Spring 2022

Unmasking Medical Fake News Using Machine Learning Techniques

Garima Chaphekar
San Jose State University

UNMASKING MEDICAL FAKE NEWS USING MACHINE LEARNING TECHNIQUES

A Thesis

Presented to

The Faculty of the Department of Computer Engineering

San José State University

In Partial Fulfillment

of the Requirements for the Degree

Master of Science

by

Garima Chaphekar

May 2022

The Designated Thesis Committee Approves the Thesis Titled

UNMASKING MEDICAL FAKE NEWS USING MACHINE LEARNING TECHNIQUES

by

Garima Chaphekar

APPROVED FOR THE DEPARTMENT OF COMPUTER ENGINEERING

SAN JOSÉ STATE UNIVERSITY

May 2022

Jorjeta Jetcheva, Ph.D.	Department of Computer Engineering
Carlos Rojas, Ph.D.	Department of Computer Engineering
Stas Tiomkin, Ph.D.	Department of Computer Engineering

ABSTRACT

UNMASKING MEDICAL FAKE NEWS USING MACHINE LEARNING TECHNIQUES

by Garima Chaphekar

Fake news has always been a critical and challenging problem in the information environment. The propagation of false news is a serious concern, especially in medical information, which can have dangerous and potentially deadly consequences. With the tsunami of online misinformation, it is crucial to fight fake medical news. In this study, we use machine learning techniques to help detect fake news related to diseases, including COVID-19, Ebola, Zika, SARS, Cancer, and Polio. To facilitate research in this space, we create a new medical dataset named MedHub. MedHub has records from two publicly available datasets on COVID and manually curated facts and myths about the other diseases. In addition, we build several different machine learning models trained on MedHub, including KNN, Naïve Bayes, SVM, Logistic regression, and MLP classifier, and present a proof-of-concept web application that uses these models to detect fake medical news. Our best-performing model, which we call Disease Myth Buster, is based on BERT and achieves an accuracy of 99%. In addition, we perform experiments to demonstrate that 1) our models perform well at identifying misinformation related to any disease even if it is not represented in the dataset, and 2) they are well optimized to identify COVID-19 specific misinformation, and 3) Disease Myth Buster can be extended for general fake news classification using Transfer learning. We create two new manually curated test datasets for the first two experiments. The first test dataset has 164 records related to Diabetes and the second test dataset has 13459 records of COVID-19 myths. We open-source all our datasets and models for future research.

ACKNOWLEDGMENTS

Foremost, I want to express my deep and sincere gratitude toward Dr. Jorjeta Jetcheva, without whom this thesis would not have been possible. I thank her for her continuous guidance and invaluable support. Besides my advisor, I also want to thank Dr. Carlos Rojas and Dr. Stas Tiomkin for joining the panel committee and providing their valuable feedback.

I want to thank my husband, grandfather, and parents for everything they have done for me. Lastly, I thank all my family and friends for their love and support.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF FIGURES

1 INTRODUCTION

In today's internet age, the term fake news is ubiquitous. The proliferating and evolving nature of fake news has given birth to its many definitions and classification. A closer review of the literature studying the problem of Fake news would reveal that the authors have used different meanings and interpretations of the term, which varies as per their research. For example, Lazer et al. [1] define "fake news" as fabricated information and classify it into two categories – misinformation (false or misleading information) and disinformation (incorrect information that is purposely spread to deceive people). Similarly, Allcott and Gentzkow [2] acknowledge fake news as intentionally, and verifiably false information meant to mislead readers. In this research, the authors focus on political fake news articles. Moreover, Waldrop [3] observes the seven different types of mis- and disinformation: satire, misleading content, imposter content, fabricated content, false connection, false context, and manipulated content. The seven different types of fake news vary in their degree of intent to deceive. Furthermore, Zubiaga et al. [4] add that fake news also encompasses rumors, which are "circulating pieces of information whose veracity is yet to be determined at the time of posting." The article also classifies rumors into long-standing and newly-emerging rumors based on their temporal characteristics. Thus, *fake news* is generally false information that looks like real news intentionally or unintentionally spread to mislead people and cannot be scientifically verified. In our research, we refer to any fact that can be scientifically verified as real news and any misinformation or rumor which cannot be scientifically verified as fake news.

Fake news has always been a problem in the information environment. Still, over the last decade, its circulation and proliferation have increased mainly due to the easy accessibility of the internet. Ease of access to social media has also contributed to its propagation. Allcott and Gentzkow explain how users rely on social media due to the absence of significant third-party filtering, fact-checking, or editorial judgment [2]. The

study also provides evidence of prominent fake news stories during the months before the 2016 election. Fake news creates more confusion and has a highly negative impact on individuals and communities. It also adversely affects the financial world, stock markets, and response to calamities. Its implications are amplified when it enters the health domain. In her article Warraich [5] writes, “Fake news threatens our democracy. Fake medical news threatens our lives.” She also mentions an interesting phenomenon of the “nocebo effect,” which describes that anticipation is the sole reason behind people experiencing adverse effects in medicine. Ultimately it leads to people discontinuing the medication. It is particularly damaging in that it undermines the credibility of the standard, doctor-recommended, and evidence-based medicines.

In the medical domain, fake news related to any newly discovered disease is produced and spread quickly. The reason is the scientific uncertainties due to the continuous evolution of knowledge and research around the disease. At the same time, people are desperate for information, and this gap is filled by misinformation; hence the situation gives a thriving platform to fake news on the Internet. For example, in the covid-19 pandemic, we came across famous fake remedies like eating garlic, consuming alcohol, or disinfectants. Some of the suggested remedies were harmful and directly threatened lives. Li et al. [6] report that 25% of the coronavirus’s most viewed videos contain misleading information. As the research on the virus and its types are still in progress, verifying information becomes even more challenging. Resistance to the COVID vaccine is fueled by the plethora of online misinformation surrounding its effectiveness and purpose. Indeed, a study by Loomba et al. [7] found that vaccine hesitancy is growing faster than ever in US and UK and is attributed to online misinformation. Similarly, the research by Sear et al. [8] reveals that the “anti-vax” community is attracting more support. Low acceptance of vaccines and their rejection might deter the goal of herd immunity, making humanity more vulnerable to future resurgence. The urgency of the situation has made

people easy prey to false news. Not only are the newly emerging diseases like Covid-19 and vaccination present on the fake news hit list, but it also persists for long-existing and chronic diseases like Cancer, Polio, Zika, HIV/AIDS, SARS, and Ebola, etc. Some top myths about cancer disease are that it is contagious, incurable, only happens to people with bad habits or addictions, and can be cured by performing rituals [9]. Likewise, Wella et al. [10] identified famous myths about HIV and AIDS in their study, which include "mosquitoes can transmit HIV," "if I shake my husband's hand I will be infected," "if my spouse is HIV positive then I am also positive," "condoms have a hole through which the virus can pass." In another article, Feuer [11] comments on the "conspiracy theories" surrounding Ebola, where the virus is alleged as a "bioweapon," "form of population control," and a "form of profit-making venture."

Such misconceptions add more fear and instill wrong beliefs about the disease, its cause, and treatment. The landscape of fake news in the health domain is much larger than misinformation in any other domain like political, economic, or social. It caters to a larger audience, proliferates much more quickly, and has deleterious consequences. Thus, it is of paramount importance to tackle the problem of misinformation in the medical field. Even though we cannot altogether remove the source and circulation of false news, we can certainly develop a system that helps people determine the veracity of the information.

The research objective of this thesis is to combat fake news related to the top diseases humankind has faced in the following ways. First, we introduce a new comprehensive labeled dataset with specific information related to the diseases, mainly COVID-19, Ebola, HIV/AIDS, Zika, Polio, Cancer, etc. The dataset is made by combining two publicly available datasets on COVID-19 and manually curating the top facts and myths on the remaining other diseases. For every disease covered in the dataset, the data is collected from different websites, official health websites, social media platforms, and journals to cover the vast forms of misinformation. Each record in the dataset is labeled "Fake" or

"Real." The dataset has both facts and myths about all the diseases to avoid class imbalance. We plan to open-source the dataset to the research community to enable future studies on preventing the spread of health-related misinformation. Second, we train five different machine learning models on the dataset to distinguish between fake medical and authentic news about the aforementioned diseases. In particular, the five models are KNN, Logistic Regression, SVM, Naïve Bayes, and MLP Classifier, achieving an accuracy of 81%, 86%, 85%, 84.5%, and 83%, respectively. Each model has two variants – one uses TF-IDF transformation, and the other uses Count Vectorizer. We also develop a proof-of-concept web application to showcase our models and how users can interact with a fake news detection system. The web application leverages the decision-making of the models to make a recommendation to the user regarding the veracity of the information. We also display relevant supporting information about the text with the results to learn more. Third, we also encourage users to give us feedback about the decisions made by the model. The credibility feedback will help us increase our dataset and improve our model accuracy over time. We introduce our best-performing model, "Disease Myth Buster," trained with 99% accuracy on our dataset. It is a deep model trained using BERT.

We also conduct learning experiments to demonstrate our model's robustness, their efficacy in identifying COVID-19 specific myths, and how the Disease myth buster can be extended for general fake news classification. For the first experiment, we test our model on a manually curated test dataset to show that our models perform well on the different types of diseases not covered in our original dataset. Our test dataset has facts and myths related to Diabetes. Diabetes disease is not covered in our original dataset. All our models perform well and achieve more than 60% accuracy on the test set. Since our models are trained on health-related information, we believe they promise any new disease misinformation detection.

