https://doi.org/10.65025/ICAIC25183m

Early detection of mental illnesses: Analysis of questionnaire and social media data to identify early signs of depression and anxiety.

Youssef Mathouri
Department of Computer Science,
Faculty of Sciences Ibn Tofail University, Morocco.
youssef.mathouri@uit.ac.ma

Korchiyne Redouan
Department of Computer Science,
Faculty of Sciences Ibn Tofail University, Morocco.
redouan.korchiyne@uit.ac.ma

Badour Mohcine
Department of Computer Science,
Faculty of Sciences Ibn Tofail University, Morocco.

Younes Chihab
Department of Computer Science,
Faculty of Sciences Ibn Tofail University, Morocco.

Boukhalfa Alaeddine
Department of Computer Science,
Faculty of Sciences Ibn Tofail University, Morocco.

Abstract
The increasing prevalence of mental health disorders, particularly depression and anxiety, highlights the pressing need for effective early detection mechanisms. Existing approaches often focus on analyzing social media data or utilizing standardized tools like the PHQ-9 and GAD-7. Notably, prior studies on mental illness detection have demonstrated significant success, achieving accuracies of 91% using vector-space word embeddings and 98% when combined with lexicon-based features. However, these methods are limited by their reliance on single-source data, which may not capture the full complexity of mental health states. In this study, we propose a novel hybrid approach that integrates structured data from standardized questionnaires with unstructured data from social media posts. Our methodology employs deep neural networks to process questionnaire responses and Natural Language Processing (NLP) techniques to extract emotional and contextual signals from social media content. This combination enables the extraction of complementary features, enhancing the model's ability to detect subtle indicators of depression and anxiety. Preliminary experiments demonstrate the effectiveness of our approach, achieving a predictive accuracy of 93%, thus bridging the gap between single-source limitations and comprehensive mental health assessments. This integrated framework opens the door to real-time mental health monitoring and personalized intervention strategies.

Keywords
Early detection; Mental illnesses; Sentiment analysis

1. Introduction

Mental illnesses, particularly depression and anxiety, represent some of the most significant public health challenges of the 21st century, affecting millions worldwide. According to the World Health Organization (WHO) [1], over 300 million individuals suffer from depression, while anxiety disorders impact approximately 264 million people globally [2]. These conditions disrupt daily lives, strain personal relationships, and reduce workplace productivity, imposing substantial burdens on public health systems and national economies. If left untreated, they can lead to severe outcomes such as chronic stress, self-harm, or even suicide. Despite the urgent need for timely intervention, many individuals remain undiagnosed or untreated due to stigma, limited awareness, or fear of social judgment [3]. This highlights the critical need for innovative approaches to early detection.

Traditional methods for diagnosing mental health conditions often rely on clinical assessments and standardized tools, such as the Patient Health Questionnaire (PHQ- 9) for depression and the Generalized Anxiety Disorder scale (GAD-7) [4][5]. While these tools are effective and validated, they depend on individuals recognizing their symptoms and seeking professional help a process that is often delayed or avoided altogether. To address this gap, researchers have turned to technology, particularly social media platforms, as an alternative source of data for early detection. Social media platforms like Twitter, Facebook, and Reddit have become ubiquitous, with billions of users sharing thoughts, emotions, and life experiences. These platforms generate vast amounts of unstructured text data that can provide valuable insights into an individual's mental state [6].

Recent studies have shown that linguistic patterns, including word choice, sentence structure, posting frequency, and emotional tone, can serve as indicators of depression and anxiety. For example, studies by De Choudhury et al. [7] and Resnik et al. [8] demonstrated how Twitter and Reddit data could predict depressive symptoms with significant accuracy. Similarly, Pedersen [9] combined vector embeddings with lexicon-based features to classify anxiety-related posts, achieving accuracies as high as 98%. Machine learning (ML) and Natural Language Processing (NLP) techniques have emerged as powerful tools for analyzing these data sources. However, challenges persist, including cultural and linguistic variations, small sample sizes, and biases in training data, which hinder the generalizability of existing models [10].

This study aims to address these challenges by integrating data from two complementary sources: traditional psychological assessments and social media content. By combining structured data (e.g., PHQ-9 and GAD-7 responses) with unstructured data (e.g., linguistic and behavioral patterns from social media), this research seeks to develop a predictive model that leverages the strengths of both approaches. The objectives of this study are threefold:

- To enhance the early detection of depression and anxiety by synthesizing multiple data types.
- To evaluate the effectiveness of this integrated approach compared to single-source models.
- To identify key linguistic and behavioral features that signal early warning signs of mental health conditions.

This work builds on previous research while addressing key gaps, such as the need for larger, more diverse datasets and improved model validation. The findings will contribute to the development of scalable tools for mental health professionals, enabling earlier interventions and better support for individuals at risk.

2. Literature review

2.1. Overview of Previous Research

The early detection of mental disorders, such as depression and anxiety, has garnered increasing attention in public health research. Traditional diagnostic tools, including the Patient Health Questionnaire (PHQ-9) and the Generalized Anxiety Disorder 7-item scale (GAD-7), are widely employed to assess symptom severity [11]. These instruments have been validated by numerous studies, underscoring their reliability and capacity to provide precise and quantifiable results [12]. Spitzer et al. (1999) demonstrated that the PHQ-9 is an effective tool for diagnosing depression in various clinical settings [13], while Kroenke et al. (2006) confirmed the efficacy of the GAD-7 scale for assessing anxiety [14].

