DEEP LEARNING TECHNIQUES FOR CATEGORIZING USER GENERATED TEXT ON THE INTERNET

A Thesis Submitted in Partial Fulfillment of the Requirements for the Degree of

DOCTOR OF PHILOSOPHY

By

Ms. ANSHU MALHOTRA
(2K16/PHDCO/09)

Under the supervision of

Prof. (Dr.) Rajni Jindal
Professor, Department of Computer Science and Engineering
Delhi Technological University, Delhi

Department of Computer Science and Engineering

DELHI TECHNOLOGICAL UNIVERSITY
(Formerly Delhi College of Engineering)
Shahbad Daulatpur, Main Bawana Road, Delhi-110042. India

July 2024

Dedicated to The Tripod of My Life:

Mummy, Papa & Bhai

@ Mummy: More than mine, it is your Achievement.

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

DECLARATION

I, Anshu Malhotra, PhD research scholar (Roll No. 2K16/PHDCO/09), hereby certify that the work which is being presented in the thesis entitled "**Deep Learning Techniques for Categorizing User Generated Text on the Internet**", in partial fulfilment of the requirements for the award of the Degree of Doctor of Philosophy, submitted in the Department of Computer Science & Engineering, Delhi Technological University, Delhi, India is an authentic record of my own bonafide research work carried out during the period from August, 2016 to July 2024 under the supervision of Prof. (Dr.) Rajni Jindal. I further declare that the work presented in the thesis has not been submitted by me for the award of any other degree of any other university or institution.

Ms. Anshu Malhotra
Roll No: 2K16/PHDCO/09
Department of Computer Science and Engineering
Delhi Technological University, Delhi, India

Date: 25th July 2024

Place: Delhi

DELHI TECHNOLOGICAL UNIVERSITY

(Formerly Delhi College of Engineering)

Shahbad Daulatpur, Main Bawana Road, Delhi-42

CERTIFICATE

This is to certify that the research work presented in the thesis entitled **"Deep Learning Techniques for Categorizing User Generated Text on the Internet",** submitted by **Ms. Anshu Malhotra** (Roll No: 2K16/PHDCO/09), for the award of the Degree of Doctor of Philosophy, from the Department of Computer Science & Engineering, Delhi Technological University, Delhi, India is an authentic work carried out by her under my supervision. The thesis embodies results of original work and studies carried out by the student herself and the contents of the thesis do not form the basis for the award of any other degree to the candidate or to anybody else from this or any other University/Institution.

Supervisor

Prof. (Dr.) Rajni Jindal
Professor
Department of Computer Science and Engineering
Delhi Technological University, Delhi, India

Date: 25th July 2024

Place: Delhi

ACKNOWLEDGEMENTS

I owe a debt of gratitude to my supervisor, Prof. (Dr.) Rajni Jindal, Professor, Department of Computer Science and Engineering, Delhi Technological University, Delhi, India, for her invaluable guidance, constant support, patience, and encouragement at every step during my PhD. Her dedication and insightful advice have helped me navigate challenges and foster a profound enthusiasm for this research work. I am truly fortunate to have her as my PhD advisor.

I would like to extend deep gratitude to the Head of Department and all the faculty members of the Computer Science & Engineering Department at Delhi Technological University, for their continual support and motivation. I am also thankful to all the staff members and research fellows for their tireless assistance and help throughout these years.

I am particularly grateful to my special friend, Dr. Pratima Sharma, who helped me stay positive and motivated throughout my PhD research. I am thankful to my fellow research scholars, especially Dr. Amrita Sisodia, Dr. Minni Jain, and Ms. Kim for their limitless support.

A special note of thanks goes to my loving brother Aakarsh, for always being there for me whenever I needed anything and helping me pull through the all-nighters needed for research work. For a part-time research scholar like me, he has been an at-home virtual research lab. He has been an indispensable part of my PhD journey, and I feel fortunate to have had him around. I am lucky to have friends like Sapna, Garima, Namita, Sameer, and Latika, who kept me sane and joyful during my difficult times.

Finally, and above all, I would like to thank my parents, Mrs. Alka Malhotra and Mr. Anil Malhotra, who are the cornerstone of my journey. They have been my most significant pillars of strength, supporting me till this day with their love, unwavering encouragement, and steadfast faith in me. Their help in balancing my personal and professional life along with my PhD research has been crucial. Completing this research would never have been possible without them. I feel truly blessed to have them as my parents.

Ms. Anshu Malhotra
Roll No: 2K16/PHDCO/09
Department of Computer Science and Engineering
Delhi Technological University, Delhi, India

ABSTRACT

The Internet of the present day, popularly known as Web 3.0, is phenomenally different from the Internet that was developed decades ago. It is no longer a one-way channel for information dissemination to the users. Today, the Internet sustains and thrives on the content provided by its users. Internet users are no longer just information consumers; rather, they are content or information producers as well. With the Internet becoming an indispensable part of our lives, today, people spend a significant amount of their time on the Internet, thereby creating a humungous amount of User Generated Content (UGC) as a by-product, e.g., product reviews, social media posts, etc. UGC content can be of a multimodal and multilingual nature. In the last decade, various research and real-world applications of UGC have been proposed and developed using Artificial Intelligence and Machine Learning, e.g., opinion mining, trend prediction, sentiment analysis, public health monitoring, etc. The objective of the research presented in this thesis is to study the applications of *Deep Learning Techniques for Categorizing User Generated Text on the Internet*. This research work presented in this thesis makes the following significant contributions.

First, we have conducted an in-depth systematic literature review to understand the state-of-the-art, highlight research gaps in existing work, and identify open challenges related to research applications of deep learning techniques for user generated content available on the Internet for various real-world social computing applications.

Second, we have reviewed, compared, and empirically evaluated all popular supervised deep neural networks to benchmark their performance for a real-world application of user generated text categorization tasks.

Third, the primary contribution of our research work is that we have proposed an explainable and interpretable system for supervised and unsupervised categorization of user generated text from the Internet by using the latest breakthrough techniques in deep learning for NLP domain, i.e., Transformer based LLMs. We have conducted extensive and in-depth experiments with six LLMs (BERT, DistilBERT, RoBERTa, MentalBERT, PsychBERT, PHSBERT) and four datasets. For explainability and interpretability (XAI) of predictions from the above deep learning models, we have used the two most recent techniques: LIME and SHAP. Next, we have demonstrated the use of the Transformer-based unsupervised topic modeling technique BERTopic to analyze large-scale unlabeled UGC datasets for deriving insights.

Fourth, we have performed Few Shot Learning and Active Learning experiments with pretrained LLMs, which can be beneficial for low resource research domains where good quality, large annotated UGC datasets are unavailable. For these scenarios, pre-trained LLMs can be trained with only a few good quality data samples annotated by experts using the above deep learning paradigms. Experiments were done with various LLMs for multiple datasets to analyze and compare their performance. We have demonstrated that it is possible to achieve high/comparable accuracy with even less than 10% of samples from the entire dataset.

At last, we have conducted preliminary work to extend our research to categorizing multimodal user generated content on the Internet by exploring the use of recent innovative advancements in the field of deep learning for other modalities, i.e., images and videos. We have proposed a deep transfer learning framework for affective analysis of multimodal user generated content from the Internet.

The review, analysis, empirical evaluations, and experimental results demonstrate the applications of proposed explainable deep learning techniques for social computing applications using text from the Internet. This thesis successfully helps advance the research related to the applications of deep learning techniques for categorizing user generated content from the Internet.

CONTENTS

ABBREVIATIONS & ACRONYMS

ACC	Accuracy
AL	Active Learning
ADODL	Average Difference in Overall Depression Levels
AE	Autoencoder
AHR	Average Hit Rate
AI	Artificial Intelligence
ANEW	Affective Norms for English Words
ANN	Artificial Neural Network
AUC	Area Under Curve
BART	Bidirectional and Auto-Regressive Transformer
BDI	Beck Depression Inventory
BERT	Bidirectional Encoder Representations from Transformers
Bi-GRU	Bidirectional Gated Recurrent Unit
Bi-LSTM	Bidirectional Long Short-Term Memory
Bi-RNN	Bidirectional Recurrent Neural Network
BOW	Bag of Words
CBOW	Continuous Bag Of Words
C-SSRS	Columbia-Suicide Severity Rating Scale
CNN	Convolution Neural Network
DBM	Deep Boltzmann Machine
DBN	Deep Belief Network
DL	Deep Learning
DNN	Deep Neural Network
DT	Decision Tree
ERDE	Early Risk Detection Error
ELMo	Embeddings from Language Models
FC	Fully Connected
FSL	Few Shot Learning

FF	Feed Forward
F1	F1 Score
GAN	Generative Adversarial Network
GCN	Graph Convolution Network
GNN	Graph Neural Network
GPT	Generative Pre-trained Transformer
GRU	Gated Recurrent Unit
HAN	Hierarchical Attention Network
HNN	Hierarchical Neural Network
LDA	Latent Dirichlet Allocation
LIME	Local Interpretable Model-agnostic Explanations
LIWC	Linguistic Inquiry and Word Count
LR	Logistic Regression
LSTM	Long Short-Term Memory
LSA	Latent Semantic Analysis
MAE	Mean Absolute Error
MH	Mental Health
ML	Machine Learning
MLP	Multi-Layer Perceptron
MTL	Multi-Task Learning
MSE	Mean Squared Error
NB	Naïve Bayes
NER	Named Entity Recognition
NLP	Natural Language Processing
NN	Neural Network
NRC	National Research Council Canada (Affect Lexicon)
OSN	Online Social Network
P	Precision
R	Recall
RBM	Restricted Boltzmann Machine
RF	Random Forest
RNN	Recurrent Neural Network
SHAP	SHapley Additive exPlanation

SoWE	Suicide-oriented Word Embeddings
SLR	Systematic Literature Review
STL	Single Task Learning
SVM	Support Vector Machines
TF-IDF	Term Frequency - Inverse Document Frequency
UMLS	Unified Medical Language System
UGC	User Generated Content
VADER	Valence Aware Dictionary and sEntiment Reasoner
VAE	Variational Autoencoder
WHO	World Health Organization
XGB	eXtreme Gradient Boosting
XAI	Explainable AI
ZSL	Zero-Shot Learning

LIST OF FIGURES

LIST OF TABLES

CHAPTER 1

INTRODUCTION

This chapter provides an introduction to the research work presented in this thesis. It begins with background details related to User Generated Content from the Internet and its applications, and also briefly introduces Deep Learning techniques for Natural Language Processing. It discusses the research motivation, problem statement, and key research contributions. The chapter concludes with the documentation of the outline and organization of this thesis.

The Internet, invented by Vint Cerf, Bob Kahn, and Tim Berners Lee back in the 1980s, has indeed come a long way and transformed into something that probably was never imagined by its creators. What started as a read-only, structured collection of hyperlinked documents over TCP/IP decades ago (popularly referred to as Web 1.0) has now become decentralized, unstructured, omnipresent, and ubiquitous Internet of today (Web 2.0 and Web 3.0). The users are no longer just information consumers but also the creators of the content available over the web. The User Generated Content (UGC) is noisy, unstructured, heterogenous, multimodal, and multilingual in nature and can comprise of text, photos, videos, audio, reels, memes, etc. This massive amount of publicly available information, if processed and analyzed efficiently, can be used to develop a plethora of innovative web applications and functionalities for Internet users.

The fundamental change in the structure of WWW applications has made it more communication-oriented and inclusive for all, where anyone can connect with anyone and freely express themselves over various online social networks and discussion forums/communities. Human communication is shifting from the real to the virtual world, and the trend has only accelerated in the last two years during the COVID-19 pandemic. A large share of human social interactions now happens over the web. With this paradigm shift over the last decade, research in the domain of social media analytics has focused on tapping this form of online communication between users to develop Artificial Intelligence (AI) and machine Learning (ML) based real-world applications for various social computing tasks, e.g., sentiment analysis, opinion mining, cyberbullying detection, etc. However, one of the most crucial of these applications is public health monitoring and surveillance on online social networks [1] [2] [3]. Hence, to conduct the research work reported in this thesis, we have focused on this real-world application of User Generated Text (UGT) available on the Internet. We focus our research on leveraging Deep Learning Techniques for Categorizing User Generated Text available on the Internet (Social Media) for Mental Healthcare Applications. We elaborate more on the chosen research theme and problem of this thesis in the following subsections.

1.1 Deep Learning Techniques

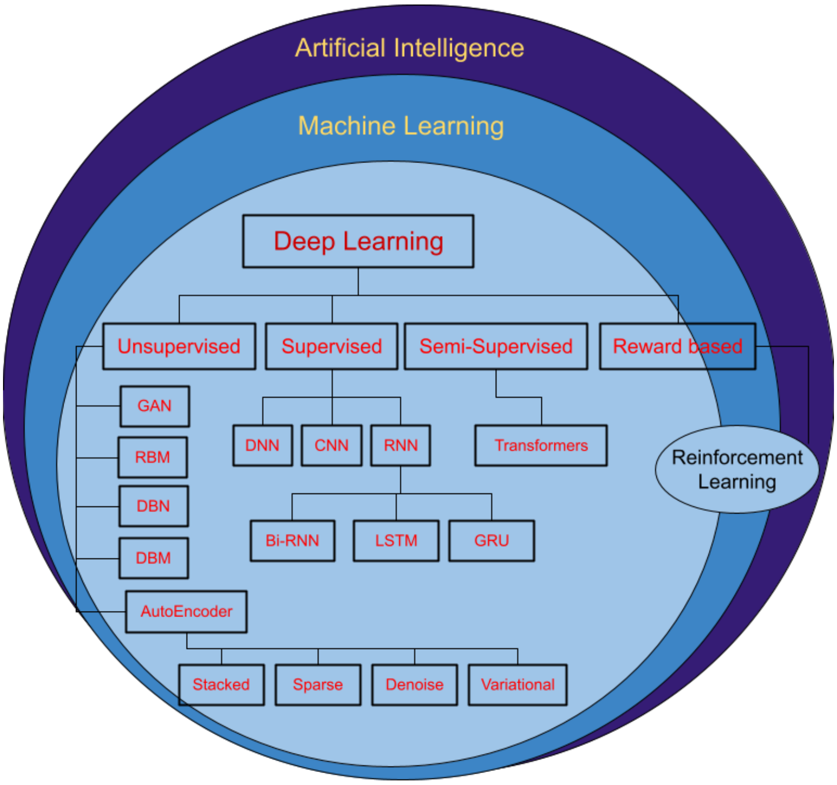


Figure 1.1 *Taxonomy of Deep Learning Techniques*

Machine learning techniques are computational algorithms that can automatically discover and learn hidden patterns from historical data via training (supervised, unsupervised, semi-supervised, reinforced [4] [5]) and are used to develop expert decision-making systems that can make predictions on new unseen data [4] [5]. Deep Learning (DL) algorithms are a subgroup of ML techniques that refer to deep neural networks that are complex, multi-layered networks made of interconnected artificial neurons (or perceptron) that mimic the functioning of the human brain. These are feed-forward neural networks with error backpropagation [6]. Typically, a DL NN consists of a single input layer, multiple hidden layers (configurable), and a single output layer. The number of neurons in the input layers is equal to the size of a single training data instance, and the number of nodes in the output layer is equal to the number of target/output classes. Neurons are the decision-making units in all layers. They transform the output of their previous layer using their activation functions and weights and pass them as input for their subsequent layer. During the training phase, the neural network tries to learn these network weights to minimize model loss and prediction errors [7] [8]. As shown in Figure 1.1, some of the most commonly used supervised DL networks are: ANN, DNN, CNN, RNN, LSTM, GRU, and their Bi-directional variants (Bi-RNN, Bi-LSTM, and Bi-GRU); whereas some of the popular unsupervised deep learning techniques are: GAN, RBM, DBN, DBM, and Autoencoders [6] [7] [8]. Deep learning techniques, especially Transformer based Large Language Models, have become state of the art for all research domains due to their phenomenal classification accuracy as they can discover and learn complex, non–linear, hierarchical abstract patterns from unstructured data (text, images, sound, videos). UGC from OSN is also unstructured data in the form of text, images, videos, and memes posted by users, and hence, that explains the growing popularity of DL for social media analytics as well. In the next chapter, we elaborately explain some of these popular deep learning techniques which have been used by related research studies for categorization of user generated content from the internet.

1.2 Research Motivation

Mental health, i.e., people's psychological and emotional well-being, is one of the most neglected public health concerns that deserves appropriate attention by government and healthcare bodies. Some of the common mental disorders are: depression, anxiety, stress, PTSD, panic attacks, bipolar and borderline personality, schizophrenia, eating and sleep problems, substance abuse, etc. [9]. Mental health disorders, particularly depression, can also trigger self-harming behavior, suicide ideation, and attempts [10]. Depression is the most common and severe mental health issue and is also the principal cause of physical disability worldwide [11]. We believe the two issues, i.e., depression and suicide/self-harm, are closely associated and are of the most importance. According to WHO, suicide is one of the leading causes of death worldwide and is considered a critical public health concern [12] [13]. Hence, timely diagnosis and treatment of mental health disorders is also equally important, like physical diseases. However, mental health issues are ignored due to multiple reasons such as lack of awareness, associated social stigma, and expensive and time-consuming clinical procedures, e.g., questionnaires, interviews, need for multiple sessions and continuous assessments with practitioners (psychologists, counselors, etc.), need for round the hour availability of emotional/mental support system. This kind of desired mental healthcare system is not very advanced or developed in most parts of the world yet. These reasons necessitate contemporary and innovative, nonintrusive technology solutions for the non-clinical diagnosis of human psychological health issues. WHO has also emphasized the need to develop such inter-disciplinary social computation web applications for mental health risk assessment and surveillance that can assist in the early detection and prevention of suicide and other mental health issues [10].

Motivated by the above reasons, for carrying out our research related to Deep learning techniques for categorizing textual UGC on the Internet, we have focused on the mental healthcare research domain and its real-world research application. We use various social media UGC datasets to demonstrate the applications of deep learning techniques for categorizing user-generated text on social media datasets to detect the following mental health disorders: depression and suicide/self-harm.

1.3 Social Media User Generated Text for Mental Healthcare Applications

Online social media platforms have become ubiquitous and the preferred medium of (instantaneous) communication, where people come together for one-to-one, one-to-many, and public interactions, predominantly to express their thoughts, emotions, feelings, and sentiments. The publicly available user-generated content from social media can be leveraged to detect early warning signs related to suicide, such as self-harm, depression, etc. Numerous research studies have proposed the use of conventional AI and ML algorithms along with the traditional NLP techniques for various mental healthcare surveillance tasks to detect mental health issues from social media users' posts [1] [2] [3] [14] [15] [16]. However, as is also evident from these studies, the predictive correctness of these conventional machine learning algorithms (e.g., LR, SVM, NB, etc.) depends heavily on the features selected for training them. It requires explicit and extensive feature engineering and selection techniques to capture relevant domain-specific knowledge and nuances. The feature engineering and selection process may not be straightforward for complex tasks such as mental health risk assessment from social media content. There are various domain-specific challenges, like: shared symptoms across various mental health disorders, variation in causes of mental health disorders across gender, age, demographics, life stage, etc., determining the difference in severity levels and deciding which user needs immediate assistance, temporal variation in user's mental state, no feedback loop or correlation with users' life events. To overcome these limitations, recent research in this domain has focused on applying deep learning techniques for mental health risk assessment from social media content, as these complex networks can detect complex patterns from big data without the need for multiple handcrafted features. Due to their multi-layered and nonlinear architecture, deep neural networks can automatically learn complex patterns and feature representations at multiple abstraction levels for such complex problems from big heterogeneous datasets. The research reported in this thesis also contributes vastly to advancing the state-of-the-art for the domain of mental healthcare applications of user-generated content from social media.

1.4 Research Objectives and Problem Statement

The main high-level objective of this research is to develop a framework for categorizing user generated text on the internet by leveraging deep learning techniques. To accomplish this objective and to demonstrate the results of our proposed research work, we have chosen the research problem of detecting mental health disorders (specifically, suicide/self-harm risk and depression) from user-generated text on online social networks. The detailed research objectives are listed below:

1. To study and implement text pre-processing techniques required for cleaning the noisy and unstructured user generated text on the internet or social media.

2. To study and implement a few conventional machine learning techniques for text categorization.

3. To study and compare various deep learning based feature representation techniques (i.e., document representation models) for real-world application of user generated text classification problems.

4. To identify and collect gold standard benchmark datasets of user-generated text for the chosen research problem in order to conduct further research and analysis.

5. To review, implement, and empirically compare the performance of various deep learning techniques for a real-world application of user generated text classification tasks.

6. To design and implement a novel framework using deep learning techniques for categorizing user generated texts for a real-world application on the Internet.

7. To extend the scope of the research by exploring the use of recent innovative techniques like: Transfer Learning, Active Learning, and Multimodal Deep Learning.

1.5 Research Contribution of the Thesis

The key research contributions of this thesis are as follows:

1. We have conducted an in-depth systematic literature review to understand the state-of-the-art related to deep learning techniques for categorizing user generated text on the Internet (specifically, from online social media for mental healthcare applications). Through this survey, we identified and enumerated gold standard datasets of user-generated content for the chosen application problem to enable future research and analysis in this domain. The survey has greatly helped in understanding the current state-of-the-art, research gaps, open challenges, and future research directions for advancing research applications of deep learning techniques to user generated content available on the Internet for various real-world social computing applications.

2. We have reviewed, compared, and empirically evaluated all popular supervised deep neural networks to benchmark their performance for a real-world application of user generated text

categorization tasks. We have used two publicly available mental healthcare UGC datasets to accomplish this research objective.

3. The primary contribution of our research work is that we have proposed an explainable and interpretable system for supervised and unsupervised categorization of user-generated text from the Internet or online social networks by using the latest breakthrough techniques in deep learning for NLP domain, i.e., Transformer-based LLMs. These are essentially black box models whose decisions are difficult to understand and explain, which limits their adoption for real-world applications. We have used surrogate, model agnostic techniques LIME and SHAP to provide post hoc explainability and interpretability to supervised models' results. Additionally, we have proposed and demonstrated the use of the unsupervised LLM model, namely, BERTopic, to derive interpretable insights from big UGC datasets that are difficult to label. Our proposed system can be used for any real-world application and with any open-source pretrained LLM checkpoint available.

4. We have performed Few Shot Learning experiments with pretrained LLMs that have already been adapted for a related real-world domain, in our case (mental) healthcare. This Transfer Learning approach can especially be useful to leverage the benefits of deep learning techniques for low-resource scenarios when only a few labeled samples are available or when it is not feasible to annotate large UGC datasets. In these scenarios, LLMs can be fine-tuned with only a few good quality samples annotated by experts.

5. We have demonstrated a proof-of-concept implementation of Deep Active Learning by training the current state-of-the-art Transformer-based LLMs with an Active Learning loop. We have demonstrated that with this training paradigm, it is possible to achieve high/comparable classification accuracy (as is obtained by training on the full available dataset) but instead with as few as 10% of the samples from the dataset. Deep Active Learning can be useful to mitigate the challenges associated with data annotation or in cases when very little labeled data is available.

6. We have conducted preliminary work on extending our research for categorizing multimodal user generated content from the Internet by using recent innovative advancements in the field of deep learning for other modalities, i.e., images and videos. We have proposed a deep transfer learning framework for the affective analysis of multimodal user-generated content.

1.6 Thesis Organization

The organization of the thesis is presented in this section, which comprises of seven chapters as listed below:

Chapter 1: Introduction

This chapter provides an introduction to the research work presented in this thesis. It begins with background details related to User Generated Content from the Internet and its applications, and also briefly introduces Deep Learning techniques for Natural Language Processing. It discusses the research motivation, problem statement, and key research contributions. The chapter concludes with the documentation of the outline and organization of this thesis.

Chapter 2: Preliminaries

This chapter provides the necessary technical details and preliminary background related to Deep Learning techniques. But first, through a proof-of-concept for public healthcare monitoring, the chapter explains and demonstrates the conventional ML and NLP techniques required for pre-processing, exploratory data analysis, and deriving insights from user generated text from the Internet.

The following paper has been published from this work:

- Jindal, R., & Malhotra, A. (2022). Efficacious Governance During Pandemics Like Covid-19 Using Intelligent Decision Support Framework for User Generated Content. In *Proceedings of Second Doctoral Symposium on Computational Intelligence: DoSCI 2021* (pp. 435-448). Springer Singapore. (DoSCI 2021, International Conference organized virtually by Institute of Engineering and Technology, Dr APJ Abdul Kalam Technical University, Lucknow, India on 06th March 2021) (Scopus)

Chapter 3: Literature Review

This chapter presents the research methodology, findings, analysis, and results from the in-depth systematic literature review that was conducted to understand the state-of-the-art related to research applications of deep learning techniques for categorizing user generated content available on the Internet for various real-world social computing applications (e.g., mental healthcare).

The following paper has been published from this work:

- Malhotra, A., & Jindal, R. (2022). Deep learning techniques for suicide and depression detection from online social media: A scoping review. *Applied Soft Computing*, *130*, 109713. (SCIE, IF: 8.7)

Chapter 4: Empirical Review & Evaluation of Deep Learning Techniques

This chapter presents the comparative results from an empirical review and evaluation of all popular supervised deep learning neural networks to benchmark their performance for a real-world UGC text categorization task using two publicly available mental healthcare datasets.

The following paper has been accepted from this work:

- Malhotra, A., & Jindal, R. (Accepted, In-Press). Social media analytics using deep neural networks for mental healthcare applications. In A. Khamparia & D. Gupta (Eds.), *Recent Advances in Computational Intelligence Applications for Biometrics and Biomedical Devices*. Elsevier. (Book Chapter). (Scopus) (Accepted, In-Press)

Chapter 5: Proposed System with Transformer-based LLMs and XAI

This chapter covers a detailed description of the proposed system for supervised and unsupervised categorization of user generated text from the Internet by using Transformer-based LLMs and explaining the model predictions using XAI techniques. It elaborately discusses the qualitative and quantitative results from experimental evaluation with multiple LLMs and user-generated text datasets.

The following paper has been published from this work:

- Malhotra, A., & Jindal, R. (2024). Xai transformer based approach for interpreting depressed and suicidal user behavior on online social networks. *Cognitive Systems Research*, *84*, 101186. (SCIE, IF: 3.9)

Chapter 6: Prototypes for Future Research Enhancements

This chapter demonstrates the preliminary research work done (prototypes / proofs-of-concept) for extending the scope of the research by exploring the use of recent innovative techniques like: Active Learning, Transfer Learning, and Multimodal Deep Learning for categorizing user generated content on the Internet.

The following papers have been published/accepted from this work:

- Malhotra, A., & Jindal, R. (2020). Multimodal deep learning based framework for detecting depression and suicidal behaviour by affective analysis of social media posts. *EAI Endorsed Transactions on Pervasive Health and Technology*, *6*(21). (Scopus) (Published)

- Malhotra, A., & Jindal, R. (2021). Multimodal deep learning architecture for identifying victims of online death games. In *Data Analytics and Management: Proceedings of ICDAM* (pp. 827-841). Springer Singapore. (ICDAM 2020 organized virtually by Jan Wyzykowski University Poland and B.M. Institute of Engineering and Technology, India on 18th June 2020) (Scopus) (Published)

- Jindal, R., & Malhotra, A. (Accepted & Presented, In-Press). Leveraging Deep Active Learning and Large Language Models for Cost Efficient Categorization of User Generated Content. In *Proceedings of Fifth International Conference on Data Analytics and Management 2024.* Springer. (ICDAM 2024 organized jointly by London Metropolitan University, London, UK on 14th -15th June 2024) (Scopus) (Accepted & Presented, In-Press)

Chapter 7: Conclusion

This final chapter discusses the summary of the work done, key takeaways, results, outcomes, limitations, conclusions, and future scope of this research work.

List of Publications from the Thesis: This section lists the papers related to this research work published/accepted/communicated in various International/National Journals/Conferences of repute.

References: This section is the list of references referred to in this research work.

CHAPTER 2

PRELIMINARIES

This chapter provides the necessary technical details and preliminary background related to Deep Learning techniques. But first, through a proof-of-concept scenario for public healthcare monitoring, the chapter explains and demonstrates the conventional ML and NLP techniques required for pre-processing, exploratory data analysis, and deriving insights from user generated text from the Internet.

2.1 ML and NLP Techniques for User Generated Text Processing

This section explains and demonstrates the conventional ML and NLP techniques required for pre-processing, exploratory data analysis, and deriving insights from user generated text from the Internet. These can be useful for data-driven decision making for various real-world use cases. We use the scenario for public healthcare monitoring as a proof-of-concept to discuss these techniques (Refer to Figure 2.1). For the purpose of this discussion, we have grouped these techniques into three categories based on their primary function: Text Pre-processing, Information Extraction, and Trend Prediction through Topic Modelling.

