclanthology.org/D19-1371.

[26] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding.

[27] EnyanDai. EnyanDai/FakeHealth; 2024. Original-date: 2020-03-29T01:58:58Z. Available from: https://github.com/EnyanDai/FakeHealth.

[28] Schwitzer G. HealthNewsReview is relevant as ever for health care interventions; 2024.

[29] HealthNewsReview. HealthNewsReview.org;. Available from: https://web.archive.org/web/20220803193934/https://www.healthnewsreview.org/.

[30] Shu K, Mahudeswaran D, Wang S, Lee D, Liu H. Fakenewsnet: A data repository with news content, social context, and spatiotemporal information for studying fake news on social media. Big data. 2020;8(3):171-88. Publisher: Mary Ann Liebert, Inc., publishers 140 Huguenot Street, 3rd Floor New â¦.

[31] Cinelli M. The echo chamber effect on social media | PNAS;. Available from: https://www.pnas.org/doi/abs/10.1073/pnas.2023301118.

[32] Sommariva S, Vamos C, Mantzarlis A, ÄÃ o LUL, Martinez Tyson D. Spreading the (Fake) News: Exploring Health Messages on Social Media and the Implications for Health Professionals Using a Case Study. American Journal of Health Education. 2018 Jul;49(4):246-55. Publisher: Routledge _eprint: https://doi.org/10.1080/19325037.2018.1473178. Available from: https://doi.org/10.1080/19325037.2018.1473178.

[33] Al Khaja KAJ, AlKhaja AK, Sequeira RP. Drug information, misinformation, and disinformation on social media: a content analysis study. Journal of Public Health Policy. 2018 Aug;39(3):343-57. Available from: https://doi.org/10.1057/s41271-018-0131-2.

[34] Garg N, Venkatraman A, Pandey A, Kumar N. YouTube as a source of information on dialysis: A content analysis. Nephrology. 2015;20(5):315-20. _eprint:

https://onlinelibrary.wiley.com/doi/pdf/10.1111/nep.12397. Available from: https://onlinelibrary.wiley.com/doi/abs/10.1111/nep.12397.

[35] Chen L, Wang X, Peng TQ. Nature and Diffusion of Gynecologic CancerâRelated Misinformation on Social Media: Analysis of Tweets. Journal of Medical Internet Research. 2018 Oct;20(10):e11515. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution: Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. Available from: https://www.jmir.org/2018/10/e11515.

[36] Kumar N, Pandey A, Venkatraman A, Garg N. Are video sharing Web sites a useful source of information on hypertension? Journal of the American Society of Hypertension. 2014 Jul;8(7):481-90. Available from: https://www.sciencedirect.com/science/article/pii/S1933171114005476.

[37] Leong AY, Sanghera R, Jhajj J, Desai N, Jammu BS, Makowsky MJ. Is YouTube Useful as a Source of Health Information for Adults With Type 2 Diabetes? A South Asian Perspective. Canadian Journal of Diabetes. 2018 Aug;42(4):395-403.e4. Available from: https://www.sciencedirect.com/science/article/pii/S1499267117303982.

[38] Heath C, Bell C, Sternberg E. Emotional selection in memes: The case of urban legends. Journal of Personality and Social Psychology. 2001;81(6):1028-41. Place: US Publisher: American Psychological Association.

[39] Jonah Berger, Katherine L Milkman. What Makes Online Content Viral? - Jonah Berger, Katherine L. Milkman, 2012;. Available from: https://journals.sagepub.com/doi/full/10.1509/jmr.10.0353?casa_token=mEns4QwGkoEAAAAA%3AgZgex7VEf-NhJK-4_aOYokG9pwQg4_oyQYBFKZVMoRqD0brr8XbpWrG0GKqPKMM4EdkZwFpZH6aFvg.

[40] Keselman A, Smith CA, Leroy G, Kaufman DR. Factors Influencing Willingness to Share Health Misinformation Videos on the Internet: Web-Based Survey. Journal of Medical Internet Research. 2021 Dec;23(12):e30323. Company: Journal of Medical Internet Research Distributor: Journal of Medical Internet Research Institution:

Journal of Medical Internet Research Label: Journal of Medical Internet Research Publisher: JMIR Publications Inc., Toronto, Canada. Available from: https://www.jmir.org/2021/12/e30323.

[41] Uscinski JE, Enders AM, Klofstad C, Seelig M, Funchion J, Everett C, et al. Why do people believe COVID-19 conspiracy theories? Harvard Kennedy School Misinformation Review. 2020 Apr;1(3). Available from: https://misinforeview.hks.harvard.edu/article/why-do-people-believe-covid-19-conspiracy-theories/.

[42] van der Linden S. Misinformation: susceptibility, spread, and interventions to immunize the public. Nature Medicine. 2022 Mar;28(3):460-7. Publisher: Nature Publishing Group. Available from: https://www.nature.com/articles/s41591-022-01713-6.

[43] Tafuri S. Addressing the anti-vaccination movement and the role of HCWs - ScienceDirect;. Available from: https://www.sciencedirect.com/science/article/pii/S0264410X13015053?casa_token=zsEu3wJ3m-MAAAAA:R2XqattxjjP3piyDnULnt_Y2zNbsLN0YDS8kbqvhwW_gYDuU-EdrlmGFeluV0w7Ua-lywncuRw.

[44] Wakefield AJ, Murch SH, Anthony A, Linnell J, Casson DM, Malik M, et al. RETRACTED: Ileal-lymphoid-nodular hyperplasia, non-specific colitis, and pervasive developmental disorder in children. The Lancet. 1998 Feb;351(9103):637-41. Publisher: Elsevier. Available from: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(97)11096-0/fulltext?_sp=c37a1f11-48d4-44c6-ab26-98cf44c923f4.1537210935599.

[45] Tafuri S, Gallone MS, Cappelli MG, Martinelli D, Prato R, Germinario C. Addressing the anti-vaccination movement and the role of HCWs. Vaccine. 2014 Aug;32(38):4860-5. Available from: https://www.sciencedirect.com/science/article/pii/S0264410X13015053.

[46] Frank Destefano. Negative association between MMR and autism - The Lancet;. Available from: https://www.thelancet.com/journals/lancet/article/PIIS0140673699001609/fulltext.

[47] DeStefano F, Thompson WW. MMR vaccine and autism: an update of the scientific evidence. Expert Review of Vaccines. 2004 Feb;3(1):19-22. Publisher: Taylor & Francis _eprint: https://doi.org/10.1586/14760584.3.1.19. Available from: https://doi.org/10.1586/14760584.3.1.19.

[48] Prof Brent Taylor, Elizabeth Miller, Paddy Farrington, Maria-Christina Petropoulous, Isabelle Favout-Mayoud, Jun Li, et al.. Autism and measles, mumps, and rubella vaccine: no epidemiological evidence for a causal association - The Lancet;. Available from: https://www.thelancet.com/journals/lancet/article/PIIS0140-6736(99)01239-8/fulltext.

[49] Tsirintani M. Fake News and Disinformation in Health Care- Challenges and Technology Tools. In: Public Health and Informatics. IOS Press; 2021. p. 318-21. Available from: https://ebooks.iospress.nl/doi/10.3233/SHTI210172.

[50] Zhou C, Xiu H, Wang Y, Yu X. Characterizing the dissemination of misinformation on social media in health emergencies: An empirical study based on COVID-19. Information Processing & Management. 2021 Jul;58(4):102554. Available from: https://www.sciencedirect.com/science/article/pii/S0306457321000583.