ENHANCING HEALTH-RELATED TWEET CLASSIFICATION: AN EVALUATION OF TRANSFORMER-BASED MODELS FOR COMPREHENSIVE ANALYSIS

by

FORAM PANKAJBHAI PATEL

THESIS

Submitted in partial fulfillment of the requirements

for the degree of Master of Science in Computer Science at

The University of Texas at Arlington

May 2023

Arlington, Texas

Supervising Committee:

Dr. Chengkai Li, Supervising Professor

Dr. Shirin Nilizadeh

Dr. Negin Fraidouni

ACKNOWLEDGEMENTS

I would like to express my sincere appreciation to my supervising professor, Dr. Chengkai Li, for providing me with invaluable support and guidance throughout the course of my research work. Dr. Li's insights and expertise have been instrumental in shaping my research and expanding my knowledge in the field. I would also like to extend my gratitude to my committee members, Dr. Shirin Nilizadeh and Dr. Negin Fraidouni, for their time, effort, and valuable feedback that has helped me to improve and refine my work.

I am grateful for the technical guidance provided by Zhengyuan, who has been an excellent mentor, always available to answer my questions and provide valuable insights. Furthermore, I would like to express my sincere gratitude to the members of the IDIR lab, who have been a great source of inspiration, encouragement, and friendship throughout my research journey. I am truly thankful to Xiao for being an amazing friend who made my time in the IDIR lab and my MS journey so much fun and fulfilling. I would like to extend my heartfelt appreciation to Nasim for always being there for me and offering help whenever I needed it. Additionally, I am grateful to Haiqi and Zeyu for their friendship and for being a part of my research journey. You have all made a significant impact on my life, and I feel blessed to have had the opportunity to work alongside such a talented and supportive group of individuals. Thank you for inspiring me, challenging me, and teaching me so much. I am grateful for the memories and experiences that we shared, and I look forward to staying in touch in the future.

Finally, I would like to thank my family, my partner and my best friends for their unwavering support, love, and encouragement, which have been crucial in helping me reach this milestone. Their steadfast belief in me has been a constant source of motivation and inspiration, and I am immensely grateful for their presence in my life.

April 24, 2023

ABSTRACT

ENHANCING HEALTH-RELATED TWEET CLASSIFICATION:
AN EVALUATION OF TRANSFORMER-BASED MODELS FOR
COMPREHENSIVE ANALYSIS

FORAM PANKAJBHAI PATEL, M.S.
The University of Texas at Arlington, 2023

Supervising Professor: Dr. Chengkai Li

The task of health-related tweet classification entails identifying whether a given tweet is health-related or not. While existing research in this area has made significant progress in classifying tweets into specific sub-domains of health, such as mental health, COVID-19, or specific diseases, there is a need for a more comprehensive approach that considers a broader range of health-related topics. This thesis addresses this need by proposing a diverse and comprehensive dataset that includes various existing health-related datasets, data collected through a keyword-based approach, and manually annotated data. However, the use of health-related keywords in a figurative or non-health context poses a significant challenge to the classification task. To overcome this challenge, the thesis explores the use of Transformer-based models, such as BERT, BERTweet, RoBERTa, and DistilBERT, which have the ability to understand the contextual meaning of words. The study experiments with these models to assess their effectiveness in classifying health-related tweets. Based on the findings

of the thesis study, Transformer-based models, including BERT, DistilBERT, and RoBERTa, had lower F1-scores of 0.882, 0.870, and 0.872, respectively when evaluated on test data. The highest F1-score of 0.900 was achieved by adding the BiLSTM layer to the BERTweet model, which was then fine-tuned on our proposed dataset and RHMD (Reddit Dataset). Additionally, an ablation analysis was conducted to highlight the significance of the BiLSTM layer and the RHMD dataset in enhancing the BERTweet model's performance for health-related tweet classification.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

LIST OF TABLES

CHAPTER 1

INTRODUCTION

1.1 Background

The increasing amount of information posted on social media platforms over the past few years has brought about a new era of healthcare information sharing. Twitter, in particular, has become a go-to platform for individuals seeking and sharing information about health and wellness [1]. Public posts on Twitter can include a diverse range of content, including updates about oneself, opinions, discussions, marketing, and political campaigning. Health-related tweets offer a wide-ranging look at health topics, covering insights from professionals, updates on research, public health news, personal stories, and experiences of dealing with medical conditions [2, 3].

Moreover, Twitter has also become a popular forum for patient advocacy and support groups. Individuals dealing with rare or chronic conditions have the opportunity through social media to connect and empathize with others navigating similar health journeys, fostering a supportive and understanding community [4]. These groups can also serve as a valuable source of information about treatments, clinical trials, and healthcare providers.

1.2 Use Cases

Analyzing large volumes of Twitter data in real-time and providing timely and relevant information to healthcare professionals and policymakers holds significant importance in healthcare domain [5]. By using health-related tweets, healthcare professionals can track the spread of diseases, identify areas where health interventions

are needed, and monitor the effectiveness of interventions in real-time. For example, during the COVID-19 pandemic, Twitter data was extensively used to track the spread of the virus, monitor public sentiment towards vaccination, and identify areas of misinformation.

In addition to monitoring public health trends, healthcare professionals with the help of patient-generated data on Twitter can even gain insights into patient needs and concerns, identify potential treatment options, and provide personalized care. Overall, the use of health-related content on Twitter has the potential to revolutionize public health research, surveillance, and patient care.

1.3 Problem Statement

Platforms like Facebook and Twitter have become popular social networking sites for individuals to share personal health information, making them a significant source of health-related data. Twitter, in particular, allows researchers to access tweets via its API, which has been utilized in various healthcare studies [6]. As a result of the growing popularity of social media platforms such as Twitter, there has been a marked increase in health-related posts. In order to glean valuable healthcare insights, it is crucial to distinguish health-related tweets from the vast volume of tweets [7]. As a result, we define a binary classification task to classify a tweet as either health-related or non-health related.