One of the desired features of the fake news classifier is a high recall rate for the ”fake” class. The models must perform well in identifying fake news than real news. In this effort, we perform the second experiment to showcase that our models are optimized for detecting fake news. For this experiment, we prepare a test dataset of 13459 records of COVID-19 specific myths. All our models achieve are able to identify more than 70% myths.

Lastly, we experiment with building new general fake news detection models leveraging the transfer learning approach. This experiment uses the Disease Myth Buster as our base model and trains another deep learning model using standard fake news datasets, namely, ISOT and LIAR datasets. The new models achieved outstanding results and demonstrated that our model, Disease Myth Buster, could be extended easily for any fake news identification. We also run parallel experiments to see the performance of the various combinations. The experiment results reveal that using Disease Myth Buster as the base model gives better results and performs well at identifying general fake news instead of using the simple BERT embedding in the neural network.

2 LITERATURE REVIEW

2.1 Analysis on Characteristics of Fake News

The problem of fake news is not new and is a global topic of interest among researchers. Over the years, the definition of this term has evolved and has acquired polysemy. Research is done to understand why fake news exists and proliferates much more quickly than real news, its structural characteristics, and its primary consumers. A number of studies have been done to classify and understand the Fake news phenomenon. Baptista and Gradim [12] state the twofold purpose behind fake news – to deceive and to disseminate quickly. They observe that the language used in the fake text is simple, emotional, and non-technical.

Similarly, Horne and Adali [13] also notice the above properties of the linguistic style of fake news and, in addition, highlight that the titles in fake news are attractive, long, and use more named entities while the actual content is repetitive. Their research also sheds light on the result that fake news consumers are more unlikely to read the actual content beyond the title, explaining why titles are lengthy. In similar research, Shrestha and Spezzano [14] reveals that fake news titles and articles have negative sentiments and emotions associated with them. The content body is less descriptive than real news and articles. These findings are helpful in the sense that while manually curating our dataset, more focus is given on short text rather than full articles. The average number of words per record in all our datasets, including MedHub and the two test sets, is 20 words.

2.2 General Fake News Detection

Researchers have made several attempts to address the problem of fake news using various techniques. Some research uses traditional machine learning models like Naïve Bayes, Support Vector Machines (SVM) [15], while others use deep learning models like CNN and BiLSTM [16]. The datasets used in most experiments are the standard fake news datasets like LIAR, BuzzFeed, ISOT, etc. While some studies focus on general fake

news detection, others focus on fake news on social media like Twitter or Facebook. For instance, Helmstetter and Paulheim [17] used weak supervised learning techniques for the automatic detection of fake news on Twitter. In recent years, in addition to the primary research on fake news identification, there have been multiple studies on seemingly peripheral topics like bot detection for fake news [18] [19] and stance detection [20]. In particular, Ferrara in [18] provides evidence of the participation of automated accounts in spreading fake news. Similarly, a study in [19] proves the presence of Bots during the 2016 U.S. presidential election campaign period, which became inactive post the elections and were active again during the 2017 French presidential election. Both the mentioned studies focus on fake news spread by Bots on Twitter. Bots are automated accounts programmed to automate normal user account activities such as tweeting, retweeting, following, and liking posts. However, misinformation in the medical domain can originate and propagate on all online platforms, including social media, and is not always orchestrated using bots. Therefore, tackling the fake news problem related to top diseases requires a comprehensive, all-inclusive solution. In our approach, we collect data from multiple diverse sources to reflect the multi-platform nature of people's media diet.

Most research on fake news detection is limited to training machine learning models, evaluating their accuracy, and comparing them with other state-of-the-art models. Even though the models achieve great results for accuracy, the literature lacks the demonstration of the models put in real-time use for fake news detection. Very little research has worked towards creating the tool to make fact-checking simple and accessible. To address this concern in our research, we present a simple proof of concept that deploys trained machine learning models in our experiment for real-time fact-checking.

2.3 Domain Specific Fake News Detection

While ample research is done on identifying general fake news, several previous works have also focused on domain-specific fake news like political, entertainment, or

healthcare. For example, the model proposed in [2] detects fake news explicitly related to US presidential elections. After the outbreak of COVID-19, researchers started working on specifically COVID-19 fake news. For example, the authors in [21] analyze COVID-19 related conversations on Italian Facebook from January to April 2020. Similarly, a study by Jouyandeh et al. [22] explores the problem of fake news detection related to COVID-19 and its vaccination on Twitter. They explore and compare the performance of the different classifiers. Following the past studies, we concentrate on only one domain – healthcare. Our research aims at detecting fake news related to the top disease humanity has ever faced. Misinformation related to these diseases poses a severe threat to public health. Verifying the truthfulness of the information related to the diseases becomes even more challenging due to its evolving nature and medical illiteracy among ordinary people. In response to the rapidly growing misinformation in medicine, we develop reliable machine learning models trained on the latest facts and myths on the diseases to help determine the veracity of the information. We also test our models on the disease not covered in our dataset. To simulate an experiment that tests how well our models can identify health or disease-related misinformation on emerging diseases, we prepare a new test dataset of around 162 records related to Diabetes –this disease is not present in our training set and serves to demonstrate the versatility of our models in detecting fake news for a range of diseases, including those that are new and have not been encountered before.

2.4 Fake News Detection in Health Care

Any pandemic or new disease is accompanied by a tsunami of misinformation related to it. A study analyzing the main types, sources, and claims of COVID-19 reports that the fact-checkers increased by 900% from January to March 2020, highlighting the existence of misinformation during the early months of the pandemic [23]. Circulation of fake news in the pandemic era is even more harmful as its scope is broad, ranging from dangerous cures, antivaccination and false conspiracy theories to altering general public opinion (for

example, a preventative health behavior like mask mandate in COVID-19 was perceived a cause for facial deformities, hypoxemia, and bad bite teeth).

Misinformation not only exists for the new disease but also for chronic diseases like cancer and diabetes. Alternative cures are the main class of fake news for such diseases. Shi et al. [24] observe the dramatic increase in online search for cannabis as a cancer cure. They also mention that the fake news stories maintaining alternative cancer treatments gained 4.26 million engagements compared to the valid stories gaining only 0.036 million engagements on social media, highlighting that misinformation has an enormous outreach and high level of influence. Furthermore, another study by Goel et al. [25] highlights the high prevalence of wrong beliefs and misconceptions among hypothyroid patients.

The scale of the crisis has led researchers to delve more into handling healthcare-related misinformation. Payoungkhamdee et al. [26] present a new Thai healthcare-specific dataset named "LIMESODA," followed by evaluating the dataset using different deep learning approaches. Ciora and Cioca [27] make another successful attempt at fake news management in healthcare. The article presents a machine learning model trained using the KNN-BSA algorithm and achieves an accuracy of 70%. The datasets used in the experiments are two publicly available datasets on COVID. Even though we need to tackle the misinformation on COVID, we should not forget that the same problem exists for other diseases like cancer, Ebola, or ZIKA. Despite advancements in the studies of these diseases, there still exists a lot of misconception among people regarding its actual facts. Our research focuses on facts and myths about diseases such as cancer, Ebola, Zika, HIV/ AIDS, and H1N1 flu along COVID-19. In our research, the machine learning models are trained on a dataset with all the latest information about the aforementioned diseases.

We develop five supervised machine learning models - KNN, SVM, Logistic Regression, MLP Classifier, and Naive Bayes, trained using two embedding techniques:

TF-IDF and Count vectorizer. We also develop Disease Myth Buster with the best classification accuracy, using pretraining the BERT model. We achieve up to 99% accuracy with our models. To show that our model performs best at identifying any health-related fake news, rumors, and myths, we experiment to test our model on specific COVID-19 related myths. For this experiment, we collect more than 13,000 titles of the fake news articles from the COVID resources section from the https://www.poynter.org/. The test set used in this experiment has only fake news. All our models can identify more than 75% of the fake news. Disease Myth Buster performs the best, achieving an accuracy of 88%.

3 METHODOLOGY

3.1 Dataset Description

One of the biggest challenges in the NLP domain is that there is an abundance of unlabeled data, but labeled data is limited. Labeled data is required for any supervised machine learning task like text classification. There are several public datasets available to address the fake news identification problem. However, a close analysis of the dataset would reveal some associated issues. For example, some datasets contain only tweets, titles, or articles from one website or social media. Some datasets are multilingual, making model learning more complex and leading to poor accuracy. Most of the existing datasets are multi-domain but are predominantly political. Thus, one of the major roadblocks to health-related fake news identification is the lack of adequately labeled medical datasets.

Furthermore, this problem aggravates due to the continuous evolution of knowledge surrounding any disease if the disease is newly discovered. As the research on the disease is ongoing, what might seem true today might be false tomorrow and vice-versa. Hence, we need an updated dataset. To address this concern, we construct a novel benchmark dataset named MedHub in our research. MedHub has 11001 records focused on health-related information.

As the platform existence of fake medical news is incredibly vast, ranging from websites to social media, it is of utmost importance that the data in the dataset should cover many aspects and be in multiple formats. The advancing research in fake news detection during the COVID-19 pandemic resulted in many COVID-19 specific datasets like COVID-19 Fake News Dataset [28], COVIDLIES [29], COVID-19 Rumor Dataset [30], and CoAID [31]. From the COVID datasets mentioned above, we include the two most comprehensive Covid datasets - COVID-19 Rumor Dataset and the CoAID dataset in MedHub. Both the datasets have claims, social media posts, and short ground truth labels (real and fake) in addition to the news themselves. The COVID-19 Rumor

Dataset collects rumors related to the COVID from different fact-checking websites and Twitter. Each record in the dataset has a veracity status of "True," "False," and "Unverified." We include only the True and the False records in our dataset as our dataset has only two labels – Fake and Real. Similarly, the CoAID Dataset collects tweets, replies, social media posts, and information on the website. It has around 2018 and 21043 fake and true claims and about 18000 and 260000 fake and true news articles. We select only 3756 records from the CoAID dataset to add to MedHub. These records are specifically claims, article titles, and article abstracts. As our primary goal is to identify misinformation related to the disease, and most myths are short and attention-grabbing, we prefer to keep short sentences in our dataset instead of lengthy articles. We retrieve only the label (real or fake) and the text column content from both datasets.