In the realm of digital analysis, pioneering studies such as those by De Choudhury et al. (2013) examined the use of platforms like Twitter to identify signs of depression through the analysis of linguistic behaviors and social interactions [15]. The findings highlighted that specific indicators, such as word choice, emotional language, and post- ing frequency, could be correlated with depressive symptoms. Guntuku et al. (2019) expanded this research to other platforms, including Facebook and Reddit, employing advanced natural language processing (NLP) techniques to detect signs of anxiety and depression [16]. These studies showcased the potential of social media data for early detection, while also revealing inherent limitations such as the lack of data standardization and the diversity of cultural contexts. Other notable studies include Resnik et al. (2015), who utilized NLP algorithms to analyze Reddit posts to detect depressive indicators. This study demonstrated that the combination of specific keywords and linguistic features could enhance the accuracy of depression detection [17]. Similarly, Shen et al. (2018) explored the use of social media data and machine learning approaches to predict episodes of depression, emphasizing the importance of analyzing long-term behavioral changes to improve detection [18].

2.2. Research Gaps

Despite significant contributions from previous studies, several gaps remain. Approaches relying solely on self-administered questionnaires are prone to bias due to their dependence on voluntary participation and subjective recognition of symptoms [19]. These methods may lack sensitivity to symptom variations over time, limiting the ability to monitor subtle changes in mental state. Conversely, while innovative, social media analysis is often criticized for its lack of clinical validation and difficulty in generalizing results to diverse populations, particularly due to cultural and linguistic differences [20]. Additionally, online publications are frequently influenced by social context and platform norms, complicating the interpretation of expressed emotions and behaviors. Few studies have explored the integration of social media data with validated psychological tools to develop hybrid models capable of overcoming these challenges. This gap is critical for improving early detection methods. For instance, studies focusing exclusively on textual content sometimes overlook other data formats, such as images and videos, which can also provide valuable insights into individuals' mental states [21]. Benton et al. (2017) highlighted the utility of multimodal data for better understanding complex behaviors associated with mental disorders [14]. Furthermore, Eichstaedt et al. (2018) utilized combined analyses of social media texts and metadata to enhance depression prediction, demonstrating the importance of a broader and more diversified research framework [22].

2.3. Study Rationale

The importance of this study lies in addressing these gaps by combining the strengths of traditional methods with social media data analysis. Integrating data from clinical questionnaires (such as PHQ-9 and GAD-7) with the analysis of social media posts could create a more holistic detection framework. This approach could improve the accuracy of early diagnoses while addressing the limitations of self-assessment and variations in online emotional expression. By employing artificial intelligence (AI) and NLP techniques, the study aims to develop a multidimensional predictive model that enhances and enriches current methods.

Studies by Park et al. (2015) demonstrated that combining textual data with questionnaires increased the accuracy of mental state predictions, although they lacked large-scale validation [23]. Other research, such as Lin et al. (2020), emphasized the need to adapt NLP models to cultural specificities to improve their global applicability [24]. These findings underscore the need for an integrated and adaptable model for the early detection of mental disorders.

2.4. Contribution of the Study

This research builds on previous efforts in early detection of mental disorders by offering an integrative perspective. Unlike studies that favor either self-administered questionnaires or social media analysis in isolation, our approach combines these two data sources to create a robust and adaptable predictive model. The goal is to provide mental health professionals with detection tools capable of identifying early warning signs with increased sensitivity, facilitating quicker and more tailored interventions. Drawing on prior research findings, our study aims to leverage advanced NLP techniques, such as deep neural network models, to analyze social media data while considering responses from clinical questionnaires. This combination could enable more precise risk profiling and improve the responsiveness of mental health interventions [25].

2.5. Critical Analysis

The analysis of existing studies indicates that while numerous social media-based models have been developed, they often lack clinical validation, limiting their applicability in medical contexts. Studies using the PHQ-9 and GAD-7, although valuable, do not always capture the temporal dynamics of symptoms and may not detect subtle variations over extended periods. Social media analysis approaches also present significant challenges, such as biases introduced by online behavior and cultural diversity. For example, Liu et al. (2018) found that NLP models tended to underperform when applied to populations with unique or unconventional linguistic practices [26]. Coppersmith et al. (2015) also demonstrated that Twitter-based models for depression detection could be effective but required adjustments to reduce false positives [27]. Nevertheless, previous studies have paved the way for innovative analysis methods while highlighting the need for an integrated approach that combines these technologies with clinically validated tools. This study aims to fill that gap by developing a comprehensive and validated method that meets current research demands.

In conclusion, integrating NLP and AI techniques with established psychological tools could transform the detection of mental disorders, enabling earlier and more personalized interventions that support the needs of modern public health systems. The combined approach would offer improved sensitivity, more accurate predictions, and adaptability to various cultural contexts, thereby enhancing the effectiveness of prevention and treatment strategies.

