Exploring BERT-Based Classification Models for Detecting Phobia Subtypes: A Novel Tweet Dataset and Comparative Analysis

Anik Das, Milton King, James Alexander Hughes
St. Francis Xavier University
4130 University Avenue, Antigonish, Nova Scotia, Canada, B2G 2W5
x2021gmg@stfx.ca, mking@stfx.ca, jhughes@stfx.ca

Abstract

Phobias, characterized by irrational fears of specific objects or situations, can profoundly affect an individual's quality of life. This research presents a comprehensive investigation into phobia classification, where we propose a novel dataset of 811,569 English tweets from user timelines spanning 102 phobia subtypes over six months, including 47,614 self-diagnosed phobia users. BERT models were leveraged to differentiate non-phobia from phobia users and classify them into 65 specific phobia subtypes. The study produced promising results, with the highest f1-score of 78.44% in binary classification (phobic user or not phobic user) and 24.01% in a multi-class classification (detecting the specific phobia subtype of a user). This research provides insights into people with phobias on social media and emphasizes the capacity of natural language processing and machine learning to automate the evaluation and support of mental health.

Keywords: Social Media, Phobia, BERT, Classification

1. Introduction

Phobia, classified as an anxiety disorder, is characterized by an intense and irrational fear of particular objects, situations, or activities, despite there being no significant real risk or only a minor threat involved. Unlike fear, which serves as an adaptive response to immediate danger, phobia generates a relentless state of anticipation and preparation (Adolphs, 2013). For example, fear might prompt an individual to escape from a genuinely perilous animal, whereas phobia induces an irrational fear associated with a specific animal, even in the absence of any genuine threat. The first edition of the American Psychiatric Association DSM in 1952 (Association et al., 2000) briefly diagnosed phobic reactions, including common fears such as syphilis, dirt, closed and open spaces, and animals. Patients typically cope by avoiding the phobic object or situation. Following this, phobia types were expanded, encompassing social phobias, agoraphobia, and specific phobias, with further subtypes including those associated with blood, injections, injury, and so on. These phobias manifest in various forms, with nearly 12.5% of U.S. adults experiencing specific phobias at some point in their lives[1]. Importantly, phobias can significantly affect one's well-being and contribute to the development of psychological disorders such as anxiety and depression (Mekhilef et al., 2012; Wardenaar et al., 2018).

Mental health research using social media is a growing field, leveraging data from platforms such as Facebook, Twitter, and Weibo to gain valuable insights into individuals' mental well-being, behaviors, and attitudes (Santos et al., 2023; Mann et al., 2020; Paul et al., 2018; Baes et al., 2022; Bakar and Nawi, 2021; Vemprala et al., 2021; Alt, 2015; Selvadass et al., 2022; Naslund et al., 2020; De Choudhury, 2013). These platforms offer plenty of information about users' daily lives, including their emotions, social interactions, and activities, which can be harnessed to uncover patterns and trends related to mental health. Researchers in this field investigate various aspects of mental health, including depression (Santos et al., 2023; Mann et al., 2020; Paul et al., 2018; Baes et al., 2022; Bakar and Nawi, 2021), anxiety (Santos et al., 2023), fear (Vemprala et al., 2021; Alt, 2015), and stress (Selvadass et al., 2022). This approach allows researchers to access a diverse population, reaching individuals who may not have traditional mental health services readily available or those who refrain from seeking help from healthcare facilities. In addition, it helps identify new risk factors, understand the impact of these conditions on individuals and communities, and develop potential interventions for mental health problems (Naslund et al., 2020; De Choudhury, 2013). In essence, social media-based mental health research holds promise in providing valuable insights and enhancing the development of more effective strategies for mental health prevention and treatment.

To our knowledge, there have been no studies that have delved into the realm of phobia investigation using social media data. Existing research focused primarily on the identification of

[1] https://www.hcp.med.harvard.edu/ncs/index.php

LREC-COLING 2024, pages 6894–6908
20-25 May, 2024.

phobias by various means, such as brain magnetic resonance imaging (MRI) (Böhnlein et al., 2021; Lueken et al., 2015), physiological signals (Zhang et al., 2016; Sharma et al., 2016; Šalkevicius et al., 2019; Sandulescu et al., 2015; Petrescu et al., 2020; Ihmig et al., 2020), and text data collected from surveys and therapeutic email communications (Intia et al., 2022; Hoogendoorn et al., 2016). Furthermore, attention given to the exploration of these phobia subtypes is insufficient, as a significant number of phobia categories remain unused in research. Many of the investigations primarily focused on recognizing social anxiety in individuals (Sharma et al., 2016; Intia et al., 2022; Hoogendoorn et al., 2016). Some studies targeted specific forms of anxiety, such as anxiety about public speaking (Zhang et al., 2016; Šalkevicius et al., 2019). Several concentrated on animal phobias, for instance, the fear of spiders (Böhnlein et al., 2021) and snakes (Lueken et al., 2015). Another study aimed at dental phobia by identifying blood injection injuries (Lueken et al., 2015). As a result, there are a multitude of phobias that require thorough investigation. The process of detecting phobias through social media entails a variety of techniques for analyzing data from social platforms, with the overarching aim of identifying individuals who might be grappling with phobias or anxiety disorders. This analysis includes examining linguistic patterns within posts and texts across social media platforms.

This research introduces a novel dataset, which is comprised of tweets related to various phobias, along with an automated approach for phobia detection. The dataset is extracted from the Twitter timelines of users and encompasses a total of 102 phobia subtypes (can be found in Table 1). We have implemented a series of preprocessing steps and established specific string search patterns to identify self-reported tweets related to phobias. Moreover, we have developed distinct models using BERT to assess their effectiveness in binary classification, distinguishing between users affected by phobias and those who are not, as well as in a multi-class classification to pinpoint the specific subtype of phobia. To the best of our knowledge, this study is the first to create such a comprehensive dataset of tweets associated with a wide range of phobia subtypes and to design automated models for phobia detection using Twitter data. The notable contributions of this research are outlined below:

- Constructed a novel phobia-related tweet dataset based on self-reported users' timelines containing 102 phobia subtypes.
- Applied systematic pre-processing and string-searching steps to identify self-reported phobia-affected users.
- Evaluated the performance of eight BERT-based models, assessing their capability to identify users with phobias and accurately classify the specific subtype of phobia associated with them.

Our dataset will be a valuable resource for investigating the impact of various phobias on individuals' mental well-being. Furthermore, the results produced by the classification model can be instrumental in the development of automated tools and systems designed to identify individuals with specific phobias. The remaining sections of our research paper are structured as follows: Section 2 provides the research context for automated methods in phobia identification. The process of data collection, pre-processing, self-diagnosis recognition, and data categorization is detailed in Section 3. Section 4 explains the BERT-based classification models. In Section 5, we present the results and engage in a comprehensive discussion of our findings. Finally, Section 6 concludes our research and outlines potential future research avenues related to our work.

2. Background and Literature Review

Several studies affiliated with phobias have focused on identifying specific brain areas using MRI data through the utilization of machine learning approaches (Böhnlein et al., 2021; Lueken et al., 2015). Böhnlein et al. (2021) emphasized the significance of incorporating the entire brain's features, obtained from fMRI, in differentiating individuals with spider phobia from healthy controls, achieving a 73% balanced accuracy using a multivariate machine learning approach. Their findings suggested that neurofunctional activity in spider phobia extends beyond specific regions of interest (ROIs) within the brain, favoring the use of whole-brain features. Lueken et al. (2015) employed Gaussian process classifiers to distinguish between healthy controls and individuals with specific phobia subtypes (e.g., snake and dental phobias), focusing on gray and white matter volumetric data from MRI recordings. These features exhibited greater accuracy in predicting these phobia subtypes. However, conducting MRI-based studies to identify various types of phobias is associated with high expenses, clinical complexities, ethical considerations, participant variability, experimental control challenges, small sample sizes, unavailability of diverse phobia subtypes, data analysis intricacies, and ethical approval requirements.