Our research focuses on identifying fake news related to the top diseases. We manually curate more than 1000 records on diseases like Ebola, Cancer, H1N1-flu, ZIKA, Polio, and SARS so that our models identify fake news related to a range of diseases. We scrape data from different websites ranging from fact-checking, official health care department websites to newspaper articles and Wikipedia to collect the data. We also refer to several scientific journals like IEEE Access, SpringerLinks' journals, and ACM Journals to include the latest research on diseases. All the gathered records, especially the myths, are manually checked and verified by cross-validating with multiple sources. For each disease covered in MedHub, we include all the related information like when the disease was discovered, its cause, symptoms, risks, number of people affected, and the latest research on it. We also cover famous myths about these diseases. All this information is gathered from the web. Including some records from existing public datasets and manually curated records from diverse sources reflect the multiplatform nature of people's media diet.

A study done by Pennycook et al. [32] reveals that one of the reasons for misinformation propagation is people do not think about the content's veracity before sharing it. The conclusion drawn in the study is that if the individuals are made to think about the content's accuracy, their level of "truth discernment" tends to increase. Following these results, we introduce a feature for the users in our web application to give feedback on whether they agree or disagree with the model's evaluation. This feature will also enable the continuous (though always moderated!) expansion of MedHub. We open-source our dataset to facilitate future research in the direction of disease misinformation.

We experiment with MedHub to include non-medical data like politics and entertainment. However, we found that the models do not perform well when trained with data from multiple domains and therefore restricted MedHub to medical data only.

Apart from Medhub, we develop two manually curated test datasets to show the model's robustness in identifying misinformation related to a range of diseases. The first test dataset has 162 hand-labeled records related to diabetes. The second test set has 13459 records crawled from the fact-checking website - poynter.org. The second is to demonstrate that our models best classify COVID-19 specific myths. We leverage the Beautiful Soup Python package for data scrapping. Table 1 shows the dataset's attributes developed in the research.

Table 1
Attributes of MedHub and the Test datasets

Dataset Name	**Total records**	**Real records**	**Fake records**	**Unique words**	**Average sentence length**
MedHub	11101	6036	5065	30789	20
Diabetes Test	162	93	69	727	13
COVID-19 specific misinformation	13459	0	13459	23459	17

3.1.1 Data Analysis

We perform data visualization to highlight the key characteristics of Medub. The dataset has two columns – one for text, and the other column indicates the label of the text. The rating scale for each record in the dataset is binary and can be classified as "Fake" or "Real." Fig. 1 shows the first five records in MedHub.

	Text	Label
0	"Spraying chlorine or alcohol on the skin kill...	Fake
1	"Only older adults and young people are at risk"	Fake
2	"Children cannot get COVID-19"	Fake
3	"COVID-19 is just like the flu"	Fake
4	"Everyone with COVID-19 dies"	Fake

Fig. 1. Dataset with text and label fields for training.

Another key feature of MedHub is that it contains short text as opposed to long articles. We do a word count for each record in the corpus by counting the number of words in the "Text" column. Fig. 2 depicts the average word length of records in MedHub. From the graph, we can see that the average word length of the record is less than 20 words.

MedHub has in total of 11001 records, out of which 6036 are real records while 5065 are fake records. Fig. 3 shows the class distribution. We also form a word cloud of the corpus. Fig. 4 shows the word cloud.

D'Ulizia et al. [33] put forward a few key characteristics of the datasets while doing a systematic comparative review of twenty-seven popular fake news datasets, which are news domain, application purpose, language, size, news content type, rating scale, and spontaneity. Table 2 describes the aforementioned properties of MedHub. MedHub covers

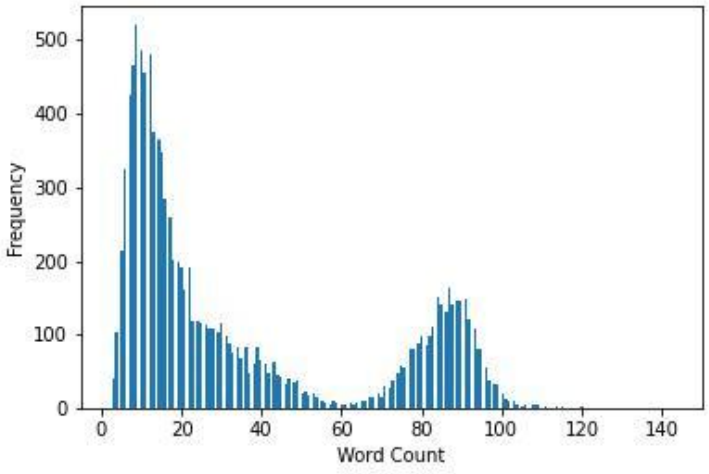

Fig. 2. A histogram of sentence length in MedHub.

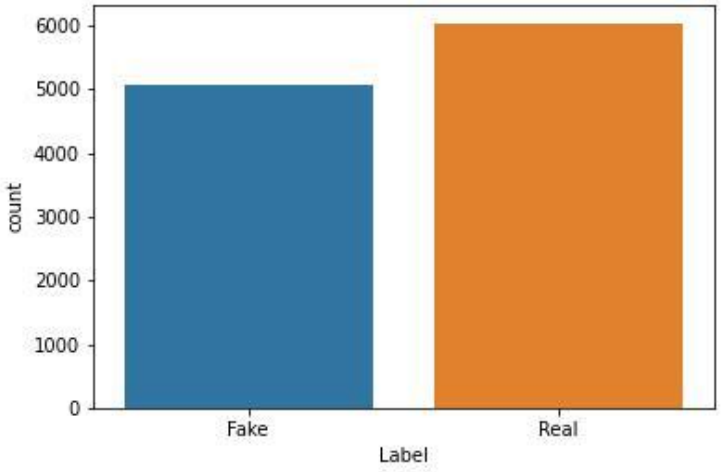

Fig. 3. Class distributions of MedHub.

the different kinds of fake news, including myths, fake news article titles, wrong beliefs, misconceptions, hoaxes, and rumors. However, we don't classify the different types of fake news as MedHub covers information from multiple media sources. We believe it is not viable to determine one's intention when the data is collected from a wide variety of sources.

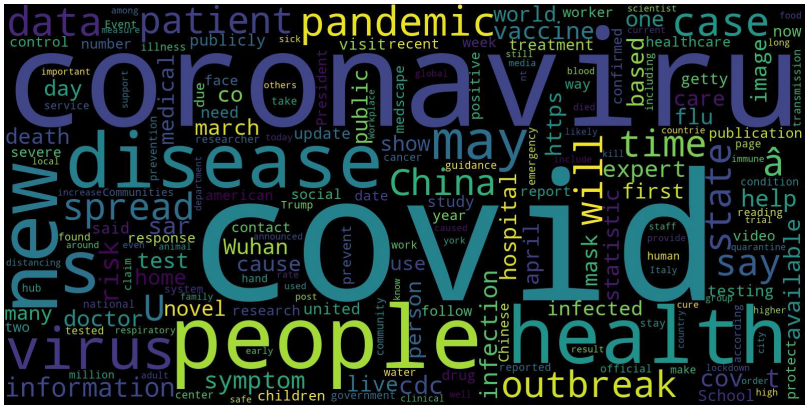

Fig. 4. Word cloud of MedHub.

Table 2
Properties of MedHub

Property of the Dataset	**Value**
Domain	Health
Application purpose	Fake news classification
Total Records	11001
Rating Scale (Labels)	Binary - “Real” and “Fake”
Language	Monolingual (English)
Media platform	Mainstream media and online social media
News Content Type	Text
Spontaneity	Spontaneous

3.2 Text Preprocessing

The first step toward model building in Natural Language Processing (NLP) is data preprocessing. It removes the text’s noise and makes learning easy for the models. The model’s performance and the validity of any experiments with and results of text classification are highly dependent on the quality of the input data. Our experiments corroborate the hypothesis that the models perform better when fed with preprocessed text as compared to the raw text. We also note that preprocessing steps sequence plays a vital role in how the data is formatted during the experiment. No text normalization techniques like Stemming or Lemmatization have been employed as we notice that it produces meaningless and different meaning words for our corpus. For example, the word

"genetically" gets converted to "genet," and both the words have totally different meanings, changing the interpretation of the sentence. It is primarily due to the presence of medical words in MedHub. Additionally, stemming and Lemmatization sometimes become a time-consuming step in text preprocessing. A similar finding is reported in [34], which indicates that the stemming process has high time complexity while achieving bare improvement in the performance of the model.

We apply the preprocessing steps outlined below.

3.2.1 Lower Case

We convert the complete text in the corpus to lower case.

3.2.2 Removing Punctuations

Punctuation marks do not add any real meaning to the text; hence, we remove all the punctuations in this step. We use the 32 punctuations in the string- string.punctuation from the string module.

3.2.3 Converting Numbers to Words

As MedHub contains all medical data, the text contains numbers too. For example, a record telling in which year was Zika discovered, different numbers of virus strains, etc. Preserving such information helps the model learn better, and hence we convert the numbers to words using the inflect library in python.

3.2.4 Expanding Contractions

As MedHub is a collection of article titles, and social media posts, the text is very likely to be informal and abbreviated. We convert the contractions in the text to the complete word for better analysis of the words.