1.3.1 Criteria for an effective health-related Tweet classification task

To develop an effective health-related tweet classification task, we need to consider several criteria. Firstly, we need a labelled dataset of health-related tweets that can be used to train and evaluate our classification model. Secondly, the dataset needs to be comprehensive and balanced, meaning that it should include tweets re-

lated to various health topics and cover a range of sentiments and opinions. This will help ensure that our model is robust and can generalize well to new and unseen data. Another important criterion is handling figurative or sarcastic language. Tweets can often contain figurative or sarcastic language, which can be challenging to classify accurately. Therefore, we need to use certain approaches to handle this type of language, such as learning contextual information. By considering these criteria, we can develop a robust and effective health-related tweet classification task that can aid in improving healthcare delivery and patient outcomes.

1.3.2 Challenges in health-related tweet classification

- Lack of comprehensive health dataset: The lack of a comprehensive health dataset is a significant challenge in health-related tweet classification, as it can make it difficult to develop a generalized approach to classifying health-related tweets.

Study	**Health Topics**
Xiong et al. [8]	Medication Intake
Lin et al. [9]	Self-reported Chronic Stress
Porvatov et al. [10]	COVID-19
Kayastha et al. [11]	Adverse Drug Effect (ADE) mentions
Ahne et al. [12]	Diabetes
Medford et al. [13]	COVID-19
Stens et al. [14]	Lupus & Reproductive Health
Arora et al. [15]	Depression, Anxiety & Mental Illness
Yuan et al. [16]	Vaccination
Oscar et al. [17]	Alzheimer's Disease

Table 1.1. Existing Studies

As shown in Table 1.1, many existing studies have focused on specific health topics, such as mental health or diabetes, which may not be representative of the broader health domain. To develop a more generalized approach, we need a large, high-quality dataset that covers a broad range of health-related topics and sentiments. This would enable us to train models that can accurately classify tweets related to any health topic, rather than just a specific subset of topics. Moreover, by creating a more comprehensive dataset, we can also address issues related to bias and ensure that our models are representative of diverse populations and perspectives.

- Non-literal usage of health-related keywords: Another challenge in health-related tweet classification is when health-related keywords are used in a non-literal manner. For example, the tweet *"The Man U fans are about to have a heart attack over this loooool https://t.co/qx6jEErKAt"* uses the health keyword "heart attack" to express the distress / disappointment. This way, health-related keywords can be used in non-health contexts, which can create ambiguity for a classification model that relies on literal usage of these keywords. This can lead to inaccurate classification of tweets as health-related or non-health related. To overcome this challenge, researchers have developed various techniques such as machine learning algorithms that can learn the context of the tweet and identify non-literal usage of health-related keywords. Additionally, using contextual embeddings like BERT[18] or RoBERTa[19] that capture the meaning of the text in context, rather than just looking at individual words, can also help in identifying non-literal usage of health-related keywords in tweets. By identifying and handling non-literal usage of health-related keywords, we can improve the accuracy of health-related tweet classification and ensure that the model is better able to distinguish between health-related and non-health-related tweets.

1.3.3 Research Objectives

The major research objectives include:

- Many existing studies concentrate on tweet datasets that primarily revolve around specific health subdomains such as Parkinson's disease, Alzheimer's, mental health, or COVID-19. Research on classifying tweets as health-related or non-health-related is comparatively limited, calling for a more expansive perspective. There is a need for a comprehensive dataset covering a broad range of health-related topics to enable the development of a health-related tweet classification model. This thesis aims to fill this gap by creating a comprehensive dataset by using different approaches which are discussed later in this thesis.
- Many tweets on Twitter containing health-related keywords may not actually pertain to health. In some cases, a tweet might include a health-related keyword in a figurative or non-health context, posing a significant challenge for tweet classification. For instance, consider the following tweets: *'I got Minecraft fever'* or *'financial aid gives me the biggest headache I give up'* [20]. Although these tweets contain health-related keywords *'fever'* and *'headache'*, the keywords are used figuratively, making the tweets non-health-related. Hence, it is critical to study various approaches that can accurately learn the contextual meaning of a given sentence. Transformers models have shown great success in learning word representations through contextual information and improving classification performance. The use of these models has become increasingly popular in natural language processing tasks, and these models have achieved state-of-the-art performance on many benchmark datasets. The thesis explores the use of transformer-based pre-trained models, including BERT[18], RoBERTa[19], DistilBERT[21], and BERTweet[22], for fine-tuning on task-specific datasets for downstream tasks such as classification.

1.4 Thesis Structure

This thesis consists of four main sections, which are outlined below:

Chapter 2: Data Collection and Preprocessing This chapter presents the process involved in data gathering, collection, and preprocessing for the formation of a new dataset.

Chapter 3: Methodology This chapter provides a detailed description of the experiment setup and methodology used in this study.

Chapter 4: Results In this chapter, the focus is on detailed experiment results obtained followed by an ablation analysis.

Chapter 5: Conclusion This chapter provides a brief summary of the entire thesis work and offers insights into possible future work.

CHAPTER 2

DATA COLLECTION AND PREPROCESSING

2.1 Data Collection

The collection of comprehensive data for health-related research can be a challenging task due to the vastness and complexity of the domain. To address this issue, we adopted a multi-faceted approach by incorporating existing datasets, gathering tweets through Twitter's Streaming API using a keyword-based approach, as well as manually annotating a subset of tweets to enhance our dataset. The objective of our study was to ensure comprehensive coverage across all health-related topics. The four main sources of data that we utilized in creating our dataset are detailed below:

1. Existing Health-related Datasets (Existing Health Data):

   We acquired existing datasets from multiple tasks organized by Social Media Mining for Health (SMM4H) in 2017, 2018, and 2022. Basically, The Social Media Mining for Health Applications (SMM4H) is a workshop that unites researchers focusing on automated methods of collecting, extracting, representing, analyzing, and validating health informatics data derived from social media platforms such as Twitter and Facebook. The details of the existing datasets we used in our study are provided below:

   - Adverse Drug Reaction (ADR) Mentions - SMM4H 2017, 2018: The dataset was developed with the aim of distinguishing Twitter posts that mention adverse drug reactions from those that do not. Both classes of tweets belong to the health category as mentioned in the dataset description and, as

such, were considered health-related for our classification task. The Table 2.1 shows the statistics of this dataset from year 2017 and 2018.