Physiological signals, such as galvanic skin response (GSR), heart rate variability (HRV), heart

rate (HR), blood volume pulse (BVP), skin temperature, and respiration, play a crucial role in detecting and classifying different levels of anxiety (Katsis et al., 2011). For instance, Zhang et al. (2016) employed GSR signals to classify anxiety levels during public speaking, achieving an average accuracy of 86.70% for high anxiety and 78.83% for calmness using back-propagation neural network algorithms. Similarly, GSR signals were also used to identify social anxiety (Sharma et al., 2016), yielding accuracy rates of 82.3% in training, 85.7% in testing, and 80% in holdout cases with neural networks and multilayer perceptron (MLP). Šalkevicius et al. (2019) predicted public speaking anxiety levels from GSR, BVP, and skin temperature data with 80.1% accuracy in a leave-one-out setup and 86.3% accuracy in 10 $\times$ 10 fold cross-validations using signal fusion-based Support Vector Machine (SVM) classifiers. Additionally, Sandulescu et al. (2015) utilized GSR and BVP signals in SVM models to classify stressful and non-stressful situations, achieving an average accuracy of nearly 79% in tests conducted on five participants. Real-time GSR and HRV measurements were used by Petrescu et al. (2020) to predict anxiety levels during exposure to heights, resulting in high accuracy levels for different anxiety categories (94.29% accuracy for low-medium levels, and 92.38% accuracy for low-medium-high levels). Ihmig et al. (2020) used GSR, HR, HRV and respiration data to implement bagged trees, achieving 89.8% and 74.4% accuracies in classifying two-level (low, and high) and three-level (low, medium, and high) anxiety levels, respectively. Nonetheless, the scope of these investigations was primarily centered on discerning various anxiety levels, and it is essential to employ precise procedural protocols for gathering physiological signals to guarantee dependable and uniform data acquisition, thereby reducing the likelihood of errors or artifacts that could potentially influence subsequent analyses.

Textual data has emerged as a prominent resource for understanding mental health conditions, including phobias. Intia et al. (2022) conducted an online survey involving agoraphobia-related questionnaires, employing various feature selection and reduction techniques as well as classification algorithms (e.g., random forest, decision tree, k-nearest neighbors, and SVM). Their study resulted in the random forest classification exhibiting an impressive accuracy rate of 98.02%, outperforming other classification methods. Meanwhile, Hoogendoorn et al. (2016) developed predictive models for identifying social anxiety disorder using sociodemographic data and the content of therapeutic emails from patients during treatment, encompassing email topics, sentiment, word usage, and writing styles. Their findings highlighted the superiority of shorter-term trends (with respect to long-term trends) in predicting the disorder, and the logistic regression model achieving the best f1-score of 84%. However, text-based studies have not comprehensively explored the multitude of phobia types, which aligns with the primary aim of our research — to identify and investigate a broad spectrum of phobia subtypes.

The studies on phobia subtypes mentioned earlier have primarily concentrated on specific fears, including spiders (arachnophobia) (Böhnlein et al., 2021), snakes (ophidiophobia) (Lueken et al., 2015), needles or injections (trypanophobia) (Lueken et al., 2015), public speaking (glossophobia) (Zhang et al., 2016; Šalkevicius et al., 2019), public spaces or crowds (agoraphobia) (Sharma et al., 2016; Intia et al., 2022; Hoogendoorn et al., 2016), and heights (acrophobia) (Petrescu et al., 2020). Nonetheless, there exists a wide array of phobia subtypes that remain unexplored and require in-depth investigation. Therefore, our research is primarily focused on exploring these undiscovered areas.

Self-disclosure of a diagnosis involves individuals openly sharing information about their mental health conditions, often through conversations, online posts, or other communication channels. This voluntary sharing contributes to a better understanding of mental health conditions (De Choudhury et al., 2013c; Tsugawa et al., 2015; MacAvaney et al., 2018), which, in turn, can assist researchers, clinicians, and support networks in providing appropriate assistance and interventions. In a similar study, De Choudhury et al. (2013c) employed crowdsourcing to collect data from Twitter users who self-reported clinical depression diagnoses. They utilized an SVM classifier to predict the likelihood of depression before its reported onset based on individual tweets, achieving a classification accuracy of 70%. To establish a basis for self-reported information, Tsugawa (Tsugawa et al., 2015) used data obtained from a web-based questionnaire designed to gauge the levels of depression in Twitter users. Later, they harnessed an SVM classifier to detect signs of depression by scrutinizing users' Twitter interactions and mapping them against self-reported data, achieving an accuracy of roughly 69%. Macavaney et al (MacAvaney et al., 2018) introduced the RSDD-Time (Reddit Self-Reported Depression Diagnosis-Time) dataset, a repository containing manually annotated self-reported depression diagnosis posts from Reddit, categorized based on temporal measures. They also examined the effectiveness of various machine learning models, including logistic regression, linear SVM, and gradient boost tree (GBT), for predicting the

recency of the diagnosis (indicating when the diagnosis occurred) and the condition state (identifying if the diagnosed condition is current or previous). Among all the models, GBT topped in diagnosing the recency classification, while logistic regression topped in condition state classification, achieving respective f1-scores of 46% and 44%, respectively. Hence, self-reported diagnoses can provide valuable insights for identifying phobias.

BERT-based models have been extensively utilized in mental health research for the automatic classification of text data related to mental health. Researchers have employed BERT-based models to analyze mood disorders, such as anxiety (Santos et al., 2023; Ameer et al., 2022), and depression (Santos et al., 2023; Senn et al., 2022; Solse et al., 2022; Anantharaman et al., 2022; Ameer et al., 2022). These models were utilized to automatically identify emotions expressed in text, such as happiness, sadness, or anger (Adoma et al., 2020; Peng et al., 2021). BERT-based models were also applied to identify other mental health-related topics, such as mental health stigma (Lee and Kyung, 2022), and suicidal tendencies (Metzler et al., 2022; Wang et al., 2021). The effectiveness of BERT in capturing contextualized representations from text data has been demonstrated in the analysis of mental health-related texts, providing valuable insights into various aspects of mental health. Furthermore, according to Park et al. (2022), the BERT model has demonstrated superior performance in long-text classification compared to alternative models, such as Longformer (a transformer-based model designed to efficiently handle lengthy documents by incorporating global attention mechanisms), To-BERT (a text-to-text framework utilizing a BERT-like architecture for various natural language processing tasks), and CogLTX (a cognitive-inspired model that integrates linguistic and visual information for text classification), which are developed for diverse natural language processing purposes. Nonetheless, BERT's effectiveness in long text classification establishes it as the preferred choice for the phobia classification task in this study.

3. Dataset Creation

3.1. Data Source

We gathered public tweets associated with phobias through the Twitter public API[2]. We employed a keyword-based search strategy to accumulate tweets related to various phobia types. This approach involved a list of 204 keywords, which is comprised of 102 phobia-type names and 102 brief descriptions of these phobia types sourced from internet searches on common phobias (depicted in Table 1). For instance, for the phobia type "acrophobia," we used both the name "acrophobia" and its brief description, "fear of heights," to retrieve relevant tweets. Our keyword-based data collection extended over six months, starting on November 29, 2022, and ending on April 29, 2023. A total of 811,569 tweets were collected during this process. Likewise, we made use of the Twitter API for collecting user timelines by searching the usernames associated with the collected phobia-related tweets. During this process, we gathered the most recent 50 tweets from each user. Some user timelines contained fewer than 50 tweets, particularly if the user had fewer than 50 'public' tweets in their timeline. Furthermore, this dataset included tweets in multiple languages and was used for self-reported phobia diagnosis, as discussed in Section 3.2.