3.2.5 Removing Links

The records taken from the COVID-19 Rumor Dataset contain a few website links along with the actual text. We remove such links from the text.

3.2.6 Removing Multiple Spaces and Unicode Characters

We remove all the extra spaces and Unicode characters from the corpus.

3.2.7 Removing Stop Words

We form a custom list of stop words by combining the Gensim and NLTK library's stop words and removing negation words like ”not,” ”cannot,” ”no,” and ”only.” Removing such words helps the models perform better at classifying sentences with negation words. For example, ”Ebola is real” is a record labeled ”Real” in the dataset. If the model is tested with the sentence ”Ebola is not real,” and the negation word ”not” is removed as part of text pre-processing, the sentence given to the model becomes ”Ebola is real.” The output for this sentence will be ”Real,” which is wrong. Also, the sentences ”Ebola is real” and ”Ebola is not real” will give the same results.

3.2.8 Feature Extraction

We employ two vector representation techniques for training each model – TF-IDF (term frequency-inverse document frequency) and Count Vectorizer.

TF-IDF is a simple yet powerful technique based on the Bag of words approach for text vectorization. It is a statistical way of calculating the relevance of a term in the document. It works by calculating the product of the weight of the term in the document called the term frequency (TF) (counting the number of times the word appears in the document divided by the total number of words in the document) and the inverse document frequency (IDF). IDF is calculated by taking a log of the total number of records or sentences in the corpus divided by the number of records or sentences containing the word. IDF increases the weight of the words, which are rare in the document corpus instead of the frequently occurring terms. The TF-IDF converts the raw text into a sparse 2D matrix of TF-IDF features of size- no of records in the corpus, total vocabulary of the corpus. The limitation of using TF-IDF is that the resultant matrix fails

to capture any semantic information. This technique can be computationally expensive in the case of vast vocabulary, as the size of the matrix depends on the number of unique words in the corpus. In mathematical terms, the TF-IDF for a word w in the sentence s from the complete recordset (MedHub) D is calculated as

$$TF - IDF(w,s,D) = TF(w,s) * IDF(w,D) \tag{1}$$

where TF (w, s) = Number of times the word w appears in the sentence s / total number of words in the sentence s and IDF (w, D) = log(number of sentences s in the corpus D / number of sentences s in the corpus D containing the word w)

Count Vectorization is another simple technique of word frequency representation. It is very similar to one-hot encoding. This technique also produces a sparse matrices like the TF-IDF vectorization. The size of the matrix depends on the number of records in the corpus and the total vocabulary. This simple technique also fails to capture the relational information of the words in the sentence.

3.3 Machine Learning Models

We train five different machine learning models: KNN (K- Nearest Neighbors), Logistic Regression, Naïve Bayes, SVM (Support Vector Machine), and MLP Classifier. Each model is trained using the two-vectorization technique mentioned above. All the models are trained using 80% of the data and tested on the remaining 20%.

3.3.1 KNN

It is a simple, non-parametric text classification algorithm. It works by finding the nearest neighbor of the vector and calculating the similarity between the closest vectors. The decision of classification of the new vector is influenced by the value of K and distance metrics. There are several attempts in the literature that employ KNN for text classification. Mladenova and Valova [35] use KNN for fake news and clickbait posts

identification on Facebook. They experiment with different values of K, distance metrics, and data scaling methods. Their best model achieves 83.5% accuracy. The main drawback of KNN for text classification is that it is sensitive to data distribution, and the time complexity is high. To identify the nearest neighbors, it has to calculate the vector's distance to all the existing vectors in the dataset. However, as MedHub has only around 11,000 records, KNN works well. We use the scikit learn library to implement KNN. We also experiment with different values of K using Grid search CV and see that an extremely small value of K, for example, K less than ten, leads to overfitting. Keeping the value of K between 10 and 100 achieves almost similar results, but as the value of K increases, the time complexity also increases. Increasing the value of K beyond 100 leads to poor outcomes. We choose the value of K to be 10. Our KNN model performs well with TF-IDF compared to Count Vectorizer and achieves training and testing accuracy of 85.5% and 79.4%, respectively.

3.3.2 Logistic Regression

It is a discriminative classifier and works well with binary classification. It is based on the concept of probability and uses a sigmoid function to limit the output between 0 and 1. A threshold is set to predict the actual class; the estimated probability then gets converted to the class depending on the threshold. Abdelminaam et al. [36] experimented with Logistic regression along with different models like Decision Tree, SVM, Random Forest, and deep models for COVID-19 specific fake news identification. The authors experiment with other TF-IDF vectorization techniques, and their best performing Logistic regression model is obtained using tri-gram with an accuracy of 81%. We implement the logistic regression model using the TF-IDF and the Count Vectorizer. Both the models perform equally well on the test set. However, the model trained with TF-IDF tends to overfit the data. Increasing the value of C also leads to overfitting, and hence we keep the default value of C as 1. Post experimenting with different values of regularization, the best results

are obtained with L2 regularization. Our best performing logistic regression model achieves training and testing accuracy of 92.85% and 86.4%, respectively.

3.3.3 Naïve Bayes

It is a probabilistic and a generative classifier. It is based on Baye's theorem. Naïve Bayes does not consider the ordering of words and hence does not retain any relationship between the words while doing classification. It is one of the first algorithms that tackled the spam filtering problem. Various versions of Naïve Bayes like Gaussian NB, Multinomial NB, and Bernoulli NB are available and have been used for text classification in the past. Granik and Mesyura [37] propose a simple approach for fake news classification using the Naïve Bayes algorithm. The study uses the 2000 records from the standard BuzzFeed news dataset, and their model achieves a test accuracy of 74%. We employ the Multinomial NB for our experiment as it is appropriate for feature frequency like word count. In our experiment, the Naïve Bayes model trained with TF-IDF performed better than the count vectorizer and achieves a training and testing accuracy of 90.7% and 86.03%.

3.3.4 Support Vector Machines

SVM is a universal binary classifier and works by finding the linear or polynomial threshold function to separate the instances of one class from the rest. It is well suited to handle the large feature space like text and performs well at text classification compared to other algorithms. The literature has evidence that shows that the SVM models outperform most of the other models for text classification tasks. For example, Abdelminaam et al. [36] show that the SVM model trained with unigrams TF-IDF performs best at identifying fake and real tweets containing COVID-19 information compared to other simple models like Decision Tree, KNN, Random Forest, and Logistic Regression. The SVM model achieves a test accuracy of 96.38%. We train the SVM model using the sklearn library. SVM provides a choice of kernels like linear, polynomial,

Gaussian, RBF, and sigmoid. A linear kernel is often a preferred choice over other non-linear kernels due to lesser parameters to hyper tune. We attempt to find the best hyperparameter combination of C, gamma, and kernel using GridSerachCV with 5-fold cross-validation. The best accuracy is achieved with RBF kernel and gamma and C values set to .1 and ten, respectively. However, this combination leads to overfitting, and hence we reduce the value of gamma further to .01. The effect of overfitting was more in the Count vectorizer model than in the TF-IDF vectorizer. Hence, we reduce the value of C to five and further reduce the gamma value to .007 in SVM using the Count vectorizer. Reducing the gamma value addresses the issue of overfitting but slightly lowers the accuracy. The low value of gamma and C leads to underfitting, and the high value leads to overfitting. The final train and test accuracy for TF-IDF and Count vectorizer SVM are 89%, 85% and 91%, 83%, respectively.

3.3.5 MLP Classifier

It is a basic feedforward neural network and has the configuration of the Input layer followed by the hidden layer followed by the output layer. We use the scikit learns' simple implementation of MLP Classifier, which has many parameters like the number of hidden layers, activation function, the alpha value for L2 regularization, iterations denoting the epochs, and optimization algorithm, and learning rate. We experiment with different values of the parameters and observe that hyper tuning the parameters can achieve almost 99% training accuracy. However, the models tend to overfit and perform a little low on the test dataset. To address the issue of overfitting, we increase the alpha value to 1.5. Alpha is the L2 regularization penalty term; increasing its value encourages small weights, resulting in a decision boundary that less fit the data. We also enable early stopping. By enabling this, the algorithm keeps aside 10% of the data for validation and stops the training process when it sees no further improvement in the validation score. We keep the learning rate to be adaptive and the default activation function (ReLU). With

these parameter values and one hidden layer with 100 neurons, the model achieves 87% and 83% train and test accuracy with TFIDF vectorization and 92%, 85% train and test accuracy with Count vectorization.

3.3.6 BERT Based Model

Google published its pre-trained BERT (Bidirectional Encoder Representations from Transformers) [38] Model in 2018. BERT is a pre-trained model, and hence its vocabulary is fixed. It is trained on the Wikipedia article and book corpus, and its vocab size is around 30,000. We have to use the BERT tokenizer to map the words to a sequence of embeddings for using BERT. BERT deploys the WordPiece Tokenizer, a subword-based tokenization algorithm. If the word does not exist in the BERT vocab dictionary, it breaks down into subwords and forms the tokens. Each of the words is mapped to a vocabulary id. The word embeddings are contextual in nature and are such that the distance between the vectors reflects their similarities. The input to the BERT model follows a fixed format. All the sentences fed to the BERT have to be of fixed length, either padded or truncated. Choice of the max length of the input impacts the training time and the accuracy. The input sentence should also start and end with the artificial ”CLS” and ”SEP” tokens, respectively. The CLS token marks the beginning of the sentence. The second input to the BERT model after input ids (words mapped to tokens) is the attention mask. It is an array of 0, and 1 of the size fixed-length differentiating between actual words and padded tokens. It helps the models to understand which words are to be prioritized and which can be ignored. BERT comes in two flavors – base and large. The difference lies in the number of transformer layers, attention heads, and hidden units.