Class / Year	**2017**	**2018**
Contain ADR mention (1)	768	1115
Do not contain ADR mention (0)	7498	12,827
Total	**8266**	**13,944**

Table 2.1. Statistics of ADR dataset

- Medication Intake - SMM4H 2017, 2018: This dataset was designed to differentiate between tweets that describe personal medication intake, tweets that suggest possible medication intake, and tweets that mention medication names but do not indicate personal intake. For our classification task, all three classes were considered health-related and the statistics for the dataset is displayed in Table 2.2.

Class / Year	**2017**	**2018**
personal medication intake (1)	1383	2563
possible medication intake (2)	2253	4071
non-intake (3)	3542	5727
Total	**7178**	**12,361**

Table 2.2. Statistics of Medication Intake dataset

- Mention of vaccination behavior - SMM4H 2018: This dataset was created to classify tweets that mention behavior related to influenza vaccination versus those that do not. The Table 2.3 presents the statistics of this dataset. All tweets in the dataset are related to vaccination, but the aim

is to distinguish tweets that indicate whether someone has received or intends to receive a flu vaccine. Both classes of tweets were considered health-related for our task.

Class / Year	**2018**
Mention vaccination behavior (1)	1360
Do not mention vaccination behavior (0)	3359
Total	**4719**

Table 2.3. Statistics of Vaccination dataset

- Covid19 Face Masks category - SMM4H 2022: The dataset provided in [23] is intended for stance detection on Twitter in relation to three health mandates related to the COVID-19 pandemic: Face Masks, Stay At Home Orders, and School Closures. After conducting a comprehensive analysis of the dataset, we specifically chose tweets from the Face Masks category for our classification task. During this process, we randomly selected 100 tweets and manually examined them. Our findings revealed that 95% of the tweets were health-related, leading us to focus on tweets predominantly associated with health. However, we acknowledge the presence of some noise in the form of tweets that may not be directly health-related. The Table 2.4 shows the statistics for this dataset.

We were able to collect a total of *53,528* tweets from the existing datasets mentioned earlier, and all of these tweets were categorized as health-related. By incorporating these datasets, we were able to obtain a diverse and comprehensive dataset that could be used to effectively train our classifier and ensure its accuracy and effectiveness.

Class / Year	**2022**
positive stance (FAVOR)	4262
negative stance (AGAINST)	2409
neutral stance (NONE)	389
Total	**7060**

Table 2.4. Statistics of Covid19 Stance detection (Face Masks Category) dataset

2. Keyword-based approach to collect tweets through Twitter's Streaming API (Twitter Streaming_keywords):

   We used a keyword-based approach to collect tweets using Twitter's Streaming API. To begin, we curated a list of 2247 health-related keywords sourced from reputable websites like the CDC, WHO, and WebMD. Basically, we extracted keywords from WebMD's Common Topics page, WHO's Health Topics page, and CDC's Most Searched Diseases & Conditions page, filtering out duplicates to generate a comprehensive keyword list.

   We employed a two-step approach using the Streaming API for collecting tweets. Initially, we collect a random set of tweets. Subsequently, we categorize each tweet based on the presence or absence of health-related keywords from our predefined list. If a tweet contains at least two of our health-related keywords, it is categorized as health-related. This method has facilitated the collection of *5000* health-related tweets. It took approximately 5-6 hours on average to obtain 100 tweets. All these 5000 tweets were collected within the specific time span of December 10, 2022, to January 11, 2023. Conversely, tweets devoid of any of our health-related keywords are identified as non-health related. Using this process, we obtained *60,000* non-health related tweets. Unlike health-related tweets, it was faster to collect non-health related tweets, taking only a

few seconds to acquire 100. All 60,000 non-health related tweets were gathered within a two-day timeframe, specifically on November 28, 2022, and February 2, 2023. To summarize, our process involves collecting a random set of tweets first and then sorting them into health-related or non-health related categories based on keyword analysis. This approach has successfully resulted in a diverse and comprehensive dataset.

3. Manually Annotated Tweets Collected via Twitter Streaming API (Twitter Streaming_annotated):

   To create a more comprehensive dataset for health-related tweet classification, we recognized the importance of incorporating tweets that contain health-related keywords but are used in non-health contexts. Instead of solely focusing on tweets that either contain or lack health-related keywords, we expanded the criteria to include tweets with a single health-related keyword. By including such tweets, we aimed to capture instances where health-related keywords are used in a non-health context. This step was necessary to enhance the dataset's diversity and provide the model with examples that require contextual understanding for accurate classification. As part of this approach, we collected 2000 tweets over a two-day period, specifically on February 16, 2023, and February 17, 2023. Each tweet was manually annotated, carefully considering the contextual information present, in order to assign accurate labels as either health-related or non-health related. During the manual annotation process, we established fixed definitions for health-related and non-health related tweets.

   Health-related: To identify health-related tweets, we first look for keyword related to health. Then, we evaluate the tone of the tweet to ensure that it does not contain sarcasm or over-exaggeration. If the tweet contains health-

related keyword and the tone is informative and trustworthy, we annotate it as health-related.

Non-Health related: To determine whether a tweet is non-health related, we evaluate tweets that contain health-related keyword by examining the tone. If the tone of the tweet is sarcastic or mocking or if the health-related keyword is being used in a figurative sense, we categorize the tweet as non-health related. After the manual annotation process applied to the 2000 tweets collected through our modified criteria, we ended up with *1545* tweets categorized as health-related and *455* categorized as non-health related, enhancing the diversity of our dataset. We will incorporate these tweets into our final dataset for further analysis.