2.1.1 Text Pre-processing Techniques

Text pre-processing is a mandatory precursor for any analysis or system that leverages user-generated content from the Internet, e.g., popular social media platforms, mainly because the user-generated content is non-standardized, is of multimodal and multilingual nature, contains heterogeneous platform-specific information, contains noise and is error prone. Hence, before utilizing machine learning and big data and text analytics techniques, the following pre-processing steps become essential [17].

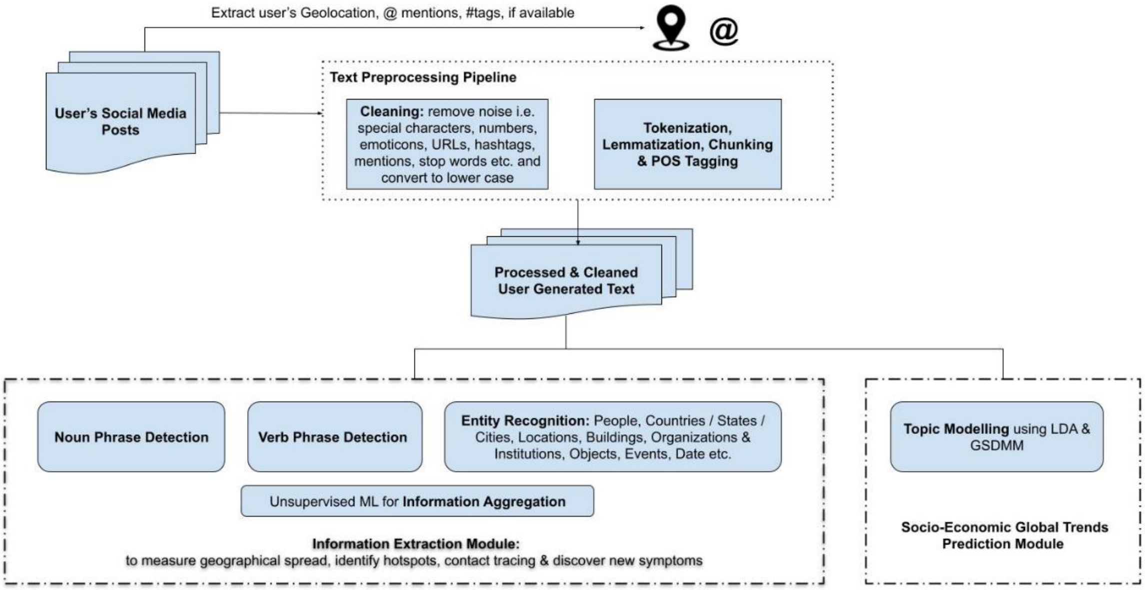

Extract user's Geolocation, @ mentions, #tags, if available
@
User's Social Media Posts
Text Preprocessing Pipeline
Cleaning: remove noise i.e. special characters, numbers, emoticons, URLs, hashtags, mentions, stop words etc. and convert to lower case
Tokenization, Lemmatization, Chunking & POS Tagging
Processed & Cleaned User Generated Text
Noun Phrase Detection
Verb Phrase Detection
Entity Recognition: People, Countries / States / Cities, Locations, Buildings, Organizations & Institutions, Objects, Events, Date etc.
Unsupervised ML for Information Aggregation
Information Extraction Module:
to measure geographical spread, identify hotspots, contact tracing & discover new symptoms
Topic Modelling using LDA & GSDMM
Socio-Economic Global Trends Prediction Module

Figure 2.1 *Public Health Monitoring from User Generated Text using ML & NLP Techniques*

1. Cleaning & Noise Removal: In order to standardize the user-generated text and enhance the data quality of input to subsequent NLP and ML algorithms that follow, it is essential to remove the noisy elements from texts like special characters, punctuations, numerics, emoticons, geolocation tags, @ mentions, # tags, and URLs. Platform specific non-textual information like: location tags, hashtags, and mentions are noise for any NLP-based system; however, in the scope of our current application, this information may be very useful for geolocation tracking, contact tracing, and identifying hotspots. Next, the stop words (like a, the, and etc.) are also removed, and at last, case conversion is done to bring uniformity as people usually are not very case-conscious while posting online.

2. Tokenization: This is a fundamental step of any NLP pipeline that is done in order to break or extract meaningful tokens from the input text document, sentence, or phrase. Tokens are the logical inputs to any NLP algorithm and can be created in 3 ways: word level, sub-word level, i.e., n-grams, or character level.

3. Lemmatization: This is the process of reducing the words from the document vocabulary to their root word from which they are derived in order to group together and analyze the different inflected forms of the same base word as a single entity. Unlike stemming, which is a very crude heuristic process that chops off the affixes of a word, lemmatization is done using proper grammatical and morphological rules and correct identification of parts of speech. This step reduces the dimensionality of the documents (in our case, user posts) and makes the feature matrix less sparse.

4. Chunking & POS Tagging: Using tokenization and lemmatization alone is not sufficient for all NLP applications. These are Bag of Words (tokens) based approaches that lead to loss of meaningful information about the semantic structure and actual meaning of the sentence. An alternate approach for text pre-processing is Chunking along with Part of Speech tagging. It basically refers to extracting phrases of words from the sentence to understand the logical sentence structure. It helps to derive various constituents from unstructured text, i.e., nouns, pronouns, adjectives, verbs, adverbs, prepositions, conjunctions, and interjections. Chunking and POS tagging are essential steps of Named Entity Recognition (NER), which is explained in the next sub-section.

2.1.2 Information Extraction Techniques

User-generated unstructured textual data (users' posts, comments, etc.) contain a vast amount of information, all of which may not be relevant. Information extraction is an NLP task done to retrieve the information of interest within the context or scope of current information needs and requirements, e.g., names of entities (person, organization names), the relationship between entities, a place/location, a date, sequence of events, actions, an idea, thought or a state of being, etc. For public healthcare monitoring, these pieces of information can be useful for measuring the geographical spread of disease/pandemic, identifying hotspots, detecting probable cases, contact tracing, and discovering new health indicators or symptoms. Using various NLP techniques explained below, these structured pieces of useful information can be extracted from free-flowing text [17].

1. Noun & Verb Phrase Detection: In any communication language, there are eight parts of speech that basically determine the grammatical role a word plays in the sentence. These are: nouns, pronouns, adjectives, verbs, adverbs, prepositions, conjunctions, and interjections. In the NLP domain, the task of determining and assigning a correct part of speech tag to each word in a sentence based on the role it plays is called POS Tagging. POS Tagging helps to understand sentence structure and build rules to identify and extract the relevant information of interest, e.g., the noun and verb phrases in the user’s posts. The POS tags can be aggregated and statistically analyzed to derive various insights from user-generated textual content on the Internet.

2. Entity Recognition: Another popular NLP and AI-based automated information extraction technique is Named Entity Recognition (NER), which can be used to augment the information retrieved from users' unstructured textual posts further. NER is the task of locating and identifying named atomic elements or entities from unstructured text and classifying them into predefined categories such as: people names, locations, organization and company names, date and time objects, quantifying measures, currencies, artifacts, etc. As per the English dictionary: an entity is defined as a thing or a concept with distinct characteristics and independent existence. Unlike POS tagging, which assigns part of speech tags to each word token, NER is able to extract entities that may be a single word or word phrases (chunks) referring to the same concept, thereby giving more meaningful learning and generating valuable insights from large volumes of unstructured text. Machine Learning models need to be trained with relevant language literature to make them learn the different entity categories and granular rules so that they can locate and identify the relevant entities from unstructured text. In the case of basic applications, one may even use a lexicon or rule-based NER system.

3. Information Aggregation: For data-driven real-time decision making using unstructured user-generated data for any real-world use case, the data streams need to be processed on a continuous basis. In order to handle the volume and velocity of the incoming unstructured data streams in real-time and for its efficient processing, it is essential to aggregate and categorize the information into meaningful buckets. For this purpose, unsupervised machine learning algorithms: K-Means Clustering, and Hierarchical Agglomerative clustering can be used to group and aggregate the semantically related and similar information extracted above. These clusters of valuable information pieces can be presented to the decision makers via a dashboard or a keyword-based search tool.

2.1.3 Trend Prediction through Topic Modelling

User generated text can be leveraged for trend prediction related to various social, economic, local, and global issues by using Topic Modelling techniques like Latent Dirichlet Allocation (LDA) and Gibbs Sampling Dirichlet Mixture Model (GSDMM) algorithm. Topic modeling is an unsupervised machine learning technique that builds a statistical topic model from raw unstructured text to discover hidden and abstract topics, themes, and ideas being discussed in them. Topic modeling is an effective technique to quickly understand and summarize large volumes of free-form text and extract meaningful insights when annotated or labeled data is not available.

LDA algorithm [18] builds a statistical model based on the distribution of words in any given input document by considering each document as a collection of topics, further where each topic is a collection of semantically related dominant keywords. LDA represents each document as a mixture (probability distribution) of topics and each topic as a mixture (probability distribution) of words and tries to infer what topics would create the probability distribution of words as seen in the documents. LDA algorithm treats documents as a bag of words and is based on the matrix factorization technique, where the aim is to convert the Document-term matrix (N, M) to two lower dimension matrices: Document-Topic matrix (N, K) and Topic-Word (K, M) matrix; K being the input parameter, i.e., the number of top K topics to extract. After an initial random assignment of a topic to documents and words to a topic, LDA optimizes the probability distribution of the lower dimension matrices by improving the assignments done in the previous steps. It iterates through each document and its each word to determine the proportion they contribute to the topic assigned to the document and the proportion in which they contribute to the overall topics in all documents, based on which they are reassigned to new topics. This way, LDA backtracks to compute the topic-word distribution that would create the topic the overall document set represents. Hyperparameters alpha and beta control document-topic and topic-word density; alpha decides the number of topics assigned to each document, and beta controls the number of words used to model a topic. GSDMM [19] is a variation of LDA for short text topic modeling; this algorithm assumes that a document consists of a single topic only instead of a mixture of topics as in the case of LDA. Pre-processed textual UGC can be used as input documents to these algorithms for trend prediction.

2.2 Deep Learning Techniques for Natural Language Processing

In this section, we briefly discuss various deep learning algorithms. A simple artificial neural network is the building block of all the advanced deep learning algorithms we describe next. Deep learning algorithms can be grouped into four broad categories based on their learning approach and the characteristics of available training data: (1) supervised, (2) unsupervised, (3) semi-supervised, and (4) reward-based learning (deep reinforcement learning) [4] [5]. The important deep learning algorithms under each category are briefly explained in Table 2.1.

2.2.1 Supervised Deep Learning Algorithms

Supervised learning algorithms use labeled data samples as input datasets for training, which makes the algorithms learn the classification boundaries [4]. For instance, to predict whether a person has cancer or not from a given set of numeric features, each of the training data samples (i.e., feature vectors) X_i are annotated and assigned labels Y_i as per the class they belong to. These (X_i, Y_i) pairs are then used to train the algorithm in a supervised manner. Some of the most popularly used supervised deep learning algorithms are: ANN, DNN, CNN, RNN, LSTM, GRU, Bi-RNN, Bi-LSTM, and Bi-GRU (Refer Table 2.1).

2.2.2 Unsupervised Deep Learning Algorithms

Unsupervised learning algorithms learn directly from the unlabelled (unclassified) training data [4]. They look for underlying structures or hidden patterns in given input data without any guidance from labels. Some of the unsupervised deep learning algorithms are: GAN, RBM, DBN, DBM, and Autoencoders (Refer to Table 2.1).

2.2.3 Semi-supervised Deep Learning Algorithms

Semi-supervised learning is a combination of supervised and unsupervised learning. Semi-supervised learning is used when only a small amount of training data is labeled; however, a large portion of the training dataset is unlabeled [4]. Semisupervised deep learning networks use a mix of discriminative and generative deep learning algorithms. Transformers [20] and Transformer based language models, namely: Google's BERT [21], OpenAI's GPT [22], and XLNet [23] are semi-supervised deep learning networks.

Transformers have proven to be a breakthrough for the NLP domain. They have become the state-of-the-art model for natural language modeling and natural language understanding tasks. They mitigate the drawbacks of RNN models. RNNs process the input text sequentially, which makes it harder to identify and learn the contextual patterns in long sequential inputs. Whereas, Transformers are highly parallelized neural network architectures with attention mechanisms (self and multi-head) for modelling the language context; this makes them efficient for training on extremely large text datasets [20].

2.2.4 Deep Reinforcement Learning

Deep reinforcement learning (or deep reward-based learning) is a combination of deep neural network algorithms described above and the traditional AI framework of reinforcement learning [5]. Reinforcement learning is a subfield of AI and ML where an intelligent system is trained through trial and error. The goal is to make the agent learn to make correct decisions for the task at hand in a real-world environment. The agent's goal is to maximize the reward function; every correct action taken by the agent that takes it closer to the target is rewarded, and the agent's incorrect actions are penalized [4]. Deep reinforcement learning has gained the limelight since it closely mimics the human brain's structure and functioning.

Table 2.1 *Brief Description of Popular Deep Learning Algorithms*

Deep Learning Algorithm	**Brief Description / Key Characteristics**
Supervised Techniques	
ANN	Artificial Neural Network is a shallow neural network with a single hidden layer of neurons (perceptrons). [7]
DNN	Deep Neural Network is composed of multiple interconnected hidden layers of neurons (called dense neural layers), where each layer operates on the output from its previous layer. It is a feed-forward architecture with error backpropagation and activation functions to introduce non-linearity. [8]
CNN	Convolutional Neural Network is a variation of DNN with layers to perform additional operations like convolution, pooling, batch-norm, and dropout. These filters or operations help to improve classification accuracy by preventing overfitting and underfitting. [24]
RNN	Recurrent Neural Network is a sequential DNN with internal memory states, using which it can process and persist input information over time. At every time instant, the network makes a prediction by using the current input and the last hidden state (i.e., the last output of its previous hidden layer). This helps in the modeling of sequential data patterns for temporal tasks. Vanishing and exploding gradients are the main drawbacks of a vanilla RNN implementation. [25]
LSTM	Long Short-Term Memory network is a popular variant of vanilla RNN with a chain-like structure. It has three additional gates within each RNN

Deep Learning Algorithm	Brief Description / Key Characteristics
	cell, viz. input, forget, and output. These gates determine the quantum of information that will be retained and allowed to flow through the network. LSTM overcomes the drawbacks of RNN and is able to learn long-range sequential patterns from long input sequences of time series data. [26]
GRU	GRU is a simpler version of LSTM, which is easy and efficient to train. GRU has only two gates: reset and update. [27]
Bi-RNN, **Bi-LSTM,** **Bi-GRU**	These are Bi-directional variants of RNN, LSTM, and GRU, respectively. They have two units: one takes the previous hidden state as input (forward direction), and the other utilizes information from future states of the sequential input data. [28]
Unsupervised Techniques	
GAN	Generative Adversarial Network has two sub-networks: a generator network and a discriminator network, where one network tries to outsmart the other. The task of the generator network is to produce a realistic data sample, and the task of the discriminator network is to learn the decision boundaries to classify actual data samples from the training dataset vs. the real looking data sample generated by the first sub-network. [29]
RBM	Restricted Boltzmann Machine is a stochastic, generative neural network with no output nodes. It has a two-layered architecture where the input and hidden layers are connected to form an undirected bipartite graph-like network. RBM learns the probability distribution of the input training data using a contrastive divergence gradient descent training algorithm. [30]
DBN	Deep Belief Network is also a generative deep neural network. It has a directed, graphical architecture that is formed by stacking multiple layers of RBMs or autoencoders. [31]
DBM	Deep Boltzmann Machine is similar to DBN, but it is undirected in nature. [32]
Autoencoder (AE)	Autoencoder is a type of feed-forward artificial neural network used for dimensionality reduction. AE is used to learn compressed, smaller dimension latent feature representations (encodings) for unlabeled input data. It has two sub-networks: encoder network and decoder network, where the former learns to encode the input training data, and the latter attempts to reconstruct the original training input from its encodings. Four frequently used variations of a vanilla AE are: Stacked, Sparse, Denoise, and Variational. [33]
Stacked AE	Stacked AE is a deep neural network with multiple AE layers stacked back to back, one after another. In the encoder network, the size of each

Deep Learning Algorithm	Brief Description / Key Characteristics
	AE layer is smaller than its precursor layer, whereas the decoder network is a mirror image of the encoder network structure, and the size of each layer is smaller than its following layer. [34]
Sparse AE	Sparse AE imposes a regularization constraint, namely, L1 sparsity, in order to produce shorter encodings. L1 sparsity creates an information bottleneck in the network by reducing the number of active nodes in each layer. [35]
Denoise AE	The input to this network is noisy/corrupt training dataset samples, where random noise has been deliberately introduced. The network is trained to learn the process of reconstructing the original, clean sample. [36]
Variational AE	It is a Bayesian deep generative AE network with an additional goal of learning the probabilistic distribution of input training data. Along with the MSE term, the training function also includes a KL divergence term. This helps in regularizing the training process so that robust, latent encodings of the input data are learned. [37]
Semi-supervised Techniques	
Transformers	Transformers have an encoder-decoder like structure, where the encoder and decoder are stacked together for layer-by-layer transformations of input. Encoder and decoder also use attention mechanisms. [20]

CHAPTER 3

LITERATURE REVIEW

This chapter presents the research methodology, findings, analysis, and results from the in-depth systematic literature review that was conducted to understand the state-of-the-art related to research applications of deep learning techniques for categorizing user generated content available on the Internet for various real-world social computing applications (e.g., mental healthcare).

In this research work, we have conducted an in-depth systematic literature review (SLR) to understand the state-of-the-art related to research applications of deep learning techniques for categorizing user generated content available on the Internet for various real-world social computing applications. Specifically, this work reviews 96 research studies that have applied deep learning techniques, either for creating deep feature representations or for training deep learning models for categorizing user-generated content available on online social networks for mental healthcare applications (depression, suicide/self-harm). We have analyzed and provided a detailed technical description of these research studies w.r.t. deep learning techniques, model architectures, neural feature representations or embeddings, datasets, modality, and novel performance metrics used by them for UGC text classification. We have identified and enumerated the gold standard UGC datasets for the chosen application problem and have discussed their characteristics in detail, which can be of great value to researchers working in this domain for their future research and analysis. In this study, we have also included the datasets in regional languages and from locally popular online social networks. This SLR has helped in identifying research gaps and open challenges related to the UGC text categorization problem and proposes solutions for future research directions for advancing research applications of deep learning techniques to user generated content available on the Internet for various real-world social computational applications.

The rest of the chapter is organized as follows: Section 3.1 describes our research methodology for conducting this SLR; Section 3.2 provides the detailed technical overview of the research studies included in this literature review; Section 3.3 reviews and reports the gold standard datasets used by these studies; Section 3.4 presents the key findings, analysis and results related to the applications of deep learning techniques for UGC text categorization; towards the end in Section 3.5, we enumerate the research gaps and open challenges identified which our work aims to address in the following chapters; and Section 3.6 summarizes the key takeaways from this survey.

3.1 Research Methodology

This systematic literature review was performed according to the Preferred Reporting Items for Systematic Reviews and Meta-analysis (PRISMA) guidelines [41]; the PRISMA flow diagram for this systematic review is shown in Figure 3.1. This systematic literature review covers articles published until June 2022 for this research problem. The literature search and selection process are explained in detail below, but first, we define the research questions we seek to answer from this systematic literature review.

3.1.1 Research Questions

Following is the list of research questions (RQs) we wish to understand through this systematic literature review:

RQ1. What deep learning techniques and model architectures have been proposed for detecting depression, self-harm, and suicide from online social media?

RQ2. What user-generated content modalities and their corresponding feature representation techniques have been used as input for training the above deep learning models?

RQ3. What are the various online social networking platforms for which deep learning models to detect depression, self-harm, and suicide have been developed?

RQ4. What human languages have been taken into consideration by researchers for building deep learning systems for this research problem?

RQ5. What are the various available benchmark datasets for conducting research related to the detection of depression, self-harm, and suicide from online social networks?

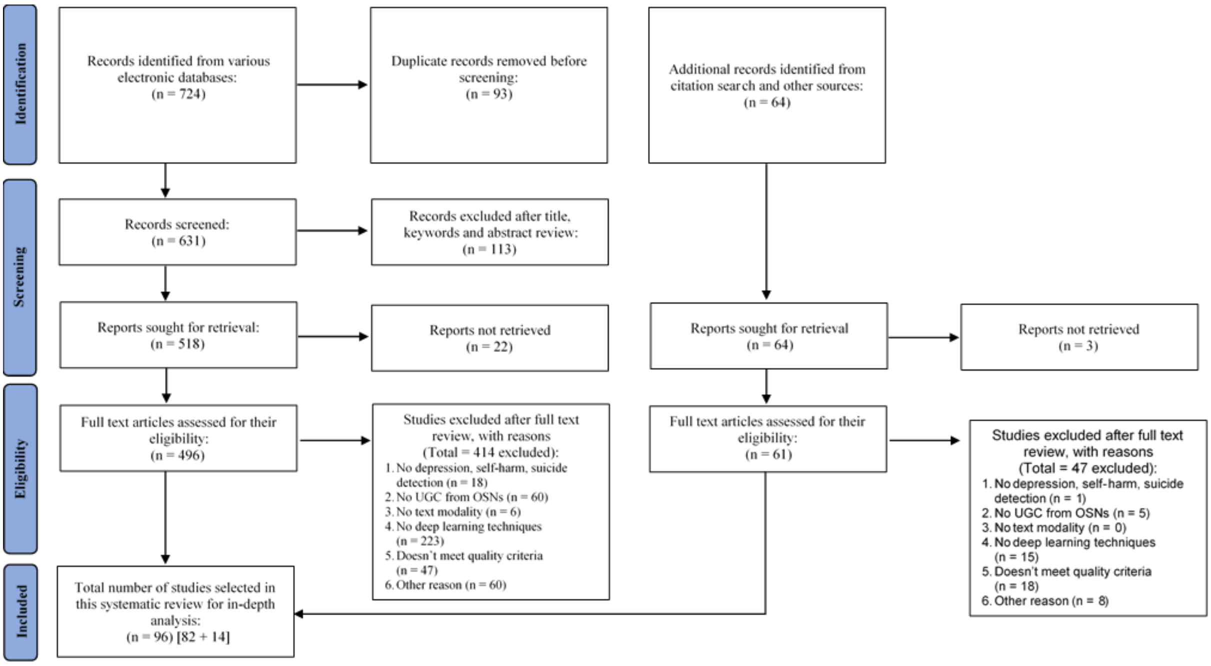

Identification
Screening
Eligibility
Included
Records identified from various electronic databases: (n = 724)
Duplicate records removed before screening: (n = 93)
Additional records identified from citation search and other sources: (n = 64)
Records screened: (n = 631)
Records excluded after title, keywords and abstract review: (n = 113)
Reports sought for retrieval: (n = 518)
Reports not retrieved (n = 22)
Reports sought for retrieval (n = 64)
Reports not retrieved (n = 3)
Full text articles assessed for their eligibility: (n = 496)
Studies excluded after full text review, with reasons (Total = 414 excluded):
1. No depression, self-harm, suicide detection (n = 18)
2. No UGC from OSNs (n = 60)
3. No text modality (n = 6)
4. No deep learning techniques (n = 223)
5. Doesn't meet quality criteria (n = 47)
6. Other reason (n = 60)
Full text articles assessed for their eligibility: (n = 61)
Studies excluded after full text review, with reasons (Total = 47 excluded):
1. No depression, self-harm, suicide detection (n = 1)
2. No UGC from OSNs (n = 5)
3. No text modality (n = 0)
4. No deep learning techniques (n = 15)
5. Doesn't meet quality criteria (n = 18)
6. Other reason (n = 8)
Total number of studies selected in this systematic review for in-depth analysis: (n = 96) [82 + 14]

Figure 3.1 *PRISMA Flow Diagram*

3.1.2 Literature Search Strategy

The literature search for this survey was performed using the following search keywords and conditions:

(depression OR self-harm OR suicide OR mental health OR mental disorder) AND (online social network OR social media OR user generated content OR deep learning OR neural network OR perceptron OR machine learning OR artificial intelligence OR social network analysis OR big data analytics OR natural language processing OR text mining OR data mining)

We retrieved research articles published between 2010 and June 2022 where deep learning and AI and ML techniques have been used for the detection of depression, self-harm, suicide ideation, and mental health issues from social media content. We used the search keywords "mental health" and "mental disorder" as well because researchers often use them as umbrella terms in their article titles and keywords. For the same reason, we also included “machine learning” and “artificial intelligence” in the search terms since deep learning algorithms are a subset of these. We used various electronic databases, namely: Web of Science, ACM Digital Library, Science Direct, Springer Link, IEEE Xplore Digital Library, Wiley, Taylor & Francis, MDPI, PubMed, Scopus. We also searched on Google Scholar and Semantic Scholar, which are useful portals that list and index research titles, abstracts, and hyperlinks to their corresponding journal or proceedings where they are published. Towards the end, we investigated the citation map, i.e., the references section of the research articles included in this survey, to include any relevant literature we may have missed during the search phase. A total of 788 research publications were identified using the search strategy, citation map, and other sources.

3.1.3 Study Selection Process

Now, we describe our study selection method. Seven hundred twenty-four records were initially identified from various electronic databases (phase 1). As a first step, 93 duplicate records were removed from these, and then the remaining 631 research records were screened using their title, keywords, and abstract to exclude irrelevant records, after which another 113 were dropped. Next, 22 records whose full text could not be retrieved had to be excluded from the SLR. At last, the remaining 496 research articles for which their full text was available were read and analyzed in-depth to determine their eligibility as per the inclusion and exclusion criteria explained below; 414 research studies had to be omitted from this survey due to various exclusion reasons:- 18 studies did not have depression/self-harm/suicide detection as their primary goal, 60 did not use UGC for detection, 6 used other input modalities from UGC except text, 223 did not apply deep learning algorithms, 47 did not meet the publication quality criteria, and another 60 were excluded due to other miscellaneous reasons. Exactly similar methodology was followed for 64 additional research records identified in the second phase through citation search and other sources. Out of the 788 (724 + 64) articles identified during the search process of Phase 1 and Phase 2, 96 research articles have been included in this systematic review. The corresponding PRISMA flow diagram for this SLR is shown in Figure 3.1.

Inclusion and Exclusion Criteria:

1. The research studies were included if at least one of their primary objectives was to detect depression or self-harm/suicide using UGC from online social media.

2. The research studies were included if only they used UGC datasets collected from any OSN websites, online forums/communities, or blogs where users socially interact with each other. They were excluded if they used datasets such as audio/video recordings of face-to-face clinical interviews, transcripts of clinical interviews, clinical notes / EHR, and suicide notes.

3. The included research publications must utilize textual UGC content from OSNs for training deep learning models. Few research studies did not utilize text modality, i.e., they ignored the text from the users' posts and did not use it as an input for training deep learning models. We exclude these studies from our survey for the following reasons: Text is the only ubiquitous modality across disparate social media networks. Multimodal content may not always be available, as not all users post multimodal content. It has been observed from the multimodal social media datasets for this domain that textual user posts are far more abundant than user posts with pictures/videos, etc., and such multimodal datasets are scarce. Only text-based classifiers are known to perform better than only image / video-based classifiers [83]. Images and videos require sophisticated processing and higher computation power and hence make it challenging to build low-cost, large-scale surveillance systems

4. The research studies were included only if they used deep learning techniques or deep neural architectures. In addition, we also survey the research studies where ANNs have been utilized since these studies serve as a baseline comparison for all other articles. This is because ANNs are the foundational building blocks of all other advanced deep neural network architectures. The studies that used only statistical techniques and shallow learning techniques, i.e., traditional machine learning algorithms, e.g., NB, DT, SVM, RF, XGB, etc., were excluded.

5. The research studies were included if they were published in quality journals with high impact factors, in proceedings of popular top-tier conferences, or if they have significant/novel research contributions supported by evaluation experiments and baseline comparisons to advance the research in this domain.

6. Some of the other miscellaneous reasons for exclusion are: Only primary research studies were included, and secondary or tertiary studies were excluded. For the studies where conference publication was extended to journal publication, we included only the corresponding journal publication, and its conference version was excluded. Unpublished thesis, keynote talks, poster presentations, and magazine or editorial articles were excluded.