Fig. 5 shows the output of the BERT tokenizer applied to one sample record. The first line in the figure shows the text. In the next step, the words in the sentence are broken down into tokens and appended with CLS and SEP at the start and the end. The output of the tokenizer is a dictionary with arrays of chosen fixed length size. The first array is the

```
ebola not real
['[CLS]', 'e', '##bol', '##a', 'not', 'real', '[SEP]']
{'input_ids': <tf.Tensor: shape=(1, 100), dtype=int32, numpy=
array([[  101,   174, 15792,  1161,  1136,  1842,   102,     0,     0,
            0,     0,     0,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0,     0,     0,     0,     0,     0,
            0,     0,     0,     0,     0,     0,     0,     0,     0,
            0]])>, 'token_type_ids': <tf.Tensor: shape=(1, 100), dtype=int32, numpy=
array([[0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]])>, 'attention_mask': <tf.Tensor: shape=(1, 100), dtype=int32, numpy=
array([[1, 1, 1, 1, 1, 1, 1, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0,
        0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0, 0]])>}
```

Fig. 5. Demonstration of BERT Tokenizer applied to one sample record.

input_ids which is an array of tokens mapped to numbers. The array starts and ends with 101 and 102, representing the CLS and the SEP token. Another array is the attention mask of 0 and 1, 0 representing the padding tokens and 1 marking the real tokens.

It is possible to leverage the existing BERT Model in many NLP tasks like text classification, name-entity recognition, part of speech tagging, and question answering, and take advantage of the pre-training rather than training a model from scratch. Several types of text classification models based on BERT are explored in the literature. Ding et al. [39] present a BERT-based mental model for fake news classification. The model achieves better results than state-of-the-art models for the LIAR dataset. Choudhary et al. [40] put forward “BerConvNet,” a hybrid deep learning model combining BERT and CNN. The study validates the performance of BerConvNet on four different benchmark datasets for fake news identification. The comparative analysis done in the study reveals that the new model outplays the existing state-of-the-art models on various performance metrics. Following the past literature, we leverage the BERT model for presenting our best-performing model, Disease Myth Buster. Disease Myth Buster is a deep learning model of five layers based on BERT that consists of two main building blocks. Fig. 6 shows the overall architecture of the model.

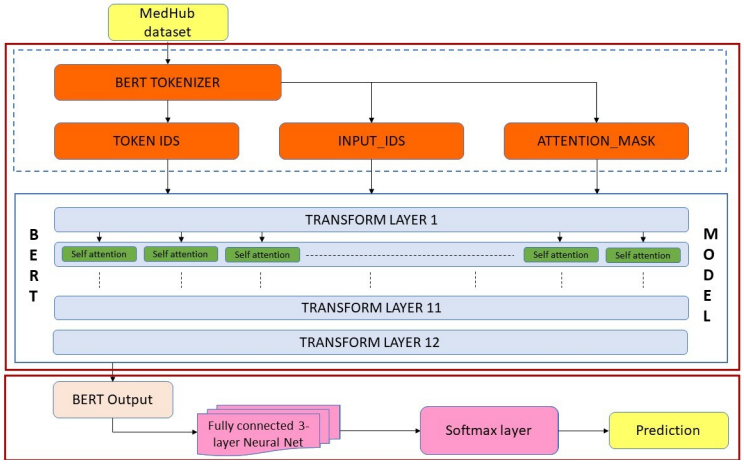

Fig. 6. Architecture of Disease Myth Buster

The first layer in the model is the BERT layer. The first block leverages the BERT model for producing the context-aware word embeddings which are then fed as input to the second part, the neural network. Fig. 7 shows the detailed parameters of each layer in the model.

Layer (type)	Output Shape	Param #	Connected to
input_ids (InputLayer)	[(None, 100)]	0	[]
attention_mask (InputLayer)	[(None, 100)]	0	[]
bert (TFBertMainLayer)	TFBaseModelOutputWithPoolingAndCrossAttentions(last_hidden_state=(None, 100, 768), pooler_output=(None, 768), past_key_values=None, hidden_states=None, attentions=None, cross_attentions=None)	108310272	['input_ids[0][0]', 'attention_mask[0][0]']
intermediate_layer_1 (Dense)	(None, 1024)	787456	['bert[0][1]']
intermediate_layer_2 (Dense)	(None, 512)	524800	['intermediate_layer_1[0][0]']
intermediate_layer_3 (Dense)	(None, 216)	110808	['intermediate_layer_2[0][0]']
output_layer (Dense)	(None, 2)	434	['intermediate_layer_3[0][0]']

Total params: 109,733,770
Trainable params: 109,733,770
Non-trainable params: 0

Fig. 7. Model Summary of Disease Myth Buster.

We use the BERT-base-case model in our research. Table 3 shows the details of the BERT model leveraged in our research. It has 12 transformer layers; each layer performs the self-attention mechanism and passes it to the next layer of the feed-forward network. The self-attention mechanism is the basic building block of the transformer in BERT. Each layer outputs a vector of size 768. As the word embedding passes through the different layers in the transformer block, it learns the context of the whole sentence. The top embeddings produced are fully contextualized and are aware of the entire sentence. We use the transformers and the TensorFlow library for using BERT. The BERT model outputs the last hidden state and the pooler output. Last_hidden_state is the hidden state sequence at the final layer of the BERT model. Pooled_output is the hidden state sequence of the CLS token run through a linear layer and a Tanh activation function. The last hidden state and the Pooler output are of the size (batch size, fixed input length, 768) and (batch size, 768), respectively. Any top embeddings can be fed into the classifier as they are fully contextualized. Still, there is a possibility that they may be localized to the meaning of a particular token. Hence, to get around this artifact, we use the CLS embedding (pooled_output) as it does not focus on any single token in the sentence.

Table 3
BERT-base architecture

Name of Parameter	**Value**
BERT Model	BERT-base-cased
Number of encoders	12
Number of Attention Heads	12
Size of embeddings produced / Hidden layer size	768
Total Number of Parameters	110M
Size of input	100
Library used for implementing BERT	Transformers
Tokenizer	BERT Tokenizer

The pooler output is fed as an input to a dense, fully connected neural network of 3 layers followed by the last layer of the sigmoid activation function. The batch size used for training is 8. Post experimenting with different attributes, we achieve 99% accuracy

for Disease Myth Buster. Table 4 shows the hyperparameter details of our best-performing model Disease Myth Buster.

Table 4
Disease Myth Buster architecture

Name of Parameter	Value
Batch Size	8
Training data size	80%
Validation data size	10%
Test data size	10%
Optimizer	Adam
Loss function	Binary Cross entropy
Epochs	10
Metrics	Accuracy, Precision, and Recall

3.4 Proof-of-Concept Web Application

Trained and high accuracy machine learning models in a vacuum are not sufficient to solve the fake news detection problem. Hence, we developed a proof of concept to identify fake medical news specific to diseases that deploy the trained machine learning models mentioned above. Our web application is a tool that helps the user to assess the credibility of medical information in real-time.

The web application lets the user enter to enter the text, select the vectorization technique (TF-IDF or Count Vectorizer) and the machine learning model. It then leverages the trained model to predict the credibility of the entered text. Since we believe that no machine learning model is entirely accurate and robust in identifying fake news, we also display the top five relevant google search links results for the entered text. We give users the option of whether they agree or disagree with the model's evaluation. This feedback serves twofold purposes. It will help in the continuous upgrade of MedHub as the entered text will be saved at the back end and improve the model's performance. Second, as more and more human-labeled data is gathered, the web application can be extended to display crowd-sourced credibility ratings of the text in the future.

The application's front end is a simple GUI written using HTML and CSS. The back end is implemented using the python web application framework, Flask. The web application is deployed on the Amazon EC2 instance. Fig. 8 and Fig. 9 shows the web application's interface.

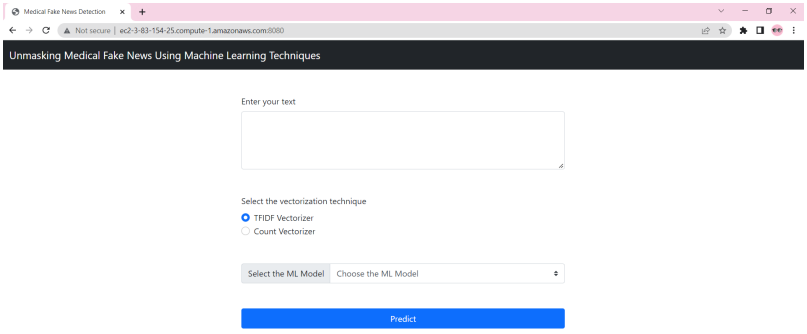

Fig. 8. Screen dump of Web application.

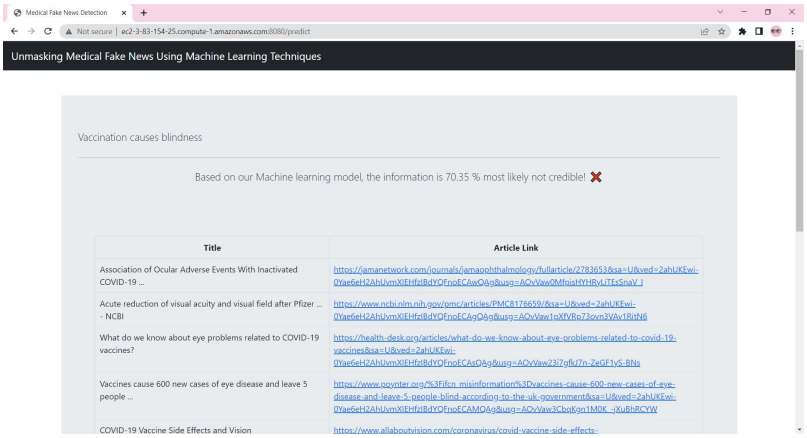

Fig. 9. Screen dump of Web application.