4. Datasets with Figurative Health Mentions:
    - HMC2019: The tweet dataset introduced in [24] consists of three classes: Figurative Mentions, Other (Non-Personal) Health Mentions, and Health (Personal) Mentions. These classes cover 10 different disease keywords, including Alzheimer's disease, cancer, depression, stroke, heart attack, Parkinson's disease, cough, fever, headache, and migraine. The diverse range of diseases covered in the dataset increases its robustness for analysis. For our study, we consider the Figurative Mentions class as non-health related, while the Other (Non-Personal) Health Mentions and Health (Personal) Mentions classes are considered health-related. This approach yields a dataset consisting of *9821* health-related tweets and *3241* non-health related tweets. A benefit of using this dataset is that it helps to address the challenge of identifying the figurative usage of health-related keywords in tweets, which is valuable for our research work.

- RHMD 2022: The paper [25] presents the Reddit Health Mention Dataset (RHMD), a dataset of multi-domain Reddit data for health mention analysis. RHMD comprises 10,015 manually labeled Reddit posts that mention 15 common disease or symptom terms, and are labeled as personal health mentions, non-personal health mentions, or figurative health mentions similar to the HMC 2019 dataset. In our study, the same categorization criteria we applied to the HMC2019 dataset is used for the RHMD, with Figurative Mentions labeled as non-health related and the other two classes identified as health-related. Overall, we obtained *6790* health-related Reddit posts and *3225* non-health related Reddit posts. Although this dataset is not included in our main Twitter dataset, we kept it aside for experimentation to evaluate its impact on classification performance.

Data Sources / Class	**Health-related**	**Non-Health related**
Existing Health Data	53,528	0
Twitter Streaming_keywords	5000	60,000
Twitter Streaming_annotated	1545	455
HMC2019	9821	3241
Our Proposed Dataset (Total)	**69,894**	**63,696**

Table 2.5. Statistics of Proposed Data

Table 2.5 contains the statistical information of the proposed dataset that has been presented for the thesis study.

2.2 Data Preprocessing

To ensure the quality and relevance of our dataset, we performed a series of data preprocessing steps that aimed to eliminate irrelevant information and reduce

noise. These steps involved removing Retweet symbols from tweets, filtering out non-English language tweets using Googletrans library, and keeping only tweets with a length of five or more words. Additionally, we replaced @mentions, #hashtags, and URLs with [mention], [hashtag], and [url] placeholders to avoid bias towards specific words or phrases and facilitate better generalization to new and unseen data. After this preprocessing, we removed any duplicate tweets and applied lowercasing to standardize the textual data.

Data Sources / Class	**Health-related**	**Non-Health related**
Existing Health Data	250	0
Twitter Streaming_keywords	250	500
Twitter Streaming_annotated	250	250
HMC2019	250	250
Test Data	**1000**	**1000**

Table 2.6. Statistics of Test Data

Data Sources / Class	**Health-related**	**Non-Health related**
Existing Health Data	41,877	0
Twitter Streaming_keywords	4217	48,334
Twitter Streaming_annotated	1126	90
HMC2019	9092	2684
Train Data	**56,312**	**51,108**

Table 2.7. Statistics of Train Data

After preprocessing the entire dataset, we created a test set of 2000 tweets from the data sources mentioned earlier. The distribution of tweets from each source is shown in Table 2.6, and we manually verified the labels of this test set to ensure

its accuracy. By manually verifying the labels, we could identify and correct any potential errors or inconsistencies in the dataset to prevent them from affecting the model's performance. To avoid bias towards health-related keyword tweets during the evaluation of the model's performance, we created a balanced test set with an equal number of positive and negative cases from each data source. This helps ensure that the model's performance is measured fairly across all subsets of data, and not just on tweets where the majority of them contain health-related keywords. The remaining part of our dataset was used for training purpose and the Table 2.7 displays the statistics of the Train data.

CHAPTER 3

METHODOLOGY

Recent advancements in Natural Language Processing (NLP) have led to the development of various language models, such as widely popular Transformer-based models, which have proven to be highly effective for working with textual data. One of the most prevalent approaches for text classification tasks is to use such models, which are pre-trained on large amounts of text data and can be fine-tuned for different downstream tasks. In light of this, we utilize the following pre-trained Transformer-based models and fine-tune them for our Health-related Tweet Classification task.

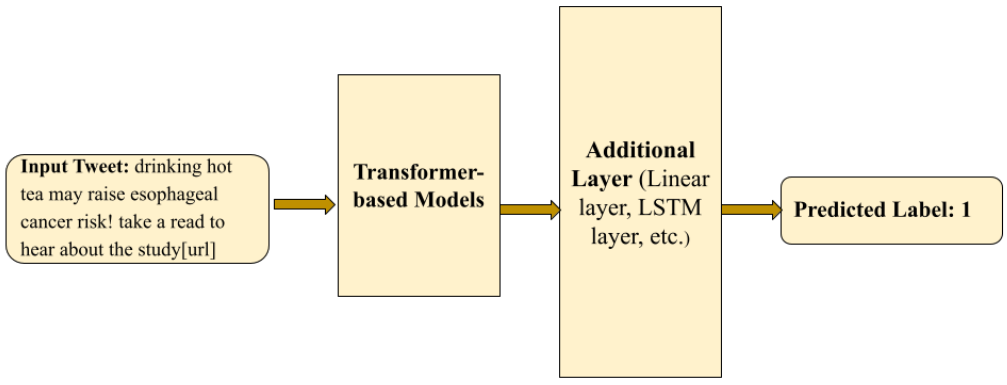

Figure 3.1. Transformer-based Model for Text Classification.

The model structure for the Health-related Tweet Classification task is illustrated in Figure 3.1. The model takes an input tweet and passes it through a Transformer-based model, a type of neural network architecture that is effective for natural language processing tasks. The output from the Transformer-based model is then processed by additional layers of either Linear or LSTM layer, which help to further refine the output. Finally, the model makes a prediction on whether the tweet is related to health or not, resulting in a binary label of 1 or 0.