3.2 Technical Overview of Research Studies

After reviewing the research publications in the scope of our work, in this section, we elucidate the deep learning techniques and methods they utilized. Table 3.1 chronologically outlines the vital information from these research studies. We provide comprehensive details about deep learning techniques used, proposed model architecture, input features, and feature representation techniques and report the performance evaluation metrics of results achieved. Statistics and attributes of the training datasets used by these studies are elucidated in next section 3.3 (Table 3.2).

The first research study we have discussed in Table 3 is the work by Mowery et al. [42]. They created handcrafted features extracted from user profile information and user tweets for training a Linear Perceptron model. Although the Linear Perceptron model is a supervised binary classifier with a linear prediction function, however, we included this study because the Linear Perceptron model is the oldest and simplest ANN (single layer). Linear Perceptron can be considered as a fundamental building block of all other deep neural networks. Therefore, the work by Mowery et al. [42] serves as a baseline comparison for all other research articles that propose the use of advanced deep neural network architectures for our research problem.

From 2017 onwards, researchers started exploring the applications of convolution and sequential deep neural networks for diagnosing mental health conditions from social media. CNNs were first used for this research domain by Gkotsis et al. [44] and by Yates et al. [45]. Similarly, some of the earliest research applications of sequential deep learning networks like RNNs, LSTMs, and GRUs were by Halder et al. [46], Trotzek et al. [47], and Sadeque et al. [48]. From 2018 onwards, training convolution and sequential neural networks with neural word embeddings became very popular [58] [59] [60] [61] [62] [63] [64] [65] [67] [84] [85] [90] [93] [127].

Around 2019, the BERT language model became very popular due to its impressive performance for various NLP tasks. BERT-based neural text embeddings were utilized for training different deep NNs to predict suicide risk, as well as for depression detection tasks [72] [73] [84] [121] [130] [136]. The recent research focus has primarily been either on Transformer based classifiers [85] [110] [115] [117] [119] [125] [126] [153] [130]; or on building complex deep neural networks with hierarchical, cascaded, fusion/hybrid, ensemble architectures. Researchers have proposed many advanced deep neural architectures, such as:

hierarchical networks (HAN) [66] [81] [96] [97] [98] [104] [106] [107] [108] [125] [126] [134] [136], ensemble and cascaded/fusion networks [93] [95] [108] [109] [116] [118] [124] [131] [132] [135], relation & cause extraction networks [99] [100], deep multimodal networks [101] [102] [103] [104] [105] [136], and multitask learning frameworks [43] [77] [86] [87] [128].

Table 3.1 *Deep Learning Techniques used by research studies included in this Systematic Literature Review and their performance evaluation results (Table Source: Our Published Paper [216])*

Reference	Research Objectives & Key Contributions	Input Modalities	Predictors & Feature Representations	Deep Learning Techniques & Architecture	Performance Metrics
Mowery et al. 2016 [42]	Classify user's posts as non-depression vs. depression suggestive	Text, User Profile Information	Syntax, N-grams, Emoticons, Sentiment, LIWC, Personality Traits, Gender, Age	Linear Perceptron (single layer FF NN)	R = 0.65, P = 0.69
Benton et al. 2017 [43]	Detect suicide ideation and attempt risk and related mental health disorders in users	Text	Pre-trained Word2Vec neural embeddings	Feed Forward NN (MLP) for STL and MTL (with shared hidden layers)	MTL NN model (best results) Depression: AUC = 0.768 Suicide: AUC = 0.848
Gkotsis et al. 2017 [44]	Distinguish mental health related posts amongst all other posts by a user, and then categorize them into eleven mental health disorders.	Text	Neural word embeddings created from the dataset text corpus	FF NN, CNN	CNN (best results) Differentiate Mental Health posts and Non-Mental Health posts: ACC = 0.9108 Depression: F1 = 0.73, R = 0.77, P = 0.70 Suicide: F1 = 0.61, R = 0.59, P = 0.62 Self-harm: F1 = 0.64, R = 0.58, P = 0.70
Yates et al. 2017 [45]	Detect users having depression and assess their self-harm risk; Contributed Reddit Self-Reported Depression Diagnosis (RSDD) dataset	Text	Post-level feature representation vectors created using CNN, then a Single user level feature representation vector is created for every user by merging all post-level representations of that user	Dense NN	Depression: F1 = 0.51, R = 0.45, P = 0.59 Self-harm (averaged metrics over all the risk levels/bands/classes i.e., green, amber, red, crisis): ACC = 0.93, F1 = 0.89,
Halder et al. 2017 [46]	Sequential modeling of the temporal progression of user's mental and emotional health (whether improving or deteriorating) through Negative Emotional Index (NMI) score for every subsequent post by the user	Text, Activity information	Fine-tuned neural text embeddings created using RNN (initialized with pre-trained GloVe neural word embeddings), Numerical features for user's activity (time since the last post, interaction counts, etc.), and NMI score.	Ensemble with two RNN components, one for text features and one for numeric features	MAE = 0.0781
Trotzek et al. 2017 [47]	Early risk prediction on the Internet for	Text	Meta linguistic features, LSA over BOW vectors, TF-IDF, N-grams,	LSTM	P = 0.61, R = 0.67, F1 = 0.64 ERDE5 = 12.82%

Reference	Research Objectives & Key Contributions	Input Modalities	Predictors & Feature Representations	Deep Learning Techniques & Architecture	Performance Metrics
	depression in users (CLEF's eRisk 2017)		Paragraph Vectors (Doc2Vec neural embeddings)		
Sadeque et al. 2017 [48]	Early risk prediction on the Internet for depression in users (CLEF's eRisk 2017)	Text	Post representations created from the occurrence frequency of popular depression lexicons and UMLS concepts	GRU, Ensemble with GRU network and SVM combined using NB classifier	Ensemble (best performing) P = 0.32, R = 0.79, F1 = 0.45, ERDE5 = 14.73%
Sadeque et al. 2018 [49]	Early detection of depression by chronologically processing users' posts; They proposed a novel metric called *F1-Latency* to measure the performance of early warning systems.	Text	BOW, Depression lexicon N-grams, UMLS concepts frequency counts, Neural word embeddings created using RNN & LSTM classifiers trained for distinguishing depression vs. non-depression indicative texts	GRU	P = 0.67 R = 0.759 F1 = 0.712
Maupomé et al. 2018 [50]	Early risk prediction on the Internet for depression in users (CLEF's eRisk 2018)	Text	30 latent topics extracted using LDA from term-document matrix created from 3000 most frequent N-grams in user's posts	FF NN (MLP) with two hidden layers	P = 0.32, R = 0.62, F1 = 0.42, ERDE5 = 10.04%
Wang et al. 2018 [51]	Early risk prediction on the Internet for depression in users (CLEF's eRisk 2018)	Text	Post level or Sentence embeddings are created using one-hot encodings generated from vocabulary of the top 300 words (i.e., those with the highest TF-IDF scores)	CNN	F1 = 0.37, R = 0.52, P = 0.29, ERDE5 = 10.81%
Paul et al. 2018 [52]	Early risk prediction on the Internet for depression in users (CLEF's eRisk 2018)	Text	fastText neural word embeddings	RNN	F = 0.21, R = 0.15, P = 0.35, ERDE5 = 9.89%
Trotzek et al. 2018 [53]	Early risk prediction on the Internet for depression in users (CLEF's eRisk 2018)	Text	GloVe and fastText neural word embeddings	CNN	F1 = ~ 0.51, ERDE5 = ~ 9.5%
Liu et al. 2018 [54]	Early risk prediction on the Internet for depression in users (CLEF's eRisk 2018)	Text	Sentence or post level embeddings created over the entire vocabulary of training data using one hot encoding technique	CNN+LSTM cascaded deep learning network	F1 = 0.29, R = 0.27, P = 0.31, ERDE5 = 10.19%
Trotzek et al. 2018 [55]	Early detection of depression by chronologically processing users' posts	Text	Set 1: Pre-trained neural word embeddings Word2Vec (CBOW, Skip-gram), fastText, GloVe, and fastText fine-tuned on the training dataset Set 2: Emotion & Sentiment related features using NRC, VADER, LIWC, Meta linguistic features [47], Readability score	Ensemble classifier combining scores from the following two classifiers: CNN trained using Features Set 1 and LR trained on Feature Set 2	F1 = 0.71, R = 0.71, P = 0.71, ERDE5 = 12.13%
Orabi et al. 2018 [56]	Detect users with depression	Text	Word2Vec embeddings (Random, Skip-gram, CBOW, and Optimized/fine-tuned for	CNN variants (Max pooling, Multi-channel, Multi-channel with Max pooling),	CNN with Max pooling and optimized neural embeddings (best results)

Reference	Research Objectives & Key Contributions	Input Modalities	Predictors & Feature Representations	Deep Learning Techniques & Architecture	Performance Metrics
			mental health care domain)	Bi-LSTM with context-aware Attention mechanism	ACC = 0.879, AUC = 0.95, F1 = 0.869, R = 0.87, P = 0.874
Shing et al. 2018 [57]	Detect suicide ideation & attempt risk in users; Contributed UMSD-V1 dataset (an anonymized, labeled dataset of Reddit users)	Text	Skip-gram neural word embeddings	CNN	F1 = 0.42
Wu et al. 2018 [58]	Identify users having depression through their posts and other external heterogeneous data sources.	Text, Profile, social network & activity information, External public heterogeneous information: traffic, weather, environment, population, living conditions	Set 1: Post level representations created using Word2Vec neural embedding; These are then stacked and encoded using LSTM to create user-level feature embeddings Set 2: Forty eight numerical features created using all other input modalities and external data sources	DNN trained using feature representation vector created by merging Feature Set 1 and Feature Set 2 vectors	P = 0.833, R = 0.714, F1 = 0.769
Cohan et al. 2018 [59]	Detect users with various mental health disorders (including depression); They have contributed SMHD (Self-reported Mental Health Diagnosis) Dataset which is a large-scale, anonymized datasets annotated for nine mental health conditions	Text	fastText neural word embeddings	FF NN with 100 hidden layers, CNN	FF NN (best results) for Depression: F1 = 0.5356, R = 0.447, P = 0.668
Sawhney et al. 2018 [60]	Detect suicide ideation & attempt risk from users' social media posts; Built a suicide lexicon of related words and phrases; Collected a suicide ideation Twitter dataset with manually labelled Tweets.	Text	Set 1: Google's Word2Vec neural embeddings Set 2: Sentence embeddings created using 1-dimension CNN from Set 1 features	RNN, LSTM (trained using Feature Set 1), C-LSTM: LSTM trained using Feature Set 2 vectors obtained from CNN (cascaded network)	C-LSTM (best results) ACC = 0.81, F1 = 0.82, R = 0.872, P = 0.78,
Du et al. 2018 [61]	Detect suicide ideation & attempt risk in users' posts, and extract psychological stressors (reasons) for suicide (cause extraction)	Text	GloVe neural word embeddings	Tweet classification: CNN, Bi-LSTM Psychological stressor extraction using Transfer Learning: RNN (with character and word level Bi-LSTM layers)	CNN (best performing model) ACC = 0.74, F1 = 0.83, R = 0.88, P = 0.78,
Coppersmith et al. 2018 [62]	Detect suicide risk & ideation in users	Text	GloVe neural word embeddings	Bi-LSTM with self-attention mechanism	AUC: ~0.94

Reference	Research Objectives & Key Contributions	Input Modalities	Predictors & Feature Representations	Deep Learning Techniques & Architecture	Performance Metrics
Ji et al. 2018 [63]	Detect suicide ideation & attempt risk in social media users from their posts; Contributed two labeled datasets of Twitter & Reddit users' posts	Text	Set 1: Lexical counts, Syntactic and POS tags, TF-IDF, LDA topics, LIWC Set 2: Word2Vec neural embeddings (Skip-gram, CBOW)	Multilayer FF NN trained with Feature Set 1, LSTM trained with Feature Set 2	Both deep learning models had comparable results (for Twitter & Twitter both) ACC = ~ 0.91 to 0.94 F1 = ~ 0.90 to 0.94 R = ~ 0.87 to 0.92 P = ~ 0.93 to 0.97
Cong et al. 2018 [64]	Detect users with depression (specifically in imbalanced datasets)	Text	Distributed neural word embeddings created using another domain-specific embedding matrix	X-A-BiLSTM cascaded model with two components: XGBoost (to handle data imbalance), followed by Bi-LSTM with Attention mechanism (for improved classification accuracy)	P = 0.69, R = 0.53, F1 = 0.60
Shen et al. 2018 [65]	Identify users with depression on a target OSN platform (sparsely labelled target domain) using heterogeneous labeled data from other OSN (source domain) through cross-domain transfer learning techniques; Contributed two benchmark, annotated datasets of Sina Weibo & Twitter users	Text, Images, Profile, social network & activity information	Seventy eight statistical features created across five categories of input modalities, sixty features shared across target and source domain, eighteen target domain-specific features Examples: emotion word and emoticon counts (text), visual features e.g. color, theme, and brightness (images), gender (profile), number of connections (network), and time of post, respectively (activity);	Deep Transfer Learning network DNN-FATC: DNN with four hidden layers and Feature Adaptive Transformation & Combination (FATC) Strategy for 60 shared features	F1 = 0.785
Song et al. 2018 [66]	Detect users with depression and identify posts that can explain the reason why a user was classified as depressed (i.e. explainable prediction results)	Text	Post level encodings or representations created using GloVe neural word embeddings. These are used as input to the following four Feature Attention Networks that create user-level feature representation vectors: 1. FF NN (MLP) with Domain knowledge features (e.g. lexicon of depression symptoms), 2. RNN with Sentiment polarity using SentiWordNet lexicon, 3. FF NN (MLP) with Response/Thinking Style (using topic modeling), 4. RNN with Writing style features (One hot encoding vectors, POS tags)	FF NN (MLP) with post-level Attention Mechanism (HAN) (XAI)	P = 0.61, R = 0.52, F1 = 0.56
Naderi et al. 2019 [67]	Early risk prediction on the Internet for Self-harm signs in users (CLEF's eRisk 2019)	Text	Set 1: TF-IDF, N-grams extracted using Mutual Information Score Set 2: Word2Vec (CBOW) neural embeddings	CNN trained using Feature Set 2, Ensemble of above CNN model and SVM trained using Feature Set 1	CNN (best performing) P = 0.12, R = 1.0, F1 = 0.22, ERDE5 = 13%, F1-Latency: 0.21
Ragheb et al. 2019 [68]	Early risk prediction on the Internet for Self-harm signs in users (CLEF's eRisk 2019)	Text	Post representation vectors were created using neural word embeddings obtained from the AWD-BiLSTM model	Two-stage deep learning model: Stage 1: ULMFiT model with Attention mechanism for mood evaluation of each post	P = 0.48, R = 0.49, F1 = 0.48

Reference	Research Objectives & Key Contributions	Input Modalities	Predictors & Feature Representations	Deep Learning Techniques & Architecture	Performance Metrics
				Stage 2: MLP with two hidden layers to detect temporal mood variations from time series Stage 1 output data	
Allen et al. 2019 [69]	Detect suicide risk & to predict the severity/degree of risk in users	Text	GloVe neural word embeddings, LIWC	CNN	F1 = 0.5

Morales et al. 2019 [70]	Detect suicide risk & to predict the severity/degree of risk in users	Text	BOW, TF-IDF, N-grams, LDA topics, POS tags, NER tags, Neural word embeddings: Skip-gram, Retrofitted Skipgram, fastText, Personality & tone features	CNN (Task A),
				LSTM (Task B)
					CNN (Task A):
					ACC = 0.52, F1 = 0.31
					LSTM (Task B):
					ACC = 0.42, F1 = 0.30

Mohammadi et al. 2019 [71]	Detect suicide risk & to predict the severity/degree of risk in users	Text	Two different post representations created using Pre-trained ELMo and GloVe neural word embeddings with Attention mechanism	An (SVM) ensemble of the following eight deep neural sub-models: CNN, Bi-RNN, Bi-LSTM, and Bi-GRU (each trained using both representation vectors)	Task A: F1 = 0.48 Task B: F1 = 0.34 Task C: F1 = 0.27
Ambalavanan et al. 2019 [72]	Detect suicide risk & to predict the severity/degree of risk in users	Text	BERT sentence and word embeddings created by fine-tuning with the training data	FF NN, Bi-LSTM with Attention, Multi-instance learning Bi-LSTM with Attention	BERT + FF NN model (best results) Task A: ACC = 0.54, F1 = 0.477 Task B: ACC = 0.36, F1 = 0.26 Task C: ACC = 0.59, F1 = 0.15,
Matero et al. 2019 [73]	Detect suicide risk & to predict the severity/degree of risk in users	Text	BERT word embeddings / neural representations, LDA topics, Big 5 personality traits, Writing style, Lexical & Statistical features	LSTM with Attention mechanism (Task A), GRU with Attention mechanism & Dual context (Task B & C)	Task A: ACC = 0.59, F1 = 0.5 Task B: ACC = 0.57, F1 = 0.5, Task C: ACC = 0.69, F1 = 0.18
Stankevich et al. 2019 [74]	Detect users with depression	Text, Profile information	N-grams (Lexical), POS tags, Syntactic and Semantic relations, Linguistic and Sentiment dictionary-based features, One hot encoding feature vectors for profile and subscription information	MLP	P = 0.36, R = 0.48, F1 = 0.41 AUC = 0.54
Tadesse et al. 2019 [75]	Classify user's posts as depression vs. non-depression suggestive	Text	N-grams (created using TF-IDF weights), LDA topics, LIWC	MLP (2 hidden layers)	ACC = 0.91, F1 = 0.93, R = 0.92, P = 0.90,
Maupome et al. 2019 [76]	Detect users with depression	Text	Skip-gram Word2Vec neural representations created from the dataset text corpus itself	RNN with continuous (at every time step) inter-document (post) averaging, and attention mechanism	P = 0.474, R = 0.728, F1 = 0.574
Buddhitha et al. 2019 [77]	Detect users with various mental health disorders (depression & PTSD) using MTL	Text	fastText & Neural word embeddings (initialized randomly), Emotion category vector (from shared multi-channel CNN layers)	CNN with multiple channels (i.e., kernels of different sizes), with shared layers for MTL	P = 0.866, R = 0.820, F1 = 0.838, ACC = 0.875
Gaur et al. 2019 [78]	Detect suicide risk & to predict the severity or degree of the risk in users; Contributed gold-standard, manually annotated dataset of users with different suicide	Text	ConceptNet neural word embeddings	FFNN, CNN	CNN (best-performing model) F1 = 0.65 Graded R = 0.60 Graded P = 0.71

Reference	Research Objectives & Key Contributions	Input Modalities	Predictors & Feature Representations	Deep Learning Techniques & Architecture	Performance Metrics
	risk severity levels; Created a suicide severity lexicon				
Sinha et al. 2019 [79]	Detect suicide ideation and attempt risk in users' posts; They have contributed a Tweets dataset (manually annotated)	Text, Social network and activity information	GloVe word representation embeddings, Graph Neural embeddings for user's social network	Stacked ensemble of the following three classifiers: Bi-LSTM with Attention mechanism for 1) current post, and 2) with temporal weighting and LR for historical context modeling from old posts,3) GCNs	F1 = 0.93, R = 0.93, P = 0.93
Mishra et al. 2019 [80]	Detect suicide ideation and attempt risk from users' posts; Extended their previous dataset to include user's historical posting behavior	Text, Social network and activity information	Text features: N-grams (with TF-IDF scores), LDA topics, POS and NRC counts, Pre-trained GloVe neural word embeddings; User's historical posting based stylistic profile; Node2Vec neural embeddings for user's social network graph; and Metadata numerical / statistical features	Feature Stacking ensemble architecture to combine: A Bi-LSTM with Attention mechanism network for text, and various machine learning classifiers trained on other handcrafted features	P = 0.68, R = 0.62, F1 = 0.65
Cao et al. 2019 [81]	Detect suicide risk and ideation in users; Created novel SoWE (Suicide-oriented Word Embeddings); Contributed an annotated large dataset of Sina Weibo users.	Text, Images, Profile & activity information	User level representation vectors were created by combining the following embeddings: Text embeddings from a LSTM network with Attention mechanism initialized with SoWE, and Images embeddings from Pre-trained ResNet model (CNN), and twelve statistical features Note: SoWE were created using LSTMs initialized with pre-trained Word2Vec / GloVe / fastText / BERT neural embeddings	LSTM with Attention mechanism (HAN)	ACC = 0.913, F1 = 0.909
Gui et al. 2019 [82]	Detect users having depression	Text	Post level embeddings created using randomly initialized CNNs & LSTMs, which are then merged to create user level representation embedding	Deep Reinforcement Learning architecture using MLP NNs for designing the depression classifier and a policy agent for selecting user's posts for training the classifier (Hierarchical)	P = 0.872, R = 0.87, F1 = 0.971, ACC = 0.87
Gui et al. 2019 [83]	Detect users having depression	Text, Images	User level representation vectors are created by combining the following embeddings using MLPs: Text features: Bi-GRU, and Image features: Pre-trained sixteen layer VGG-Net model (CNN)	Multi-agent Multimodal Deep Reinforcement Q-Learning network by using GRU & MLP NNs for designing the post selector policy agents, as well as the depression classifier (Hierarchical)	ACC = 0.90, F1 = 0.90, P = 0.90, R = 0.90,
Wang et al. 2019 [84], 2020 [85]	Detect depression risk from users' posts, and predict the severity or degree of risk; Contributed dataset of annotated Sina Weibo posts	Text	Character embeddings, Input sequence feature representation generated using pre-trained BERT-based tokenizers	CNN, LSTM, BERT, RoBERTa, XLNET	BERT (best-performing model) F1 = 0.856

Reference	Research Objectives & Key Contributions	Input Modalities	Predictors & Feature Representations	Deep Learning Techniques & Architecture	Performance Metrics
An et al. 2020 [86]	Detect users with depression	Text, Images	Text: Fine-tuned BERT neural word representation embeddings Images: Pre-trained VGG (CNN) embeddings	MTL framework: Primary task (Depression detection): LSTM network with fused text & image embeddings as input Auxiliary task (Multimodal topic modeling): CNN	ACC = 0.809, F1 = 0.811, R = 0.809, P = 0.814,
Ophir et al. 2020 [87]	Detect suicide risk & ideation in users	Text	ELMo neural word embeddings	STL model: using DNN, MTL framework: using multiple DNNs Primary task: Detecting Suicide risk & ideation, Auxiliary tasks: Detecting personality traits, psychosocial risks, psychiatric disorders	MTL framework (best performing): AUC = 0.746
Sawhney et al. 2020 [88]	Detect suicide risk & ideation from users' posts	Text, Posting activity information	For creating Tweet representation vectors: Sentence level neural BERT embeddings, Historical Context using past Tweets: BERT neural embeddings finetuned with EmoNet dataset (that is annotated with Pluchtik emotion labels)	Time Aware LSTM + DNN cascaded network	R = 0.81, F1 = 0.799, ACC = 0.851
Lee et al. 2020 [89]	Detect suicide attempt risk and ideation from users' posts in low resource languages by using cross-lingual SoWE	Text	SoWE created for three languages: English, Chinese, and Korean, using the approach followed proposed by Cao et al. [81]	Dense NN Ensemble of 3 LSTM with Attention mechanism models (created for three languages: Chinese, English, and Korean with their respective SoWE)	P = 0.8757, R = 0.8741, F1 = 0.8749, ACC = 0.875
Kim et al. 2020 [90]	Detect users with various mental health disorders (including depression)	Text	Fine-tuned Word2Vec neural embeddings (CBOW)	CNN	ACC = 0.751, F1 = 0.795, R = 0.717, P = 0.891
Alabdulkreem at al. 2020 [91]	Detect users with depression	Text	Word2Vec and GloVe neural word embeddings	RNN+LSTM cascaded deep learning network	ACC = 0.72, F1 = 0.69, R = 0.68, P = 0.71
Carvalho et al. 2020 [92]	Detect suicide attempt and ideation risk from users' posts	Text	Pre-trained Word2Vec neural embeddings (Skip-gram method)	LSTM, BERT	BERT (best-performing model) ACC = 0.788, F1 = 0.787, P = 0.79, R = 0.788
Tadesse et al. 2020 [93]	Detect suicide attempt risk and ideation from users' posts	Text	Pre-trained Word2Vec representation embeddings	LSTM, CNN, LSTM+CNN cascaded deep learning network	LSTM+CNN cascaded model (best results) ACC = 0.938, P = 0.932, R = 0.941, F1 = 0.934,
Yao et al. 2020 [94]	Detect suicide risk & ideation in opioid users' posts	Text	Char2Vec character embeddings, fastText and GloVe neural word embeddings, Domain knowledge features	CNN, RNN, Bi-RNN with Attention mechanism	CNN (best performing model) F1 = 0.961, ACC = 0.954 P = 0.968, R = 0.953,
Rao et al. 2020 [95]	Detect users with depression (specifically in imbalanced datasets)	Text	Knowledge triples infused BERT neural word embeddings (KFB)	BiGRU with Attention mechanism (KFB-BiGRU-Att), KFB-BiGRU-Att-AdaBoost hierarchical ensemble model with two components: Bi-GRU with Attention mechanism, followed by AdaBoost (to handle data imbalance) (HAN)	KFB-BiGRU-Att-AdaBoost (best performing model) F1 = 0.56, R = 0.54, P = 0.58,
Sekulic et al. 2020 [96]	Detect users with various mental health disorders (e.g. depression)	Text	Pre-trained GloVe neural word embeddings	HAN with two sets of BiGRUs with Attention mechanism: the first network learns post-level	Depression: F1 = 0.6828

Reference	Research Objectives & Key Contributions	Input Modalities	Predictors & Feature Representations	Deep Learning Techniques & Architecture	Performance Metrics
				features, and the second learns user-level features	
Jiang et al. 2020 [97]	Identify users with mental health disorders (e.g. depression)	Text	BERT sentence embeddings (post level), averaged to create user-level representations	RNN with Attention mechanism (HAN)	F1 = 0.843, ACC = 0.833
Rao et al. 2020 [98]	Detect users with depression	Text	One hot encoding vector representing users' posts	Two HNN architectures of CNNs with temporal gated convolution units: Single Gated LeakyReLU CNN (SGL-CNN) and Multi Gated LeakyReLU CNN (MGL-CNN)	MGL-CNN on CLEF eRisk 2017 dataset (best performing model results): F1 = 0.60, R = 0.57, P = 0.63
Ji et al. 2020 [99]	Detect suicide ideation & attempt risk and various mental health disorder in users (including depression), also infer their relation with risk factors	Text	Post feature representation vectors created using Bi-LSTM network initialized with GloVe neural word embeddings, LDA topics, Sentiment lexicon features	Relation Network (ANN which can infer relations), with Attention mechanism	Combined Twitter Dataset (best results): P: 0.8381, R = 0.8385, F1 = 0.8377, ACC = 0.838
Liu et al. 2020 [100]	Detect suicide ideation and risk from users' posts as well as Suicide Ideation Cause Extraction (SICE); Created and Contributed the first SICE dataset	Text	Character level embeddings created using Bi-LSTM network, Various word representation neural embeddings used: Pre-trained Word2Vec, Fine-tuned BERT and ELMo	Bi-LSTM with Conditional Random Field (CRF)	P = 0.871, R = 0.723, F1 = 0.790
Mann et al. 2020 [101]	Detect users with depression, & predict the degree/severity of risk; Contributed an annotated multimodal dataset	Text, Images	Feature representation vector created by fusing: Text: TF-IDF, fastText and ELMo neural word embeddings Images: ResNet (18, 34, 50), ResNeXt (CNN)	FC DNN	Best performing (212 days' observation period, ELMo, ResNet34): P = 0.69, R = 0.92, F1 = 0.79
Lin et al. 2020 [102]	Detect users possibly having depression	Text, Images	Feature vectors created by combining: BERT sentence embeddings for Text and CNN representation vectors for Images	DNN	P = 0.903, R = 0.870 F1 = 0.963, ACC = 0.884
Ramírez-Cifuentes et al. 2020 [103]	Detect suicide ideation and attempt risk in social media users; They have created a new clinically annotated dataset for this domain.	Text, Images, Social network graph/connections and user activity information	Text: N-grams, Neural word representation embeddings for Spanish language Images: Prediction result probability score from ResNeXt (CNN) + another CNN joint model (pre-trained using ImageNet dataset, fine-tuned with another small dataset collected from Instagram) Other numerical and statistical features using LIWC, network graph, behavior, sentiment, etc.	CNN trained with only Neural embeddings for text, MLP NN trained with all feature groups/modalities	MLP NN with all features and modalities (best performing): P = 0.85, R = 0.92, F2 = 0.88, ACC = 0.88, AUC = 0.92
Cao et al. 2020 [104]	Detect suicide risk and ideation in users.	Text, Images, Profile, social network and activity information	LSTM for creating user-level representation embeddings by fusing following multimodal post encodings: Text: Pre-trained BERT embeddings Image: CNN based Pre-trained ResNet-34 model	GNN with Attention mechanism (HAN)	Best results obtained for Sina Weibo Dataset: ACC = 0.937, F1 = 0.937, P = 0.937, R = 0.936,