3.2. Recognizing Self-reported Phobia Tweets

In line with self-reported diagnosis-focused research (De Choudhury et al., 2013c; Tsugawa et al., 2015; MacAvaney et al., 2018), it is established that self-disclosure of a diagnosis can deliver invaluable insights into an individual's mental health condition. For instance, if an individual explicitly states, "I have acrophobia", it reasonably implies that the person is personally affected by acrophobia. Conversely, if an individual states, "My friend has acrophobia," the tweet would not be categorized as acrophobia, as it refers to another person. This underscores the importance of accurate annotation.

Prior to identifying self-diagnosed phobia tweets from the Twitter obtained during the keyword search process, several pre-processing steps were implemented on these tweets. Non-English phobia-related tweets were excluded, resulting in 791,078 tweets. The Python 'langid' library[3] facilitated non-English language detection, allowing for the identification of texts in English ('en') among other languages. Subsequently, duplicate tweets were removed based on identical usernames and texts in phobia-related tweets (for instance, if a particular user posted the same tweet multiple times, we eliminated the redundant copies of that tweet). This process yielded 741,080 tweets. Finally, two string search patterns were employed to detect self-reported phobia tweets. The first pattern starts with "I" and optionally includes an auxiliary verb. The main verb is then used, followed by an optional preposition. Additionally, individuals may optionally include indefinite articles. Finally,

[2] https://developer.twitter.com/en/docs/twitter-api

[3] https://github.com/saffsd/langid.py

Table 1: Phobia subtypes with definition

Name	Definition	Name	Definition	Name	Definition	Name	Definition
achluophobia	fear of darkness	acrophobia	fear of heights	aerophobia	fear of flying	algophobia	fear of pain
alektorophobia	fear of chickens	agoraphobia	fear of public spaces or crowds	aichmophobia	fear of needles or pointed objects	ailurophobia	fear of cats
amaxophobia	fear of riding in a car	androphobia	fear of men	anginophobia	fear of angina or choking	anthophobia	fear of flowers
anthropophobia	fear of people or society	aphenphosmphobia	fear of being touched	arachnophobia	fear of spiders	arithmophobia	fear of numbers
astraphobia	fear of thunder and lightning	ataxophobia	fear of disorder or untidiness	atelophobia	fear of imperfection	atychiphobia	fear of failure
autophobia	fear of being alone	bacteriophobia	fear of bacteria	barophobia	fear of gravity	bathmophobia	fear of stairs or steep slopes
batrachophobia	fear of amphibians	belonephobia	fear of pins and needles	bibliophobia	fear of books	botanophobia	fear of plants
cacophobia	fear of ugliness	catagelophobia	fear of being ridiculed	catoptrophobia	fear of mirrors	chionophobia	fear of snow
chromophobia	fear of colors	chronomentrophobia	fear of clocks	cibophobia	fear of food	claustrophobia	fear of confined spaces
coulrophobia	fear of clowns	cyberphobia	fear of computers	cynophobia	fear of dogs	dendrophobia	fear of trees
dentophobia	fear of dentists	domatophobia	fear of houses	dystychiphobia	fear of accidents	entomophobia	fear of insects
ephebiphobia	fear of teenagers	equinophobia	fear of horses	gamophobia	fear of marriage or commitment	genuphobia	fear of knees
glossophobia	fear of speaking in public	gynophobia	fear of women	heliophobia	fear of the sun	hemophobia	fear of blood
herpetophobia	fear of reptiles	hydrophobia	fear of water	hypochondria	fear of illness	iatrophobia	fear of doctors
insectophobia	fear of insects	koinoniphobia	fear of rooms full of people	leukophobia	fear of the color white	lilapsophobia	fear of tornadoes and hurricanes
lockiophobia	fear of childbirth	mageirocophobia	fear of cooking	megalophobia	fear of large things	melanophobia	fear of the color black
microphobia	fear of small things	mysophobia	fear of dirt and germs	necrophobia	fear of death or dead things	noctiphobia	fear of the night
nosocomephobia	fear of hospitals	nyctophobia	fear of the dark	obesophobia	fear of gaining weight	octophobia	fear of the number 8
ombrophobia	fear of rain	ophidiophobia	fear of snakes	ornithophobia	fear of birds	papyrophobia	fear of paper
pathophobia	fear of disease	pedophobia	fear of children	philophobia	fear of love	phobophobia	fear of phobias
podophobia	fear of feet	pogonophobia	fear of beards	porphyrophobia	fear of the color purple	pteridophobia	fear of ferns
pteromerhanophobia	fear of flying	pyrophobia	fear of fire	samhainophobia	fear of Halloween	scolionophobia	fear of school
selenophobia	fear of the moon	sociophobia	fear of social evaluation	somniphobia	fear of sleep	tachophobia	fear of speed
technophobia	fear of technology	thalassophobia	fear of deep water	tonitrophobia	fear of thunder	trypanophobia	fear of needles or injections
trypophobia	fear of clustered patterns of holes	venustraphobia	fear of beautiful women	verminophobia	fear of germs	wiccaphobia	fear of witches and witchcraft
xenophobia	fear of strangers or foreigners	zoophobia	fear of animals				

the sentence is completed with phobia keywords. The second pattern starts with “my” and includes phobia keywords, to indicate that they are describing a personal experience or condition. If any of these specified string patterns are found within a tweet, the tweet is categorized as a self-reported phobia tweet. It is worth noting that various search strings were considered during this phase, but the presented string patterns proved to be the most effective for identifying self-reported tweets based on our dataset. As a result, a total of 23,807 self-reported diagnosis-based tweets from unique Twitter users were identified through this process.

3.3. Data Classes and Post-processing

We aimed to evaluate the effectiveness of BERT in automatically distinguishing both binary phobia detection and phobia subtype identification using user tweet timelines. For binary classification, we defined two classes: ‘non-phobia’ and ‘phobia’. The 23,807 Twitter users with self-reported diagnoses, as identified in the previous step, were labeled as the ‘phobia’ class. Subsequently, we randomly selected an equal number of Twitter users (23,807 in total) from the dataset. These users did not have self-reported phobia tweets but had phobia-related keywords in their texts, as identified through keyphrase searches. We categorized them as the ‘non-phobia’ class. The Twitter users selected for the non-phobia class were exclusively distinct users. An essential part of this selection process was the meticulous examination of all 50 tweets for each non-phobia user to ensure that these tweets did not contain any self-reported diagnosis-based content. We performed this verification by using the string patterns detailed in Section 3.2. This rigorous selection process ensured a clear separation between the non-phobia and phobia classes in our final dataset, comprising a total of 47,614 Twitter users, evenly split between the non-phobia and phobia classes (23,807 each).

To prepare the final set of 47,614 user tweets for automated classification, post-processing steps were applied to the 50 tweets collected for each user. Tweets from users identified through keyphrase searches, which already contained self-diagnosis texts, were excluded focusing solely on the user timeline of tweets. Later, the removal of duplicate tweets within user timelines was conducted, ensuring that the dataset consisted of unique tweets. Subsequently, newline characters within each tweet were systematically eliminated to prevent potential formatting issues and ensure that the text was treated as a continuous, unbroken stream of words. For consistency and compatibility with prior research, instances of usernames and URLs mentioned in tweets were replaced with standardized tokens e.g., [user] and [url]. In cases where two or more consecutive [url] or [user] tokens occurred, these were streamlined to a single instance. Hashtags and emojis were removed at this stage. These meticulous post-processing measures refined the dataset, rendering it suitable for further analysis and classification, while eliminating any extraneous elements.