3.1 Transformer Models

3.1.1 BERT

BERT is designed to learn deep bidirectional representations of natural language text, which means it can understand the context and meaning of words in a sentence both before and after the word [18]. BERT is pre-trained on large amounts of text data using a masked language modeling (MLM) and next sentence prediction (NSP) objectives. The MLM task involves masking certain words in a sentence and training the model to predict the masked words based on the context of the sentence. The NSP task involves training the model to predict whether two sentences are consecutive in the original text or not. The pre-training of BERT allows it to learn a wide range of linguistic features and representations from text data, making it highly effective for a variety of natural language processing tasks, including text classification, named entity recognition, question answering, and more. To use BERT for a specific task, such as text classification, the pre-trained model is fine-tuned on a smaller labeled dataset that is specific to the task.

3.1.2 DistilBERT

DistilBERT is a smaller and faster version of the pre-trained BERT language model [21]. DistilBERT achieves similar performance to the original BERT model while being much faster and requiring less memory, making it well-suited for use in resource-constrained environments. Specifically, DistilBERT has fewer layers, smaller hidden state dimensions, and uses a modified attention mechanism, which results in a model that is 40% smaller and 60% faster than the original BERT model.

3.1.3 RoBERTa

Like BERT, RoBERTa is pre-trained on large amount of text data and longer training epochs than BERT, and it also removes the NSP objective to focus solely on MLM [19]. Additionally, RoBERTa uses dynamic masking during pre-training, which means that different parts of each sequence are randomly masked during each epoch of training. Overall, RoBERTa is a highly effective pre-trained language model that has been shown to outperform BERT on many natural language processing tasks.

3.1.4 BERTweet

BERTweet is a pre-trained language model that is based on the same architecture as BERT, but it is tailored for analyzing language on Twitter [22]. BERTweet uses a pre-training procedure that is similar to RoBERTa, which involves training the model on a large dataset of English tweets. To pre-train BERTweet, the model was trained on a dataset of over 850 million tweets using a masked language modeling (MLM) objective, which involves masking random tokens in the input text and training the model to predict the masked tokens based on the surrounding context. The pre-training process also involved additional tasks such as next sentence prediction

and tweet-level and user-level classification. Overall, BERTweet is a highly effective pre-trained language model specifically designed for analyzing language on Twitter.

3.2 Experimental Setup

The experimental setup involved fine-tuning four transformer-based models, namely BERT, DistilBERT, RoBERTa, and BERTweet, on the proposed dataset for Health-related Tweet Classification. Hyper-parameter tuning was performed with 10 epochs and a learning rate of $1e{-}5, 5e{-}5$, using the AdamW optimizer. Additionally, a modified BERTweet model was presented, which included an additional BiLSTM layer fine-tuned on the proposed dataset and the RHMD dataset [26]. The performance of all models was evaluated on the test dataset, and the F1-score and Accuracy were used as the evaluation metrics. This experimental setup aimed to determine the effectiveness of the different transformer-based models and to evaluate the performance of the modified BERTweet model, which incorporated additional layer and diverse datasets for improved performance in Health-related Tweet Classification.

CHAPTER 4

RESULTS

4.1 Results

The Table 4.1 provides a comparison of the performance of different transformer-based models on our proposed dataset for health-related tweet classification. The F1-scores and accuracies of BERT, DistilBERT, RoBERTa, BERTweet, and BERTweet + BiLSTM (P.D. + RHMD) models are presented. The F1-score measures the harmonic mean of precision and recall, while the accuracy measures the proportion of correct predictions among all predictions.

Model	**F1-Score**	**Accuracy**
BERT	0.882	0.883
DistilBERT	0.870	0.872
RoBERTa	0.872	0.874
BERTweet	0.887	0.887
BERTweet + BiLSTM (P.D. + RHMD)	**0.900**	**0.900**

Table 4.1. Model Performance on Test Data

The results show that the modified BERTweet model with an additional BiLSTM layer fine-tuned on the proposed dataset (P.D.) and RHMD dataset achieved the highest F1-score and accuracy of 0.900. This indicates that the BiLSTM layer helps in improving the performance of the BERTweet model in health-related tweet classification. The model outperformed the other transformer-based models by a significant margin, indicating that the proposed modification is effective in enhancing

the performance of transformer-based models. Among the other models, BERTweet achieved the second-best performance with an F1-score of 0.887 and an accuracy of 0.887. DistilBERT and RoBERTa achieved lower F1-scores ranging from 0.870 to 0.882 and accuracies ranging from 0.872 to 0.883. BERT performed slightly better than the other two models with an F1-score of 0.882 and an accuracy of 0.883.

The attention weights generated by the best performing model: Bertweet + BiLSTM (P.D. + RHMD) during finetuning have been visualized in Figure 4.1 using the Transformers Interpret library [27]. These visualizations illustrate the words that are given more importance by the model while making classification decisions.

Sr. No	Ground Truth	Prediction	Word Importance
1	Health-related	Health-related	no, they need a healthy diet and regular exercise, medication and surgery should be last resort efforts. [url]
2	Non-Health related	Non-Health related	my suggestion for those mental health bunnies...but drawing it was a mistake, because now i need it [url]
3	Non-Health related	Health related	me: just trying to work my cat: oh its elbow biting time

Figure 4.1. Visualization showing the words that are focused by BERTweet + BiLSTM (P.D. + RHMD) for making a prediction.

The visualization of the attention weights for the first tweet shows that the model focuses on words like *no*, *they*, *and*, *exercise*, *medication*, and *resort* which leads to the correct classification of the tweet as health-related. Similarly, for the second tweet, the model gives more importance to words like *suggestion*, *bunnies*, *but*, and *mistake* allowing it to correctly classify the tweet as non-health-related. However, for the third tweet, the model misclassifies the tweet as health-related despite focusing on words like *to*, *work*, *cat*, and *time*. The ground truth of the tweet is actually non-health

related. This suggests that while keyword attention is important for the model's performance, contextual information plays a significant role as well. Overall, these visualizations demonstrate that the attention mechanism in the model is effectively capturing relevant information for the health-related tweet classification task. They also suggest that incorporating more contextual information may help improve the model's accuracy in certain cases.

The evaluation results in Figure 4.2 indicate that the transformer-based models perform fairly similarly on the first two subsets, Existing Health Data and Twitter Streaming_keywords, achieving accuracies above 0.960.