Reference	Research Objectives & Key Contributions	Input Modalities	Predictors & Feature Representations	Deep Learning Techniques & Architecture	Performance Metrics
			Suicide-oriented Knowledge Graph: created from user's network, profile, interactions, activity, etc.		
Chiu et al. 2020 [105]	Detect users with depression	Text, Images, Activity information	Post representations created using AdaBoost ensemble of following models: Text: Bi-LSTM network with Word2Vec embeddings Images: AlexNet (CNN) Behavior: Random Forest for 6 handcrafted features	LSTM with temporal weighting and day-based aggregations to obtain the final classification score for a user	P = 0.895, R = 0.782 F1 = 0.835
Bagherzadeh et al. 2020 [106]	Early risk prediction on the Internet for Self-harm signs in users (CLEF's eRisk 2020)	Text	Pre-trained Word2Vec neural embeddings	SVM ensemble of following networks at post level: CNN, LSTM, and SVM, along with Attention mechanism for final classification at the user level (HAN)	F1-Latency: 0.601 F1 = 0.625, R = 0.625, P = 0.625, ERDE5 = 26.8%,
Achilles et al. 2020 [107]	Early risk prediction on the Internet for Self-harm signs in users (CLEF's eRisk 2020)	Text	BERT neural word embeddings to create post representations which are then combined using CNN for creating the user's feature representation vector	LSTM (Hierarchical)	P = 0.27, R = 0.942, F1 = 0.42, ERDE5 = 40% F1-Latency: 0.367
Uban et al. 2020 [108]	Early risk prediction on the Internet for Self-harm signs in users (CLEF's eRisk 2020)	Text	Fine-tuned GloVe word embeddings, BOW, NRC, POS, LIWC	4 DL models: Bi-LSTM with Attention mechanism, HAN with CNN (post level) followed by LSTM with Attention mechanism (user level), Pre-trained BERT model, Ensemble of all of above	BERT (best performing): F1 = 0.546, R = 0.654, P = 0.469, F1-Latency = 0.462 ERDE5 = 29.1%,
Madani et al. 2020 [109]	Early risk prediction on the Internet for Self-harm signs in users (CLEF's eRisk 2020)	Text	Word2Vec (Skip-gram) neural embeddings	Ensemble of: CNN, Bi-LSTM	AHR = 0.3497 ADODL = 0.793
Castano et al. 2020 [110]	Early risk prediction on the Internet for Self-harm signs in users and Estimate user's depression severity level by predicting their responses for BDI questionnaire (CLEF's eRisk 2020)	Text	Pre-trained BERT-based tokenizers for creating Tokenized sequence for posts	Various Pre-trained BERT-based classifiers (with Softmax classification layer): BERT, DistillBERT, RoBERTa, XLM-RoBERTa	XLM-RoBERTa (best performing model): Depression: AHR = 0.37, ADODL = 0.81 Self-harm: F1 = 0.75, F1-Latency = 0.476, R = 0.692, P = 0.828, ERDE5 = 25%,
Maupome et al. 2020 [111] 2021 [112]	Early risk prediction on the Internet for Self-harm signs in users and Estimate user's depression severity level by predicting their responses for BDI questionnaire (CLEF's eRisk 2020)	Text	Authorship features like Textual Productions represented by One hot encoding vectors & LDA topics	Deep Averaging FF NNs, RNN with Attention mechanism	Self-harm: F1 = 0.525, F1-Latency = 0.513, R = 0.846, P = 0.381, ERDE5 = 26%, Depression: ADODL = 0.823, AHR = 0.386

Reference	Research Objectives & Key Contributions	Input Modalities	Predictors & Feature Representations	Deep Learning Techniques & Architecture	Performance Metrics
Sawhney et al. 2021 [113]	Detect suicide ideation & attempt risk in users and to predict severity or degree of risk	Text	Logformer neural sentence embeddings	Bi-LSTM with temporal Attention mechanism and Ordinal Regression	F1 = 0.64 Graded R = 0.61 Graded P = 0.66
Sawhney et al. 2021 [114]	Detect suicide ideation and attempt risk rom users' posts	Text, Posting activity & social network information	Tweet Representations: BERT word embeddings fine-tuned using Emotion dataset [88], Context: Temporal emotions aggregated from historical Tweet representations using Hawkes process and Social network Graph embeddings	Hyperbolic Graph Convolution Neural Network (HGCN)	R = 0.818, F1 = 0.792
Uban et al. 2021 [115]	Early risk prediction on the Internet for Self-harm and depression signs in users	Text	BOW, POS, LIWC, NRC, Pre-trained GloVe neural word embeddings	3 DL models: Bi-LSTM with Attention mechanism, HAN with CNN (post level) followed by LSTM with Attention mechanism (user level), Pre-trained Transformer-based Language Models: BERT. RoBERTa, AlBERT	HAN (best performing) Depression: AUC = 0.83, F1 = 0.45 Self-harm: AUC = 0.87, F1 = 0.65
Ren et al. 2021 [116]	Classify posts as depression vs. non-depression indicative	Text	Pre-trained GloVe neural word embeddings for representing text, and emotion words extracted from text	Hybrid / Fusion network consisting of 3 concatenated Bi-LSTM with Attention networks for learning: text semantics, positive emotions & negative emotions extracted from text	P = 0.919, R = 0.961, F1 = 0.939, ACC = 0.913
Ragheb et al. 2021 [117]	Detect users with depression and at risk of Self-harm & Suicide	Text	Pre-trained BERT-based tokenizers for creating Tokenized sequence for posts	Ensemble of negatively correlated noisy base learners: BERT, RoBERTa, XLNet (Pre-trained & Fine-tuned Transformer based Language Models)	RoBERTa-based ensemble model gave best results: Suicide: F1 = 0.79 Depression: F1 = 0.61 Self-harm: F1 = 0.52
Zogan et al. 2021 [118]	Detect users with depression	Text, user's Social network information	Cascaded DL model BERT-BART (BERT-KMeans-DistilBART) for extracting Text Summarization features, LDA topics, ANEW emotions, Domain Lexicon features, Network features	Hybrid / Fusion network with the following two cascaded DL networks: CNN-BiGRU with Attention mechanism for Text features, and Stacked Bi-GRUs for all other features	ACC = 0.901, F1 = 0.912, R = 0.904 P = 0.909,
Murarka et al. 2021 [119]	Detect various mental health disorders (e.g. depression) from users' social media posts	Text	Pre-trained BERT-based tokenizers for creating Tokenized sequence for input posts	Pre-trained BERT-based classifiers (with Softmax classification layer): BERT, RoBERTa	RoBERTa (best-performing model): Depression: F1 = 0.84, R = 0.88, P = 0.81,
Gollapalli et al. 2021 [120]	Detect suicide ideation & attempt risk	Text	Emotion enriched GloVe neural word embeddings, Temporal changes in LDA topics, and Self-harm latent topics	LSTM	Subtask 1: F1 = 0.61 Subtask 2: F1 = 0.70
Morales et al. 2021 [121]	Detect suicide ideation & attempt risk	Text	Tokenized input sequence using pre-trained BERT-based tokenizers	BERT-based classifiers (with Softmax classification layer)	Subtask 1: F1 = 0.571
Wang et al. 2021 [122]	Detect suicide ideation & attempt risk	Text	Doc2Vec neural word and post embeddings	CNN with Attention mechanism	Subtask 1: F1 = 0.69 Subtask 2: F1 = 0.737

Reference	Research Objectives & Key Contributions	Input Modalities	Predictors & Feature Representations	Deep Learning Techniques & Architecture	Performance Metrics
Bayram et al. 2021 [123]	Detect suicide ideation & attempt risk	Text	Pre-trained BERT neural embeddings	LSTM, Bi-GRU for Tweet level classification, Majority voting over Tweet scores is used for final prediction at the user level	Bi-GRU (best results) Subtask 1: F1 = 0.812 Subtask 2: F1 = 0.745
Renjith et al. 2021 [124]	Detect suicide ideation risk from users' posts and predict severity or degree of risk	Text	Word2Vec neural embeddings	LSTM-Attention+CNN cascaded deep learning model	ACC = 0.903, F1 = 0.926, R = 0.937, P = 0.916
Basile et al. 2021 [125]	Early risk prediction on the Internet for Self-harm signs in users and Estimate user's depression severity level by predicting their responses for BDI questionnaire (CLEF's eRisk 2021)	Text	Pre-trained GloVe word embeddings, and conventional NLP features like: BOW, Sentiment polarity, LIWC, Emotion using NRC lexicon	Pre-trained BERT-based classifiers (BERT, RoBERTa, DistilRobERTa), LSTM-based Hierarchical Attention Network (HAN)	Depression: AHR = 0.341, ADODL = 0.824 Self-harm: F1 = 0.433, F1-Latency = 0.426, R = 0.77, P = 0.301, ERDE5 = 8.9%,
Inkpen et al. 2021 [126]	Estimate user's depression severity level by predicting their responses for BDI questionnaire (CLEF's eRisk 2021)	Text	Tokenized input sequence using pre-trained BERT-based tokenizers called BigBirdTokenizer	Designed a HAN using Zero-shot transfer learning with Pre-trained BERT based classifier models like Sentence-BERT, and Sentence-RoBERTa	Depression: AHR = 0.284, ADODL = 0.789
Lopes et al. 2021 [127]	Early risk prediction on the Internet for Self-harm signs in users (CLEF's eRisk 2021)	Text	Word2Vec neural embeddings	LSTM, CNN	CNN (best-performing) F1-Latency = 0.206, F1 = 0.207, R = 1, P = 0.116, ERDE5 = 11.3%,
Ghosh et al. 2021 [128]	Detect users showing signs of possible depression	Text (not tweets but his bio/description), Images	Textual: Neural embeddings created from user's bio/description using GloVe embeddings and BiGRU sequence encoding layer Others: Sentiment polarity and Emotions extracted from user's bio/description text using IBM Watson Images: CNN	MTL framework using FC Dense NN with Attention Mechanism Primary task: Depression Detection Auxiliary task: Emotion Recognition	F1 = 0.69, ACC = 0.699
Uban et al. 2021 [129]	Detect users with various mental health disorders and with explainable prediction results	Text	Pre-trained GloVe neural word embeddings, BOW, LIWC, NRC, Emotions, Sentiment polarity	HAN with two LSTM networks with Attention (one for post level, followed by one for user level) (XAI)	Depression: F1 = 0.77, AUC = 0.81 Self-harm: F1 = 0.51, AUC = 0.83
Farruque et al. 2021 [130]	Classify user posts as depression vs. non-depression symptomatic and with explainable prediction results	Text	Pre-trained Neural word representations (GloVe, Word2Vec, Skip-gram), Pre-trained RoBERTa sentence level embeddings, Universal Sentence Encoder embeddings	Facebook's BART with ZSL-Centroid Method for transfer learning (XAI)	F1 = 0.783
Zogan et al. 2022 [131]	Detect users with depression symptoms and also provide explainability of the prediction results by determining user's	Text, Activity & Social network information	Set 1: Text representation encodings created using BiGRU network with Attention Mechanism (HAN) Set 2: Feature encodings / vectors created using MLP network for	Hybrid / Fusion NN with sigmoid classification (XAI)	ACC = 0.895 F1 = 0.893, R = 0.892, P = 0.902

Reference	Research Objectives & Key Contributions	Input Modalities	Predictors & Feature Representations	Deep Learning Techniques & Architecture	Performance Metrics
	posts that can explain why the user was classified as depressed.		Emotions, LDA topics, Domain specific Lexicon, user's activity, and network information		
Kour et al. 2022 [132]	Detect users with depression	Text	Neural word embeddings	RNN, CNN, CNN+Bi-LSTM cascaded deep learning network	CNN+Bi-LSTM cascaded network (best results) ACC = 0.943, F1 = 0.948, P = 0.969, R = 0.927,
Ahmed et al. 2022 [133]	Detect users with depression symptoms and provide explainable prediction results.	Text	GloVe neural word embeddings enhanced with domain knowledge emotion lexicon (transfer learning)	FFNN, LSTM, Bi-LSTM, Bi-LSTM with Attention Mechanism followed by cosine similarity based clustering (XAI)	Bi-LSTM with Attention (best performing) P = 0.90, R = 0.89
Naseem et al. 2022 [134]	Detect & estimate users' depression severity level	Text	Post representations are created using TextGCN feature embeddings, then post level vectors are combined using Bi-LSTM with Attention mechanism to create final user level feature embedding vector	FC Dense NN with Ordinal Regression classification layer (HAN)	F1 = 0.95 Graded R = 0.95 Graded P = 0.95
Ansari et al. 2022 [135]	Detect users with depression	Text	Set 1: Pre-trained GloVe neural word embeddings Set 2: Sentiment polarity and lexicon feature vectors using SenticNet, NRC, MPQA, AFINN.	Ensemble classifier by averaging scores (bagging) from the following two classifiers: LSTM with Attention network trained using Features Set 1 and LR trained using Feature Set 2	P = 0.648, R = 0.646, F1 = 0.646, ACC = 0.646
Cheng et al. 2022 [136]	Detect users with depression symptoms and provide explainable prediction results (i.e. identify user posts which can explain why a user was classified as depressed); Created and manually annotated a dataset of Instagram users	Text, Images, Posting activity information	User level feature representation vector created by combining following post-level feature embeddings: Images: Pre-trained InceptionResNetV2 (CNN) embeddings Text: Pre-trained BERT representation embeddings User activity: One hot encoding vector for time of post, time interval between subsequent posts etc.	Time Aware LSTM with Attention + FC cascaded NN (HAN) (XAI)	P = 0.95, R = 0.963, F1 = 0.956
Zeberga et al. 2022 [153]	Classify posts as depression vs. non-depression indicative	Text	Neural word embeddings: fastText, Word2Vec, GloVe, BERT	Bi-LSTM, BERT with Knowledge distillation layers (Distilled BERT) for transfer learning	Distilled BERT (best performing) ACC = 0.97

3.3 Datasets

In this section, we present a systematic review and comparison of datasets available for conducting research in this domain (Refer to Table 3.2). We also enumerate the statistics and attributes of these available training datasets in Table 3.2. Researchers have used these datasets to train their deep learning models discussed in the previous section 3.2 (Table 3.1). This comprehensive review of available datasets for this domain was done as a response to ***RQ 5: What are the various available benchmark datasets for conducting research related to the detection of depression, self-harm, and suicide from online social networks?***

Although all the datasets collected for research are already tabulated in Table 3.2, here we succinctly highlight the benchmark datasets for this research domain. Some of the benchmark datasets that have been extensively used in literature are: Reddit Self-Reported Depression Diagnosis (RSDD) dataset [45] (depression), RSDD-Time dataset [137] (depression), CLPsych 2015 dataset [138] (depression), SAD dataset [3] [139] (depression), eRisk Lab 2017 to 2020 datasets [140] [141] [142] [143] (depression & self-harm), CLPsych 2016 Triage dataset [144] (self-harm), UMSD V1 dataset [57] (suicide), CLPsych 2019 dataset (a.k.a. UMSD V2 dataset) [145] (suicide), CLPsych 2021 Shared Task dataset [146] (suicide), and SMHD Dataset (Self-reported Mental Health Diagnosis) for nine mental health conditions (including depression) [59].

Table 3.2 *Key characteristics of datasets used by research studies included in this Systematic Literature Review (RQ5) (Table Source: Our Published Paper [216])*

Reference	Dataset's Key Characteristics (Research Question 5)				
	OSN Platform	**Language**	**Mental Health Disorder**	**Cohort Size**	**Modalities**
Mowery et al. 2016 [42]	Twitter	English	Depression	Refer SAD dataset [3]	Text, Profile information
Benton et al. 2017 [43]	Twitter	English	Suicide	Dataset of total 9611 users with avg. 3521 tweets per user created from past datasets by Coppersmith et al. [14] [138] [147]	Text
Gkotsis et al. 2017 [44]	Reddit	English	Depression, Self-harm, Suicide	Non-MH: 476388 posts MH: 538272 posts Out of the above 538272 MH posts: Depression: 197436 posts Self-harm: 17102 posts Suicide: 90518 posts	Text
Yates et al. 2017 [45]	Reddit, ReachOut	English	Depression	RSDD (Depression) dataset: Positive: 9000 users Control: 10700 users Users' all Reddit posts from 2006 Jan to 2016 Oct collected	Text
			Self-harm	Used CLPsych 2016 Triage dataset [144]	
Halder et al. 2017 [46]	HealthBoards	English	Depression, Self-harm	29708 posts made by 1364 users across various discussion forums related to 24	Text,

Reference	Dataset's Key Characteristics (Research Question 5)				
	OSN Platform	Language	Mental Health Disorder	Cohort Size	Modalities
				mental health conditions, e.g., depression, self-harm, stress, anxiety, etc.	Activity information
Trotzek et al. 2017 [47]	Reddit	English	Depression	Refer eRisk Lab 2017 dataset [141]	Text
Sadeque et al. 2017 [48]	Reddit	English	Depression	Refer eRisk Lab 2017 dataset [141]	Text
Sadeque et al. 2018 [49]	Reddit	English	Depression	Refer eRisk Lab 2017 dataset [141]	Text
Maupomé et al. 2018 [50]	Reddit	English	Depression	Refer eRisk Lab 2018 dataset [142]	Text
Wang et al. 2018 [51]	Reddit	English	Depression	Refer eRisk Lab 2018 dataset [142]	Text
Paul et al. 2018 [52]	Reddit	English	Depression	Refer eRisk Lab 2018 dataset [142]	Text
Trotzek et al. 2018 [53]	Reddit	English	Depression	Refer eRisk Lab 2018 dataset [142]	Text
Liu et al. 2018 [54]	Reddit	English	Depression	Refer eRisk Lab 2018 dataset [142]	Text
Trotzek et al. 2018 [55]	Reddit	English	Depression	Refer eRisk Lab 2018 dataset [142]	Text
Orabi et al. 2018 [56]	Twitter	English	Depression	Refer CLPsych 2015 dataset [138] and Bell Let's Talk dataset [148]	Text
Shing et al. 2018 [57]	Reddit	English	Suicide	UMSD V1 dataset (Suicide): Positive: 865 users, 5008 posts Control: matched to the positive class Users' all Reddit posts from 2008 Jan to 2015 Aug collected	Text
Wu et al. 2018 [58]	Facebook	Chinese	Depression	Positive: 430 users Control: 846 users with a total of 873524 posts and 26 million action/interaction events	Text, Profile, social network & activity information, External public heterogeneous information: traffic, weather, environment, population, living conditions
Cohan et al. 2018 [59]	Reddit	English	Depression	SMHD dataset (total size): Positive (any MH issue): 36948 users (avg. 160 posts per user) Control: 335952 users (avg. 310 per user) Users' all Reddit posts from 2006 to 2017 collected Out of the above: Depression Positive: 14139 users	Text
Sawhney et al. 2018 [60]	Twitter	English	Suicide	Positive: 822 Tweets Control: 4391 Tweets	Text
Du et al. 2018 [61]	Twitter	English	Suicide	Positive: 623 Tweets Control: 2640 Tweets	Text
Coppersmith et al. 2018 [62]	Twitter	English	Suicide	Positive: 418 users, 197615 Tweets Control: 418 users, 197615 Tweets	Text
Ji et al. 2018 [63]	Reddit	English	Suicide	Positive: 3549 posts Control: matched to the positive class	Text
	Twitter	English	Suicide	Positive: 594 Tweets Control: 9694 Tweets	
Cong et al. 2018 [64]	Reddit	English	Depression	Refer RSDD dataset [45]	Text
Shen et al. 2018 [65]	Twitter	English	Depression	Positive: 1402 users, 292564 posts Control: 1402 users, 1120893 posts This sample is drawn out of their master dataset (Refer [149])	Text, Images, Profile, social network & activity information
	Sina Weibo	Chinese	Depression	Positive: 580 users, 45461 posts Control: 580 users, 30920 posts	
Song et al. 2018 [66]	Reddit	English	Depression	Refer RSDD dataset [45]	Text
Naderi et al.	Reddit	English	Self-harm	Positive: 49845 posts	Text

Reference	Dataset's Key Characteristics (Research Question 5)				
	OSN Platform	Language	Mental Health Disorder	Cohort Size	Modalities
2019 [67]				Control: 128243 posts	
Ragheb et al. 2019 [68]	Reddit	English	Self-harm	They used the eRisk Lab 2018 depression & anorexia dataset [142]	Text
Allen et al. 2019 [69]	Reddit	English	Suicide	Refer UMSD V2 dataset [145]	Text
Morales et al. 2019 [70]	Reddit	English	Suicide	Refer UMSD V2 dataset [145]	Text
Mohammadi et al. 2019 [71]	Reddit	English	Suicide	Refer UMSD V2 dataset [145]	Text
Ambalavanan et al. 2019 [72]	Reddit	English	Suicide	Refer UMSD V2 dataset [145]	Text
Matero et al. 2019 [73]	Reddit	English	Suicide	Refer UMSD V2 dataset [145]	Text
Stankevich et al. 2019 [74]	VKontakte	Russian	Depression	Positive: 148 users, 10693 posts Control: 239 users, 20706 posts	Text, Profile information
Tadesse et al. 2019 [75]	Reddit	English	Depression	Refer dataset by Pirina et al. [151]	Text
Maupome et al. 2019 [76]	Reddit	English	Depression	Positive: 214 users, 90222 posts Control: 1493 users, 986360 posts This dataset is a sample drawn out of the eRisk Lab 2018 dataset [142]	Text
Buddhitha et al. 2019 [77]	Twitter	English	Depression	Refer CLPsych 2015 dataset [138]	Text
Gaur et al. 2019 [78]	Reddit	English	Suicide	Positive: 2181 users (total), avg. 31.5 posts per user Out of the above: 500 users were manually annotated into five risk severity levels or bands using the C-SSRS scale	Text
Sinha et al. 2019 [79]	Twitter	English	Suicide	Total: 34306 Tweets (across 32558 users) Positive: 3984 Tweets Control: 30322 Tweets	Text, Social network & activity information
Mishra et al. 2019 [80]	Twitter	English	Suicide	Extended the dataset by Sinha et al. [79] to include all historical posts of 32558 users (min. 100, avg. 748 historical Tweets per user)	Text, Social network & activity information
Cao et al. 2019 [81]	Sina Weibo	Chinese	Suicide	Positive: 3652 users, 252901 posts Control: 3677 users, 491130 posts	Text, Images, Profile & activity information
Gui et al. 2019 [82]	Twitter	English	Depression	Positive: 1402 users, 292564 posts Control: 1402 users, 556033 posts This sample is drawn out of a prior dataset by Shen et al. (Refer [149])	Text
Gui et al. 2019 [83]	Twitter	English	Depression	The users are from a prior dataset by Shen et al. (Refer [149]) for whom images in their Tweets (if available) were crawled. Positive: 1402 users, with 251834 textual posts & 40730 text + image posts Control: 1402 users sampled from 5160 users, with 3303366 textual posts & 650817 text + image posts	Text, Images
Wang et al. 2019 [84], 2020 [85]	Sina Weibo	Chinese	Depression	Positive: 2158 posts (with different severity levels ranging from 1 to 3) Control: 11835 posts (severity level 0)	Text
An et al. 2020 [86]	Twitter	English	Depression	Refer dataset by Gui et al. [83]	Text, Images
Ophir et al. 2020 [87]	Facebook	English	Suicide	Total: 1002 users, 83292 posts Positive: 361 users Control: 641 users	Text
Sawhney et al. 2020 [88]	Twitter	English	Suicide	Refer dataset by Sinha et al. [79] and Mishra et al. [80]	Text, Posting activity information
Lee et al. 2020 [89]	Naver Café	Korean	Suicide	Positive: 10000 posts Control: 21723 posts Date for creating SoWE: 6093 Sina Weibo posts (Chinese), 2410 Naver Café posts (Korean), Reddit dataset (English) [78]	Text

Reference	Dataset's Key Characteristics (Research Question 5)				
	OSN Platform	Language	Mental Health Disorder	Cohort Size	Modalities
Kim et al. 2020 [90]	Reddit	English	Depression	Total: 248537 users, 633385 posts for 6 MH disorders Out of the above: Depression Positive: 136506 users, 258496 posts	Text
Alabdulkreem et al. 2020 [91]	Twitter	Arabic	Depression	Positive: 200 users, 10000 posts	Text
Carvalho et al. 2020 [92]	Twitter	Portuguese	Suicide	Positive: 1181 posts Control: 1265 posts	Text
Tadesse et al. 2020 [93]	Reddit	English	Suicide	Refer dataset by Ji et al. [63]	Text
Yao et al. 2020 [94]	Reddit	English	Suicide	Positive: 51366 posts	Text
Rao et al. 2020 [95]	Reddit	English	Depression	Refer RSDD dataset [45]	Text
Sekulic et al. 2020 [96]	Reddit	English	Depression	Refer SMHD dataset [59]	Text
Jiang et al. 2020 [97]	Reddit	English	Depression	Positive: 3183 users, 1585000 posts	Text
Rao et al. 2020 [98]	Reddit	English	Depression	Refer RSDD dataset [45] and Refer eRisk Lab 2017 dataset [141]	Text
Ji et al. 2020 [99]	Reddit, Twitter	English	Suicide & Depression	Three datasets used for experiments: UMSD V1 [57], Combined Twitter dataset [63][138], and 54412 Reddit posts collected by authors	Text
Liu et al. 2020 [100]	Sina Weibo	Chinese	Suicide & SICE	Positive: 5994 posts annotated with SIC Control: 7019 posts with no SIC	Text
Mann et al. 2020 [101]	Instagram	Portuguese	Depression	Total: 221 users (students) at different severity levels as per BDI score, or at no/minimal risk, ~ 26 posts per user in the last 212 days	Text, Images
Lin et al. 2020 [102]	Twitter	English	Depression	The users are from a prior dataset by Shen et al. (Refer [149]) for whom images in their Tweets (if available) were crawled.	Text, Images
Ramírez-Cifuentes et al. 2020 [103]	Twitter	Spanish	Suicide	Total: 252 users, 1214474 posts, 3056387 images 84 users in each class: positive, control & focused control (users who used suicide lexicon but weren't positive themselves)	Text, Images, Social network & activity information
Cao et al. 2020 [104]	Sina Weibo	Chinese	Suicide	Refer to their prior dataset [81] of Sina Weibo users	Text, Images, Profile, social network & activity information
	Reddit	Chinese	Suicide	Reddit users' dataset (each user has min. 100 posts]: Positive: 392 users and Control: 108 users	
Chiu et al. 2020 [105]	Instagram	English, Chinese	Depression	Positive: 260 users, 9458 posts Control: 260 users, 22286 posts	Text, Images, Activity information
Bagherzadeh et al. 2020 [106]	Reddit	English	Self-harm	Refer eRisk Lab 2020 dataset [143]	Text
Achilles et al. 2020 [107]	Reddit	English	Self-harm	Refer eRisk Lab 2020 dataset [143]	Text
Uban et al. 2020 [108]	Reddit	English	Self-harm	Refer eRisk Lab 2020 dataset [143]	Text
Madani et al. 2020 [109]	Reddit	English	Depression	Refer eRisk Lab 2020 dataset [143]	Text
Castano et al. 2020 [110]	Reddit	English	Self-harm	Positive: 120 users, 1346 posts Control: 875 users, 5585 posts	Text
			Depression	Refer eRisk Lab 2020 dataset [143]	
Maupome et al. 2020 [111] 2021 [112]	Reddit	English	Depression & Self-harm	Refer eRisk Lab 2020 dataset [143]	Text
Sawhney et al. 2021 [113]	Reddit	English	Suicide	500 users were selected randomly from a prior dataset by Gaur et al. [78] & manually	Text