4 RESULTS

We train all our classifiers on 80% data of MedHub and test in on the remaining 20%. To evaluate the performance of the models, we use the commonly used metrics for assessing classifiers – Accuracy, Precision, Recall, and F1 score. Precision is the percentage of the predicted classes that are correct. The recall is the percentage of the positive class predicted correctly. F1 score is the harmonic mean of recall and precision.

Table 5 and Fig. 10 shows the evaluation metrics and the confusion matrix for all models using TF-IDF vectorization. SVM, Logistic Regression, and Naive Bayes perform the best and are almost similar in their performance, followed by MLP Classifier and then KNN.

Table 5
Classification results of models using TF-IDF Vectorizer

Model Name	Training Accuracy	Test Accuracy	Precision	Recall	F1 Score
Naïve Bayes	90.7	84.5	85	85	84
SVM	89	85	85	85	85
Logistic Regression	92.5	86.4	86	86	86
MLP Classifier	87.09	83	83	83	83
KNN	83.3	81.9	84	82	82

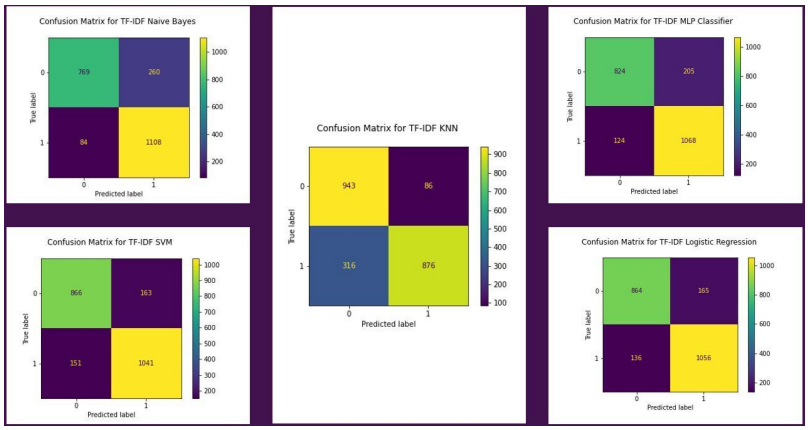

Fig. 10. Confusion matrix for all models using TF-IDF Vectorization.

Table 6 and Fig. 11 shows the evaluation metrics and the confusion matrix for all models using Count Vectorizer. All the models perform well and achieve more than 80% test accuracy except for KNN. One of the desirable properties of the Fake news detection system is a high recall rate for the Fake class; it should be able to identify fake records well compared to real records. Even though the accuracy of KNN is low, its recall rate for the Fake class is very high. It has a low recall rate for the real class, and hence its weighted recall comes down to 65%. It works well in identifying fake news but does a poor job at classifying real records.

Table 6
Classification results of models using Count Vectorizer

Model Name	**Training Accuracy**	**Test Accuracy**	**Precision**	**Recall**	**F1 Score**
Naïve Bayes	87.6	83.3	83	83	83
SVM	91	83.25	85	83	83
Logistic Regression	97.5	87.2	88	87	87
MLP Classifier	92	85	85	85	85
KNN	73.4	65.2	76	65	63

Table 7 shows the evaluation metrics for our best performing BERT based model, Disease Myth Buster. It achieves a training and testing accuracy of 99%.

Table 7
Classification results of BERT based Disease Myth Buster

Model Name	**Training Accuracy**	**Test Accuracy**	**Precision**	**Recall**	**F1 Score**
BERT Disease Myth Buster	99	99	99.5	99.5	99

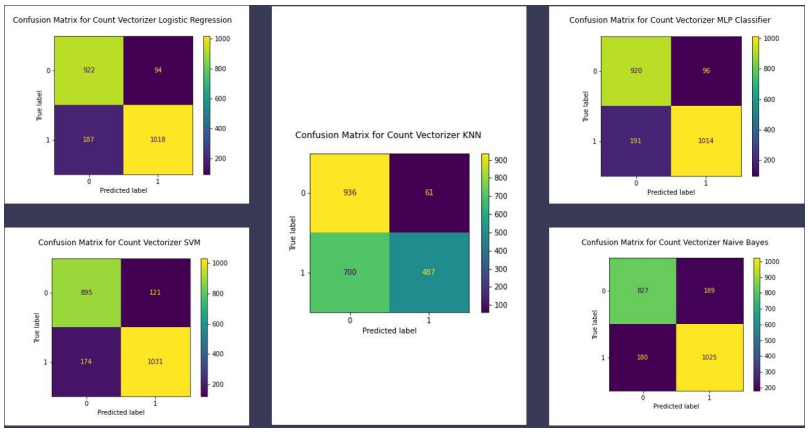

Fig. 11. Confusion matrix for all models using Count Vectorization.

5 EXPERIMENTS

We perform various learning experiments to assess the viability of our trained models. In particular, the experiments are conducted with the aim to show

- its robustness and efficacy in identifying misinformation related to a range of diseases,
- Since Medhub has more records related to COVID-19 and all our models have a high recall rate for the Fake class (one of the desirable features of the Fake news detection system), our models are well optimized to identify any misinformation specific to COVID and,
- How Disease Myth buster can be extended with Transfer learning for general fake news detection.

For the above-mentioned first two experiments, we create two new manually curated labeled datasets. We use the DataComPy python package to demonstrate that there is no overlap between MedHub and the new test datasets. We compare the datasets using this library to obtain a statistical report depicting the data frames' similarities and dissimilarities.

5.1 Testing on Diabetes Dataset

We experiment to show that all our models perform well in detecting fake news related to any disease not covered in the dataset. We form a new diabetes test dataset of 162 records for this experiment. All the records for the test dataset are manually curated from the web. Diabetes is not covered in MedHub. We test all our models with the new test dataset. The best classification test accuracy is achieved with our BERT-based Disease Myth Buster model. It reaches an accuracy of 70%. All other models achieve accuracy greater than 60% except for the KNN count vectorizer. The low performance of Count Vectorizer KNN could be attributed to the lower train and test accuracy of the model. Table 8 shows the evaluation metrics for all models tested on the test dataset. This

experiment shows that the models trained on our dataset can be used to classify misinformation related to any new disease.

Table 8
Classification results of all models on the Diabetes Test dataset

TF-IDF Vectorizer					
Model Name	**Naïve Bayes**	**SVM**	**Logistic Regression**	**MLP Classifier**	**KNN**
Accuracy	64	67	67	70	62
Count Vectorizer					
Model Name	**Naïve Bayes**	**SVM**	**Logistic Regression**	**MLP Classifier**	**KNN**
Accuracy	61	65	65	63	47
BERT Based Model					
Model Name	**Disease Myth Buster**				
Accuracy	70				

5.2 Testing on COVID-19 Specific Misinformation

In this experiment, we demonstrate that all our models perform excellently well in identifying COVID-19 specific myths as our dataset, MedHub has more COVID-19 related data than other diseases. In the effort of this experiment, we prepare a new hand-labeled evaluation dataset for the testing. All the records in the dataset are only COVID-19 related rumors, myths, or misinformation. All the records are extracted from the COVID-19 misinformation section in https://www.poynter.org/. This section is a repository of different classes of articles ranging from False, misleading, and not true to no evidence. For each of the claims, support evidence as articles is provided. We crawl the website and collect all titles and labels for the articles. Our test dataset has 13,459 records. Table 1 provides further insights into the dataset's attributes.

We evaluate all our models on the test dataset. Table 9 shows the performance result of all models. All the models except Naive Bayes achieve more than 75% accuracy. The recall for the "Fake" class for the Naive Bayes TF-IDF and Count vectorizer model is low, and hence the models achieve an accuracy of 63.2 and 71.9 in this experiment. The

lowest-performing Count vectorizer KNN model performs well in this experiment achieving an accuracy of 88% as its recall rate for the ”fake” class is high. The best accuracy is obtained by Disease Myth Buster, which can identify 88% of COVID-19 specific myths.

Table 9
Classification results of all models on the Poynter Test dataset

TF-IDF Vectorizer					
Model Name	**Naïve Bayes**	**SVM**	**Logistic Regression**	**MLP Classifier**	**KNN**
Accuracy	63.2	76.3	76.27	75.16	82.9
Count Vectorizer					
Model Name	**Naïve Bayes**	**SVM**	**Logistic Regression**	**MLP Classifier**	**KNN**
Accuracy	71.9	79.2	83.69	84.8	88.1
BERT Based Model					
Model Name	**Disease Myth Buster**				
Accuracy	88				

5.3 Transfer Learning

One of the significant bottlenecks in NLP domain applications like fake news detection is the lack of adequately labeled available datasets. Recently, this problem has been addressed by the concept of Transfer learning. Transfer learning is a technique in which the previously learned skills and knowledge is applied to a novel task. The authors in [41] leverage the transfer learning approach to address the problem of COVID-19 misinformation using the datasets from the pre-COVID-19 era. The study makes three datasets – Dataset 1 has three publicly available datasets, including the political dataset, Kaggle dataset, and fakeoreal dataset, Dataset 2 has the articles from the LIAR dataset, and Dataset 3 is a combination of datasets 1 and 2. Models are trained on the three datasets and tested on Coronavirus-related information. The models trained on non-Covid data worked well with the COVID data.