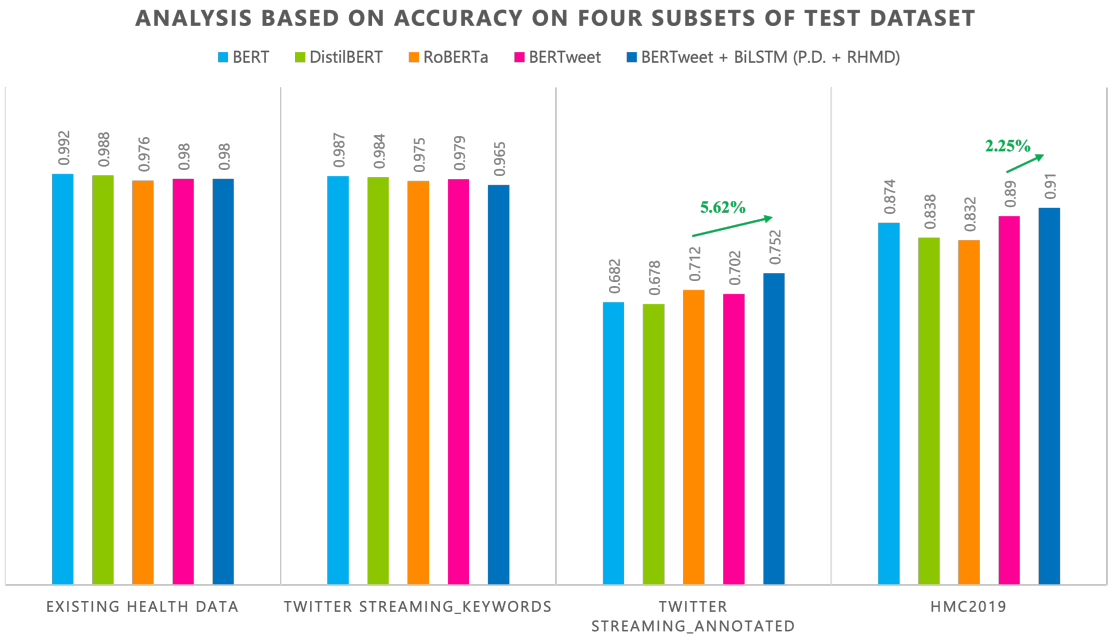

Figure 4.2. Model Performance on subsets of Test Data.

However, The BERTweet + BiLSTM (P.D. + RHMD) model outperforms other models on subsets that require more contextual understanding of the tweets, such as the Twitter streaming_annotated subset and HMC2019 subset. The model achieves

an accuracy of 0.752, which is 5.62% higher than the accuracy achieved by the second-best performing model, RoBERTa. This indicates that the model is effective in accurately classifying health-related tweets from the tweets that contain figurative or non-literal use of health keywords. Furthermore, the model achieves an accuracy of 0.910 on the HMC2019 subset, which is 2.25% higher than the accuracy achieved by the second-best performing model, BERTweet. These results demonstrate that the proposed modified BERTweet model with an additional BiLSTM layer is highly effective in accurately classifying health-related tweets, especially those that require more contextual understanding. The insights gained from the performance evaluation can inform the development of more effective transformer-based models for health-related tweet classification in the future.

4.2 Ablation Analysis

Model	**F1-Score**	**Accuracy**
BERTweet + BiLSTM (P.D. + RHMD)	0.900	0.900
BERTweet (P.D. + RHMD)	0.871	0.872
BERTweet + BiLSTM	0.895	0.895
BERTweet	0.877	0.879

Table 4.2. Ablation Analysis on Best Model: BERTweet + BiLSTM (P.D. + RHMD)

The table 4.2 provides the results of an ablation study conducted on the best-performing model, BERTweet + BiLSTM (P.D. + RHMD), to investigate the contribution of the BiLSTM layer and the RHMD dataset to the performance of the model. The full model achieved an F1-score and accuracy of 0.900, which is the highest among all the models evaluated in the study. This indicates that the combination

of the BERTweet model, BiLSTM layer, and RHMD dataset is effective in achieving high accuracy in health-related tweet classification. The removal of the BiLSTM layer resulted in a significant drop in F1-score to 0.871 and accuracy to 0.872. This suggests that the BiLSTM layer is crucial in achieving high accuracy in health-related tweet classification using transformer-based models. The BiLSTM layer helps the model capture the temporal relationships between words in the input sequence, which is important in analyzing text data. On the other hand, the removal of the RHMD dataset resulted in a slight drop in F1-score to 0.895 and accuracy to 0.895. This indicates that while the RHMD dataset contributes to the performance of the model, it is not as critical as the BiLSTM layer. The RHMD dataset contains health-related tweets collected from various sources, which helps the model learn health-related concepts and improve its performance in health-related tweet classification. Finally, the removal of both the BiLSTM layer and the RHMD dataset resulted in the low F1-score and accuracy of 0.877 and 0.879, respectively. This confirms that the BiLSTM layer and the RHMD dataset are essential components of the BERTweet + BiLSTM model for achieving high accuracy in health-related tweet classification.

Overall, the results of the ablation study highlight the importance of the BiLSTM layer and the RHMD dataset in the BERTweet + BiLSTM model for health-related tweet classification. The study can provide valuable insights for researchers and practitioners in optimizing transformer-based models for health-related tweet classification and can be utilized in analyzing health-related social media data for gaining insights into public health trends, identifying potential disease outbreaks, monitoring the effectiveness of health campaigns, and supporting patient care.