To develop a dataset for the multi-class classification system for phobia subtypes, the initial collection of 23,807 Twitter users categorized as ‘phobia’ class was used for further labeling with specific phobia subtypes. This process utilized the phobia-related keywords from our data collection, as described in Section 3.1. However, it was evident that certain self-diagnosed phobia subtypes had either very limited or no Twitter users, leading to an imbalanced data distribution. In response, we decided to exclude classes with fewer than ten Twitter users. This action was taken to address the issue of imbalanced data distribution in multi-class classification, which can introduce prediction bias and reduce accuracy for classes with less data. Following this procedure, we acquired a refined dataset encompassing a total of 23,700 Twitter users, spanning 65 distinct phobia subtypes. The distribution of data for each class is visually

presented in Figure 1, where class 1 represents the 'achluophobia' category, and the other classes correspond to the phobia subtypes (class 0 is disregarded, representing the non-phobia class). Phobia subtypes such as 'acrophobia,' 'trypophobia,' 'arachnophobia,' and 'claustrophobia' were more readily available in terms of data availability, making them easier to find.

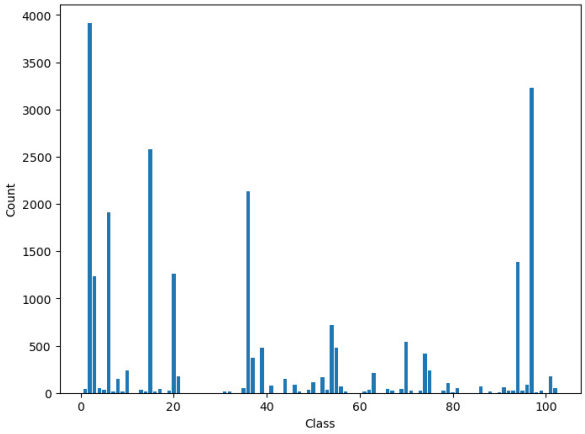

Figure 1: Amount of data in multi-class dataset. Here in Class-axis, '1' represents 'achluophobia' class, '2' represents 'acrophobia' class, and so on ('0' is disregarded, representing the non-phobia class)

3.4. Word cloud Visualization

We utilized word cloud visualization to depict the 47,614 user timeline dataset comprising self-diagnosed non-phobia and phobia classes. Figure 2 showcases the word cloud visualization, with Sub-figure 2(a) displaying the entire dataset, Sub-figure 2(b) illustrating the non-phobia class, and Sub-figure 2(c) representing the phobia class. Notably, in the top 10 frequent words in both the non-phobia and phobia classes, common words such as 'like', 'people', 'one', 'dont' (don't), 'im' (I'm), 'get', 'know', and 'time' dominate both classes, reflecting frequent usage to convey personal experiences, opinions, or share information across various topics. Additionally, 'us' (US or United States), and 'new' emerge as unique words in the non-phobia class, suggesting discussions related to collective groups (often including the speaker), US-related issues and recent events, discoveries, products, or experiences. Conversely, 'love', and 'good' stand out as unique words in the phobia class, indicating users' succinct expression of positive emotions and sentiments within the tweet's limited character space.

(a) Whole dataset

(b) Non-phobia class (c) Phobia class

Figure 2: Word cloud visualization of the dataset; (a) word cloud of the whole dataset, (b) world cloud for non-phobia class, and (c) word cloud for phobia class

4. BERT Embedding and Classification

BERT (Devlin et al., 2018) stands as a cutting-edge deep learning model widely employed for natural language processing tasks. In many implementations, the 'transformers' library, developed by Hugging Face, is commonly used to implement BERT and other transformer-based models. This architecture, purposefully designed to capture long-range dependencies and relationships between words in a text, equips BERT for a wide range of natural language understanding tasks. However, it is worth noting that BERT's input is limited to 512 tokens. To address this limitation, tweets were reordered in various model configurations during aggregation (described in the next Section 4.1) to ensure essential data was included as input to the models even after truncation (Hou et al., 2022; Zhao and Yu, 2021).

4.1. Models

For the classification task, we organized the data in several ways to prepare it for input into BERT, resulting in the creation of distinct classification models. Each model received an aggregated sentence comprising of 50 tweets from each user, except for the final model. Moreover, each model adopted unique strategies for arranging the sentences or tweets within the text, with the exception of the last model. As part of the aggregation process for user timeline tweets, we applied the same post-processing steps detailed in Section 3.3. The models are outlined as follows:

Model 1 (M1) – User Timeline: In M1 (user timeline), the 50 tweets were consolidated in the usual way, following the sequence of tweets from the user's timeline, into a single sentence.

Model 2 (M2) – TextRank with Sentence: Within M2 (TextRank with sentence), we employed an efficient unsupervised sentence ranking algorithm, specifically the 'TextRank' method (Mihalcea and Tarau, 2004), to arrange the sentences. TextRank assesses the importance of phrases within a given input text by constructing a graph of phrase relationships based on co-occurrence, iteratively updating phrase scores, and extracting key phrases for summarization or analysis. Similarly, it ranks sentences by considering each sentence as a node in a graph, and it assesses their significance based on relationships and co-occurrences between sentences, resembling its approach to ranking keywords or phrases. Sentences that are more interconnected with other sentences, especially those identified as significant, are assigned higher ranks, facilitating the identification of key sentences for summarization or other natural language processing tasks.

After consolidating the 50 tweets into a single sentence, we applied TextRank to select key sentences, and these were positioned at the beginning of the aggregated sentence. For our implementation, we utilized PyTextRank[4] integrated into the spaCy pipeline[5] with default settings.

Model 3 (M3) – TextRank with Phrases: Utilizing TextRank's phrase ranking capabilities mentioned earlier, in M3 (TextRank with phrases), phrases from each of the 50 tweets were individually ranked. The highest-ranked phrase determined the ranking of the individual tweets, which were subsequently reordered during aggregation into a sentence based on this ranking.

Model 4 (M4) – Phobia Keywords: In M4 (phobia keywords), we searched phobia keywords within the 50 tweets of each user. When any of these keywords were found, the corresponding tweets were placed at the beginning of the aggregated sentence, while the remaining tweets retained their original order. It is essential to note that while the keywords were present in the tweets, they were not employed for self-reported diagnosis.

Model 5 (M5) – Random Tweets: M5 (random tweets) involved the random aggregation of the 50 tweets from each user's timeline into a single sentence.

Model 6 (M6) – Individual Tweets: A different approach was adopted for M6 (individual tweets). For each user, all 50 tweets were individually categorized as either 'non-phobia' or 'phobia' based on their primary annotation. As a result, the dataset for this specific model configuration comprised a total of 165,278 individual tweets, including 84,177 individual tweets classified as non-phobia and 81,101 individual tweets categorized as phobia. Therefore, it represents a distinct classification task compared to the previously discussed models.

In M1 through M5, the users' tweets were arranged in unique ways to compile them into a single sentence before BERT tokenization. Therefore, the text lengths remained consistent across these models, demonstrating an average of 3788.85, a median of 3614, and a standard deviation of 1893.90. Conversely, in M6, where each user's 50 tweets were individually considered, the text segments were notably shorter in length, with an average of 97.22, a median of 91, and a standard deviation of 61.89.

4.2. Hyperparameters

The model is comprised of three hidden layers, with a dropout rate of 0.1 applied between these hidden layers, following the recommendation by Devlin et al. (2018). The Adam optimizer was employed, using a learning rate of $1e{-}5$. Throughout all experiments, the hardware configuration was an Intel Core i7-11800H processor, an NVIDIA GeForce RTX 3060 GPU with 6 GB of memory, and 16 GB of RAM.