3. Methodology

Social media has become an increasingly popular data source for detecting mental illnesses through text. For example, De Choudhury [28] built a corpus of more than 2 million Twitter posts, including a ‘depression’ class with tweets from 476 highly active users self-identified as clinically diagnosed with depression. To identify depression, they used feature vectors that included engagement with the Twitter platform, the social graph of user Twitter activity, emotional and linguistic style using Linguistic Inquiry and Word Count (LIWC) (Pennebaker et al., 2015) and a depression lexicon including antidepressant names. De Choudhury et al. (2013) used a mix of text features and metadata features and achieved 70% accuracy in predicting depression in tweets. Another major data set of tweets labelled for depression was generated by Coppersmith et al. (2015) and contained 3 million tweets from about 2000 Twitter users, including 600 self-identified clinically depressed users. From this data set, Nadeem (2016) achieved 86% accuracy with a naive Bayes unigram classifier. Resnik et al. (2015) used the same data set with latent Dirichlet allocation (LDA) and supervised LDA techniques to predict the likelihood of target classes based on topics. Their supervised LDA techniques included the associated labels of documents as priors for topic modeling. This approach modified an unsupervised learning method and achieved a precision of

0.648 at a recall of 0.5. Preotiuc-Pietro et al. (2015) also participated in the shared task and applied a range methods including: LDA, word vector embeddings, GloVe vector embeddings, and unigrams in order to generate word clusters and then feature vectors based on said word clusters. The same data set has also been used to identify patients with post-traumatic stress disorder (PTSD) in social media in the Copper- smith shared task (Coppersmith et al., 2015). Using this data set, Pedersen (2015) used lexical decision lists with N-grams (N between 1 and 6) and achieved a classification accuracy of 74.2% in classifying tweets from people with PTSD. While Twitter data are available in large volumes, tweets are limited in length and can restrict the potential for contextual processing. By contrast, LiveJournal is a platform for people to discuss common interests and has also been studied to identify community posts by people with depression. Nguyen et al. (2014) found that affective word features from the Affective Norms for English Words (ANEW) and mood tags posted by users gave lower coverage than LIWC features and LDA Topic modeling. Using LIWC and LDA as features for classification, they achieved 93% accuracy. Psychopathology researchers have investigated social anxiety in the context of social media. For example, Fernandez et al. (2012) studied profile information and usage patterns of Facebook users. They concluded that social anxiety was significantly negatively correlated with the number of Facebook friends and positively correlated with the number of completed sections of a Facebook profile.

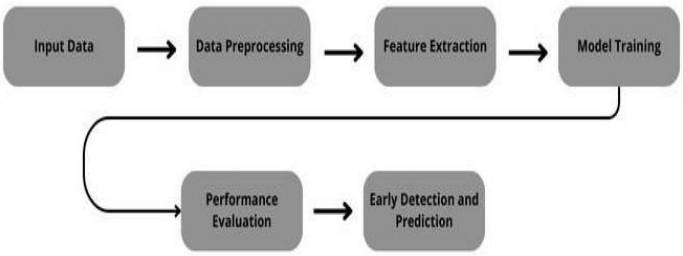

Fig. 1: Early Detection of Depression and Anxiety: Process Flow

Similar to LiveJournal and Facebook, Reddit offers relatively rich bodies of text from users in the context of self-assembled communities. Red dit is a social website for news aggregation, content rating, and discussion. Reddit allows posts up to 40,000 characters per comment, compared to the 140-character limit of Twitter. Each month, 234 million unique users contribute 75.15 million posts and 725.85 comments to the site 1. The website contains more than 1 million subpages, called subreddits, each focusing on its own topic, many of which involve sharing personal stories and experiences in order to seek or give advice. The subreddits concerning depression and anxiety both involve over 100,000 community members 2. De Choudhury and De (2014) studied mental health disclosure on Reddit and concluded that users share their experiences and challenges with mental illnesses as well as the impacts of their illnesses on their work, lives, and relationships. They also found that users use the platform not only for self-expression, but also for seeking diagnosis and treatment information for their conditions. Kumar et al. (2015) studied the r/Suicide Watch community on Reddit after celebrity suicides and found increased posting activity and increased suicidal ideation in post content, by using linguistic measures, N-gram comparison, and topic modeling.

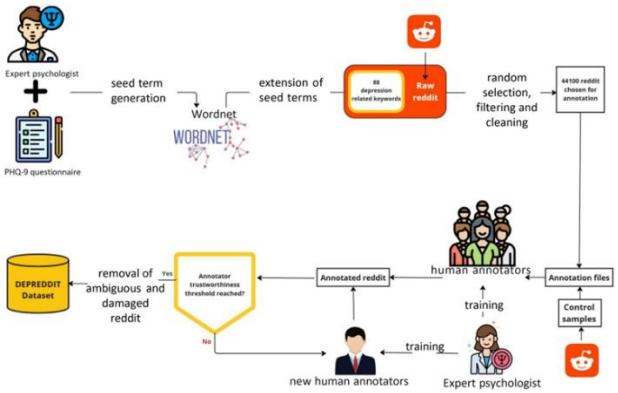

Fig. 2: Overview of the dataset creation process

Prior qualitative research on anxiety on social media and previous success in identifying depression on social media point to the possibility of identifying anxiety and anxious behavior on social media. In this paper, we use a variety of linguistic features to attempt the first time to identify Reddit posts related to anxiety. As opposed to LIWC and N-gram models, we specifically examine how well vector-space representations and LDA features separate texts pertaining to anxiety from more ordinary texts.

3.1. Data Collection

Direct access to posts by subreddit is made possible by the comprehensive Reddit API. Over the course of three months, we gathered 22,808 Reddit posts for this experiment. These posts consist of 12,837 general posts titled ”Control” and 9971 posts titled ”Anxiety”. The majority of posts about anxiety are gathered from the r/anxiety subreddit; posts for the Anxiety class are also mined from three other anxiety-related subreddits: r/panicparty, r/healthanxiety, and r/socialanxiety. We also gathered posts for the Control class from a range of subreddits because the majority of posts about anxiety are written in the first person.