The ablation study was conducted to analyze the impact of the BiLSTM layer and the RHMD dataset on the performance of the BERTweet model. The performance of this model was compared to the performance of models that lacked one or

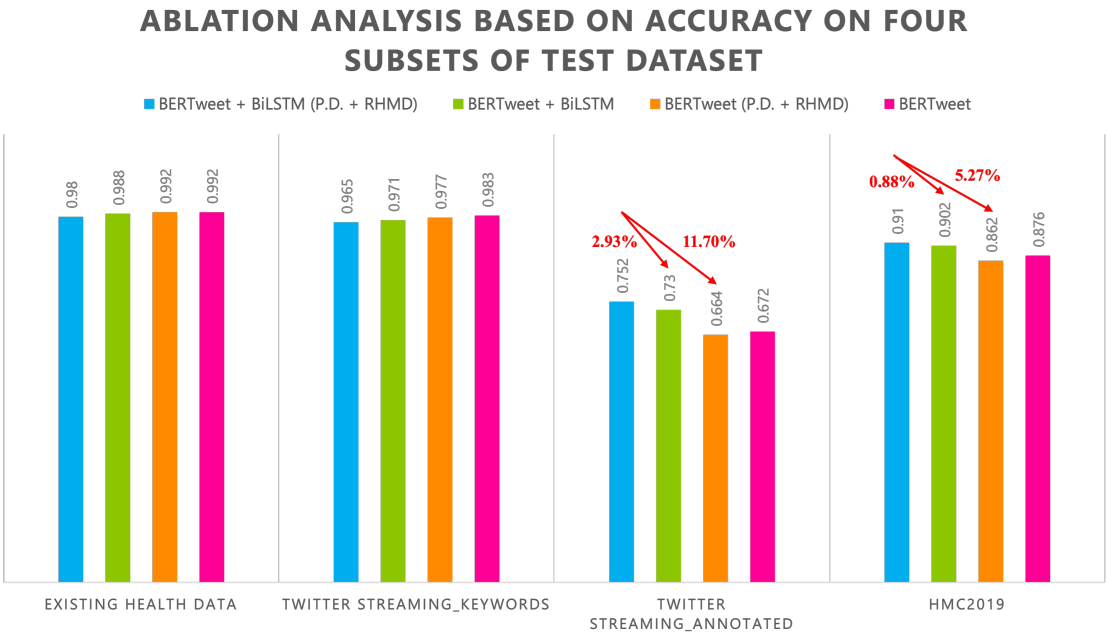

Figure 4.3. Ablation Analysis: Model Performance on subsets of Test Data.

both components. In Figure 4.3, For the Twitter Streaming_annotated subset, the ablation study shows that the removal of the BiLSTM layer resulted in a significant decrease in accuracy, dropping from 0.752 to 0.664, which is a decrease of 11.70%. Similarly, for the HMC2019 subset, the removal of the BiLSTM layer also resulted in a decrease in accuracy, dropping from 0.91 to 0.862, which is a decrease of 5.27%. This demonstrates that the BiLSTM layer plays a crucial role in the BERTweet + BiLSTM (P.D. + RHMD) model's performance on these subsets, and its removal results in a substantial drop in accuracy. On the other hand, when the RHMD dataset was removed, there was a drop in accuracy, but the decrease was not as significant, indicating that the BiLSTM layer plays a more critical role in improving model performance. For other subsets, such as Existing Health Data and Twitter Streaming_keywords, there was not much difference or change in accuracy observed

when any component was removed, implying that the overall model performance for these subsets was not significantly impacted by the absence of any component.

4.3 Application

Wildfire Tool The health-related tweet classification has a major application in the Wildfire platform created by our IDIR lab as a part of research. Wildfire is designed for social sensing tasks, which involve collecting and analyzing data from social media platforms like Twitter to gain insights into various phenomena such as public health, disaster response, and political events. Seed collection in Wildfire involves retrieving tweets based on a query or set of keywords using Twitter's APIs, while expansion collection involves exploring new accounts and tweets using a customized ranking function. The health-related tweet Classifier is a task-specific classifier used in Wildfire that provides a relevance score for each account and tweet related to health. This score is used in the customized ranking function to prioritize the exploration of relevant accounts and tweets, thereby accelerating the data collection process for health-related data. Overall, Wildfire and the health-related tweet Classifier enable researchers and analysts to efficiently collect and analyze large amounts of social media data, with a particular focus on the health domain.

CHAPTER 5

CONCLUSION

In conclusion, this thesis presented a comprehensive dataset for Health-related Tweet Classification, and evaluated various transformer-based models on it. The findings of the study highlight the importance of incorporating additional layers and diverse datasets to improve the performance of transformer-based models in Health-related Tweet Classification. The results indicate that BERTweet + BiLSTM (P.D. + RHMD) achieved the highest F1-score of 0.900, which underscores the significance of the BiLSTM layer and the RHMD dataset for the BERTweet model. Furthermore, this study contributes to the development of Health-related Tweet Classification, an important area of research in health communication. The study's results suggest that BERTweet + BiLSTM (P.D. + RHMD) can accurately classify health-related tweets, which could be beneficial for monitoring public health concerns and identifying emerging health trends. The findings could also assist health professionals in addressing health-related misinformation and promoting evidence-based practices on social media platforms.

5.1 Future Work

Future work could expand upon the findings of this study by including additional models, such as BioBERT and ClinicalBioBERT, for a broader comparison of transformer-based models in Health-related Tweet Classification. Additionally, utilizing emojis and sentiment learning could be explored to enhance the model's performance by incorporating additional contextual information. Incorporating more

data from multiple social networks could also improve model performance by increasing the diversity of the dataset. Contrastive learning could be applied to improve the utilization of labeled data and enhance the model's ability to distinguish between different health-related tweet categories [28]. Overall, further research in these areas could contribute to the advancement of machine learning techniques for analyzing health-related social media data and lead to more accurate and efficient classification models.

REFERENCES

[1] L. Donelle and R. G. Booth, "Health tweets: an exploration of health promotion on twitter." *Online journal of issues in nursing*, vol. 17, no. 3, 2012.

[2] L. Sinnenberg, A. M. Buttenheim, K. Padrez, C. Mancheno, L. Ungar, and R. M. Merchant, "Twitter as a tool for health research: a systematic review," *American journal of public health*, vol. 107, no. 1, pp. e1–e8, 2017.

[3] K. A. Alnemer, W. M. Alhuzaim, A. A. Alnemer, B. B. Alharbi, A. S. Bawazir, O. R. Barayyan, and F. K. Balaraj, "Are health-related tweets evidence based? review and analysis of health-related tweets on twitter," *Journal of medical Internet research*, vol. 17, no. 10, p. e246, 2015.