4.3. Data Split

For all the models mentioned in Section 4.1, the data was initially divided into training, validation, and test sets, using an 8:1:1 ratio before executing the classification model. These models underwent 100 epochs of training, incorporating early stopping with a patience value of 5. Additionally, during the training process, the best model weights were automatically saved based on the performance achieved on the validation data. Later, the test data was assessed using the best-saved models, and their performances were compared based on the f1-scores. Furthermore, to provide a benchmark for comparison, the majority class baseline (MCB) was calculated. The MCB predicts the class with the highest number of instances in the dataset for all cases. It serves as a straightforward baseline for assessing more sophisticated models, although it may not fully capture the intricacies of the data, which advanced models are designed to do.

5. Result and Discussion

5.1. Binary Classification Results

Table 2, and Table 3 provides a summary of the classification performance for distinguishing be-

[4] https://github.com/DerwenAI/pytextrank

[5] https://www.bibsonomy.org/bibtex/2616669ca18ac051794c0459373696942/rerry

tween non-phobia and phobia using various models. Notably, all models completed training up to the seventh epoch (except M2, which finished at the sixth epoch), and interestingly, the best epoch identified was the first epoch for all models. M5 (random tweets) achieved the highest f1-score of 75.35%. Employing a basic data organization method, such as randomly ordering the tweets, could yield superior results, resonating with the findings of Park et al. (2022) in their study on long document classification. The second-highest f1-score, 75.31%, was attained by M3 (TextRank with phrases), which was close to the highest score. The other three models, M1 (user timeline), M4 (phobia keywords), and M2 (TextRank with sentence), achieved f1-scores of 74.09%, 71.67%, and 69.20%, respectively. Conversely, the lowest f1-score, 68.48%, was recorded for M6 (individual tweets). Indeed, for M6 (individual tweets), the classification model encountered challenges in distinguishing the classes when the user timeline of tweets was individually considered instead of being aggregated together. BERT's architecture is designed to process longer texts by dividing them into chunks or tokens. Individual tweets are typically very short, limiting BERT's capacity to comprehend the entire context and resulting in an information loss.

Table 2: Binary classification results for M1 to M5 (here, 'P', 'R', 'F1', 'Acc' means precision, recall, f1-score, and accuracy respectively)

Models	Epoch	Results			
		P	R	F1	Acc
M1 (user timeline)	1	73.23	74.97	74.09	73.52
	2	70.93	86.24	77.84	75.20
M2 (TextRank with sentence)	1	67.86	70.60	69.20	68.30
	2	-	-	-	-
M3 (TextRank with phrases)	1	73.49	77.22	75.31	74.84
	2	71.18	87.36	**78.44**	76.14
M4 (phobia keywords)	1	74.89	68.72	71.67	72.43
	2	70.91	86.72	78.02	75.20
M5 (random tweets)	1	72.86	78.02	75.35	74.11
	2	70.92	85.51	77.54	74.86
MCB_{M1-M5}	-	50	100	66.67	50

Table 3: Binary classification results for M6 (here, 'P', 'R', 'F1', 'Acc' means precision, recall, f1-score, and accuracy respectively)

Models	Epoch	Results			
		P	R	F1	Acc
M6 (individual tweets)	1	68.05	68.91	68.48	68.61
	2	67.35	72.64	69.90	69.05
MCB_{M6}	-	50.93	100	67.49	50.93

Further, since the best models were automatically saved, the weights of the second epoch were also saved for all models (except for M2), even though it was not the best epoch based on the train-validation accuracy curves during model training. Intriguingly, when we evaluated the test dataset using these saved weights, we observed higher f1-scores for all the models, but with notable differences in precision and recall results. The top-performing model was M3 (TextRank with phrases), achieving an f1-score of 78.4%. The other models, M4 (phobia keywords), M1 (user timeline), and M5 (random tweets), achieved f1-scores of 78.02%, 77.84%, and 77.54%, respectively. As in the previous results, the lowest f1-score was recorded for M6 (individual tweets) at 69.05%.

Overall, the f1-scores of the models surpassed both the f1-scores of MCB_{M1-M5} and MCB_{M6} (66.67% and 67.49%, respectively), indicating that the models outperformed the majority class predictions for all instances. This demonstrates the models' effectiveness in distinguishing between different classes while maintaining a balance between precision and recall. The highest f1-score, 78.44%, was achieved by M3 (TextRank with phrases) in its second epoch. The capability of TextRank to rank tweets with phrases proved to be effective in identifying the most crucial tweets for each user. The confusion matrix for the model is depicted in Figure 3. Notably, the model faced more challenges in predicting the 'non-phobia' class compared to the 'phobia' class. Users who were misclassified, along with their timelines of tweets, predominantly contained content unrelated to phobia, including daily life updates, engaging topics, personal thoughts, and more. This inherent variability in tweet content posed a challenge for the model in accurately distinguishing between non-phobia and phobia users.

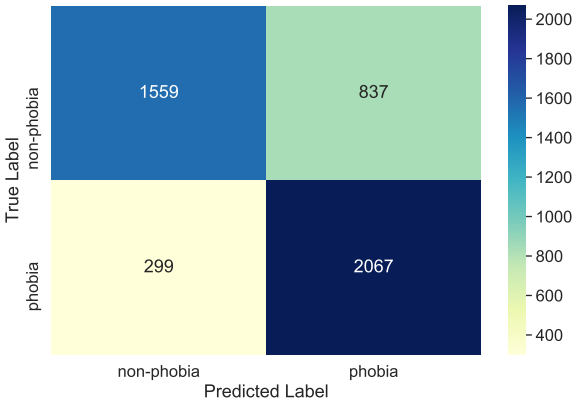

Figure 3: Confusion matrix for the M3 (TextRank with phrases) on its second epoch in binary classification

5.2. Multi-class Classification Results

For the multi-class classification of phobia subtypes, we chose M1 (user timeline) and M3 (TextRank with phrases), as well as the best settings identified in the binary classification (as mentioned in Section 5.1). Both models completed training up to the eighth epoch, with the second epoch yielding the best results. M1 (user timeline) outperformed M3 (TextRank with phrases) with an f1-score of 22.62%, while M3 achieved an f1-score of 21.12%. Similar to previous experiments, the best weights were automatically saved, including the weights of the third epoch for both models, despite it not being the best epoch based on the train-validation accuracy curves during model training. In this case, both models demonstrated reasonable performance, with M1 (user timeline) achieving the highest f1-score of 24.01% and M3 (TextRank with phrases) obtaining a 23.16% f1-score. In this context, the f1-scores for all the models fell below the f1-score of the MCB (28.40%). This implies that the model performances were not as good as simply predicting the majority class for all instances. The models encountered difficulties in accurately classifying instances from the minority class, resulting in lower precision and recall values in the predictions. This could be considered an inadequate overall performance for multi-class classification.

The results are depicted in Table 4. Further analysis was conducted on the highest score achieved in multi-class classification for M1 (user timeline) during its third epoch, and evaluation scores were computed for each class. This analysis suggests that the model faced challenges in accurately predicting most of the phobia classes, particularly those with limited data. Furthermore, only 10 out of 65 classes demonstrated measurable F1 scores, while the remaining 55 classes did not yield satisfactory results, indicating that the classifier encountered difficulties in correctly classifying instances within these classes. Classes with fewer than 542 instances were particularly challenging for the model to classify accurately due to the limited data available. Also, compared to binary classification, multi-class classification typically experiences lower performance as the number of classes increases, posing greater challenges in accurately classifying each class, as evidenced by the evaluation scores computed for individual classes in the multi-class scenario.