Table 1: Subreddits used for data collection

Anxiety subreddits	Control subreddits	
r/anxiety	r/askscience	r/parenting
r/healthanxiety	r/writingprompts	r/Christianity
r/socialanxiety	r/writing	r/jokes
r/panicparty	r/atheism	r/writing
	r/showerthoughts	r/talesfromretail
	r/lifeprotips	r/talesfromtechsupport
	r/personalfinance	r/talesfromcallcenters
	r/theoryofreddit	r/fitness
	r/randomkindness	r/frugal
	r/books	r/youshouldknow
	r/askdocs	r/relationships
	r/teaching	r/nostupidquestions
	r/legaladvice	

That use first-person accounts. Additionally, using a diverse mix of subreddits reduces the influence of words specific to a particular community. The subreddits included in each data category are listed in Table 1. The average post length in the Anxiety collection was 171.83 words (869.14 characters). The average length of the posts in the Control group was 164.82 words (846.28 characters). These counts show how many kens were processed, with HTML tags, URLs, and punctuation removed. By eliminating stop words and lemmatizing word tokens, we apply additional preprocessing.

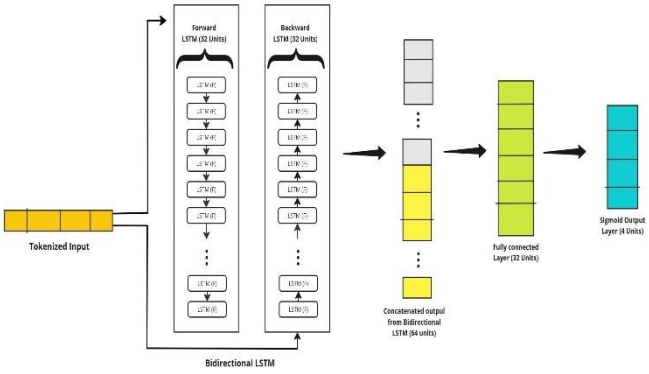

Fig. 3: Overview of the dataset creation process

3.2. Feature Generation

3.2.1. Vector space embeddings: Word2Vec and Doc2Vec

Mikolov et al. (2013) [29] introduced an efficient estimation of words in vector space for both skip gramandcontinuous bag-of-words (CBOW)models. With all training examples, we constructed a CBOW model with a window size of 5 words between current and predicted words in the sentence and use the mean of the context word vec tors. For training, we make 5 iterations over the corpus and use negative down sampling to draw 5 noise words to speed up training. We empirically select an embedding dimension of 300. With the CBOW model, we constructed feature vectors by taking the

mean of all tokens in each training ex ample. Intuitively, this corresponds to finding the center of the cluster of words in the vector space belonging to the target label category.

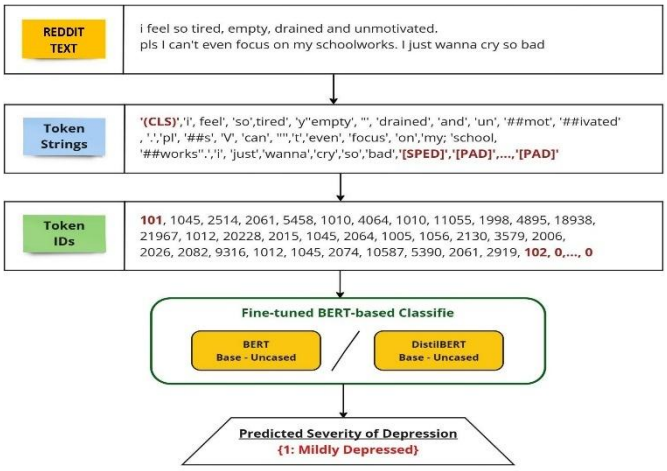

Fig. 4: Severity of depression prediction from a simple tweet

Paragraph context can be used to further improve predictive models. A distributed memory model with paragraph vectors (PV-DM) was presented by Le and Mikolov (2014). In addition to mapping each word to a distinct vector, each paragraph vector was mapped to a distinct vector. The current work uses gradient descent to infer para- graph vectors with every new training example, in addition to word vector updates, during feature-generation model training. The feature vector of the post is constructed using the paragraph vectors in addition to the word vectors. A sliding window over the paragraph is used to calculate contexts of fixed length. When predicting the next word, the contexts generate paragraph information that serves as a memory component to provide history.

3.2.2 LDA topic modeling

According to Blei et al. (2003), latent Dirichlet allocation (LDA) is a Bayesian generative technique that models textual bodies as a mixture of underlying latent topics, each of which is defined by a distribution over individual word. First, we create two LDA models for the Control and Anxiety classes, respectively, using the training set. We produce the latent unlabeled topics for every class after the LDA model has been trained. The ten subjects with the greatest information gain for both groups are listed in Table 2. Each training example is represented by a 20-dimensional array of likelihoods generated by the top 10 topics for each of the Anxiety and Control LDA models.