Reference	Dataset's Key Characteristics (Research Question 5)				
	OSN Platform	Language	Mental Health Disorder	Cohort Size	Modalities
				annotated into five risk severity levels or bands using the C-SSRS scale	
Sawhney et al. 2021 [114]	Twitter	English	Suicide	Refer dataset by Sinha et al. [79] and Mishra et al. [80]	Text, Posting activity & social network information
Uban et al. 2021 [115]	Reddit	English	Depression & Self-harm	Combined all available datasets from CLEF's eRisk Labs (2017 to 2020) for both tasks Refer [140][141][142][143]	Text
Ren et al. 2021 [116]	Reddit	English	Depression	Refer dataset by Pirina et al. [151]	Text
Ragheb et al. 2021 [117]	Reddit	English	Depression & Self-harm	Datasets from CLEF's eRisk Labs Refer [142] [152] [143]	Text
			Suicide	Refer UMSD dataset [57][145]	
Zogan et al. 2021 [118]	Twitter	English	Depression	Positive: 2159 users, 447856 posts Control: 2049 users, 1349447 posts This sample is drawn out of a prior dataset by Shen et al. (Refer [149])	Text, Social network information
Murarka et al. 2021 [119]	Reddit	English	Depression	Total: 17159 posts (for 5 MH disorders and control class) Out of these posts: Depression: 3062 posts Control: 2478 posts	Text
Gollapalli et al. 2021 [120]	Twitter	English	Suicide	Refer CLPsych 2021 Shared Task dataset [146]	Text
Morales et al. 2021 [121]	Twitter	English	Suicide	Refer CLPsych 2021 Shared Task dataset [146]	Text
Wang et al. 2021 [122]	Twitter	English	Suicide	Refer CLPsych 2021 Shared Task dataset [146]	Text
Bayram et al. 2021 [123]	Twitter	English	Suicide	Refer CLPsych 2021 Shared Task dataset [146]	Text
Renjith et al. 2021 [124]	Reddit	English	Suicide	Refer UMSD V2 dataset [145]	Text
Basie et al. 2021 [125]	Reddit	English	Depression & Self-harm	Refer eRisk Lab 2021 dataset [150]	Text
Inkpen et al. 2021 [126]	Reddit	English	Depression	Refer eRisk Lab 2021 dataset [150]	Text
Lopes et al. 2021 [127]	Reddit	English	Self-harm	Refer eRisk Lab 2021 dataset [150]	Text
Ghosh et al. 2021 [128]	Twitter	English	Depression	Refer dataset by Shen et al. [149]	Text (not tweets but his bio/description, Images
Uban et al. 2021 [129]	Reddit, Twiter	English	Depression	Refer eRisk Lab datasets [141][142][143][140], CLPsych 2015 dataset [138], and dataset by Shen et al. [149]	Text
			Self-harm	Self-harm: Refer eRisk Lab 2020 dataset [143]	
Farruque et al. 2021 [130]	Twitter	English	Depression	Positive: 255 posts	Text
Zogan et al. 2022 [131]	Twitter	English	Depression	Positive: 5899 users, 508786 posts Control: 5160 users, 2299106 posts This sample is drawn out of a prior dataset by Shen et al. (Refer [149])	Text, Activity & Social network information
Kour et al. 2022 [132]	Twitter	English	Depression	Refer dataset by Shen et al. [149]	Text
Ahmed et al. 2022 [133]	Not Mentioned	English	Depression	15044 posts	Text
Naseem et al. 2022 [134]	Reddit	English	Depression	Refer eRisk Lab 2020 dataset [143]	Text
Ansari et al. 2022 [135]	Reddit, Twiter	English	Depression	Refer CLPsych 2015 dataset [138], eRisk Lab 2018 dataset [142] [140], and dataset by Pirina et al. [151]	Text
Cheng et al. 2022 [136]	Instagram	English	Depression	Positive: 526 users, 20618 posts Control: 528 users, 23772 posts	Text, Images, Posting activity information
	Twitter	English	Depression	Refer dataset by Gui et al. [83]	
Zeberga et al. 2022 [153]	Reddit, Twitter	English	Depression	Reddit: 75000 posts Twitter: 25000 posts	Text

Reference	Dataset's Key Characteristics (Research Question 5)				
	OSN Platform	Language	Mental Health Disorder	Cohort Size	Modalities
Coppersmith et al. 2014 [154]	Twitter	English	Depression	Positive: 441 users, 1 million tweets Control: 5728 users, 13.7 million tweets	Text
CLPsych 2015 dataset by Coppersmith et al. [138]	Twitter	English	Depression	Positive: 477 users, minimum 25 most recent tweets per user (maximum 3000) Control: matched to the positive class	Text
Coppersmith et al. 2015 [14]	Twitter	English	Depression	Positive: 393 users, 546000 tweets Control: matched to the positive class	Text
Coppersmith et al. 2015 [147], 2016 [155]	Twitter	English	Suicide	Positive: 125 users, minimum 100 most recent tweets per user (maximum 3200) Control: matched to the positive class	Text
SAD dataset by Mowery et al. 2016 [3], 2017 [139]	Twitter	English	Depression	Total 9300 tweets Positive: 2471 tweets Control: 6829 tweets	Text, Profile information
CLPsych 2016 Triage dataset by Milne et al. [144]	ReachOut	English	Self-harm	Total: 65024 posts during 2012 Jul - 2015 Jun Positive: 1227 posts	Text
RSDD-Time dataset by MacAvaney et al. 2018 [137]	Reddit	English	Depression	Temporal annotations for 598 posts that were selected from the RSDD dataset [45] in order to include information related to time aspects, e.g., time of diagnosis, time span or duration it lasted, if it is still present, etc.	Text (along with its time series annotation information)
CLEF 2016 Test Collection dataset by Losada et al. [140] a.k.a. eRisk Lab 2017 dataset [141] or eRisk Lab 2018 dataset [142]	Reddit	English	Depression	Positive: 137 users, 49580 posts Control: 755 users, 481873 posts (posts are chronologically ordered)	Text (along with its chronological order related information)
Bell Let's Talk Dataset by Jamil et al. 2017 [148]	Twitter	English	Depression	Positive: 53 users Control: 101 users (avg. 3864 words per user)	Text
Shen et al. 2017 [149]	Twitter	English	Depression	Positive: 1402 users, 292564 Tweets Control: 300 million users, 10 billion Tweets Possible Depression Candidate Class for experiments: 36993 users, 35076677 Tweets	Text, Images, User's Profile information, Social network, Activity & interaction information
CLPsych 2019 dataset by Zirikly et al. a.k.a. UMSD V2 dataset [145]	Reddit	English	Suicide	Positive: 621 users, Control: 621 users with total posts: 1105 (Task A), 66625 (Task B), and 70327 (Task C) This dataset is a sample drawn out of the UMSD V1 dataset by Shing et al. [57]	Text
Pirina et al. 2018 [151]	Reddit	English	Depression	Positive: 1200 users Control: 641 users	Text
CLEF's eRisk Lab 2020 & 2021 dataset by Losada et al. 2020 [143] 2021 [150]	Reddit	English	Task 1: Self-harm	Positive: 41 users, 6927 posts Control: 299 control users, 163506 posts	Text (along with its chronological order related information)
	Reddit	English	Task 2: Depression	Positive: 20 users, 10941 posts, and their responses to the BDI questionnaire (from eRisk Lab 2019 Task 3 [152])	
CLPsych 2021 Shared Task dataset by MacAvaney et al. [146]	Twitter	English	Suicide	Subtask 1 (Prior 30 days' data): 136 users, Average of 24 posts per user (68 Positive, 68 Control) Subtask 2 (Prior 6 months' data): 194 users, Average of 102 posts per user (97 Positive, 97 Control) Both the datasets are a sample from their prior dataset by Coppersmith et al. (Refer [62] in Table 4)	Text

3.4 Review Findings, Analysis & Results

In this section, we present the key findings, analysis, and results from the systematic literature review of 96 research publications included in this survey. This has helped us understand the answers to the Research Questions (RQs) we defined in section 3.1. In order to discern the answers to our RQs, we inspect the technical aspects of the deep learning architectures proposed by the 96 research articles, which are also summarized in Table 3.1 and Table 3.2. We now discuss our key findings, analysis and summarize the results pertinent to our RQs. For further details, the readers may refer to these tables.

3.4.1 Research Question 1

RQ1. What deep learning techniques and model architectures have been proposed for detecting depression, self-harm, and suicide from online social media?

We first enumerate the popular deep learning techniques used by various research studies reviewed in this SLR. As shown in Figure 3.2, CNN, LSTM, and BERT (including its variants like RoBERTa, DistilBERT, etc.) are the most preferred and commonly used deep learning techniques in this area. In addition to these, Bi-Directional LSTM networks and feed-forward deep neural networks were also prominent among the research articles for this domain.

Next, we focus on the deep learning model architectures proposed in the literature (Refer to Figure 3.3). Many researchers have proposed complex and advanced deep neural architectures by combining multiple deep learning techniques. 15% of included research articles have created ensembles of various deep networks, 20% of studies have proposed a deep hierarchical network, and 16% of articles have designed cascaded or hybrid/fusion deep networks. Additionally, 38% of research articles have also utilized Attention layers in their proposed deep neural network. However, as is evident from Figure 3.3, significant research has not been done to utilize multitask learning (5%), transfer learning (6%), reinforcement learning (2%), and XAI (6%) architectures for this research domain.

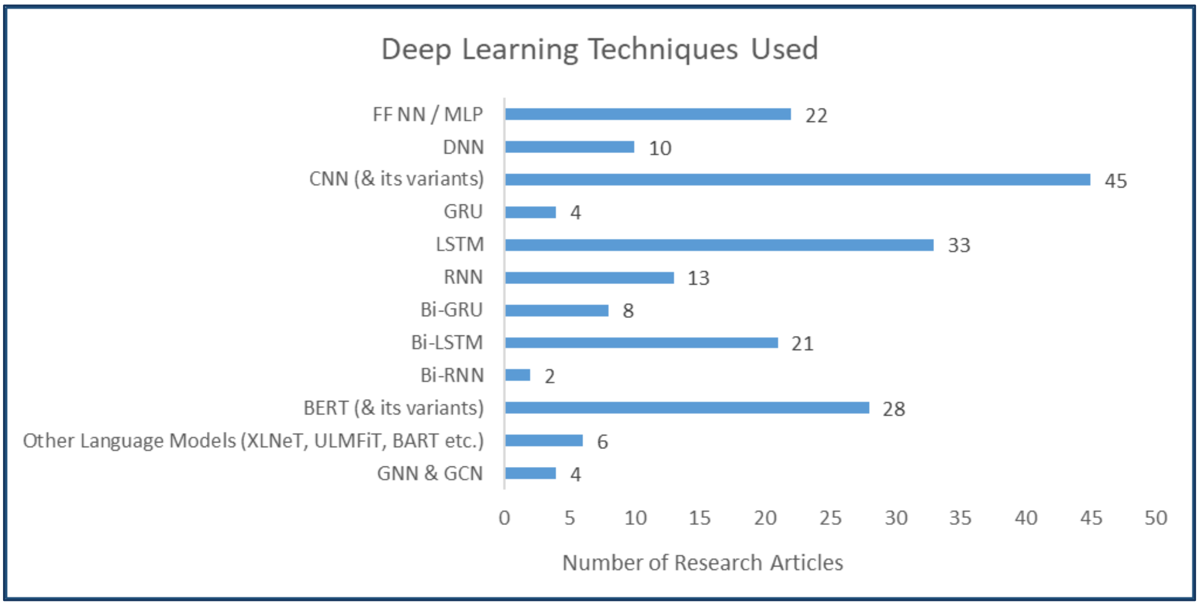


***Figure 3.2** Deep Learning Techniques used by publications included in this SLR (RQ1)*

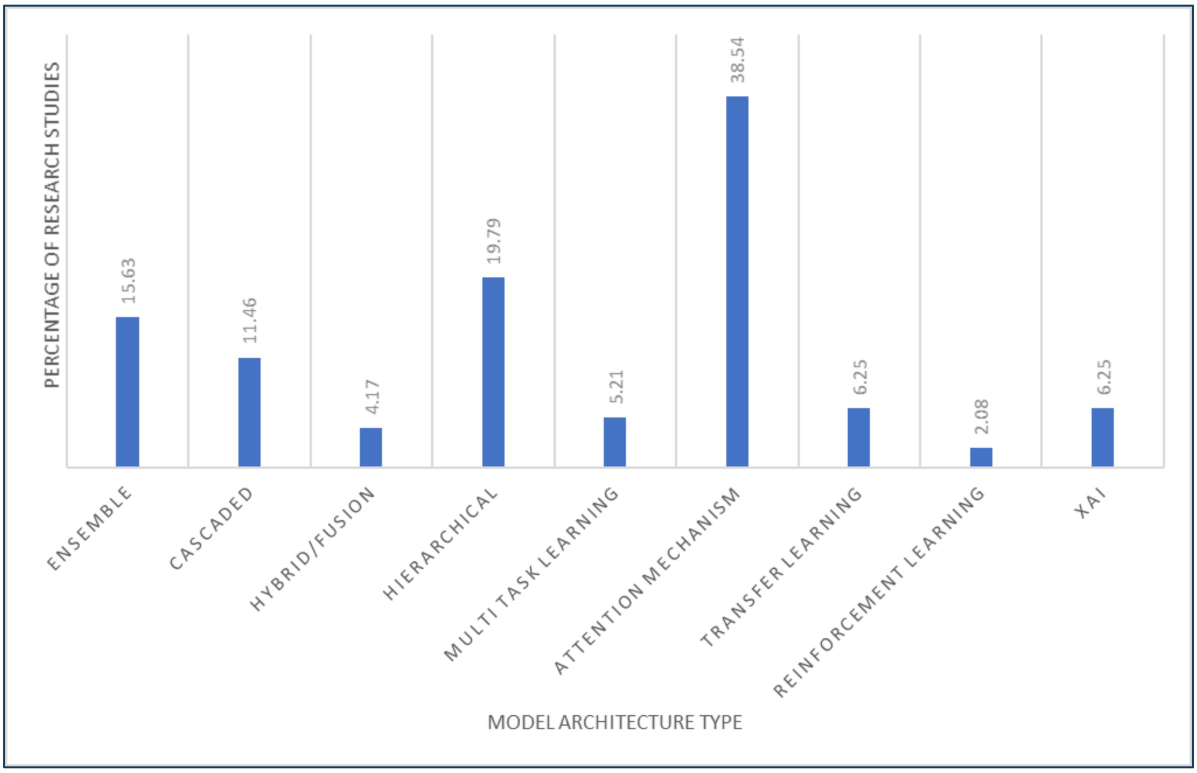


***Figure 3.3** Deep Learning Model Architectures proposed by research studies included in this SLR (RQ 1)*

3.4.2 Research Question 2

RQ2. What user-generated content modalities and their corresponding feature representation techniques have been used as input for training the above deep learning models?

Next, we summarize the key learnings and inferences related to the training of the deep learning algorithms mentioned in subsection 3.4.1 above. We inspected the UGC dataset modalities that were used. We examine their corresponding feature generation techniques that were used to create the input feature representation vectors for training the deep learning classifiers mentioned above. As demonstrated in Figure 3.4, the majority of the research articles used only unimodal textual datasets for model development. Only 22% of research articles selected for this literature survey have attempted to develop multimodal deep neural network models for this research problem. Amongst them, five research studies have utilized text and images from UGC datasets [83] [86] [101] [102] [128]. Ten publications have made use of available user information (such as that related to the user's profile, behavior, network, activity, etc.) in addition to the text posted by these users [42] [46] [58] [74] [79] [80] [88] [114] [118] [131]. Only six research articles made use of all the modalities, i.e., text, images, and available user information [65] [81] [103] [104] [105] [136].

For the textual content in UGC datasets, the common feature extraction and representation techniques used to create the input feature vector are indicated in Figure 3.5. Neural embeddings are the most popular choice since they outweigh the other handcrafted features in terms of performance. In addition to Char2Vec / Doc2Vec / Word2Vec embeddings that are obtained by training neural networks on relevant datasets, various pre-trained or fine-tuned neural embeddings have also been utilized. Some of the most popular pre-trained embeddings are: Word2Vec (CBOW, Skip-gram), fastText, GloVe, and BERT. CNNs were used to create feature representations for images in the UGC datasets collected from OSNs. Several handcrafted numerical and statistical features were computed for users' profile, network, activity, behavioral information, and other meta-data. The most important and commonly used feature extraction and representation techniques by the 96 research studies are briefly explained in Table 3.3.

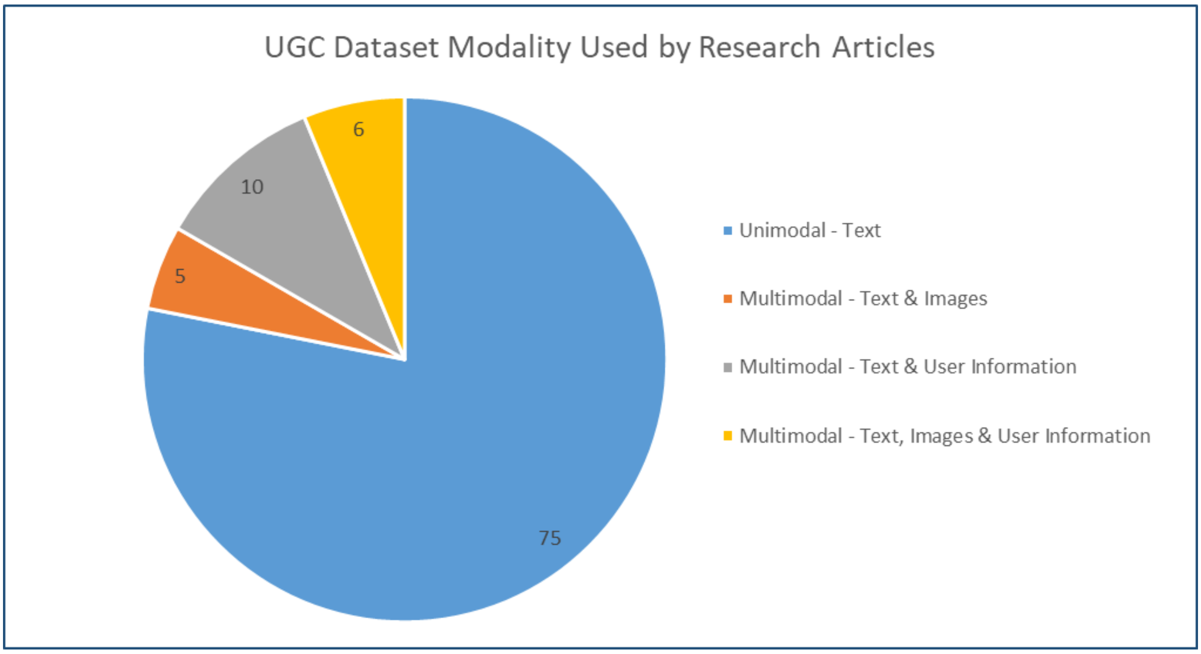


Figure 3.4 *Distribution of UGC Dataset Modalities used by publications included in this SLR (RQ 2)*

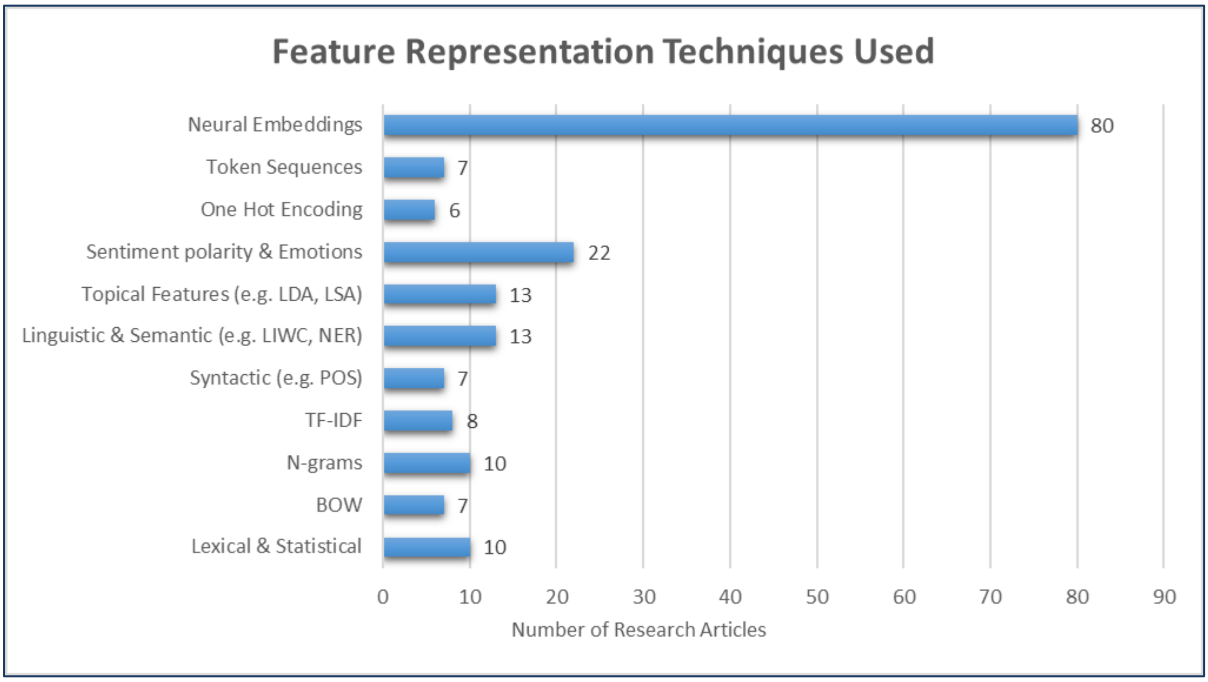


Figure 3.5 *Popular Feature Representation Techniques used by publications included in this SLR (RQ 2)*

Table 3.3 *Brief Description of Feature Extraction & Representation Techniques used by research studies included in this Systematic Literature Review*

Feature Extraction Technique	**Brief Description / Key Characteristics**
Lexical & Statistical features	These are usually frequency or count based features relating to the lexicon/vocabulary of the language. They are created by tokenizing text into meaningful chunks (words or phrases). Examples: Total number of special characters/punctuations/sentences / stop words in the input text, Frequency of occurrence of domain lexicons, e.g., counts of depression/suicide related words. The three most commonly used lexical features: BOW, N-grams, and TF-IDF are explained next. Count Vectorizer is used for creating these features. [156]
BOW	It is the simplest of all language representation models, where input text is treated as a multiset of words. The word order and grammatical/structural information are discarded, and the feature representation vector is built only based on the number of occurrences of the word token. [156]
N-grams	It is based on the frequency of occurrence of N continuous tokens or grams, hence the name N-gram. This model preserves some level of structural information. BOW explained above is a unigram (1-gram) model. Bigram (2-gram) and trigram (3-gram) models are the popularly used ones. [156]
TF-IDF	Term Frequency – Inverse Document Frequency score assigns weights to a term based on how much significant information its occurrence provides. Terms that are frequent across all documents are given less weightage, and the occurrence of rare terms is given higher weights. It is a product of the frequency of a term within a document, inversely scaled with the percentage of the total number of documents it appears in. [156]
Syntactic features	These are features related to the syntax or grammar of any language. Examples include POS tagging, counts of verbs/nouns/adjectives, or grammatical rules. [156]
Linguistic & Semantic features	These features help understand the semantics / linguistic relevance of words or phrases in the input text. Examples: NER, Word sense disambiguation, Relationship extraction, LIWC, LSA, Authorship features, e.g., writing style (active voice / passive voice, etc.) [156]
Topical features	These features are related to the abstract topics/themes/concepts within a text document. Example: LDA algorithm builds a statistical model to understand key topics within a document by identifying the words related to a topic within a document. [156]

Feature Extraction Technique	Brief Description / Key Characteristics
Sentiments & Emotion features	These features are used to measure the sentiment polarity and identify the different emotions expressed in the text, e.g., joy, sadness, etc. Emoticons / Emojis used within the text and various emotion-related lexicon databases, e.g., ANEW, NRC, VADER, are used for this purpose. [156]
One Hot Encoding	One hot encoding is used to generate vector representations for text and categorical variables. Every word/character is assigned a unique vector comprising of only 0s and 1s. These vectors are then used to create final vector representations for the sequence of texts. [157]
Token Sequences	This feature representation technique is used for building input feature vectors for Transformer based classifiers. E.g., BERT-based classifiers use WordPiece tokenizers to convert input text to a sequence of tokens with special markers. [158]
Neural Word Embeddings	Word embeddings are dense, low dimension vector representations for words/tokens in the input text. The vector consists of floating point numbers that are learned by training a dense neural network. Words with similar meanings are encoded with similar vectors. These learned vectors are then used as lookup tables to create document/sentence representations for NLP tasks. Some of the frequently used techniques for creating word embeddings are explained next. [159]
Word2Vec	It is a statistical technique to learn word embeddings from a large text corpus using shallow neural networks. Two network architectures were proposed: CBOW and Skip-gram. [160]
GloVe	It is an unsupervised technique to learn word embeddings based on statistical co-occurrence patterns of words in the training tech corpus. [161]
fastText	Extension of Word2Vec technique for learning word representations. Word embeddings are learned for N-grams of the word, which are later combined to get the word representations. This helps the model understand prefix/suffice patterns and smaller words. Using learned N-gram embeddings, the model can construct embeddings even for the words not in the training corpus. [162]
BERT embeddings	These are contextualized word embeddings learned using BERT-based models. Embeddings learned for one task using BERT-based classifiers are saved and then later used to create word representations during another related machine learning task. Hence, these are also called pre-trained embeddings. Output from any of the 12 transformer layers in the BERT encoder can be used as embeddings for the word token. [163]

3.4.3 Research Questions 3 & 4

RQ 3. What are the various online social networking platforms for which deep learning models to detect depression, self-harm, and suicide have been developed?

RQ 4. What human languages have been taken into consideration by researchers for building deep learning systems for this research problem?

We answer these research questions by examining the characteristics of datasets used for training the deep learning models in the 96 research studies included in our SLR. Refer to Figure 3.6 and Figure 3.7 below for RQ 3 and RQ 4, respectively. As can be inferred from these figures, the past research has focused primarily on English language user-generated content, collected mainly from Twitter and Reddit platforms. Other popular OSNs like Facebook and Instagram, as well as regional OSNs like Sina Weibo (in China), Vkontakte (in Russia), and Naver Café (in Korea), have not been given sufficient attention. Enough datasets are unavailable to carry out similar research for other regional languages and regional / less popular social networks. Significant research has not been conducted to develop multi-lingual deep learning models to detect even the commonest of all mental health issues (depression, self-harm, and suicide).

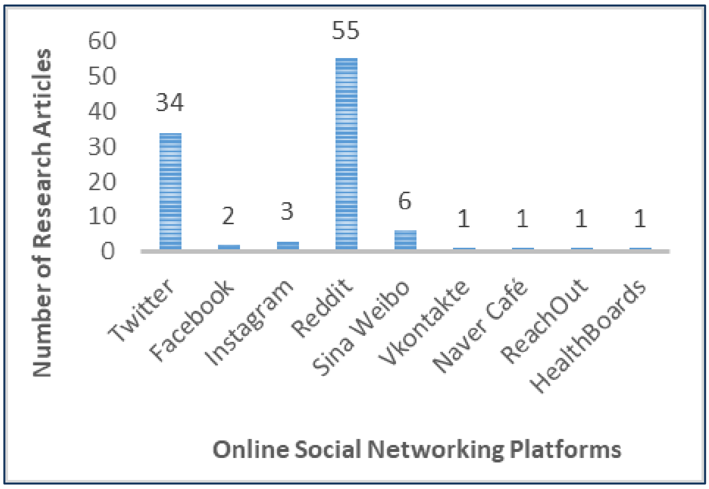


Figure 3.6 *Online Social Network Datasets used for training deep learning models for mental health assessment tasks (RQ3)*

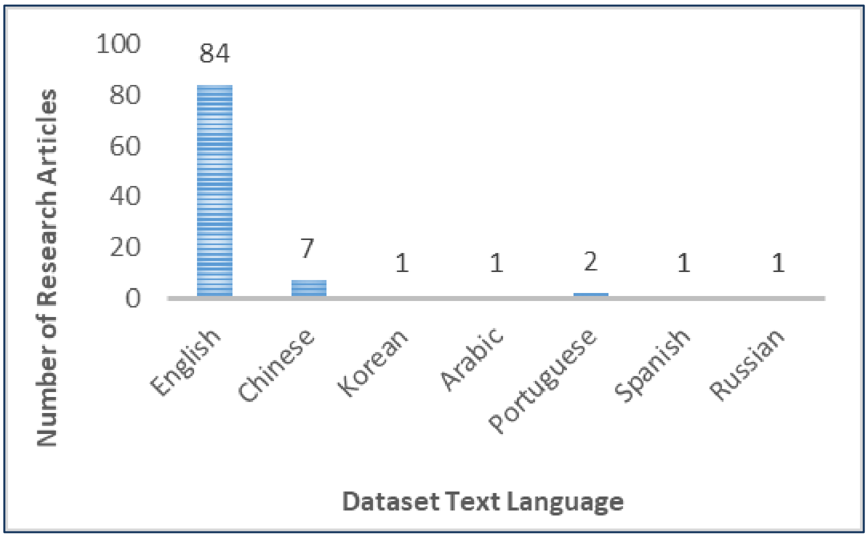


Figure 3.7 *Language Distribution of various UGC datasets used for training deep learning models for mental health assessment tasks (RQ4)*

3.5 Research Gaps

The following research gaps and open research challenges were identified after conducting the systematic literature review:

1. The existing research studies have mainly focused on improving the correctness of the classification decision by using complex deep learning networks that largely remain like a black box to the user. Their decisions are difficult to explain and interpret by humans, which hinders their wide adoption for real-world use cases. In the real world, model explainability/interpretability is equally important as model correctness to build the user's trust in an AI system, especially for critical applications like healthcare. To overcome this major drawback, we have focused our research primarily on designing explainable and interpretable supervised and unsupervised deep learning models using state-of-the-art LLMs.