Table 4: Multi-class classification results (here, 'P', 'R', 'F1', 'Acc' means precision, recall, f1-score, and accuracy respectively)

Models	Epoch	Results			
		P	R	F1	Acc
M1 (user timeline)	2	24.00	27.30	22.62	27.30
	3	27.13	28.19	**24.01**	28.19
M3 (TextRank with phrases)	2	23.04	25.27	21.12	25.27
	3	23.70	26.75	23.16	26.75
MCB	-	16.51	100	28.40	16.51

6. Conclusions

In this research, we proposed a novel dataset containing 811,569 tweets and the user timeline of tweets of 102 phobia subtypes (collected over six months). Detailed pre-processing and string-searching methods were applied to find 47,614 self-diagnosed phobia users. Finally, different BERT models were evaluated for binary classification of non-phobia and phobia, and multi-class classification of 65 phobia subtypes (102 classes were reduced to 65 classes by excluding classes with fewer than ten Twitter users to minimize imbalanced data distribution) using the user timeline of tweets. The best f1-scores achieved were 78.44% for the binary classification and 24.01% for the multi-class classification. In the context of multi-class classification, the model was unable to correctly predict 55 out of 65 classes that represent distinct phobia subtypes, which is one of the limitations of this research.

In the future, an in-depth exploration of multi-class classification will be undertaken, particularly in response to the observed misclassifications incurred by the investigated multi-class classification models in this study. To enhance the BERT model's performance in multi-class classification, both the inclusion of a diverse and balanced dataset and the utilization of advanced data augmentation techniques will be employed to improve its discriminatory capabilities across various phobia subtypes. This study holds practical relevance, as it paves the way for the creation of valuable applications such as a chatbot for phobia detection or web-based tools. These applications, if developed, could have a substantial real-world impact on individuals' mental well-being by providing accessible resources for comprehending and addressing their phobias.

7. Ethical Considerations

In self-reported phobia classification research, ethical concerns include privacy and consent. Analyzing public social media data for self-reported phobia tweets risks intruding on individuals' privacy, potentially causing them to feel exposed or vulnerable. Researchers must prioritize obtaining proper consent and ensuring data anonymity to mitigate these privacy concerns.

8. References

Ralph Adolphs. 2013. The biology of fear. *Current biology*, 23(2):R79–R93.

Acheampong Francisca Adoma, Nunoo-Mensah Henry, and Wenyu Chen. 2020. Comparative analyses of bert, roberta, distilbert, and xlnet for text-based emotion recognition. In *2020 17th International Computer Conference on Wavelet Active Media Technology and Information Processing (ICCWAMTIP)*, pages 117–121. IEEE.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2017. 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*, pages 127–138. Springer.

Israa Alghanmi, Luis Espinosa Anke, and Steven Schockaert. 2020. Combining bert with static word embeddings for categorizing social media. In *Proceedings of the sixth workshop on noisy user-generated text (w-nut 2020)*, pages 28–33.

Dorit Alt. 2015. College students' academic motivation, media engagement and fear of missing out. *Computers in human behavior*, 49:111–119.

Iqra Ameer, Muhammad Arif, Grigori Sidorov, Helena Gòmez-Adorno, and Alexander Gelbukh. 2022. Mental illness classification on social media texts using deep learning and transfer learning. *arXiv preprint arXiv:2207.01012*.

Karun Anantharaman, S Angel, Rajalakshmi Sivanaiah, Saritha Madhavan, and Sakaya Milton Rajendram. 2022. Ssn_mlrg1@ lt-edi-acl2022: Multi-class classification using bert models for detecting depression signs from social media text. In *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion*, pages 296–300.

American Psychiatric Association et al. 2000. Diagnostic and statistical manual of mental disorders. *Text revision*.

Arnie Mae M Baes, Aunhel John M Adoptante, John Carlo A Catilo, Patrick Kendrex L Lucero, Janice F Peralta Peralta, and Anton Louise Pernez de Ocampo. 2022. A novel screening tool system for depressive disorders using social media and artificial neural network. *International Journal of Intelligent Systems and Applications in Engineering*, 10(1):116–121.

Fahem Abu Bakar and Nazri Mohd Nawi. 2021. Predicting depression using social media posts. *Journal of Soft Computing and Data Mining*, 2(2):39–48.

David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.

Joscha Böhnlein, Elisabeth J Leehr, Kati Roesmann, Teresa Sappelt, Ole Platte, Dominik Grotegerd, Lisa Sindermann, Jonathan Repple, Nils Opel, Susanne Meinert, et al. 2021. Neural processing of emotional facial stimuli in specific phobia: An fmri study. *Depression and Anxiety*, 38(8):846–859.

Nina Cesare, Olubusola Oladeji, Kadija Ferryman, Derry Wijaya, Karen D Hendricks-Muñoz, Alyssa Ward, and Elaine O Nsoesie. 2020. Discussions of miscarriage and preterm births on twitter. *Paediatric and perinatal epidemiology*, 34(5):544–552.

Lan-lan Chen, Yu Zhao, Peng-fei Ye, Jian Zhang, and Jun-zhong Zou. 2017. Detecting driving stress in physiological signals based on multimodal feature analysis and kernel classifiers. *Expert Systems with Applications*, 85:279–291.

Yong Cheng, Lu Jiang, Wolfgang Macherey, and Jacob Eisenstein. 2020. Advaug: Robust adversarial augmentation for neural machine translation. *arXiv preprint arXiv:2006.11834*.

Claude Coulombe. 2018. Text data augmentation made simple by leveraging nlp cloud apis. *arXiv preprint arXiv:1812.04718*.

Munmun De Choudhury. 2013. Role of social media in tackling challenges in mental health. In *Proceedings of the 2nd international workshop on Socially-aware multimedia*, pages 49–52.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013a. Predicting postpartum changes in emotion and behavior via social media. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 3267–3276.

Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013b. Social media as a measurement tool of depression in populations. In *Proceedings of the 5th annual ACM web science conference*, pages 47–56.

Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013c. Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media*, volume 7, number 1, pages 128–137.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Xin Luna Dong, Yaxin Zhu, Zuohui Fu, Dongkuan Xu, and Gerard de Melo. 2021. Data augmentation with adversarial training for cross-lingual nli. 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)*, pages 5158–5167.

IV Dsm et al. 1994. Diagnostic and statistical manual of mental disorders. *Washington, DC: American psychiatric association*, 968:33–36.

Gobind Simran Kaur Gahir and Jignyasa B Sanghavi. 2020. Covid-19 tweets textual analytics using machine learning classification for fear sentiment. *International Journal*, 9(5).

Santiago González-Carvajal and Eduardo C Garrido-Merchán. 2020. Comparing bert against traditional machine learning text classification. *arXiv preprint arXiv:2005.13012*.

Hongyu Guo, Yongyi Mao, and Richong Zhang. 2019. Augmenting data with mixup for sentence classification: An empirical study. *arXiv preprint arXiv:1905.08941*.

Alfons O Hamm, Bruce N Cuthbert, Jutta Globisch, and Dieter Vaitl. 1997. Fear and the startle reflex: Blink modulation and autonomic response patterns in animal and mutilation fearful subjects. *Psychophysiology*, 34(1):97–107.

Mark Hoogendoorn, Thomas Berger, Ava Schulz, Timo Stolz, and Peter Szolovits. 2016. Predicting social anxiety treatment outcome based on therapeutic email conversations. *IEEE journal of biomedical and health informatics*, 21(5):1449–1459.

Le Hou, Richard Yuanzhe Pang, Tianyi Zhou, Yuexin Wu, Xinying Song, Xiaodan Song, and Denny Zhou. 2022. Token dropping for efficient bert pretraining. *arXiv preprint arXiv:2203.13240*.

Frank R Ihmig, Frank Neurohr-Parakenings, Sarah K Schäfer, Johanna Lass-Hennemann, and Tanja Michael. 2020. On-line anxiety level detection from biosignals: Machine learning based on a randomized controlled trial with spider-fearful individuals. *Plos one*, 15(6):e0231517.

Israt Jahan Intia, Md Mehedi Hasan, Khairul Alam, and Khondoker Sangida Ferdous. 2022. Prediction of agoraphobia disease based on machine learning. In *2022 13th International Conference on Computing Communication and Networking Technologies (ICCCNT)*, pages 1–5. IEEE.