Table 2: Topics from the LDA model with the highest information gain (IG) in bits

Topic	Topic Words	IG
Ctrl. T1	like, want, know, go, said	0.1764
Anxi. T4	try,drive, look, car, walk	0.0850
Anxi. T2	drink, smoke, alcohol, weed, draw	0.0792
Ctrl. T2	pay,money, will, can, account	0.0699
Ctrl. T8	year,will, time, work, school	0.0691
Anxi. T3	school, class, year, college	0.0654
Anxi. T8	game,help, eat, food, play	0.0654
Anxi. T5	work,job, get, call, time	0.0654
Anxi. T10	people, feel, social, think	0.0654
Ctrl. T1	doctor, pain, medic, feel, feel	0.0654

3.2.3. N-gram language models

Another standard method used to extract features from text is to calculate the probability of a document within a language model. In our experiments, we use four different corpora to calculate probabilities of unigrams and bigrams. We build the first two models using the Anxiety and Control training examples, respectively. We build the third model using 100,000 unlabelled tweets from the Sentiment140 dataset (Go et al., 2009) and use the NLTK Brown corpus (Bird, 2006) for the fourth model. To generate feature vectors, we calculate the log-probability of each input sentence as uni grams and as bigrams. For each Reddit post, the associated feature subvector contains a unigram and bigram probability, with Laplace smoothing for each model, i.e., 8 dimensions in total.

3.2.4. Learning embeddings and topics

The type of model from which we extract features can be built using any corpus. Here, we compare using in-domain training examples with using an other corpus for building the word vector (with word2vec), document vector (with doc2vec), and topic (LDA) models. Here, we choose Twitter as a suitable candidate, since it constitutes a similar social media platform, and since previous literature used Twitter data. To compare, 100,000 tweets from Sentiment140 (Go et al., 2009) were used to build word2vec, doc2vec and LDA topic models. These models were further used to generate training and test feature vectors. Table 3 summarizes the different accuracies of using our Red- dit training set compared to using Twitter data to build the feature generation models. Higher accuracies were achieved when the models were trained with Reddit examples rather than with the 100,000 tweets for word2vec and LDA features. However, the Twitter-trained document vector model generated more effective feature vectors than the equivalent model from Reddit data. This result is likely due to the larger number of training examples used to build the Twitter doc2vec model. While word vectors are shared between documents, document vectors are always unique in each new document (Le and Mikolov, 2014). Compared to our Reddit corpus, this Twitter corpus includes a higher number of training documents but each document is shorter in length. Thus using the Twitter corpus in training vector representations may increase the complex- ity of the doc2vec model more than the word2vec model. To achieve higher accuracy while maintaining consistency, we hereafter use training examples to build models for feature vector generation.

Table 3: Accuracies from feature vectors generated by models trained with Reddit data vs Twitter data

	SVM		NN	
Project	REDDIT	TWITTER	REDDIT	TWITTER
Word2vec	0.906	0.813	0.900	0.786
Doc2vec	0.772	0.803	0.797	0.823
LDA	0.868A	0748	0.846	0.721

4. Results

4.1. Frequency

We use the entire labelled data set of 22,808 Reddit posts to compare lexicon differences. We calculate each unigram's frequency across the Anxiety and Control sets. To identify differentiating subsets, the top 200 unigrams for each category are sorted, and the unigrams that show up in both lists are eliminated. To identify the most common bigrams, we employ the same procedure. The top 10 most common bigrams and the top 15 most common unigrams for each group are compiled in Table 4.

Table 4: Frequent N-grams (N=1, 2) in each class

Category	Most frequent n-grams
Anxiety unigrams	anxiety, myself, anyone, social, panic, friends, feeling, having anxious, else, talk, bad, thought, better, felt
Control unigrams	our, call, us, edit, old, tell, phone, use, give, same, customer, post money, let, reddit
Anxiety bigrams	(my anxiety), (social anxiety), (my life), (anxiety and), (anxiety i) (anyone else), (talk to), (panic attacks), (panic attack), (where i)
Control bigrams	(we are), (from the), (we have), (she was), (he was), (thank you) (that the), (and he), (and she), (what is)

The bigrams and unigrams on anxiety specifically address anxiety and disorders associated with it, including panic attacks and social anxiety. The words ”feeling,” ”thought,” ”felt,” and ”bad” are among the most common anxiety unigrams. Unigrams and bigrams in the control data, on the other hand, contain Reddit-specific vocabulary (e.g., edit, post, Reddit). Certain customer and phone-related terms (such as call, phone, customer) that are not commonly found in the Anxiety group data are present in the control group data from r/talesfromcallcenters, r/talesfromtechsupport, and r/talesfromretail. Additionally, compared to the Anxiety set, the Control set usually has a higher number of third-person and first-person plural pronouns. However, there are more first-person singular pronouns in the Anxiety set's most common unigrams and bigrams.

5.4. Collocations

Studying collocations captures how groups of words are combined to produce meaning beyond the sum of individual component words. While N-gram frequencies in the previous section reveal how often words appear, identifying collocations can reveal important topics mentioned within a corpus. To find the collocations in both the Control and Anxiety posts, we again analyze the entire data set. Using the NLTK collocation library, we filter collocations by empirically selecting a minimum frequency of 100 for bigrams and 75 for trigrams. We then extract the 30 most collocated N- grams ranked by pointwise mutual information (Manning et al., 1999) from each of the Anxiety and Control sets. We also remove collocations that appear in both the Anxiety and Control collocation lists. Table 5 summarizes the top 10 most collocated bigrams and trigrams for both groups.