2. One of the common limitations across almost all the research publications reviewed is that they have trained their proposed neural networks with imbalanced datasets, where positive class samples are much less than control class samples due to various challenges associated with data collection and annotation. Though collection of UGC from the Internet may still be

feasible via available Web APIs, however, annotating large UGC for any research problem still remains a challenge. We have proposed the use of Few Shot Learning using pre-trained LLMs (Transfer Learning) and Deep Active Learning to overcome this challenge for low resource research domains or domains where data annotation is difficult.

3. Most of the research studies have focused on unimodal UGC, i.e., Text. Very few studies have focused on multimodal deep learning techniques. We have conducted preliminary work on extending our research to the categorization of multimodal user generated content from the Internet.

4. The existing state-of-the-art research studies have not prioritized the applications of deep learning techniques for multilingual or code-mixed user-generated content classification and have largely focused on only English language content.

5. Sufficient research focus has not been given to temporal and ordinal classification research problems related to user-generated content on the Internet.

Through the research work presented in the following chapters, we have tried to mitigate and address some of the above research gaps.

3.6 Discussion & Summary

Through this literature review, we have focused on understanding the current state-of-the-art, research gaps, open challenges, and future research directions for advancing research applications of deep learning techniques for categorizing user generated content available on the Internet for various real-world social computing problems. The survey has helped vastly in learning about the most recent deep learning techniques and model architectures for text categorization and for creating deep neural feature representations/embeddings.

Most of the surveyed articles employed variants of neural embeddings for feature representation, and convolution or sequential deep learning networks for the primary classification task. Survey analysis revealed that ensemble, hybrid, and cascaded deep neural network architectures attained higher classification performance by combining different neural architectures and benefitting from all. It was observed from survey findings that various pre-trained neural word embeddings, e.g., GloVE, fastText, etc., used for creating text feature representations improve the overall classification accuracy. The network is initialized with these pre-trained embeddings or weights, which are then fine-tuned during network training. This transfer learning approach has proven to be beneficial since large, annotated UGC datasets are not easily available.

Deep learning techniques are known to improve the classification performance for unstructured data and alleviate the challenges related to handcrafted feature engineering required for training machine learning algorithms. This systematic review has shown that deep learning techniques have wide applications for innovative social computing applications using user generated content from the Internet. A feedback loop that can validate the predictions made by a real-world data-driven decision making system is essential to gauge and ascertain the utility of deep learning techniques for real-world use cases. Future research demands the development of cross-platform, multi-lingual, multimodal, multitasking, explainable/interpretable social computation systems with privacy, ethics, fairness, governance related principles enforced by design.

CHAPTER 4

EMPIRICAL REVIEW & EVALUATION OF DEEP LEARNING TECHNIQUES

This chapter presents the comparative results from an empirical review and evaluation of all popular supervised deep learning neural networks to benchmark their performance for a real-world UGC text categorization task using two publicly available mental healthcare datasets.

In this chapter, we review, compare, and empirically evaluate all popular supervised deep neural networks to benchmark their performance for a real-world UGC text categorization task using two publicly available mental healthcare datasets. This was essential to do in order to gain more insights about various deep learning neural networks since the existing research studies we surveyed have used different datasets that are often not publicly available and have focused on different tasks. The datasets used by these studies in the SLR conducted above were dissimilar as they were collected from different sources and were in varied languages. Also, researchers used diverse feature representation and feature selection techniques in combination with different neural networks. They have not necessarily compared all supervised deep networks on the same dataset using similar feature representations. We also observed that, at times, the researchers had reported the performance of their proposed techniques using different evaluation metrics. All these nonuniformity factors make it difficult to compare and infer which deep neural network architecture is best suited for a given real work application for UGC text categorization.

The rest of the chapter is organized as follows: in Section 4.1, we compare the network architecture, strengths, limitations, and applications of the most popular supervised deep learning algorithms; Section 4.2 mentions the characteristics of the datasets used in this research; in Section 4.3 we discuss the experimental details and presents the results of the empirical evaluation of these deep learning techniques; the chapter ends with a summary of key takeaways from this research in Section 4.4

4.1 Comparison of Supervised Deep Learning Algorithms

In this sub-section, we discuss in-depth key characteristics of the two most important categories of supervised deep learning algorithms: convolutional neural networks and sequential neural networks. We analyze and compare their network architectures, strengths, and limitations to better understand their applications w.r.t. NLP domain. In this study, we have also included a lesser known variation of convolutional networks: Temporal Convolutional Networks (TCNs) which are as effective as sequential networks but are much more efficient to train like convolutional networks.

4.1.1 Convolutional Neural Networks

Convolutional Neural Network is a specific type of dense feed-forward neural network that was initially proposed for computer vision tasks such as object recognition and proved to be a major research breakthrough for the domain [6] [24] [38]. Hence, their 1-D version was used for NLP classification tasks. In addition to the dense neural layers, typically, a CNN includes multiple convolution layers, pooling layers, ReLu, batch-norm, and dropout layers. Unlike dense neural layers that learn global patterns, the convolutional layer consists of multiple filters or kernel functions that operate over small spatial regions of the training input instance to learn local patterns. These learned abstractions are stacked and then passed to the next layer. These local patterns capture the neighborhood context or low-level surrounding information around features. ReLu or rectified linear unit layers are added to introduce non-linearity in the learning process. Pooling, batch-norm, and dropout layers are added to the network for dimensionality reduction to prevent overfitting and reduce computation steps. The major drawbacks of CNN are that it takes fixed-size input and can't handle variable length sequences, fails to capture positional or orientation information, and is not spatially invariant to the data; hence, it is not very suitable for sequence classification tasks. Hence, they don't perform very well for long or variable-length text classification, time series data, and audio and video data streams. For example, to predict the Nth word of a sentence, the prediction should be a function of all the previous (N-1) words of the sequence and not just the last (N-1)th word. Sequential or Recurrent

Neural Networks discussed next are the deep learning networks designed for tasks requiring sequential or temporal input data modeling.

4.1.2 Sequential Neural Networks

A recurrent neural network is a sequential deep learning network with internal memory cells or states to persist and process historical information over time. Unlike ANN or CNN, where inputs and outputs are considered to be independent of each other, RNN has hidden layers with shared weights and a feedback loop where current output is fed back as an input to be used in computation at the next time step. An RNN layer can have as many computation steps that are equivalent to multiple unfolded hidden layers, and the last output is compared with the target output or is passed as an input to the next layer of a deep neural network. At every computation step or time instant, an RNN unit in a network makes prediction Y'_t using the current input X_t and the internally stored hidden state $H_{t,}$ which is basically the last output of the unfolded previous hidden layer. With an internal hidden memory state that is rewritten at every time instant, the prediction error is backpropagated through time (BPTT algorithm). RNN can process variable-length input sequences, and the model size does not increase with the input size [6] [25].

Theoretically, RNN was designed to learn long-range sequential patterns; however, its practical implementation suffers from short-term memory issues, which makes it difficult to make accurate predictions if the previous states that influence the current state are not in its recent past context. Vanishing and exploding gradients are the main drawbacks of vanilla RNN, which is why it cannot remember historical/contextual information over long-range sequences or multiple computation time steps. When the gradient of the model loss becomes extremely small, the weights updates to the model parameters become insignificant (vanishing gradients), whereas when the error gradients accumulate, the gradient grows exponentially and results in large updates to model weights during training (exploding gradients). These factors lead to low performance or accuracy and increased training time.

Exploding gradients problem can be easily solved in practice by gradient clipping technique where the maximum value of the gradient is capped at a threshold. The vanishing gradients issue can be prevented in multiple ways, like using the ReLu activation function, Multilevel hierarchical networks, Residual networks, or by using the gated variants of RNN, namely: LSTM and GRU. These are a special type of RNN that can remember and retain relevant information for longer time periods so that it can be used later for computation. Their default behavior or cell structure helps to handle long-term input dependencies.

These RNN variants have additional gates or units within each RNN cell to determine the information flow through the network that is required for computations/predictions at each step and accordingly update the network weights. LSTM extends a typical vanilla RNN layer with three additional memory gates: forget, input, and output [26]. Forget gate determines how much past data should be remembered by looking at the current input (e.g. if there is any change in context), and it deletes/omits the rest. Input gate decides what information should be allowed for use in the current time step based on its level of importance w.r.t. current step. Output gate chooses what information from the current cell state should be passed on as the output to the following layers. GRU is a simplified version of LSTM with fewer gates and parameters, making it memory efficient and faster to train as compared to LSTM [27]. Instead of separate internal cell states like in LSTM, GRU has hidden states that are selectively updated by two gates: Reset and Update. The Reset gate decides how much of the past information, i.e., the previous hidden state, should be forgotten by the network; it is equivalent to a combination of Forget and Input gates of LSTM. The Update gate determines the quantum of new input information to use for updating the hidden state, which will then be used to compute the GRU layer output and flow as future information. All these sequential networks described above have corresponding Bi-directional architectures as well. In addition to using the previous hidden states or past information, they also make use of future information states from the input sequence to make predictions at the current time step [28]. For this reason, they have two hidden layers connected to a single output layer to flow information in both directions: forward and backward, which helps to improve classification accuracy.

4.1.3 Temporal Convolutional Networks

In this section, we explain Temporal Convolutional Networks [39]. Because of the limitations of CNNs discussed above and the ability of RNNs to model long-range dependencies within the input sequence, canonical RNNs or other RNN-based architectures are used extensively for sequence modeling or text classification tasks. TCNs are a simpler and equivalent alternative network to RNNs, that combine the advantages of both: RNNs and CNNs. They are able to capture long-term dependencies like RNNs; and unlike CNNs they can handle variable length input sequences, they don't suffer from vanishing or exploding gradient problems, and they are faster to train due to their parallel architecture [39].

A TCN is designed using causal, dilated convolutions and residual blocks. A TCN can be best described as:

$$TCN = 1\ Dimensional\ Fully\ Convolution\ Network + Dilated\ Causal\ Convolutions + Residual\ Connections \qquad \text{(Eq. 4.1)}$$

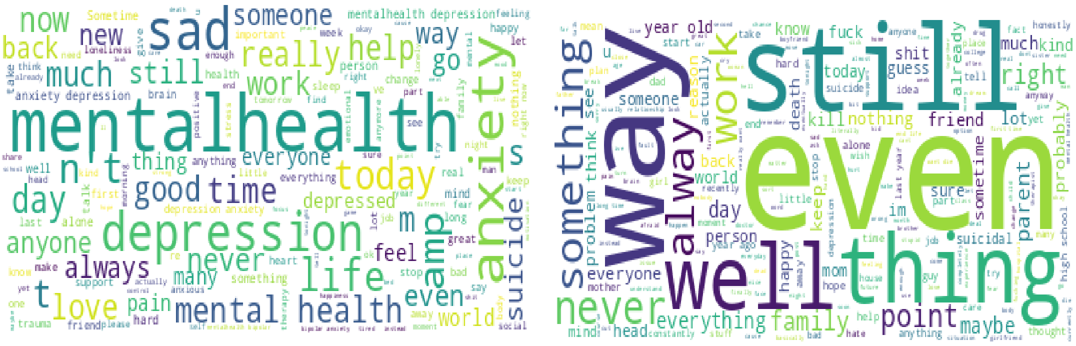
TCN uses 1D FCN with padding to keep each hidden layer the same in size as the input layer so that it can take variable-length inputs and produce an output of the same length as the input given. Causal convolutions ensure that future state information doesn't influence past decisions; hence, convolutions at any given time instant T are done using information from the previous hidden layers only up to time instant T. Simple causal convolutions can remember historical information in a linear proportion of the depth of the network. To retain and model long-range dependencies within a sequence, dilations are used that help expand the network's receptive field and increase the effective historical memory size of each network layer. Using a higher dilation factor, downstream or top-level output layers can represent abstract information from a wider input sequence length. Residual blocks stabilize deep neural networks with large receptive fields as they help the network learn changes to identity mapping rather than carry out multiple transformations across layers. A residual connection takes the output of one convolution layer and connects it as an input to another layer later within the block [40]. A typical TCN residual block consists of dilated causal convolution, ReLU (for non-linearity), weight normalization, and spatial dropout layers (for regularisation).

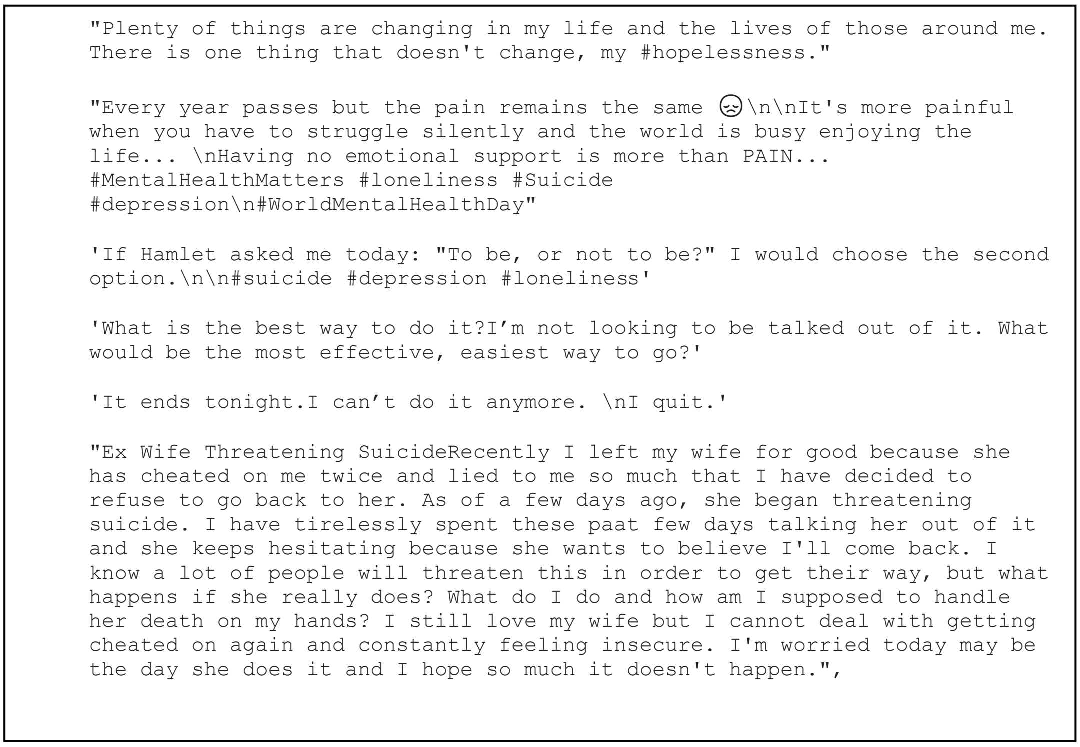
4.2 Datasets

In this section, we discuss the characteristics of two publicly available mental healthcare UGC datasets that we used to conduct the empirical evaluation of supervised deep learning networks. The first dataset was collected from Twitter using its API and Tweepy library; it contains around twenty thousand Tweets related to depression [195]; we refer to this as the *Depression Dataset* in this chapter. The second dataset was collected from Reddit using its Pushshift API and consists of more than two hundred thousand posts made on SuicideWatch and Depression subreddits; we refer to this as the *Suicide Dataset* in the following sections [196]. Both these datasets were balanced; hence, we did not apply any undersampling or oversampling techniques (Refer to Table 4.1). Some of the positive samples from these datasets for depression and suicide classes and word clouds of commonly occurring words in these classes are shown in Figure 4.1. The control class covers other day-to-day online conversations on a variety of topics.

***Table 4.1** Dataset Characteristics*

Dataset	Positive Class	Control Class
Depression Dataset [195]	11466 tweets	12054 tweets
Suicide Dataset [196]	116015 posts	115952 posts




```
"Plenty of things are changing in my life and the lives of those around me.
There is one thing that doesn't change, my #hopelessness."

"Every year passes but the pain remains the same ☹\n\nIt's more painful
when you have to struggle silently and the world is busy enjoying the
life... \nHaving no emotional support is more than PAIN...
#MentalHealthMatters #loneliness #Suicide
#depression\n#WorldMentalHealthDay"

'If Hamlet asked me today: "To be, or not to be?" I would choose the second
option.\n\n#suicide #depression #loneliness'

'What is the best way to do it?I’m not looking to be talked out of it. What
would be the most effective, easiest way to go?'

'It ends tonight.I can’t do it anymore. \nI quit.'

"Ex Wife Threatening SuicideRecently I left my wife for good because she
has cheated on me twice and lied to me so much that I have decided to
refuse to go back to her. As of a few days ago, she began threatening
suicide. I have tirelessly spent these paat few days talking her out of it
and she keeps hesitating because she wants to believe I'll come back. I
know a lot of people will threaten this in order to get their way, but what
happens if she really does? What do I do and how am I supposed to handle
her death on my hands? I still love my wife but I cannot deal with getting
cheated on again and constantly feeling insecure. I'm worried today may be
the day she does it and I hope so much it doesn't happen.",
```

***Figure 4.1** Positive class samples of Depression and Suicide related user generated content from social media datasets used in this research [195] [196]*

Text Pre-processing & Cleaning: As discussed previously in Chapter 2, User-generated textual content collected from online social networks (Web 2.0, Web 3.0) is noisy and error-prone as users write in a free-form language and don't proofread their content for spelling and grammatical mistakes like it is done for professionally published content (Web 1.0). They also use abbreviations, emoticons, and social media specific features like hashtags / @mentions, etc. All of these badly impact the classification accuracy of any ML or DL algorithm, and hence, it is essential to pre-process, clean, and standardize the user-generated text. To enhance the text quality, we apply the following NLP techniques: lowercase conversion, removing hashtags/emoticons/@ mentions, removing URLs, fixing broken Unicodes using Python's FTFY library for correct text interpretation, expanding commonly used text contractions, e.g., ain't / he'd, removing punctuations / special characters/numerics, tokenizing and removing common stop words, and at last lemmatizing the remaining word tokens to their base or root form from which they are derived. These pre-processing steps required for UGC text have already been discussed in detail in Section 2.1.

4.3 Experiments & Results

This section presents the detailed performance metrics from our empirical evaluation of various supervised deep neural networks discussed above using the UGC text datasets mentioned in Section 4.2. Additionally, we have also included Temporal Convolutional Networks, which can be considered a functional equivalent of canonical recurrent neural networks. TCN has demonstrated longer effective historical memory than an RNN with a similar network design [39]. Not many research studies in the past have used TCNs for UGC text classification tasks. We use generic TCNs with simple architecture and minimal tuning to evaluate their effectiveness for UGC text classification and to benchmark their performance with other popular supervised deep learning networks.

For our benchmarking experiments, we have kept the network design of all the supervised deep neural networks similar so that they have an equivalent number of trainable parameters or network weights. All networks were designed using a similar number of hidden layers, ReLu layers (for non-linearity) and Dropout layers (to prevent overfitting). This design decision was essential to estimate and compare their classification performance and computational requirements correctly. Uniform embedding layer was used for all networks with maximum vocabulary/feature size, output embedding dimension size, maximum input sequence length, and training batch size being kept as model constants for all networks. Keras TextVectorization was used for creating dense token sequence representations for the input text. All the networks were trained up to a maximum of three epochs for uniformity and also because validation loss started to increase since models started to overfit the training data. Keras with Tensorflow as the backend was used to implement all networks. Binary Crossentropy was used as the loss function, and Adam optimization was chosen. In all networks, the output or last layer was afully connected dense layer with the Sigmoid classifier. CNN and TCN networks had additional 1D convolutional and pooling layers. TCN layer was implemented using the Python Keras TCN library, and the dilation parameter was specified as [1,2,4]. The training and testing classification metrics for both datasets and the training time taken are mentioned in Table 4.2 (Depression dataset) and Table 4.3 (Suicide dataset). We also experimented with an ensemble or fusion architecture with a combination of different layers stacked together: TCN-CNN.

Table 4.2 *Depression Dataset Performance Metrics*

Deep Neural Network	**Training Performance**					**Testing Performance**					
	Time	*Loss*	*P*	*R*	*ACC*	*Loss*	*P*	*R*	*F1*	*ACC*	*AUC*
ANN/MLP	101 s	0.14	0.97	0.92	0.94	0.30	0.87	0.85	0.86	0.87	0.87
RNN	260 s	0.17	0.93	0.93	0.93	0.32	0.84	0.89	0.86	0.87	0.87
BiRNN	585 s	0.16	0.94	0.93	0.94	0.32	0.84	0.88	0.86	0.86	0.86
LSTM	948 s	0.14	0.94	0.94	0.94	0.30	0.84	0.89	0.86	0.87	0.87
BiLSTM	2253 s	0.13	0.97	0.93	0.95	0.31	0.87	0.84	0.85	0.87	0.86
GRU	803 s	0.14	0.94	0.95	0.95	0.34	0.83	0.89	0.86	0.87	0.87
BiGRU	1936 s	0.13	0.96	0.94	0.95	0.31	0.85	0.87	0.86	0.87	0.87
CNN	192 s	0.14	0.96	0.94	0.95	0.30	0.86	0.86	0.86	0.87	0.87
TCN	728 s	0.15	0.92	0.96	0.94	0.31	0.82	0.90	0.86	0.86	0.87
TCN-CNN	758 s	0.15	0.97	0.92	0.95	0.29	0.90	0.84	0.87	0.88	0.89

Table 4.3 *Suicide Dataset Performance Metrics*

Deep Neural Network	**Training Performance**					**Testing Performance**					
	Time	*Loss*	*P*	*R*	*ACC*	*Loss*	*P*	*R*	*F1*	*ACC*	*AUC*
ANN/MLP	740 s	0.06	0.98	0.98	0.98	0.20	0.93	0.93	0.93	0.93	0.93
RNN	2665 s	0.23	0.95	0.86	0.91	0.25	0.95	0.86	0.90	0.91	0.91
BiRNN	6097 s	0.19	0.94	0.92	0.93	0.21	0.94	0.91	0.92	0.92	0.92
LSTM	9964 s	0.15	0.94	0.95	0.94	0.21	0.87	0.97	0.92	0.91	0.91
BiLSTM	15361 s	0.15	0.96	0.93	0.95	0.16	0.95	0.90	0.92	0.93	0.93
GRU	8137 s	0.16	0.92	0.93	0.92	0.22	0.93	0.92	0.92	0.92	0.93
BiGRU	11976 s	0.15	0.96	0.92	0.94	0.17	0.95	0.91	0.93	0.93	0.93
CNN	1940 s	0.16	0.94	0.94	0.94	0.16	0.94	0.93	0.94	0.94	0.94
TCN	7654 s	0.14	0.94	0.96	0.95	0.16	0.93	0.95	0.94	0.94	0.94
TCN-CNN	7983 s	0.12	0.96	0.95	0.96	0.14	0.96	0.94	0.95	0.95	0.95

4.4 Discussion & Summary

Traditional machine learning algorithms require extensive feature engineering and selection effort based on domain knowledge to create feature representation vectors. On the other hand, deep learning techniques are known to automatically discover and learn complex, hidden patterns from large, unstructured (big) data. In this research work, we have reviewed and empirically evaluated the most popular supervised deep neural networks using two publicly available user-generated content datasets related to mental health (depression and suicide) from various social networking websites. We compare their performance using standard machine learning metrics and the time required for training them on these datasets. Additionally, we have also included a lesser-known variation of convolutional networks: Temporal Convolutional Networks, in our study. With novel architectural components like dilations, causal convolution, and residual connection block, TCN has strengthened the rudimentary CNN and adapted it for sequence classification tasks. TCN can be considered a functional equivalent to recurrent networks and requires much less training time in comparison to them. The results show that TCNs outperform, or at least benchmark, the recurrent neural networks that are widely used for sequence or text classification tasks.

To extend this research, in the future, we will also compare the performance of TCN and all the other deep neural networks mentioned above with Transformer-based Large Language Models, which have now captured the attention of researchers in the NLP domain. We would want to explore the performance of TCN when trained with pre-trained weights or pre-trained word embeddings that are currently popular e.g., BERT, and GPT.

CHAPTER 5

PROPOSED SYSTEM WITH TRANSFORMER BASED LLMs & XAI

This chapter covers a detailed description of the proposed system for supervised and unsupervised categorization of user generated text from the Internet by using Transformer-based LLMs and explaining the model predictions using XAI techniques. It elaborately discusses the qualitative and quantitative results from experimental evaluation with multiple LLMs and user-generated text datasets.

In this chapter, we discuss our proposed explainable and interpretable system for supervised and unsupervised categorization of user generated text from the Internet by using the latest breakthrough techniques in deep learning for NLP domain, i.e., Transformer based LLMs. Transformer-based Large Language Models have now become state-of-the-art for most natural language processing and computational linguistic tasks due to their unmatched prediction accuracy. However, unlike conventional machine learning algorithms, these deep neural networks are opaque black box architectures due to their complex internal structure, which makes it difficult to understand and explain their decisions. Clearly, there is a trade-off between model performance and model interpretability/explainability. However, model explainability is equally essential as model performance for real-world use cases, especially in crucial domains like healthcare.

The key objective of this research is to provide explainability and interpretability to classification decisions of pretrained LLMs (Transformers) trained for various UGC text categorization tasks. To achieve this, we have used the two most recent model agnostic, post hoc surrogate XAI techniques: LIME and SHAP. We have conducted extensive and in-depth experiments with six pretrained LLMs (BERT, DistilBERT, RoBERTa, MentalBERT, PsychBERT, PHSBERT) and finetuned them with four social media UGC datasets.

Next, we have demonstrated the use of the Transformer-based unsupervised topic modeling technique BERTopic to analyze large-scale unlabeled UGC datasets for deriving insights. We believe using LLMs in an unsupervised approach like the above can be useful for big data analytics for UGC on the Internet when supervised training of LLMs is not feasible due to dataset availability and annotation challenges.

At last, we have performed Few Shot Learning, which can be beneficial for low resource research domains, e.g., healthcare, where good quality, large annotated UGC datasets are unavailable or difficult to obtain. For these scenarios, pre-trained LLMs can be fine-tuned with only a few good quality data samples annotated by experts. We have done multiple Few Shot Learning (N-way K-shot) experiments with domain-adapted LLMs using various mental health-related UGC datasets to analyze and compare their performance.

The rest of the chapter is organized as follows: Section 5.1 provides detailed discussion around LLMs and XAI; Section 5.2 demonstrates our proposed system for UGC text categorization and explains the techniques we have used; Section 5.3 presents the results from our various experiments along with the details about the UGC datasets used for these experiments; in Section 5.4 we discuss the results of our Few Shot Learning with LLMs experiments; and at last in Section 5.5 we summarize the key takeaways from the research work presented in this chapter.

5.1 Introduction

This section provides a detailed discussion around the two key concepts in our proposed framework for user generated text categorization: Transformer based LLMs and XAI.