Vasileios Iosifidis and Eirini Ntoutsi. 2018. Dealing with bias via data augmentation in supervised learning scenarios. *Jo Bates Paul D. Clough Robert Jäschke*, 24:11.

Xiaoqi Jiao, Yichun Yin, Lifeng Shang, Xin Jiang, Xiao Chen, Linlin Li, Fang Wang, and Qun Liu. 2019. Tinybert: Distilling bert for natural language understanding. *arXiv preprint arXiv:1909.10351*.

Michał Jungiewicz and Aleksander Smywiński-Pohl. 2019. Towards textual data augmentation for neural networks: synonyms and maximum loss. *Computer Science*, 20.

Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Eduard Hovy. 2018. Adventure: Adversarial training for textual entailment with knowledge-guided examples. *arXiv preprint arXiv:1805.04680*.

Christos D Katsis, Nikolaos S Katertsidis, and Dimitrios I Fotiadis. 2011. An integrated system based on physiological signals for the assessment of affective states in patients with anxiety disorders. *Biomedical Signal Processing and Control*, 6(3):261–268.

Sosuke Kobayashi. 2018. Contextual augmentation: Data augmentation by words with paradigmatic relations. *arXiv preprint arXiv:1805.06201*.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. Albert: A lite bert for self-supervised learning of language representations. *arXiv preprint arXiv:1909.11942*.

Min Hyung Lee and Richard Kyung. 2022. Mental health stigma and natural language processing: Two enigmas through the lens of a limited corpus. In *2022 IEEE World AI IoT Congress (AIIoT)*, pages 688–691. IEEE.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*.

Kun Li, Chengbo Chen, Xiaojun Quan, Qing Ling, and Yan Song. 2020. Conditional augmentation for aspect term extraction via masked sequence-to-sequence generation. *arXiv preprint arXiv:2004.14769*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy,

Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Yun Liu and Siqing Du. 2018. Psychological stress level detection based on electrodermal activity. *Behavioural brain research*, 341:50–53.

Samuel Louvan and Bernardo Magnini. 2020. Simple is better! lightweight data augmentation for low resource slot filling and intent classification. *arXiv preprint arXiv:2009.03695*.

Ulrike Lueken, Kevin Hilbert, Hans-Ulrich Wittchen, Andreas Reif, and Tim Hahn. 2015. Diagnostic classification of specific phobia subtypes using structural mri data: a machine-learning approach. *Journal of Neural Transmission*, 122(1):123–134.

Sean MacAvaney, Bart Desmet, Arman Cohan, Luca Soldaini, Andrew Yates, Ayah Zirikly, and Nazli Goharian. 2018. Rsdd-time: Temporal annotation of self-reported mental health diagnoses. *arXiv preprint arXiv:1806.07916*.

Paulo Mann, Aline Paes, and Elton H Matsushima. 2020. See and read: detecting depression symptoms in higher education students using multimodal social media data. In *Proceedings of the International AAAI Conference on Web and social media*, volume 14, pages 440–451.

Jocelyn Mazarura and Alta De Waal. 2016. A comparison of the performance of latent dirichlet allocation and the dirichlet multinomial mixture model on short text. In *2016 Pattern Recognition Association of South Africa and Robotics and Mechatronics International Conference (PRASA-RobMech)*, pages 1–6. IEEE.

Saad Mekhilef, Rahman Saidur, and Masoud Kamalisarvestani. 2012. Effect of dust, humidity and air velocity on efficiency of photovoltaic cells. *Renewable and sustainable energy reviews*, 16(5):2920–2925.

MentalHealth.gov. [Online]. Phobias. https://www.mentalhealth.gov/what-to-look-for/anxiety-disorders/phobias. Accessed: 2022-12-06.

Hannah Metzler, Hubert Baginski, Thomas Niederkrotenthaler, and David Garcia. 2022. Detecting potentially harmful and protective suicide-related content on twitter: machine learning approach. *Journal of medical internet research*, 24(8):e34705.

R Mihalcea and P Tarau. 2004. Textrank: Bringing order into text. in proceedings of the 2004 conference on empirical methods in natural language processing, pp. 404–411.

Junghyun Min, R Thomas McCoy, Dipanjan Das, Emily Pitler, and Tal Linzen. 2020. Syntactic data augmentation increases robustness to inference heuristics. *arXiv preprint arXiv:2004.11999*.

Sebastien Montella, Betty Fabre, Tanguy Urvoy, Johannes Heinecke, and Lina Rojas-Barahona. 2020. Denoising pre-training and data augmentation strategies for enhanced rdf verbalization with transformers. *arXiv preprint arXiv:2012.00571*.

Marzieh Mozafari, Reza Farahbakhsh, and Noel Crespi. 2020. A bert-based transfer learning approach for hate speech detection in online social media. In *Complex Networks and Their Applications VIII: Volume 1 Proceedings of the Eighth International Conference on Complex Networks and Their Applications COMPLEX NETWORKS 2019 8*, pages 928–940. Springer.

Usman Naseem, Imran Razzak, and Peter W Eklund. 2021. A survey of pre-processing techniques to improve short-text quality: a case study on hate speech detection on twitter. *Multimedia Tools and Applications*, 80(28):35239–35266.

John A Naslund, Ameya Bondre, John Torous, and Kelly A Aschbrenner. 2020. Social media and mental health: benefits, risks, and opportunities for research and practice. *Journal of technology in behavioral science*, 5:245–257.

Paco Nathan. 2016. Pytextrank, a python implementation of textrank for phrase extraction and summarization of text documents. *[Online] https://github.com/DerwenAI/pytextrank*.

Nathan Ng, Kyunghyun Cho, and Marzyeh Ghassemi. 2020. Ssmba: Self-supervised manifold based data augmentation for improving out-of-domain robustness. *arXiv preprint arXiv:2009.10195*.

Sosuke Nishikawa, Ryokan Ri, and Yoshimasa Tsuruoka. 2020. Data augmentation for learning bilingual word embeddings with unsupervised machine translation. *arXiv preprint arXiv:2006.00262*.

Hyunji Hayley Park, Yogarshi Vyas, and Kashif Shah. 2022. Efficient classification of long documents using transformers. *arXiv preprint arXiv:2203.11258*.

Sayanta Paul, Sree Kalyani Jandhyala, and Tanmay Basu. 2018. Early detection of signs of anorexia and depression over social media using effective machine learning frameworks. In *CLEF (Working notes)*.

Sancheng Peng, Rong Zeng, Hongzhan Liu, Guanghao Chen, Ruihuan Wu, Aimin Yang, and Shui Yu. 2021. Emotion classification of text based on bert and broad learning system. In *Web and Big Data: 5th International Joint Conference, APWeb-WAIM 2021, Guangzhou, China, August 23–25, 2021, Proceedings, Part I 5*, pages 382–396. Springer.

Livia Petrescu, Cătălin Petrescu, Oana Mitruț, Gabriela Moise, Alin Moldoveanu, Florica Moldoveanu, and Marius Leordeanu. 2020. Integrating biosignals measurement in virtual reality environments for anxiety detection. *Sensors*, 20(24):7088.

Jonathan K Quijas. 2017. *Analysing the effects of data augmentation and free parameters for text classification with recurrent convolutional neural networks*. The University of Texas at El Paso.

Mehdi Regina, Maxime Meyer, and Sébastien Goutal. 2020. Text data augmentation: Towards better detection of spear-phishing emails. *arXiv preprint arXiv:2007.02033*.

Ryan Robert Rosario. 2017. *A data augmentation approach to short text classification*. University of California, Los Angeles.

Justas Šalkevicius, Robertas Damaševičius, Rytis Maskeliunas, and Ilona Laukienė. 2019. Anxiety level recognition for virtual reality therapy system using physiological signals. *Electronics*, 8(9):1039.