Table 5: Highly collocated N-grams (N=2, 3)

Category	Most frequent n-grams
Anxiety unigrams	anxiety, myself, anyone, social, panic, friends, feeling, hav ing, anxious, else talk, bad, thought, better, felt
Control unigrams	our, call, us, edit, old, tell, phone, use, give, same, cus tomer, post, money let, reddit
Anxiety bigrams	(my anxiety), (social anxiety), (my life), (anxiety and), (anxi ety i) (anyone else), (talk to), (panic attacks), (panic attack), (where i)
Control bigrams	(we are), (from the), (we have), (she was), (he was), (thank you), (that the) (and he), (and she), (what is)

4.3 Classification

For binary classification, Table 7 presents the 10-fold cross-validated accuracy and precision rates of using different feature types across logistic regression (LR), a linear kernel support vector machine (SVM), and a neural network (NN). SciKit-Learn was used to implement the LR and SVM classifiers (Pedregosa et al., 2011). We constructed a unique two-layer neural network with sigmoid activations and 256 hid- den units per layer. We experimentally employed a batch size of 500 and a learning rate of 0.01 for 200 iterations during optimization. Overall, all features are useful in classifying anxiety-related posts on Reddit. For single source features, we achieve the best results, of 91% accuracy, through word-vector embeddings (word2vec), and through N-gram probabilities. The performance of word2vec is slightly better than the word-vector techniques used by PreotiucPietro et al. (2015) on the Coppersmith Twitter corpus (Coppersmith et al., 2015). By contrast, using N-gram probabilities achieve an over all slightly better precision (92% with NN) than word2vec (91% with SVM). The LDA topic features also perform better than previous results using LDA to detect depression on Twitter (Resnik et al., 2015). Whether topic modelling is more ap propriate for long-form posts, as in our data, is the subject of future work.

Since our data did not include meta-data, we implemented content-based features from De Choudhury et al. (2013)

including emotion, linguistic style (from LIWC 2007), and an anxiety lexicon. In addition, we combined LIWC and LDA features from Nguyen et al. (2014). The accuracies and precisions of these implementations, as well as the aggregate features, are summarized in Table 8. For combined meth- ods, our neural network classifier consistently produces the best results. We achieve

the highest accuracy of 91% by combining LIWC with N-gram probabilities and by combining word-vector embeddings (word2vec) with LIWC using this classifier. We improve classification accuracy by 1% over only using word2vec and by 6% over the LIWC-only baseline. Also, N-grams+LIWC (92%) achieves slightly higher precision than word2vec+LIWC (91%), which is consistent with the difference in N-grams-only and word2vec-only results. Combined models, specifically word2vec+N-gram probabilities, word2vec+LDA, and LIWC+LDA (Nguyen et al., 2014), achieve comparable results with 90%, 87%, and 88% precision, respectively.

For all accuracy and precision values in Table 6, the associated recall was high, ranging between 75% and 91% depending on the classifier. The neural network classifier consistently produced recall values above 85% with variances in the order of 10−4. The SVM classifier produced the lowest recall (75%-87%) with larger variances in the order of 10−2. This fluctuation may be due to using a linear kernel, which has a lower representational power than a non-linear kernel.

Table 6: Performance comparison of different features across models (LR, SVM, NN).

Feature 2-7	LR		SVM		NN	
	Accuracy	Precision	Accuracy	Precision	Accuracy	Precision
word2vec	0.87	0.89	**0.91**	**0.91**	0.90	0.91
doc2vec	0.78	0.75	0.77	0.76	0.89	0.78
LDA	0.80	0.80	0.87	0.87	0.85	0.87
LIWC	0.85	0.85	0.71	0.81	0.82	0.83
N-grams	0.90	0.89	0.85	0.86	**0.91**	**0.92**

BERT outperforms DistilBERT across all evaluated metrics. Specifically, BERT achieves a precision of 93.5%, recall of 92.8%, and an F1-score of 93.1%, making it the more accurate model overall. In comparison, DistilBERT achieves a slightly lower precision of 91.2%, recall of 90.5%, and F1-score of 90.8%. The performance gap between BERT and DistilBERT can be attributed to the trade-offs inherent in DistilBERT's design, which aims to reduce model size and computational complexity while maintaining acceptable performance. The difference of approximately 2.3% in precision, 2.3% in recall, and 2.3% in F1-score demonstrates that while DistilBERT offers faster inference and lower resource requirements, it sacrifices some accuracy compared to the full BERT model.

Table 7: Comparison of Performance Metrics between BERT and DistilBERT

Model	Precision (%)	Recall (%)	F1-score (%)
BERT	93.5	92.8	93.1
DistilBERT	91.2	90.5	90.8

The performance comparison in Table 8 highlights the strengths and trade-offs across different models.

Logistic Regression (LR) achieves a precision of 89.0%, with its best performance observed when using n-grams. This result demonstrates the model's effectiveness in leveraging simpler features for classification tasks.

The Support Vector Machine (SVM) model outperforms LR slightly, achieving a precision of 91.0%, with its best performance attained using word2vec embeddings. This highlights the SVM's ability to effectively utilize dense vector representations for improved accuracy.

The Neural Network (NN) model achieves a precision of 92.0%, excelling when combined with n-grams. Its high precision indicates the model's capability to generalize well when handling textual data.

BERT delivers the best overall performance, with a precision of 93.5% and recall of 92.8%. Its superior results across all metrics showcase its state-of-the-art contextual representation capabilities.