5.1.1 Transformer based Large Language Models

Transformers-based language models have become state-of-the-art for natural language modeling and understanding or interpretation tasks. They are pre-trained deep learning networks that use a mix of discriminative and generative deep learning techniques for Seq2Seq transformations. The TLMs are pre-trained on large text corpus in an unsupervised manner and can later be fine-tuned for downstream tasks by (supervised or unsupervised) training on smaller domain-specific datasets. Recurrent neural networks like LSTM are slow to train as they process the inputs sequentially to learn the context and suffer from long-range dependency

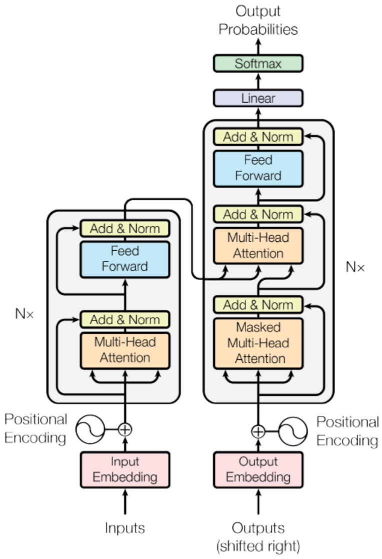


Figure 5.1 *Transformer Architecture (Image source: Vaswani et al. [20])*

and vanishing/exploding gradients issues [25] [26]. In convolutional networks, the length of context or long-range dependencies a network can learn depends on the kernel size and number of convolution layers used, and increasing either of them increases the computations required [24]. Transformers mitigate the drawbacks of these popular deep learning networks for learning the context and associations within the input sequence. Transformers are highly parallelized neural network architectures designed by stacking together multiple encoder and decoder layers, along with additional multi-head and self-attention mechanism layers [20] (Refer Figure 5.1). Attention mechanism generates attention vectors for each token to capture the relevance of other tokens towards it and models their contextual relations; this helps to decide what parts to focus on (give attention to) while processing a word token. The self-attention mechanism helps in modeling the relevant language context around each word without the need for recurrent or convolutional layers; the multi-head attention layer overcomes the vanishing gradient issue to model longer-range associations within the sequence. These attention layers are faster than recurrent layers, and the number of operations required is not dependent on the input size as is in the case of later. This network design makes them extremely efficient for training on large text datasets for seq2seq learning tasks [20]. This is the reason why Transformers are the foundation or backbone of all the recent deep contextualized language models, such as Google's BERT [21], OpenAI's GPT [22], and XLNet [23], where all of these have been pre-trained on large generic corpora. TLMs have been pre-trained on unlabeled, enormous general-purpose text readily available on the Web, like Wikipedia, to gain generic natural language understanding and knowledge about *worldly topics*. These pre-trained Transformer based Language Models can be used out of the box for transfer learning for various NLP/NLU tasks like text classification, sentiment analysis, summarization, named entity recognition, question answering, etc. They can be fine-tuned for these downstream NLP tasks using smaller domain-specific labeled datasets without the need for training from scratch, thus saving computation time and resources.

Out of all the state-of-the-art TLMs, the most popular and widely used is Google's open-source Bidirectional Encoder Representations from Transformers (BERT) model [21]. BERT does not use decoders and consists of only 12 (BERTbase) to 24 (BERTlarge) deeply stacked bidirectional Transformer encoder layers. BERT representations outperform other context-free word representations like word2vec and GloVe, as well as other unidirectional or shallowly bidirectional contextual representations like ELMo and ULMFit. BERT learns bidirectional contextual representations for sequence word tokens during the self-supervised pretraining stage for Masked Language Modelling (MLM) and Next Sentence Prediction (NSP) tasks. MLM is like a fill-in-the-blank task, where the BERT network tries to predict 15% randomly masked word tokens in unlabeled text sequences using the joint context from other words in left and right directions both in all the Transformer layers, in the process learning the (bidirectional) relationships between the words. NSP task enables BERT to learn sentence-level relationships between consecutive sentence pairs (A and B), where the goal is to predict whether sentence B is the actual sentence following sentence A or just any other random sentence from within the unlabelled monolingual text corpus (50% correct and incorrect sentence pairs each were used). BERT was trained on 64 TPUs for about four days, using 3.3 Billion words from Wikipedia and Google's BookCorpus for 1 million weight update steps [21] [164] [165].

5.1.2 Explainable AI (XAI)

In the AI and ML research domain, currently, there seems to be a trade-off between model performance (accuracy, precision, recall) vs. model interpretability and explainability. Conventional machine learning algorithms, e.g., Linear Regression, SVM, and Decision Trees, are white box or open models where it is easier to explain their predicted outcome due to the simpler mathematical operations/computations involved in arriving at the outcome [166]. However, their linearity (as with regression, SVMs) and their simpler mathematical transformation functions (like kernels in SVM) limit their model performance for tasks that involve training on (large) unstructured data, e.g., text, speech, images, and videos due to less expressivity. In deep learning literature, it has been widely studied that non-linearity assists in performance boost by helping in learning low-level and high-level abstract patterns and their hierarchical relations from unstructured data. Hence, recent research has focused on leveraging deep neural networks with complex architectures (cascade/ensemble/hierarchical) and nonlinear activation functions or various Transformer based Large Language Models to improve the classification accuracy and other metrics for unstructured data. TLMs have become state-of-the-art for natural language, speech processing, and computer vision tasks due to their unmatched competitive performance on unstructured data. However, due to their stacked or layered architecture, deep neural networks abstract the complex, nonlinear mathematical operations they perform to arrive at their decision. For such black-box deep learning networks, it is challenging to interpret and explain their decision-making process (interpretability, i.e., how the decision was made) and the logic or reasoning behind their predicted decision (explainability, i.e., why the decision was made). The best-performing deep neural network AI models require training billions of hidden parameters, which are not directly interpretable by humans. Additionally, there have been cases where these models gave correct predictions, but their decision was based on non-relevant parameters or metadata of the training data, e.g., the hardware used [167] [168]. However, model explainability should not be compromised for model performance, and alternate ways should be found to prioritize both.

Most of the existing research for user-generated content classification has focused on improving classification decision correctness through the use of complex predictive algorithms, with model decisions being largely inexplicable or difficult to interpret. However, for the increased adoption of these social computational systems in the real world for various applied machine intelligence tasks, model explainability is crucial for multiple reasons, like: to build user trust in the model decisions by increased transparency, objectivity, and reliability, prevent biases and discriminations, meet ethical, compliance and government regulatory requirements (e.g., prevent discrimination, Fair and unbiased treatment, GDPR's Right to Explanation), reduce physical danger, mitigate legal risks, and lastly, for improving model's performance further by understanding the false positives/negatives [169] [170] [171]. For sensitive and high-risk domains, e.g., healthcare, only a black box decision or outcome (however accurate it may be) without any explainability is not sufficient [172]. Users of the systems may not need to know the decision-making process (how the decision was made, i.e., interpretability), but they must at least understand why the decision was made by the system (explainability).

Due to these reasons, it is essential to develop eXplainable AI (XAI) systems where we can at least understand why behind what the systems have predicted. XAI systems are those that use techniques like: attention weight analysis, post-hoc methods (e.g., surrogate models), etc., to provide interpretability and explainability for the AI system so that humans can understand the decisions and decision-making process of the system [166] [169] [173] [174]. In our research, we have proposed and demonstrated the use of surrogate post hoc model agnostic techniques for LLMs that are not intrinsically interpretable so as to provide some degree of explainability to these LLM-based systems for UGC text categorization.

5.2 Proposed System

This section discusses our proposed explainable Transformer-based system for real-world UGC text categorization tasks, e.g., mental healthcare risk assessment. We propose a three-pronged approach using LIME [175], SHAP [176], and TopicBERT [177] techniques to interpret and explain the user's social media posts in order to understand their mental state, emotions, and behavior better. The key components of our proposed system are shown in Figure 5.2 and are explained in detail in the corresponding subsections below. After text extraction and pre-processing, supervised training, i.e., fine-tuning of multiple pretrained BERT-based LLMs, is done using various UGC mental health datasets. Then, LIME and SHAP are used to provide post hoc explainability to the decisions of these fine-tuned LLMs. Lastly, we demonstrate the applications of the transformer-based unsupervised topic modeling technique BERTopic, which can be useful for UGC text categorization for deriving insights from big UGC datasets when supervised training of LLMs is not feasible due to dataset availability and annotation challenges. The proposed methodology can also assist in causal analysis, knowing the topics, themes, issues, and concerns the users discuss online. Our proposed approach demonstrates a prototype of XAI Transformer-based language models for understanding, interpreting, and explaining users' mental health from their social media posts.

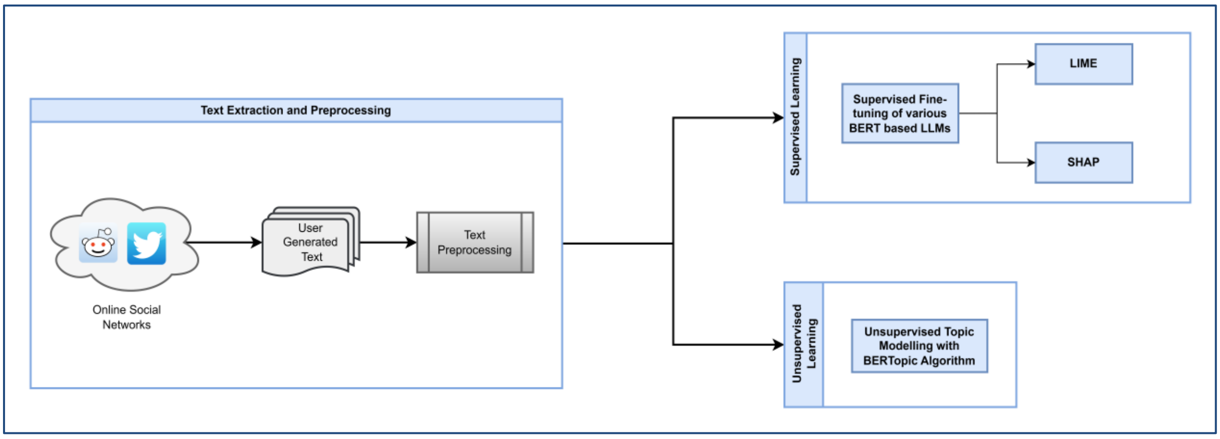

Text Extraction and Preprocessing
Online Social Networks
User Generated Text
Text Preprocessing
Supervised Learning
Supervised Fine-tuning of various BERT based LLMs
LIME
SHAP
Unsupervised Learning
Unsupervised Topic Modelling with BERTopic Algorithm

***Figure 5.2** Key Components of Proposed System for Explainable Transformer-based UGC Text Categorization*

5.2.1 Text Preprocessing and Supervised Training of Transformer Language Models

The first step is to clean and preprocess the textual user-generated content extracted from the various UGC datasets used in this research study (Refer to Section 5.3.1 for the details about the datasets used). Text written by users online is noisy and free-form text with lexical and grammatical errors and often contains abbreviations, special characters, and URLs. In order to clean and standardize the text before using it for training the various BERT-based LLMs, the following natural language preprocessing steps are done: conversion to lowercase, removal of emoticons/hashtags/@ mentions, removing URLs, fixing broken Unicodes with Python's FTFY library [178] to correct text interpretation, expanding commonly used text contractions, e.g., ain't / we'd, removing punctuations / special characters/ numerals, tokenizing and stopwords elimination, and finally lemmatizing the remaining tokens to their base or root word form from which they are derived. Text preprocessing is crucial while training conventional machine learning algorithms and is usually not required for BERT-based classifiers. However, we have done text cleaning and preprocessing primarily because: firstly, we have used uncased models (bert-base-uncased, distilbert-base-uncased, mental/mental-bert-base-uncased, PHS-BERT which uses uncased BERT); secondly, mnaylor/psychbert-cased is pretrained from the bert-base-cased checkpoint, and roberta-base model is case-sensitive [179]; hence we wanted to standardize the input text across all the LLMs for uniformity and consistency; and lastly, the pre-trained domain adapted LLMs (MentalBERT [180], PsychBERT [181], PHSBERT [182]) may have been fine-tuned differently and hence input text was standardized for uniformity and consistency. Also, case information of input text is mainly useful for NLP tasks like Named Entity Recognition and Part of Speech Tagging.

The cleaned and preprocessed text is then used to train and fine-tune various pre-trained BERT-based models (classifiers). Using the same cleaned datasets, we also retrained some other Transformer based language models proposed in recent research studies that have already been domain-adapted or fine-tuned with healthcare and mental health domain datasets. We have used the popular open-source Hugging Face Python library [179] to build the following models: BERT, DistilBERT, RoBERTa, MentalBERT [180], PsychBERT [181], and PHSBERT [182].

5.2.2 XAI for Transformer Language Models with LIME and SHAP

The XAI goal can be achieved in two ways: intrinsic or post hoc. Intrinsic explainability is achieved by restricting model complexity and using models with simpler structures like decision trees and linear models, which have the inbuilt capability to interpret their outcome through techniques like: Model weights analysis, Attention weight analysis, Feature analysis (like feature importance, pairwise feature interaction strength, features' partial dependence plots). Post hoc methods are attached as a surrogate component to explain an opaque model after it has been trained, and they do not have access to the model's internal architecture or parameter weights. These techniques use different interpretation methods, like: feature perturbations for counterfactual explanations and local/global approximations with simpler interpretable models. Further, the post-hoc XAI techniques may provide local explanations related to the prediction of a single instance or global explanations to understand the model's overall behavior and functioning or decision-making process. Lastly, post-hoc XAI techniques can be model-specific or model-agnostic. Model agnostic post hoc methods can be applied to any black-box model by providing input features and target output pairs [166].

We have used LIME and SHAP techniques to build a surrogate XAI framework for explaining and interpreting the predictions of various domain-adapted Transformer Language Models trained above using various UGC text datasets. These are the two most recent and reliable XAI techniques. LIME and SHAP are post-hoc XAI models that can be attached to complex, black box deep learning models that do not have an intrinsic characteristic of explaining or interpreting their predictions due to the large number of model parameters and non-linearity. Some other XAI techniques are: Attention mechanism, Anchors, DeepSHAP, DeepLIFT, and CXplain.

Local Interpretable Model-agnostic Explanations (LIME) is an XAI technique that has local fidelity and is model agnostic [166] [169] [175]. This means that LIME can be applied to any complex, black-box model with non-linear decision boundaries. Instead of interpreting the global model behavior, it localizes its interpretations around the single instance for which an explanation is needed. In the local vicinity of the instance, the classifier decision boundaries can be explained with linear models trained using perturbed data samples with a similar distribution as the instance for which prediction is needed. LIME generates 5000 perturbed data instances randomly and uniformly sampled, and then it obtains the target prediction decision for these samples from the black-box model. Then, a simpler, linear model is trained using these 5000 samples and the corresponding decision labels/target values above. Feature selection techniques (like Lasso) are then used to extract important features from perturbed instance feature vectors scaled with weights based on their distance from the original samples. However, there is an interpretability vs. fidelity trade-off, i.e., these explanations will not be valid universally for all predictions by the model. The process of generating LIME explanation scores explained intuitively above, can be mathematically represented by the following equation [166]:

$$\text{explanation}(x) = \arg\min_{g\in G} L(f, g, \pi_x) + \Omega(g) \qquad \text{(Eq. 5.1)}$$

where g is the explanation model (e.g., linear regression, decision tree) for instance x, f is the original model whose decision needs to be explained, π_x is the size of the neighborhood that was considered for generating model explanations, π_x is basically a proximity measure that is used to assign weightage to perturbed instances according to their distance from x, $\Omega(g)$ term determines the complexity of the explanation model. The goal of the explanation model is to minimize the Loss term L, which measures the closeness or correctness of the prediction to the prediction from the original model f. In simpler words, for the text classification domain, LIME generates perturbed samples by randomly removing words or tokens from the text sample for which an explanation is needed; the presence or absence of word tokens from the original text is depicted by 1 and 0 in the perturbed data samples generated. The original black-box model is then used to make predictions for these text variants, and the predicted probability for these is weighted by their closeness to the original text sequence (calculated as: 1 minus the number of words removed). This synthetic data is then used to train the linear explanation model by

LIME, and feature importance is computed by observing how much prediction error removing a word from the original text sequence has caused [166].

SHapley Additive exPlanation (SHAP) is another surrogate, post hoc, model-agnostic XAI technique [166] [169] [176] [183]. Unlike LIME, it does not build a simpler, linear model for local Interpretability; instead, it utilizes the concept of Shapley values from game theory [184]. In cooperative multiplayer games, optimal credit allocation of the pay-out of the game outcome is done based on the contribution made by each player towards achieving that outcome. Shapley values represent the marginal contribution of each player to the game outcome by computing their average contribution from all possible permutations of player orderings [185]. SHAP extends this game theory concept for XAI to explain the predictions of any machine learning model. In the context of machine learning, feature vectors are the players that are trying to generate the prediction for a data sample or instance. So, SHAP computes the Shapley values for each feature to measure its impact on the model's decision by calculating its contribution to the difference between the actual predicted probability and the mean prediction probability (model error) [184]. SHAP algorithm can also provide global explainability about the model's behavior by aggregating Shap values of features for all individual sample instances to compute their overall global importance [166] [176] [183]. The above concept can be explained using a simplified mock example inspired from Molnar et al. [166]. Suppose a machine learning model is trained to predict house prices using three features: area, floor, and amenities available. Let’s say for houses, each with 100 sq m. area and on the 3rd floor, but having different amenities such as air conditioning, furnishing, water heating, laundry, etc., the model predicts different prices. In this case, it is important to justify the pricing difference due to amenities to the customer. So here, the SHAP algorithm would compute the average price for a house on the 3rd floor with 100 sq m. area using all such samples from training data. Then, it will use this average price as a base value to compute the difference with the price of houses with various permutations of amenities, using which it will compute the average change in house price any amenity causes. We urge the reader to refer to Molnar et al. for a detailed mathematical calculation of this example [166].

LIME and SHAP are both widely acceptable methods for providing post hoc model explainability, and either or both of them may be used; but it is essential to understand that due to their algorithmic differences, occasionally they may give slightly different interpretations of the trained model. SHAP is theoretically more sound and rigorous, and also provides mathematical guarantees for the consistency and accuracy of model explanations it generates

[170]. Because of this, SHAP is more computationally expensive than LIME as it needs to compute Shapley values for all permutations of high-dimensional feature vectors in the dataset. This makes the practical implementations of SHAP (like KernelExplainer and TreeExplainer) slower, even with optimization and approximations. Hence, LIME is a faster and equally reliable alternative for model explainability. Thus, we have demonstrated the applications of both of these techniques is our research.

5.2.3 Unsupervised BERT Topic Modelling

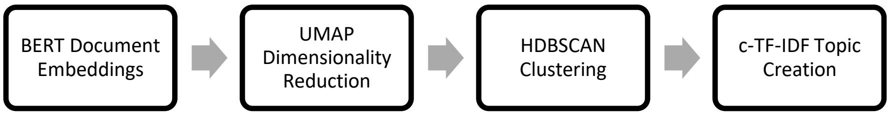


Figure 5.3 *BERTopic Algorithm*

Topic modeling is an unsupervised machine learning technique that can discover latent (hidden) topics (i.e., themes, subjects, issues) from unstructured text or documents. Thus, topic modeling can be a helpful tool to understand what users are expressing or discussing online when labeled data is not readily available for supervised text categorization. Some of the popularly used conventional topic modeling techniques are: Latent Dirichlet Allocation (LDA) [18] [186], Latent Semantic Analysis (LSA) [186] [187], and Non-Negative Matrix Factorization (NMF) [186] [188]. However, all of these are probabilistic techniques based on the bag-of-words model for document representation that ignore the semantic, positional, and contextual relationship between words of a document. They represent a document as a mixture of latent topics using the probability distribution of words in the document vector space. Additionally, the quality of the topics generated by these techniques depends on predefining the optimal number of topics (hyperparameter) and how the text preprocessing was done (stop words removal, stemming, lemmatization). Top2Vec technique for topic modeling was proposed recently to overcome these limitations [189]. Top2Vec utilizes distributed word and document neural embeddings (word2vec and doc2vec) that help in capturing semantic relationships between words, and semantically similar documents are also closer to each other in the vector space. BERTopic algorithm is a modified Top2Vec approach using the contextual pre-trained Transformer-based BERT embeddings and class-based TF-IDF scores for identifying interpretable topic representations [177] (Refer Figure 5.3). The first step is to generate document embeddings using the pre-trained language model DistilBERT Sentence Transformer, followed by a dimensionality reduction step using the UMAP algorithm [190] [191] to compress these high-dimension embeddings. Next, these reduced document embeddings are clustered using the HDBSCAN algorithm [192] [193] to group semantically related documents to ensure documents with similar topics are together in vector space. At last, cluster-level topic representations are extracted using a class-based TF-IDF approach (c-TF-IDF). We propose using the BERTopic modeling algorithm as a practical unsupervised alternative to analyze user-generated content from the Internet when data annotation or training supervised LLMs is not feasible.

5.3 Experiments & Results

Since BERT is the most frequently used state-of-the-art TLM, we have used various BERT-based TLMs in our experiments to demonstrate the results of our proposed approach for explainable Transformer based categorization of user generated text from the Internet (for the chosen research problem of mental health risk assessment). Other TLMs can also be used in a similar way by using their open-source model checkpoints [194] and can be finetuned with other domain UGC datasets.

5.3.1 Datasets

To conduct various experiments described next in the following subsections, we have used four publicly available, anonymized, class-balanced social network datasets related to depression and suicide detection that have been collected and made available for research purposes by other researchers working in this domain [195] [196] [119] [197] (Refer Table 5.1). The first dataset contains approximately eleven thousand Tweets related to depression or suicide scraped from Twitter using its API and Tweepy library [195]. Various depression and suicide-related keywords were used by the researchers to retrieve Tweets by mapping these keywords to the Tweet's metadata (Hashtags) and for annotating the Tweet as a positive class (Label 1) sample. The control class (Label 0) of equivalent size was built using random tweets from

Table 5.1 *Datasets Used*

	Class Name	***Cohort Size***
Dataset 1 [195]	Depression/Suicide Positive (Label 1)	11466 tweets
	Control (Label 0)	12054 tweets
Dataset 2 [196]	Depression/Suicide Positive (Label 1)	116015 posts
	Control (Label 0)	115952 posts
Dataset 3 [119]	Depression Positive (Label 1)	3062 posts
	Control (Label 0)	2478 posts
Dataset 4 [197]	Suicide (Label 1)	980 posts
	Depression (Label 0)	915 posts

Kaggle datasets. The control class may contain any other generic conversations on a variety of day-to-day topics. The dataset does not contain identifying information about the user. We refer to this as *Dataset 1* in this thesis/chapter.

The second dataset was collected from Reddit using its Pushshift API and consists of more than two hundred thousand posts made by users on Reddit's r/SuicideWatch, r/Depression, and r/Teenagers subreddits [196]. We have the 14th version of the dataset releases, in which the posts scraped from r/Teenagers are labeled as non-suicidal (Control class, Label 0), and all other posts are labeled as suicidal (Positive class, Label 1). The dataset does not contain identifying information about the user. We refer to this as Dataset 2 in this thesis/chapter.

The third dataset we have used is an anonymized Reddit dataset made publicly available by Murarka et al. [119]. A total of 17,159 posts were crawled from thirteen different subreddits. Five of these subreddits were associated with mental health: bipolar, ADHD, anxiety, depression, PTSD; and the other eight subreddits chosen by the authors covered a wide range of common day-to-day topics: music, travel, India, politics, English, datasets, mathematics, and science. The posts collected from these eight subreddits were annotated as class none, and the posts from five mental health-related subreddits were assigned class labels corresponding to the subreddit name. For our research, we selected the classes with labels: depression (Positive class, Label 1) and none (Control class, Label 0) from this dataset. We refer to this as *Dataset 3* in this thesis/chapter.

The fourth and last dataset we have used is the de-identified Reddit dataset released by Haque et al. in their recent research study related to differentiating between depression and severe suicidal tendencies on social networks [197]. They have scraped data of 1895 posts from two subreddits: r/SuicideWatch and r/Depression, and assigned class labels accordingly based on the subreddit the post belongs to. We refer to this as *Dataset 4* in this thesis/chapter.

5.3.2 XAI for Transformer Language Models with LIME and SHAP

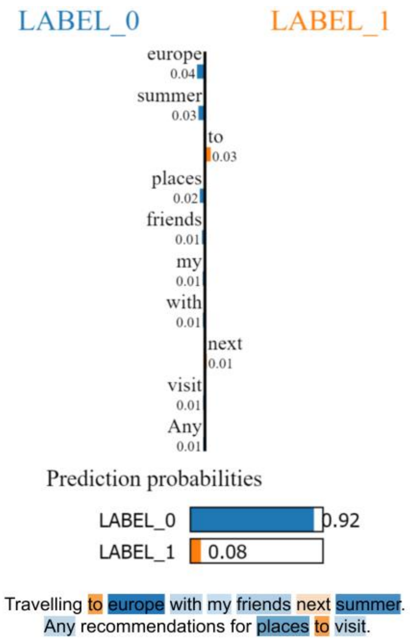
In this section, we present the qualitative analysis and results of our proposed explainable Transformer based language model detectors for depression and suicide ideation identification from social media posts.

Table 5.2 *Classification Performance Evaluation of LLMs*

LLM	**Dataset 1**				**Dataset 2**			
	Precision	*Recall*	*F1*	*Accuracy*	*Precision*	*Recall*	*F1*	*Accuracy*
BERT	0.886	0.846	0.866	0.873	0.967	0.963	0.965	0.965
DistilBERT	0.882	0.853	0.867	0.878	0.96	0.963	0.962	0.961
RoBERTa	0.903	0.858	0.879	0.883	0.962	0.966	0.964	0.963
MentalBERT	0.909	0.862	0.885	0.888	0.968	0.964	0.966	0.966
PsychBERT	0.906	0.856	0.88	0.885	0.964	0.962	0.963	0.963
PHSBERT	0.89	0.875	0.883	0.888	0.97	0.964	0.967	0.967

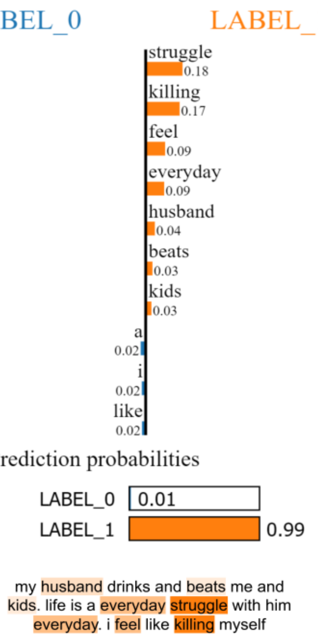
We train and fine-tune three domain-independent pre-trained TLMs (BERT, DistilBERT, RoBERTa) and three domain-adapted pre-trained TLMs (MentalBERT, PsychBERT, PHSBERT) using Dataset 1 and Dataset 2 mentioned above, and then explain their predictions for positive and control class text samples using LIME and SHAP. The goal is to understand, interpret, and explain why a post was classified as a positive class, i.e., depression/suicide positive (LABEL_1) vs. control class, i.e., neutral (LABEL_0) OR vice versa. Though the focus of our research study is qualitative analysis related to the explainability and interpretability of these LLMs, but before that, we first present their quantitative classification performance evaluation metrics in Table 5.2 above.

The following Figures 5.4 to 5.9 demonstrate the outputs of the LIME explainer for classification decisions of the transformer-based language model trained with the above datasets for sample user posts. In the graphical outputs, the classes are color-coded: LABEL_1 as orange and LABEL_0 as blue. The darker shades of these colors signify that the word had a higher contribution in the prediction probability score computed (using softmax) for their respective classes. The feature's numerical contribution to the softmax probability score is also indicated. This makes it intuitive to understand and explain the classification output of black box TLMs. For example, the words: Europe, summer, friends, and traveling have helped in the correct classification of a true negative sample (Fig 5.4a and Fig 5.4d). Also, as can be seen from these figures, for BERT trained on Dataset 2, LIME has assigned these words a much higher score as compared to BERT trained on Dataset 1, which hints about the possible data quality issues with Dataset 1 during annotation. Whereas words like struggle, killing, feeling, and everyday help in the correct identification of a true positive sample (Fig 5.4b). At the same time, the word “emotion” has contributed heavily, with a 0.44 softmax probability out of a 0.79 predicted score for a sample, which has led to a false positive classification (Fig 5.4c). The remaining figures for the other five LLMs can also be explained in a similar way. From these LIME explanations, we can observe some of the words that have contributed towards the correct identification of true positive class are: struggle, pain, sadness, husband, and killing. Likewise, some of the top contributing word features for the true negative class are: excited, snow, wonderland, NYC, and Europe. We have used dummy examples or sample text for demonstration purposes only. However, in real life, the above analysis with the UGC can help healthcare professionals in better diagnosis of mental health disorders.



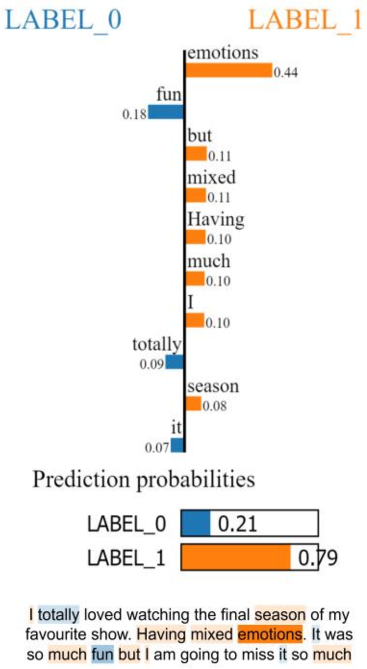
Travelling to europe with my friends next summer. Any recommendations for places to visit.