We experiment with transfer learning with two standard publicly available datasets – ISOT [42], and LIAR [43]. The ISOT dataset is a collection of real and fake news articles, primarily specific to the political and world news domain. The real articles are fetched from Reuter.com, and the fake articles are crawled from Politifact and Wikipedia. There are, in total, 44,919 records in the ISOT dataset. Each record has a title, text, subject, and label that tell whether it is real or fake. We only take the title and the label column for our testing experiment. Training the dataset with the BERT model has considerable time complexity, so we experiment with the sampled 10000 records of the ISOT dataset. Out of the 10000 records, 5248 are fake, and 4752 are real records. The LIAR dataset is a collection of human-labeled short sentences. The total size of the dataset is 12836 records, with around 10000 records in the training set and the remaining equally divided between the validation and the test set. For our experiment, we take 11,552 records, with 6499 real and 5053 fake records from the LIAR dataset.

This experiment demonstrates how our best-performing model can be extended for general fake news classification. We leverage Disease Myth Buster as the base model and build another simple neural network over it. We fine-tune our existing model and optimize it for general fake news classification. All the layers except the last one of the Disease Myth Buster are loaded. The output from this model is fed into another simple neural network of two dense layers. Then, the new model is trained using both the ISOT and the LIAR dataset. Both the models achieve high-performance accuracy compared to the existing state-of-the-art models. The model trained with the ISOT dataset reaches a training accuracy of 99% in five epochs. The model trained with the LIAR dataset attains a training accuracy of 96.48% in ten epochs. Both the models gain a test accuracy of 99%. Fig. 12 shows the model summary of the new model.

In this experiment, we also evaluate different models like the training and testing of ISOT and LIAR using BERT Tokenizer. Through the results, we prove that using Disease

Myth Buster as the base model with a simple neural network can outperform a simple neural network trained with BERT embeddings.

Layer (type)	Output Shape	Param #	Connected to
input_ids (InputLayer)	[(None, 100)]	0	[]
attention_mask (InputLayer)	[(None, 100)]	0	[]
bert (TFBertMainLayer)	TFBaseModelOutputWithPoolingAndCrossAttentions(last_hidden_state=(None, 100, 768), pooler_output=(None, 768), past_key_values=None, hidden_states=None, attentions=None, cross_attentions=None)	108310272	['input_ids[0][0]', 'attention_mask[0][0]']
intermediate_layer_1 (Dense)	(None, 1024)	787456	['bert[0][1]']
intermediate_layer_2 (Dense)	(None, 512)	524800	['intermediate_layer_1[0][0]']
intermediate_layer_3 (Dense)	(None, 216)	110808	['intermediate_layer_2[0][0]']
new_op_layer1 (Dense)	(None, 32)	6944	['intermediate_layer_3[0][0]']
new_op_layer2 (Dense)	(None, 2)	66	['new_op_layer1[0][0]']

Total params: 109,740,346
Trainable params: 109,740,346
Non-trainable params: 0

Fig. 12. Model Summary using Disease Myth Buster as the base model.

6 LIMITATIONS AND FUTURE WORK

In this section, we discuss some open issues and future research directions. Our research does have a limitation as the design decision of the proof-of-concept web application is significantly influenced by prioritizing cost. The web application is currently hosted on the free tier of Amazon EC2 instance and hence is not suited for large-scale usage. The free tier provides limited storage space and computational capability. Additionally, the hosting service will only allow the server to be up and running for only 750 hours per month for one year. In the future, the web application can be set up on a more scalable hosting server, where deep models like Disease Myth Buster can also be deployed. Increasing the scalability of the web application will also allow seamless continuous updates of Medhub through feedback gathering. Currently, our application takes the complete text entered by the user and gives the recommendation of whether it is credible or not. If the user enters a long paragraph, the models work on deciding the overall credibility of the article. The decision made by the model can be made more informative by labeling which part of the text is fake and which is real. The functionality provided by the web application can also be integrated into the browser making the real-time fact-checking tool more accessible.

Additionally, the feature that allows users to give feedback could be vulnerable to a coordinated attack that deceptively adds biased information to MedHub. The web application can be enhanced in the future to incorporate user authentication and requires users to be logged in to give feedback. It will also prevent the user from entering and providing feedback about the information more than once.

One of the interesting future research project works that could be explored with the web application is the possibility of displaying the crowd-sourced labeling of the data. As the application is used more by the public and more feedback is gathered, the web application can be enhanced to provide a multi-channel credibility check for the text. A

similar idea is implemented by Kolluri and Murthy [44]. Their seminal work discusses the development of the web application, "CoVerifi," which aims to detect whether the information is robot-generated or human-generated and display the veracity by using the credibility classifier trained on human feedback.

Our dataset, MedHub, comprises real and fake information about the top disease like Cancer, Ebola, Zika, Covid-19, polio, etc. From a dataset perspective, our future goal is to make MedHub more comprehensive and large-scale, covering more general health-related information. Currently, Medhub contains short sentences and titles as opposed to long articles. Along this line, one exciting experiment is to include essays and long paragraphs in MedHub and see how the model performance varies with this change.

The best-performing model, Disease Myth Buster, is trained using the BERT base case model. The choice of the BERT model is influenced by the training time and the hardware, like the GPU availability for faster processing. Future experiments can involve focusing on other versions of BERT like BERT-large.

7 CONTRIBUTIONS

In this section, we highlight the key contributions of our study to address the problem of fake news in healthcare. In particular, our work focuses on identifying fake news related to the top diseases humankind has faced. In this effort, we create a novel hand-labeled disease-specific dataset, MedHub. MedHub covers fake and real news related to top diseases like COVID-19, Ebola, Zika, SARS, Cancer, HIV/AIDS, and Polio. It is a collection of 11001 records collected from mainstream and online social media. We include records from two publicly available datasets on COVID-19 misinformation. For the other diseases, the facts and the myths are manually curated and collected from a wide variety of sources. We open-source MedHub to the research community with the belief that it will serve as a good starting point for future research in health fake news detection.

Our research also highlights five different machine learning models trained on MedHub: KNN, SVM, Logistic Regression, MLP Classifier, and Naive Bayes. We train the models using two feature extraction techniques, TF-IDF and Count Vectorizer. Our work also features our best performing model, Disease Myth Buster, achieving an accuracy of 99%. It is a deep model based on BERT. This model outperforms other models with the powerful ability to capture the context and semantic dependencies in a sentence. We deploy the models on a proof-of-concept web application that demonstrates the real-time use of the trained models for fake news detection. The web application encompasses additional features like displaying relevant google search results for the text, allowing the users to give feedback on the recommendation made by the models.

Other primary highlights of our work are the learning experiments that showcase the efficacy of our models. The first experiment demonstrates that our models effectively identify fake news related to any disease not covered in MedHub. The second experiment illustrates that our models perform exceptionally well in identifying any misinformation specific to COVID-19 since all our models achieve a high recall rate for the Fake class.

For these experiments, we create two new manually curated test datasets. The first dataset has 162 fake and real records on Diabetes and the second dataset has 13459 records of COVID-19 misinformation. The third class of learning experiments focuses on displaying how Disease Myth Buster can be extended for general fake news classification using transfer learning. For this experiment, we use the two standard fake news datasets, ISOT, and the LIAR dataset. We use the Disease Myth Buster as the base model and form another simple neural network over it. The experiment results reveal that using our model as the base model instead of using a simple neural network or just the BERT embedding improves the model's performance substantially.

8 CONCLUSIONS

Easy internet accessibility and technological advancement have led to the easy proliferation of fake news. People consume and forward information without verifying its truthfulness. The problem of fake news in the medical domain is more severe and worsens if related to any disease. For example, the COVID-19 pandemic has witnessed an explosion of fake news and misinformation, posing a massive threat to global health and the community. Verifying the veracity of the information can help minimize the damage. Past literature is rich in terms of work for fake news detection but is limited to political and mostly COVID-19 specific misinformation. In this effort in our research, we aim to tackle the problem of fake news related to diseases like COVID-19, Ebola, Zika, SARS, Cancer, HIV/AIDS, and polio.

To address the problem of the lack of adequately labeled datasets for medical fake news detection, we present a new comprehensive dataset, MedHub, with the latest facts and myths about the diseases. We train five machine learning models: KNN, Logistic regression, SVM, Naïve Bayes, and MLP classifier. All the models achieve more than 85% accuracy. We deploy the models on a proof-of-concept web application for real-time fake news detection. The web application encompasses additional features like displaying the relevant results for the text, giving feedback about the recommendation made by the model, upgrade MedHub as the platform is used. We also present our best-performing model with an accuracy of 99%, Disease Myth Buster. It is a deep model based on BERT.

We perform learning experiments to exhibit that our models are optimized for identifying misinformation on any disease not covered in the dataset and particularly perform excellently on detecting COVID-19 specific myths. For these experiments, we prepare two manually curated datasets. Furthermore, we demonstrate how Disease Myth Buster can be extended for general fake news classification through transfer learning through one of the experiments.

We open-source all our datasets and models to enable future research. We intend that the dataset we prepared for the study provides other researchers with a good starting point for general and health-specific fake news detection.

Literature Cited

[1] D. M. J. Lazer, M. A. Baum, Y. Benkler, A. J. Berinsky, K. M. Greenhill, F. Menczer, M. J. Metzger, B. Nyhan, G. Pennycook, D. Rothschild, M. Schudson, S. A. Sloman, C. R. Sunstein, E. A. Thorson, D. J. Watts, and J. L. Zittrain, "The science of fake news," *Science*, vol. 359, no. 6380, pp. 1094–1096, 2018.