DistilBERT, while slightly behind BERT, still achieves strong results with a precision of 91.2% and recall of 90.5%. As a lighter and faster alternative to BERT, it provides a practical trade-off between computational efficiency and performance.

Overall, the choice of model depends on the task requirements. For maximum accuracy, BERT is the optimal choice, whereas DistilBERT offers a balance of performance and efficiency. Simpler models like SVM and NN remain competitive in scenarios with limited computational resources or smaller datasets.

Table 8: Detailed Performance Comparison of Models

Combined Model	Precision (%)	Recall (%)	Comments
LR	89.0	-	Best precision obtained with n-grams.
SVM	91.0	-	Best performance with word2vec.
NN	92.0	-	Excellent precision and recall with n-grams.
BERT	93.5	92.8	Superior performance in all metrics.
DistilBERT	91.2	90.5	Lighter version of BERT with slightly lower performance.

6. Discussion

The LIWC 2015 dictionary provides sufficient coverage of anxiety-related word usage to successfully classify Anxiety and Control Reddit posts. However, by combining LIWC features with N-gram probabilities or unsupervised feature generation tech- niques (i.e., vector space embeddings and LDA Topic modeling), we can elevate the classification accuracy to 98%. Moreover, we find correlations between

anxiety and specific LDA topics such as school and alcohol (and drug) consumption (see Table 2).

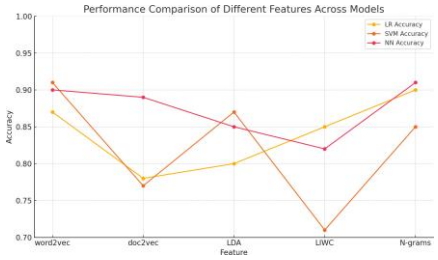

Fig. 5: Performance Comparison of Different Features Across Models

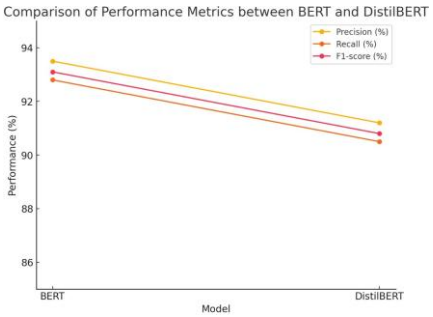

Fig. 6: Comparison of Performance Metrics between BERT and Distil- BERT

This could be an effective method of identifying topics that people with anxiety or other mental illnesses discuss on line. By counting unigram and bigram frequency, we also find lexicons relating to feelings and first person, singular pronouns predominantly represented in the Anxiety group. Furthermore, studying frequent collocations suggests that authors of anxiety-related posts are looking to find other people sharing similar experiences with anxiety. Due to the relatively recent popularity in the plat- form, little work has involved the linguistic aspects of Reddit, compared to Twitter. The lengths of posts and community organization of the web site suggests considerable potential for sophisticated methods of feature extraction as well as qualitative analysis. Despite the wide prevalence of anxiety disorders, few attempts have been made to create models capable of automatically detecting the disorder.

Acknowledgment

I would like to express my deepest gratitude to my supervisor, Professor Korchiyne Redouan, for his invaluable guidance, support, and encouragement throughout this research. His expertise and constructive feedback were essential to the success of this work. Finally, we appreciate the contributions of all participants and stakeholders involved in the study.

7. Conclusion

This study presents a novel approach that integrates validated psychological tools, such as the PHQ-9 and GAD-7, with advanced techniques for analyzing social media data to improve the early detection of mental health disorders, specifically depression and anxiety. By bridging the gap between traditional diagnostic methods and the potential of digital data analysis, this research highlights the advantages of combining self-reported assessments with behavioral and linguistic patterns derived from online activity. Such integration provides a richer and more nuanced understanding of an individual's mental state, enabling early identification of at-risk individuals and facilitating timely interventions.

The findings suggest that natural language processing (NLP) and deep neural net- works, when applied to textual and behavioral data from social media platforms, can uncover subtle mental health signals often missed by traditional methods. This dual approach not only enhances the sensitivity and specificity of detection models but also addresses the growing need for scalable, non-invasive, and real-time mental health assessment tools. However, the study also underscores the challenges inherent in this field, including the need for robust clinical validation, the handling of cultural and linguistic variability, and the risk of data biases that may impact the generaliz- ability of the proposed models.

Future directions for this research should focus on expanding the dataset to include diverse populations and multimodal data sources, such as images, videos, and meta- data, which could further enrich the predictive capabilities of the models. Additionally, the ethical implications of using social media data for mental health detection must be carefully addressed, including privacy concerns and ensuring informed consent from data subjects. Rigorous validation through collaboration with healthcare professionals and clinical trials will be crucial to transforming these theoretical models into practical tools for widespread adoption.

By advancing the intersection of artificial intelligence, psychology, and public health, this study contributes to the ongoing effort to create proactive and accessible mental health care systems. The proposed methodology has the potential to revolutionize early detection, offering healthcare providers powerful tools to mitigate the societal and personal burdens of mental illness through timely, personalized, and effective interventions.