[4] M. Paul and M. Dredze, "You are what you tweet: Analyzing twitter for public health," in *Proceedings of the International AAAI Conference on Web and Social Media*, vol. 5, no. 1, 2011, pp. 265–272.

[5] A. Ali, W. Magdy, and S. Vogel, "A tool for monitoring and analyzing healthcare tweets," in *HSD workshop, SIGIR 2013*. Citeseer, 2013.

[6] S. Kuang and B. D. Davison, "Learning word embeddings with chi-square weights for healthcare tweet classification," *Applied Sciences*, vol. 7, no. 8, p. 846, 2017.

[7] K. Srinivasulu, "Health-related tweets classification: a survey," in *Proceedings of International Conference on Recent Trends in Machine Learning, IoT, Smart Cities and Applications: ICMISC 2020*. Springer, 2021, pp. 259–268.

[8] S. Xiong, V. Batra, L. Liu, L. Xi, and C. Sun, "Detecting personal medication intake in twitter via domain attention-based rnn with multi-level features," *Computational Intelligence and Neuroscience*, vol. 2022, 2022.

[9] T.-M. Lin, C.-Y. Chen, Y.-W. Tzeng, and L.-H. Lee, "Ncuee-nlp@ smm4h'22: Classification of self-reported chronic stress on twitter using ensemble pre-trained transformer models," in *Proceedings of The Seventh Workshop on Social Media Mining for Health Applications, Workshop & Shared Task*, 2022, pp. 62–64.

[10] V. Porvatov and N. Semenova, "Transformer-based classification of premise in tweets related to covid-19," *arXiv preprint arXiv:2209.03851*, 2022.

[11] T. Kayastha, P. Gupta, and P. Bhattacharyya, "Bert based adverse drug effect tweet classification," in *Proceedings of the Sixth Social Media Mining for Health (# SMM4H) Workshop and Shared Task*, 2021, pp. 88–90.

[12] A. Ahne, F. Orchard, X. Tannier, C. Perchoux, B. Balkau, S. Pagoto, J. L. Harding, T. Czernichow, and G. Fagherazzi, "Insulin pricing and other major diabetes-related concerns in the usa: a study of 46 407 tweets between 2017 and 2019," *BMJ Open Diabetes Research and Care*, vol. 8, no. 1, p. e001190, 2020.

[13] R. J. Medford, S. N. Saleh, A. Sumarsono, T. M. Perl, and C. U. Lehmann, "An "infodemic": leveraging high-volume twitter data to understand early public sentiment for the coronavirus disease 2019 outbreak," in *Open forum infectious diseases*, vol. 7, no. 7. Oxford University Press US, 2020, p. ofaa258.

[14] O. Stens, M. H. Weisman, J. Simard, K. Reuter, *et al.*, "Insights from twitter conversations on lupus and reproductive health: Protocol for a content analysis," *JMIR Research Protocols*, vol. 9, no. 8, p. e15623, 2020.

[15] P. Arora and P. Arora, "Mining twitter data for depression detection," in *2019 international conference on signal processing and communication (ICSC)*. IEEE, 2019, pp. 186–189.

[16] X. Yuan, R. J. Schuchard, and A. T. Crooks, "Examining emergent communities and social bots within the polarized online vaccination debate in twitter," *Social media+ society*, vol. 5, no. 3, p. 2056305119865465, 2019.

[17] N. Oscar, P. A. Fox, R. Croucher, R. Wernick, J. Keune, and K. Hooker, "Machine learning, sentiment analysis, and tweets: An examination of alzheimer's disease stigma on twitter," *Journals of Gerontology Series B: Psychological Sciences and Social Sciences*, vol. 72, no. 5, pp. 742–751, 2017.

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

[19] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019.

[20] R. Biddle, A. Joshi, S. Liu, C. Paris, and G. Xu, "Leveraging sentiment distributions to distinguish figurative from literal health reports on twitter," in *Proceedings of the web conference 2020*, 2020, pp. 1217–1227.

[21] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter," *arXiv preprint arXiv:1910.01108*, 2019.

[22] D. Q. Nguyen, T. Vu, and A. T. Nguyen, "Bertweet: A pre-trained language model for english tweets," *arXiv preprint arXiv:2005.10200*, 2020.

[23] K. Glandt, S. Khanal, Y. Li, D. Caragea, and C. Caragea, "Stance detection in covid-19 tweets," in *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 2021, pp. 1596–1611.

[24] R. Biddle, A. Joshi, S. Liu, C. Paris, and G. Xu, "Leveraging sentiment distributions to distinguish figurative from literal health reports on twitter," in *Proceedings of The Web Conference 2020*, ser. WWW '20. New York,

NY, USA: Association for Computing Machinery, 2020, p. 1217–1227. [Online]. Available: https://doi.org/10.1145/3366423.3380198

[25] U. Naseem, J. Kim, M. Khushi, and A. G. Dunn, "Identification of disease or symptom terms in reddit to improve health mention classification," in *Proceedings of the ACM Web Conference 2022*, 2022, pp. 2573–2581.

[26] G. Liu and J. Guo, "Bidirectional lstm with attention mechanism and convolutional layer for text classification," *Neurocomputing*, vol. 337, pp. 325–338, 2019.

[27] C. Pierse, "Transformers Interpret," Feb. 2021. [Online]. Available: https://github.com/cdpierse/transformers-interpret

[28] C.-Y. Chuang, J. Robinson, Y.-C. Lin, A. Torralba, and S. Jegelka, "Debiased contrastive learning," *Advances in neural information processing systems*, vol. 33, pp. 8765–8775, 2020.

BIOGRAPHICAL STATEMENT

Foram Pankajbhai Patel is a computer science professional from India, who obtained her Bachelor's degree in Computer Engineering from Gujarat Technological University. She then pursued her academic goals by earning a Master of Science in Computer Science degree from The University of Texas at Arlington in 2023. During her time at UTA, Foram worked at the IDIR lab and collaborated with Ph.D. students on her thesis research in the field of Deep Learning and Fact-checking.