Virginia Sandulescu, Sally Andrews, David Ellis, Nicola Bellotto, and Oscar Martinez Mozos. 2015. Stress detection using wearable physiological sensors. In *International work-conference on the interplay between natural and artificial computation*, pages 526–532. Springer.

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *arXiv preprint arXiv:1910.01108*.

Wesley Ramos dos Santos, Rafael Lage de Oliveira, and Ivandré Paraboni. 2023. Setembrobr: a social media corpus for depression and anxiety disorder prediction. *Language Resources and Evaluation*, pages 1–28.

Samer Muthana Sarsam, Hosam Al-Samarraie, Ahmed Ibrahim Alzahrani, Chit Su Mon, and Abdul Samad Shibghatullah. 2022. Characterizing suicide ideation by using mental disorder features on microblogs: A machine learning perspective. *International Journal of Mental Health and Addiction*, pages 1–14.

Harvard Medical School. [Online]. National comorbidity survey (ncs). https://www.hcp.med.harvard.edu/ncs/index.php. Accessed: 2022-12-06.

Salomi Selvadass, P Malin Bruntha, and K Priyadharsini. 2022. Stress analysis in social media using ml algorithms. In *2022 4th International Conference on Smart Systems and Inventive Technology (ICSSIT)*, pages 1502–1506. IEEE.

Saskia Senn, ML Tlachac, Ricardo Flores, and Elke Rundensteiner. 2022. Ensembles of bert for depression classification. In *2022 44th Annual International Conference of the IEEE Engineering in Medicine & Biology Society (EMBC)*, pages 4691–4694. IEEE.

Vivek Sharma, Neelam Rup Prakash, and Parveen Kalra. 2016. Eda wavelet features as social anxiety disorder (sad) estimator in adolescent females. In *2016 International Conference on Advances in Computing, Communications and Informatics (ICACCI)*, pages 1843–1846. IEEE.

Vandita Singh, Bhupendra Kumar, and Tushar Patnaik. 2013. Feature extraction techniques for handwritten text in various scripts: a survey. *International Journal of Soft Computing and Engineering (IJSCE)*, 3(1):238–241.

Durga Solse, Anuja Magar, Priyanka Harde, Neeta Palve, and MT Jagatap. 2022. Depression detection by analyzing social media post in machine learning using bert algorithm. *Int. Res. J. Modernization Eng. Technol. Sci.*, 4(4).

Chi Sun, Xipeng Qiu, Yige Xu, and Xuanjing Huang. 2019. How to fine-tune bert for text classification? In *China national conference on Chinese computational linguistics*, pages 194–206. Springer.

Symeon Symeonidis, Dimitrios Effrosynidis, and Avi Arampatzis. 2018. A comparative evaluation of pre-processing techniques and their interactions for twitter sentiment analysis. *Expert Systems with Applications*, 110:298–310.

Tooru Takahashi. 1989. Social phobia syndrome in japan. *Comprehensive Psychiatry*, 30(1):45–52.

Nandan Thakur, Nils Reimers, Johannes Daxenberger, and Iryna Gurevych. 2020. Augmented sbert: Data augmentation method for improving

bi-encoders for pairwise sentence scoring tasks. *arXiv preprint arXiv:2010.08240*.

Sven Tönnies, Mats Mehrstedt, and Idis Eisentraut. 2002. Die dental anxiety scale (das) und das dental fear survey (dfs) zwei messinstrumente zur erfassung von zahnbehandlungsängsten. *Zeitschrift für Medizinische Psychologie*, 11(2):63–72.

Øivind Due Trier, Anil K Jain, and Torfinn Taxt. 1996. Feature extraction methods for character recognition-a survey. *Pattern recognition*, 29(4):641–662.

Sho Tsugawa, Yusuke Kikuchi, Fumio Kishino, Kosuke Nakajima, Yuichi Itoh, and Hiroyuki Ohsaki. 2015. Recognizing depression from twitter activity. In *Proceedings of the 33rd annual ACM conference on human factors in computing systems*, pages 3187–3196.

Twitter. [Online]. Twitter api documentation | docs | twitter developer platform. https://developer.twitter.com/en/docs/twitter-api. Accessed: 2023-03-15.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Naga Vemprala, Paras Bhatt, Rohit Valecha, and HR Rao. 2021. Emotions during the covid-19 crisis: A health versus economy analysis of public responses. *American Behavioral Scientist*, 65(14):1972–1989.

Rui Wang, Bing Xiang Yang, Yujun Ma, Peilin Wang, Qiao Yu, Xiaofen Zong, Zhen Huang, Simeng Ma, Long Hu, Kai Hwang, et al. 2021. Medical-level suicide risk analysis: A novel standard and evaluation model. *IEEE Internet of Things Journal*, 8(23):16825–16834.

William Yang Wang and Diyi Yang. 2015. That's so annoying!!!: A lexical and frame-semantic embedding based data augmentation approach to automatic categorization of annoying behaviors using# petpeeve tweets. In *Proceedings of the 2015 conference on empirical methods in natural language processing*, pages 2557–2563.

Yequan Wang, Minlie Huang, Xiaoyan Zhu, and Li Zhao. 2016a. Attention-based lstm for aspect-level sentiment classification. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 606–615.

Zhibo Wang, Xiaohui Cui, Lu Gao, Qi Yin, Lei Ke, and Shurong Zhang. 2016b. A hybrid model of sentimental entity recognition on mobile social media. *EURASIP Journal on Wireless Communications and Networking*, 2016(1):1–12.

KJ Wardenaar, CCW Lim, AO Al-Hamzawi, and J Alonso. 2018. Corrigendum: The cross-national epidemiology of specific phobia in the world mental health surveys (psychological medicine (2017) 47 10 (1744-1760)). *Psychological Medicine*, 48(5):878–878.

Jason Wei and Kai Zou. 2019. Eda: Easy data augmentation techniques for boosting performance on text classification tasks. *arXiv preprint arXiv:1901.11196*.

Qizhe Xie, Zihang Dai, Eduard Hovy, Thang Luong, and Quoc Le. 2020. Unsupervised data augmentation for consistency training. *Advances in neural information processing systems*, 33:6256–6268.

Jia Xue, Junxiang Chen, Chen Chen, Chengda Zheng, Sijia Li, and Tingshao Zhu. 2020a. Public discourse and sentiment during the covid 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter. *PloS one*, 15(9):e0239441.

Jia Xue, Junxiang Chen, Ran Hu, Chen Chen, Chengda Zheng, Yue Su, and Tingshao Zhu. 2020b. Twitter discussions and emotions about the covid-19 pandemic: Machine learning approach. *Journal of medical Internet research*, 22(11):e20550.

Xiaohui Yan, Jiafeng Guo, Yanyan Lan, and Xueqi Cheng. 2013. A biterm topic model for short texts. In *Proceedings of the 22nd international conference on World Wide Web*, pages 1445–1456.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2015. Character-level convolutional networks for text classification. *Advances in neural information processing systems*, 28.

Xu Zhang, Wanhui Wen, Guangyuan Liu, and Hui Hu. 2016. Recognition of public speaking anxiety on the recurrence quantification analysis of gsr signals. In *2016 Sixth International Conference on Information Science and Technology (ICIST)*, pages 533–538. IEEE.

Yi Zhang, Tao Ge, and Xu Sun. 2020. Parallel data augmentation for formality style transfer. *arXiv preprint arXiv:2005.07522*.

Anping Zhao and Yu Yu. 2021. Knowledge-enabled bert for aspect-based sentiment analysis. *Knowledge-Based Systems*, 227:107220.

Luyi Zou and William Wei Song. 2016. Lda-tm: A two-step approach to twitter topic data clustering. In *2016 IEEE International Conference on Cloud Computing and Big Data Analysis (ICCCBDA)*, pages 342–347. IEEE.