7. References

[1] Organization, W.H.: Depression and Other Common Mental Disorders: Global Health Estimates. World Health Organization (2017)

[2] Organization, W.H.: The Global Burden of Anxiety Disorders. World Health Organization (2018)

[3] Corrigan, P.W., Watson, A.C.: Understanding the impact of stigma on people with mental illness. World Psychiatry (2002)

[4] Kroenke, K., Spitzer, R.L., Williams, J.B.: The phq-9. Journal of General Internal Medicine 16(9), 606–613 (2001)

[5] Spitzer, R.L., Kroenke, K., Williams, J.B., L¨owe, B.: A brief measure for assessing generalized anxiety

disorder: The gad-7. Archives of Internal Medicine (2006)
[6] Moreno, M.A., Goniu, N., Moreno, P.S., Diekema, D.: Ethical issues in social media research. Pediatrics (2013)
[7] De Choudhury, M., Gamon, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: Proceedings of the International Conference on Web and Social Media (ICWSM) (2013)
[8] Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.A., Boyd- Graber, J.: Beyond lda: Using vector space representations to identify traits of depression. In: Proceedings of the Workshop on Computational Linguistics and Clinical Psychology (CLPsych) (2015)
[9] Pedersen, T.: Screening for anxiety disorders using social media text. Computa- tional Linguistics (2015)
[10] Calvo, R.A., Milne, D.N., Hussain, M.S., Christensen, H.: Natural language processing in mental health applications using social media data. Biomedical Informatics (2017)
[11] Kroenke, K., Spitzer, R.L., Williams, J.B.W.: The phq-9: Validity of a brief depression severity measure. Journal of General Internal Medicine 16(9), 606–613 (2001)
[12] L¨owe, B., Kroenke, K., Herzog, W., Gr¨afe, K.: Measuring depression outcome with a brief self-report instrument: Sensitivity to change of the phq-9. Journal of Affective Disorders 81(1), 61–66 (2004)
[13] Spitzer, R.L., Kroenke, K., Williams, J.B.W., L¨owe, B.: A brief measure for assessing generalized anxiety disorder: The gad-7. Archives of Internal Medicine 166(10), 1092–1097 (1999)
[14] Kroenke, K., Spitzer, R.L., Williams, J.B.W., Monahan, P.O., L¨owe, B.: Anxiety disorders in primary care: Prevalence, impairment, comorbidity, and detection. Annals of Internal Medicine 146(5), 317–325 (2007)
[15] De Choudhury, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM), pp. 128–137 (2013)
[16] Guntuku, S.C., Yaden, D.B., Kern, M.L., Ungar, L.H., Eichstaedt, J.C.: Detecting depression and mental illness on social media: An integrative review. Current Opinion in Psychology 36, 43–48 (2019)
[17] Resnik, P., Armstrong, W., Claudino, L., Nguyen, T., Nguyen, V.A., Boyd- Graber, J.: Beyond lda: Using vector space representations to identify traits of depression. In: Proceedings of the NAACL Workshop on Computational Linguistics and Clinical Psychology, pp. 99–107 (2015)
[18] Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., Zhu, T.: Depression detec- tion via harvesting social media: A multimodal dictionary learning approach. In: Proceedings of the 27th ACM International Conference on Information and Knowledge Management (CIKM), pp. 383–392. ACM, ??? (2018)
[19] Kroenke, K., Spitzer, R.L., Williams, J.B.W.: The phq-9: Validity of a brief depression severity measure. Journal of General Internal Medicine 16(9), 606–613 (2001)
[20] De Choudhury, M., Counts, S., Horvitz, E.: Predicting depression via social media. In: Proceedings of the 7th International AAAI Conference on Weblogs and Social Media (ICWSM), pp. 128–137 (2013)
[21] Benton, A., Mitchell, M., Hovy, D.: Multitask learning for mental health con- ditions with limited social media data. In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics, pp. 152–162 (2017)
[22] Eichstaedt, J.C., Smith, R.J., Merchant, R.M., Ungar, L.H., Crutchley, P., Preo¸tiuc-Pietro, D., Schwartz, H.A.: Facebook language predicts depression in medical records. Proceedings of the National Academy of Sciences 115(44), 11203–11208 (2018)
[23] Park, M., McDonald, D.W., Cha, M.: Perception differences between the depressed and non-depressed users in twitter. In: Proceedings of the 9th Inter- national Conference on Weblogs and Social Media (ICWSM), pp. 244–253 (2015)
[24] Lin, H., Jia, J., Guo, Q., Xue, Y., Li, Q., Huang, J., Feng, L.: User-level psycho- logical stress detection from social media using deep learning. Journal of Medical Internet Research 22(3), 15723 (2020)
[25] Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K.: From adhd to sad: Analyzing the language of mental health on twitter through self-reported diag- noses. In: Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology (CLPsych), pp. 1–10 (2015)
[26] Nadeem, M.: Identifying depression on twitter. In: Proceedings of the ACL Student Research Workshop, pp. 1–8 (2016)
[27] Preo¸tiuc-Pietro, D., Guntuku, S.C., Ungar, L.H.: Discovering user attributes from social media text using word embeddings. In: Proceedings of the NAACL-HLT, pp. 1–10 (2015)
[28] Fernandez, K.C., Levinson, C.A., Rodebaugh, T.L.: Profiling: Facebook status updates reveal the presence of social anxiety. Social Psychological and Personality Science 3(6), 706–713 (2012)
[29] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed repre- sentations of words and phrases and their compositionality. In: Advances in

Neural Information Processing Systems, vol. 26, pp.
3111–3119. Curran Associates, Inc., (2013)