[2] H. Allcott and M. Gentzkow, "Social media and fake news in the 2016 election," *Journal of Economic Perspectives*, vol. 31, no. 2, pp. 211–36, 2017.

[3] M. M. Waldrop, "The genuine problem of fake news," *Proceedings of the National Academy of Sciences*, vol. 114, no. 48, pp. 12631–12634, 2017.

[4] A. Zubiaga, A. Aker, K. Bontcheva, M. Liakata, and R. Procter, "Detection and resolution of rumours in social media: A survey," *ACM Comput. Surv.*, vol. 51, no. 2, p. 36, 2018.

[5] H. Warraich, "Dr. google is a liar," *The New York Times*, p. 19, Dec 2018.

[6] H. O.-Y. Li, A. Bailey, D. Huynh, and J. Chan, "Youtube as a source of information on covid-19: a pandemic of misinformation?," *BMJ Global Health*, vol. 5, no. 5, 2020.

[7] S. Loomba, A. de Figueiredo, S. J. Piatek, K. de Graaf, and H. J. Larson, "Measuring the impact of covid-19 vaccine misinformation on vaccination intent in the uk and usa," *Nature human behaviour*, vol. 5, no. 3, pp. 337–348, 2021.

[8] R. F. Sear, N. Velásquez, R. Leahy, N. J. Restrepo, S. E. Oud, N. Gabriel, Y. Lupu, and N. F. Johnson, "Quantifying covid-19 content in the online health opinion war using machine learning," *IEEE Access*, vol. 8, pp. 91886–91893, 2020.

[9] D. Valiyaveettil, M. Malik, D. Joseph, and S. Ahmed, "Myths and misconceptions about cancer among patients attending a tertiary care center in a developing country: A cause for concern," *Annals of Oncology*, vol. 28, p. 147, 2017.

[10] K. Wella, S. Webber, and P. Levy, "Myths about hiv and aids among serodiscordant couples in malawi," *Aslib Journal of Information Management*, vol. 69, no. 3, pp. 278–293, 2017.

[11] A. Feuer, "The ebola conspiracy theories," *The New York Times*, p. 5, Oct 2014.

[12] J. P. Baptista and A. Gradim, "Understanding fake news consumption: A review," *Social Sciences*, vol. 9, no. 10, p. 185, 2020.

[13] B. Horne and S. Adali, "This just in: Fake news packs a lot in title, uses simpler, repetitive content in text body, more similar to satire than real news," in *Proceedings of the international AAAI conference on web and social media*, vol. 11, pp. 759–766, 2017.

[14] A. Shrestha and F. Spezzano, "Textual characteristics of news title and body to detect fake news: a reproducibility study," in *European Conference on Information Retrieval*, pp. 120–133, Springer, 2021.

[15] A. Jain, A. Shakya, H. Khatter, and A. K. Gupta, "A smart system for fake news detection using machine learning," in *2019 International Conference on Issues and Challenges in Intelligent Computing Techniques (ICICT)*, vol. 1, pp. 1–4, IEEE, 2019.

[16] T. Jiang, J. P. Li, A. U. Haq, and A. Saboor, "Fake news detection using deep recurrent neural networks," in *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pp. 205–208, IEEE, 2020.

[17] S. Helmstetter and H. Paulheim, "Weakly supervised learning for fake news detection on twitter," in *2018 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining (ASONAM)*, pp. 274–277, IEEE, 2018.

[18] E. Ferrara, "What types of covid-19 conspiracies are populated by twitter bots?," *First Monday*, vol. 25, no. 6, 2020.

[19] E. Ferrara, "Disinformation and social bot operations in the run up to the 2017 french presidential election," *CoRR*, 2017. doi: http://arxiv.org/abs/1707.00086.

[20] M. Umer, Z. Imtiaz, S. Ullah, A. Mehmood, G. S. Choi, and B.-W. On, "Fake news stance detection using deep learning architecture (cnn-lstm)," *IEEE Access*, vol. 8, pp. 156695–156706, 2020.

[21] S. Guarino, F. Pierri, M. Di Giovanni, and A. Celestini, "Information disorders during the covid-19 infodemic: The case of italian facebook," *Online Social Networks and Media*, vol. 22, p. 100124, 2021.

[22] F. Jouyandeh, S. Sadeghi, B. Rahmatikargar, and P. M. Zadeh, "Fake news and covid-19 vaccination: a comparative study," in *Proceedings of the 2021 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, pp. 525–531, 2021.

[23] J. S. Brennen, F. M. Simon, P. N. Howard, and R. K. Nielsen, *Types, sources, and claims of COVID-19 misinformation*. Ph.D. dissertation, Dept. Social Sci.,University of Oxford, Oxford, UK, 2020.

[24] S. Shi, A. R. Brant, A. Sabolch, and E. Pollom, "False news of a cannabis cancer cure," *Cureus*, vol. 11, no. 1, 2019.

[25] A. Goel, C. Shivaprasad, A. Kolly, A. Pulikkal, R. Boppana, and C. Dwarakanath, "Frequent occurrence of faulty practices, misconceptions and lack of knowledge among hypothyroid patients," *Journal of clinical and diagnostic research: JCDR*, vol. 11, no. 7, p. OC15, 2017.

[26] P. Payoungkhamdee, P. Porkaew, A. Sinthunyathum, P. Songphum, W. Kawidam, W. Loha-Udom, P. Boonkwan, and V. Sutantayawalee, "Limesoda: Dataset for fake news detection in healthcare domain," in *2021 16th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pp. 1–6, IEEE, 2021.

[27] R. A. Ciora and A. L. Cioca, "Fake news management in healthcare," in *2021 International Conference on e-Health and Bioengineering (EHB)*, pp. 1–4, IEEE, 2021.

[28] P. Patwa, S. Sharma, S. Pykl, V. Guptha, G. Kumari, M. S. Akhtar, A. Ekbal, A. Das, and T. Chakraborty, "Fighting an infodemic: Covid-19 fake news dataset," in *International Workshop on Combating Online Hostile Posts in Regional Languages during Emergency Situation*, pp. 21–29, Springer, 2021.

[29] T. Hossain, *COVIDLIES: Detecting COVID-19 Misinformation on Social Media*. Ph.D. dissertation, Dept. Comp. Sci., Univ. of California, Irvine, 2021.

[30] M. Cheng, S. Wang, X. Yan, T. Yang, W. Wang, Z. Huang, X. Xiao, S. Nazarian, and P. Bogdan, "A covid-19 rumor dataset," *Frontiers in Psychology*, vol. 12, p. 644801, May 2021.

[31] L. Cui and D. Lee, "Coaid: Covid-19 healthcare misinformation dataset," *CoRR*, 2020. doi: https://arxiv.org/abs/2006.00885.

[32] G. Pennycook, J. McPhetres, Y. Zhang, J. G. Lu, and D. G. Rand, "Fighting covid-19 misinformation on social media: Experimental evidence for a scalable accuracy-nudge intervention," *Psychological science*, vol. 31, no. 7, pp. 770–780, 2020.

[33] A. D'Ulizia, M. C. Caschera, F. Ferri, and P. Grifoni, "Fake news detection: a survey of evaluation datasets," *PeerJ Computer Science*, vol. 7, p. 518, 2021.

[34] A. Rusli, J. C. Young, and N. M. S. Iswari, "Identifying fake news in indonesian via supervised binary text classification," in *2020 IEEE International Conference on Industry 4.0, Artificial Intelligence, and Communications Technology (IAICT)*, pp. 86–90, 2020.

[35] T. Mladenova and I. Valova, "Analysis of the knn classifier distance metrics for bulgarian fake news detection," in *2021 3rd International Congress on Human-Computer Interaction, Optimization and Robotic Applications (HORA)*, pp. 1–4, IEEE, 2021.

[36] D. S. Abdelminaam, F. H. Ismail, M. Taha, A. Taha, E. H. Houssein, and A. Nabil, "Coaid-deep: an optimized intelligent framework for automated detecting covid-19 misleading information on twitter," *IEEE Access*, vol. 9, pp. 27840–27867, 2021.

[37] M. Granik and V. Mesyura, "Fake news detection using naive bayes classifier," in *2017 IEEE first Ukraine conference on electrical and computer engineering (UKRCON)*, pp. 900–903, IEEE, 2017.

[38] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," *CoRR*, 2018. doi: http://arxiv.org/abs/1810.04805.

[39] J. Ding, Y. Hu, and H. Chang, "Bert-based mental model, a better fake news detector," in *Proceedings of the 2020 6th international conference on computing and artificial intelligence*, pp. 396–400, 2020.

[40] M. Choudhary, S. S. Chouhan, E. S. Pilli, and S. K. Vipparthi, "Berconvonet: A deep learning framework for fake news classification," *Applied Soft Computing*, vol. 110, p. 107614, 2021.

[41] S. Bojjireddy, S. A. Chun, and J. Geller, "Machine learning approach to detect fake news, misinformation in covid-19 pandemic," in *DG. O2021: The 22nd Annual International Conference on Digital Government Research*, pp. 575–578, 2021.

[42] H. Ahmed, I. Traore, and S. Saad, "Detection of online fake news using n-gram analysis and machine learning techniques," in *International conference on intelligent, secure, and dependable systems in distributed and cloud environments*, vol. 10618, pp. 127–138, Springer, 2017.

[43] W. Y. Wang, "liar, liar pants on fire: A new benchmark dataset for fake news detection," *CoRR*, 2017. doi: http://arxiv.org/abs/1705.00648.

[44] N. L. Kolluri and D. Murthy, "Coverifi: A covid-19 news verification system," *Online Social Networks and Media*, vol. 22, p. 100123, 2021.