Figure 5.4(a) *LIME explanation for control class (LABEL 0) classification by BERT trained on Dataset 1*



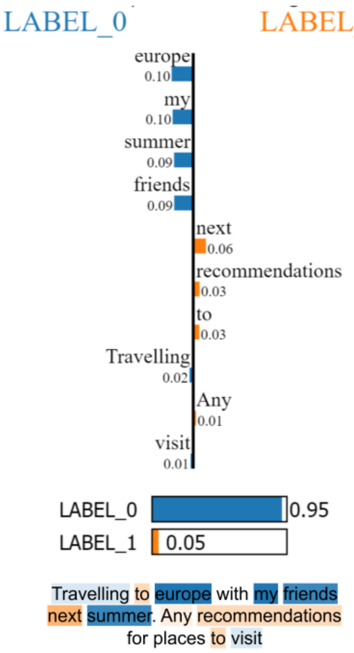
my husband drinks and beats me and kids. life is a everyday struggle with him everyday. i feel like killing myself

Figure 5.4(b) *LIME explanation for positive class (LABEL 1) classification by BERT trained on Dataset 1*



I totally loved watching the final season of my favourite show. Having mixed emotions. It was so much fun but I am going to miss it so much

Figure 5.4(c) *LIME explanation for misclassified sample by BERT trained on Dataset 1 (False Positive)*



Travelling to europe with my friends next summer. Any recommendations for places to visit

Figure 5.4(d) *LIME explanation for control class (LABEL 0) classification by BERT trained on Dataset 2*

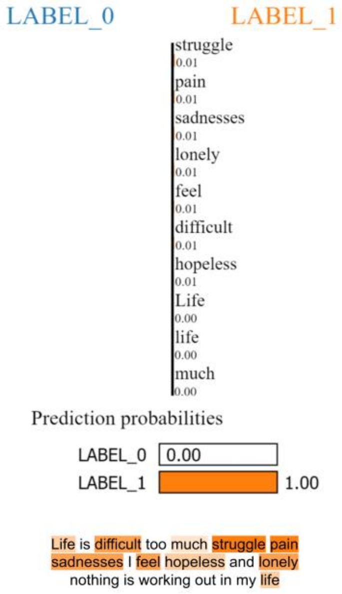


Figure 5.5(a) *LIME explanation for positive class (LABEL 1) classification by DistilBERT trained on Dataset 1*

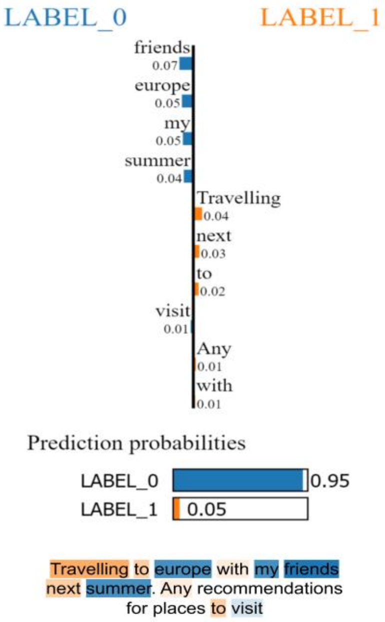


Figure 5.5(b) *LIME explanation for control class (LABEL 0) classification by DistilBERT trained on Dataset 2*

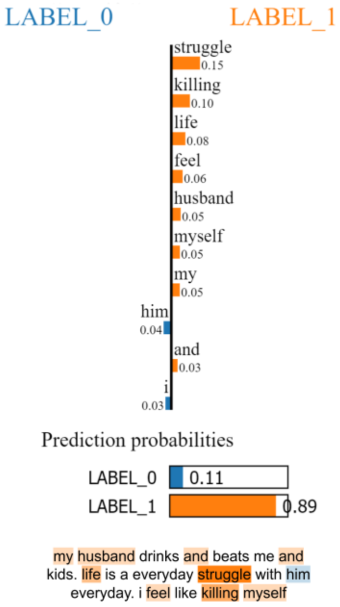


Figure 5.6(a) *LIME explanation for positive class (LABEL 1) classification by RoBERTa trained on Dataset 1*

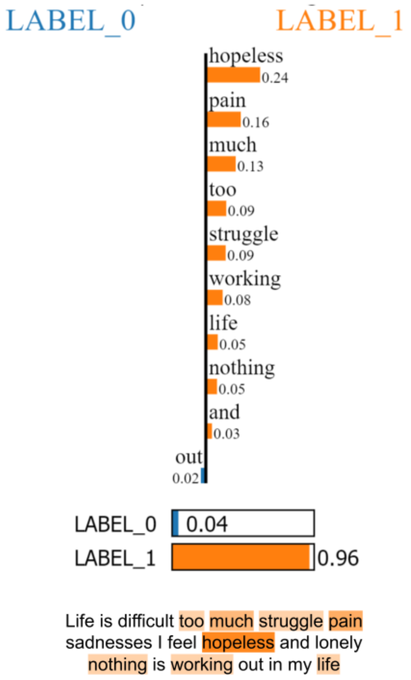
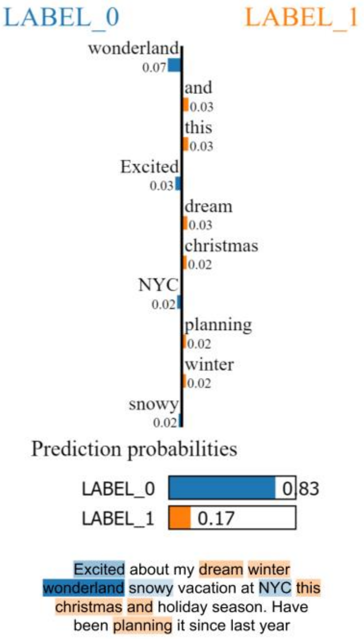
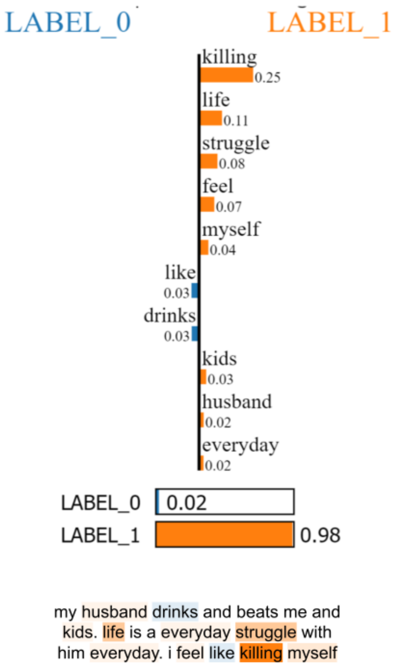


Figure 5.6(b) *LIME explanation for positive class (LABEL 1) classification by RoBERTa trained on Dataset 2*



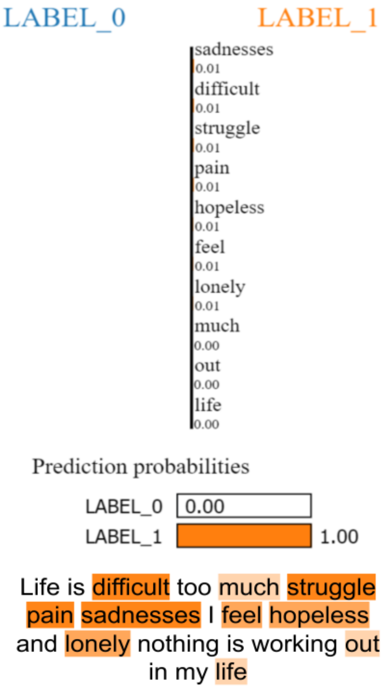
Excited about my dream winter wonderland snowy vacation at NYC this christmas and holiday season. Have been planning it since last year

Figure 5.7(a) *LIME explanation for control class (LABEL 0) classification by MentalBERT trained on Dataset 1*



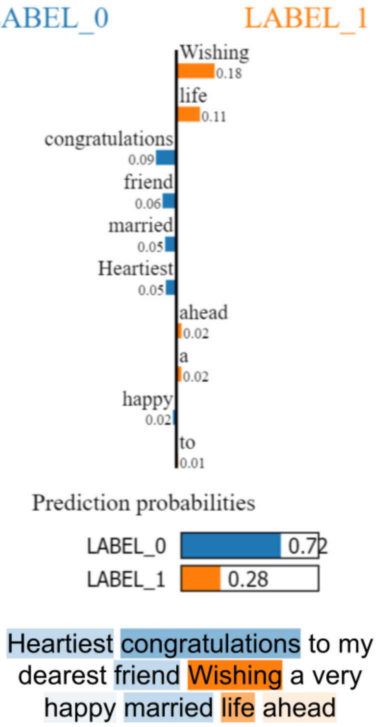
my husband drinks and beats me and kids. life is a everyday struggle with him everyday. i feel like killing myself

Figure 5.7(b) *LIME explanation for positive class (LABEL 1) classification by MentalBERT trained on Dataset 2*



Life is difficult too much struggle pain sadnesses I feel hopeless and lonely nothing is working out in my life

Figure 5.8(a) *LIME explanation for positive class (LABEL 1) classification by PsychBERT trained on Dataset 1*



Heartiest congratulations to my dearest friend Wishing a very happy married life ahead

Figure 5.8(b) *LIME explanation for control class (LABEL 0) classification by PsychBERT trained on Dataset 2*

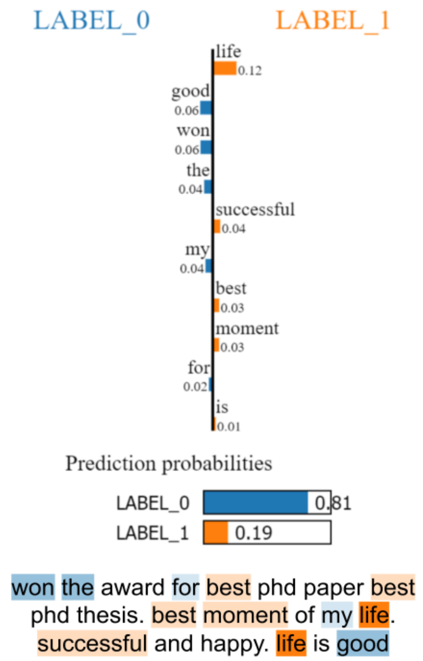


Figure 5.9(a) *LIME explanation for control class (LABEL 0) classification by PHSBERT trained on Dataset 1*

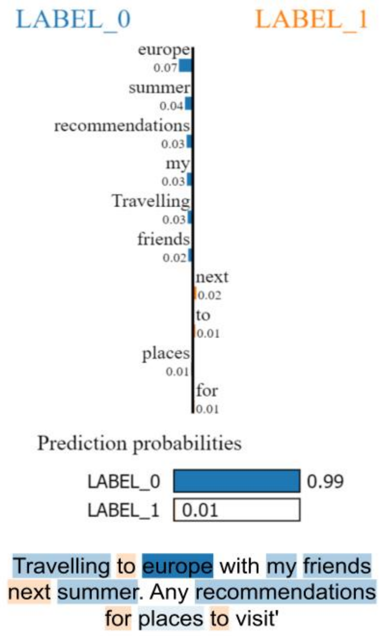


Figure 5.9(b) *LIME explanation for control class (LABEL 0) classification by PHSBERT trained on Dataset 2*

Text Sample 1: Depression/Suicide Positive Class (Label 1)

I lost my wife and child in a car accident last month. Unable to accept the loss, bear the pain, the fear of life that lies ahead for me. Feel so lonely and hopeless all the time. Not able to focus on my work at all. Taking anti-depressant pills. Every day is a struggle. Can’t live this life anymore. Life seems so meaningless without them. I feel like killing myself and committing suicide.

Text Sample 2: Control/Neutral Class (Label 0)

I am happy to share that I have been awarded as the best performer for the year and promoted to the position of Director. It’s a dream come true. Feeling happy and blessed. Special thanks to my awesome colleagues for their contribution in successfully completing our team’s annual goals. Excited for the next professional phase. Forever grateful to my friends and family for their unconditional love and support.

Figure 5.10 *Text samples for Positive class (LABEL 1) and Control / Neutral class (LABEL 0) used in SHAP experiments*

Next, we explain and interpret the output predictions of these six TLMs using various plotting techniques provided by SHAP's Python library (Refer to Figures 5.11 to 5.16) for positive and control class text samples (Figure 5.10). We obtain their predicted class using different TLMs mentioned above and use SHAP to gain insights into their decision-making. As explained previously, the SHAP algorithm computes Shapley values for each feature to indicate whether that feature has contributed positively or negatively to the model decision or output towards each class. This corresponds to a positive or a negative integer, respectively, and it is computed for every feature w.r.t. each class. Positive Shap values are color-coded as red, whereas negative Shap values are indicated in blue, where the color intensity is proportionate to their absolute magnitude. In the figures below, for every class, the words/features pushing the model output above the base value are shown in red, and those pushing the model output lower are shown in blue. The base value is the average model output learned from the training data. Since the same text samples have been used across all LLMs, and their Shapley values are somewhat logically similar, hence, in the figure captions below, we interpret these Shapley values for some of the LLMs to avoid redundancy.

XAI techniques can help give insights around data sanity and quality issues, e.g., in the experiments below, it can be seen that all false positives are mostly on LLMs trained on Dataset 1, whereas the same LLM when trained on the other dataset doesn’t give false positive. They can also help in understanding what are the aspects or word features the model is focusing on that has led to false positives and why the model predictions are incorrect. E.g., PsychBERT and PHSBERT are domain-adapted models, yet they have false positives. Looking at the heat maps, we can see which words are pushing the score down for the label 0 control class sample. We can see that some of the positive sentiment words are also contributing negatively, driving the probability score lower for class label 0, which is incorrect. Such insights can help improve the training process by ensuring data quality standards of datasets used for pretraining or finetuning LLMs, thereby developing robust and accurate models. Such explanations can also be helpful for the validation of algorithms, e.g., in the context of sarcasm. Here, the health practitioner can decide to ignore or disregard the predicted diagnosis if the prediction is localizing or based on the wrong aspect or word features.

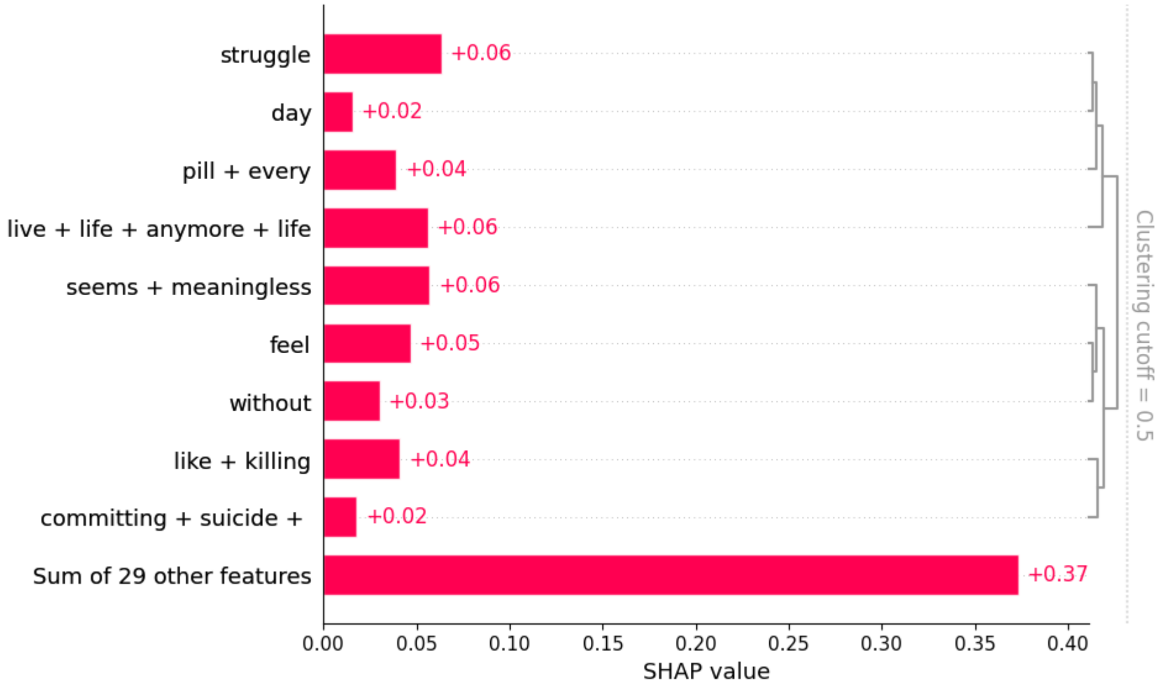


Figure 5.11(a) *Interpreting classification decision of BERT model trained with Dataset 1 for Text Sample #1*

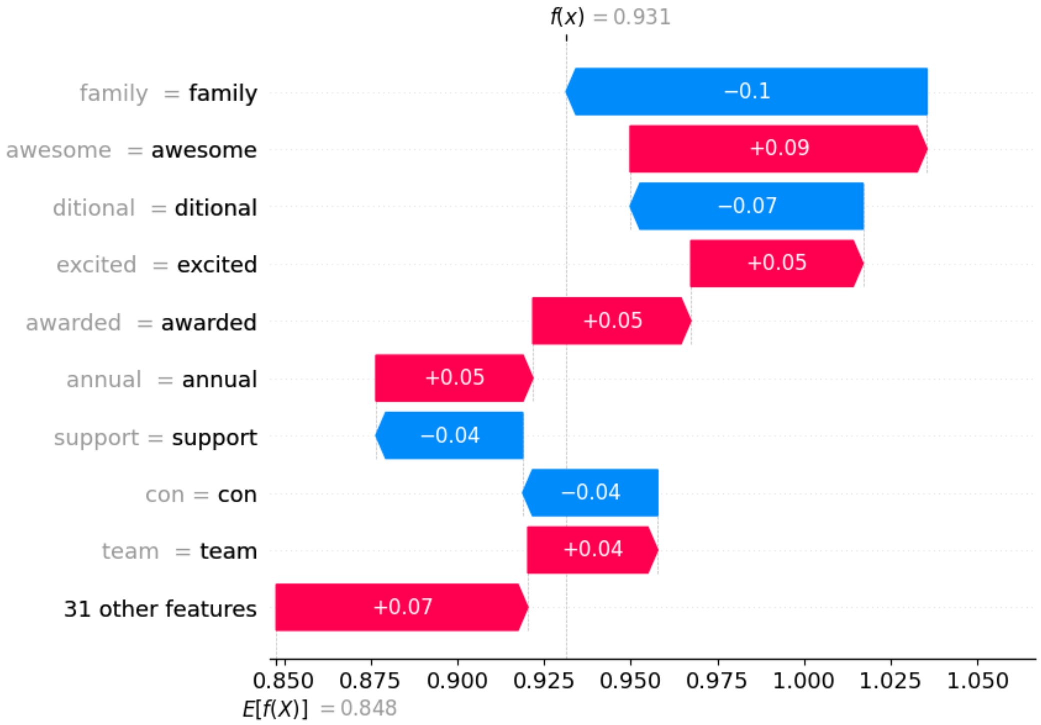


Figure 5.11(b) *Interpreting classification decision of BERT model trained with Dataset 2 for Text Sample #2. SHAP values here explain that for the correct, true negative classification by the BERT model, the words: awesome, excited, and award positively contributed to pushing the model output towards the correct class. Whereas, words family and support have pushed the model output towards label 1.*

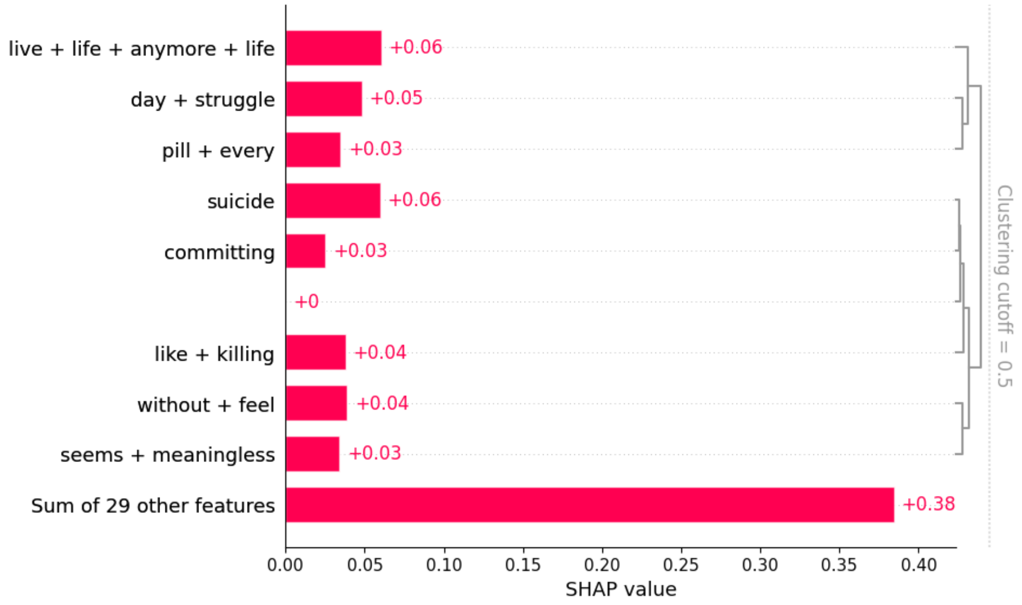


Figure 5.12(a) *Interpreting classification decision of DistilBERT model trained with Dataset 1 for Text Sample #1. The word features and their additive Shapley values are shown above for a true positive class sample.*

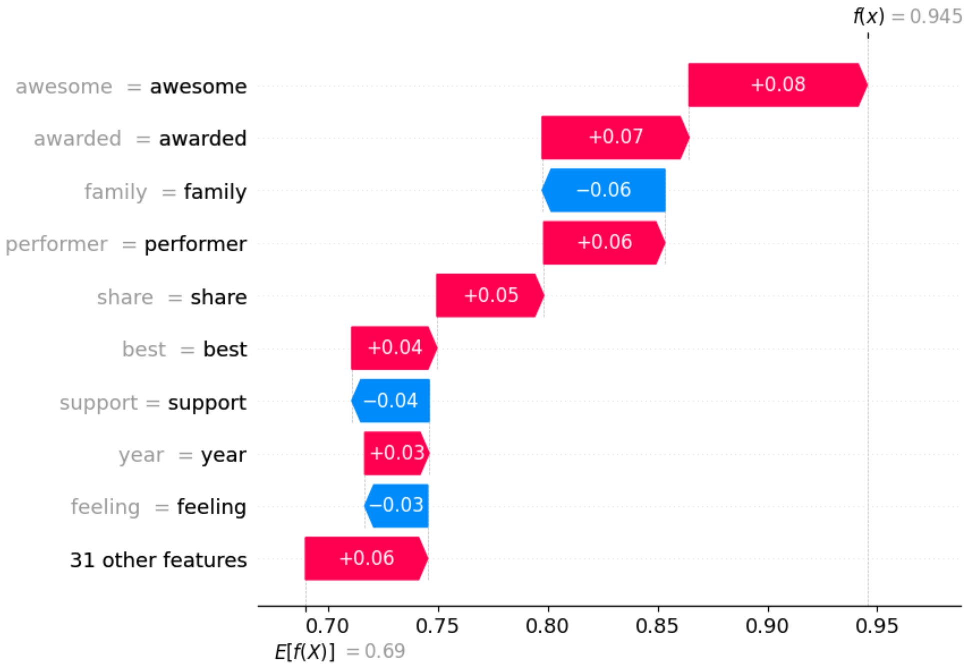


Figure 5.12(b) *Interpreting classification decision of DistilBERT model trained with Dataset 2 for Text Sample #2. In this example, even though the classification decision is correct, but it can bee seen that the words: family, support, feeling push the score of the model lower w.r.t. class label 0. This is because mental health positive users often talk about the need for family support and their feelings.*

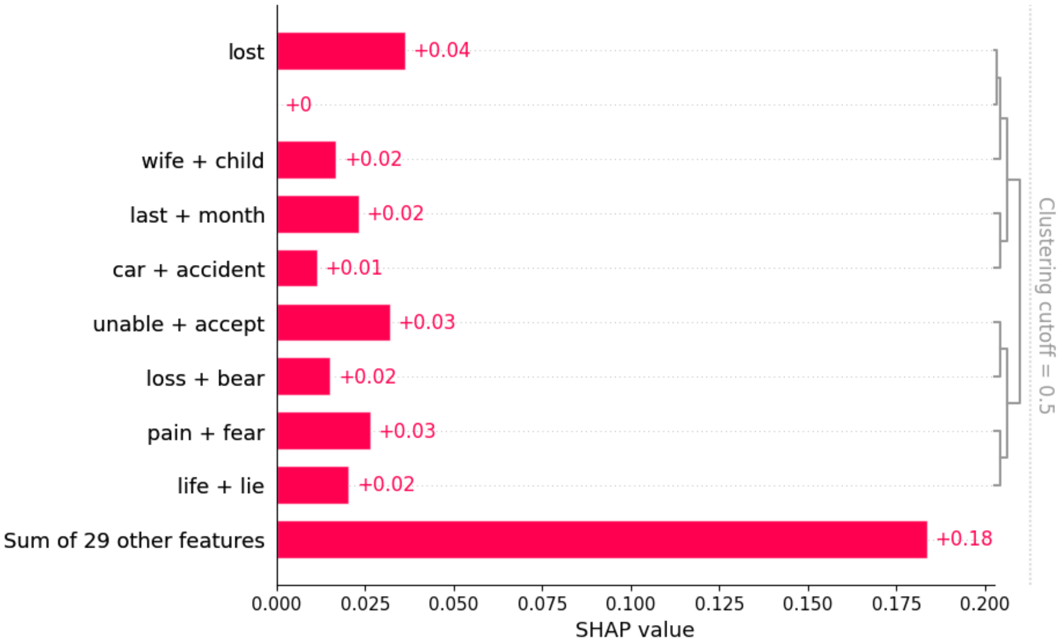


Figure 5.13(a) *Interpreting classification decision of RoBERTa model trained with Dataset 1 for Text Sample #1*

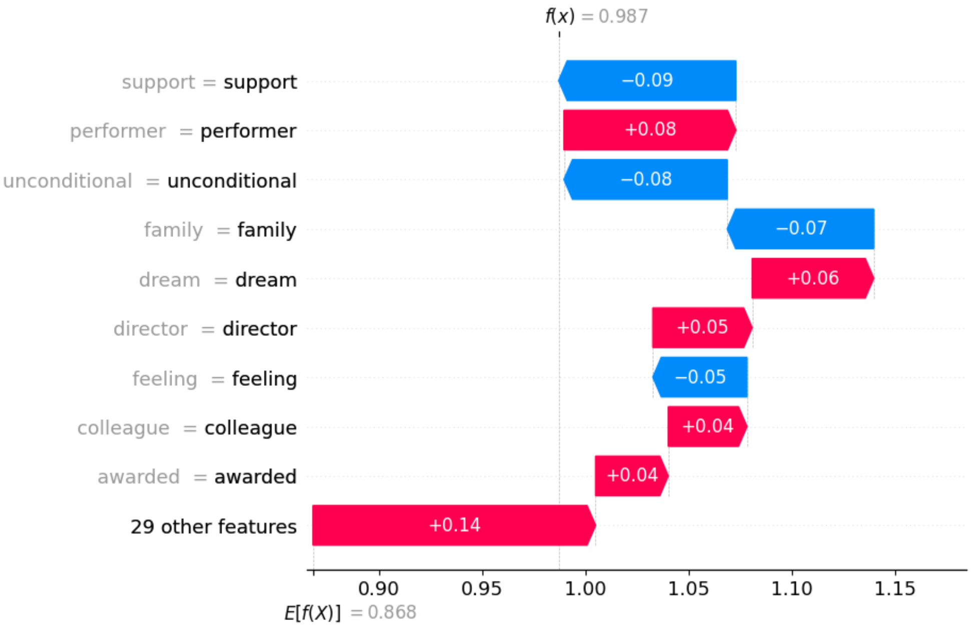


Figure 5.13(b) *Interpreting classification decision of RoBERTa model trained with Dataset 2 for Text Sample #2*

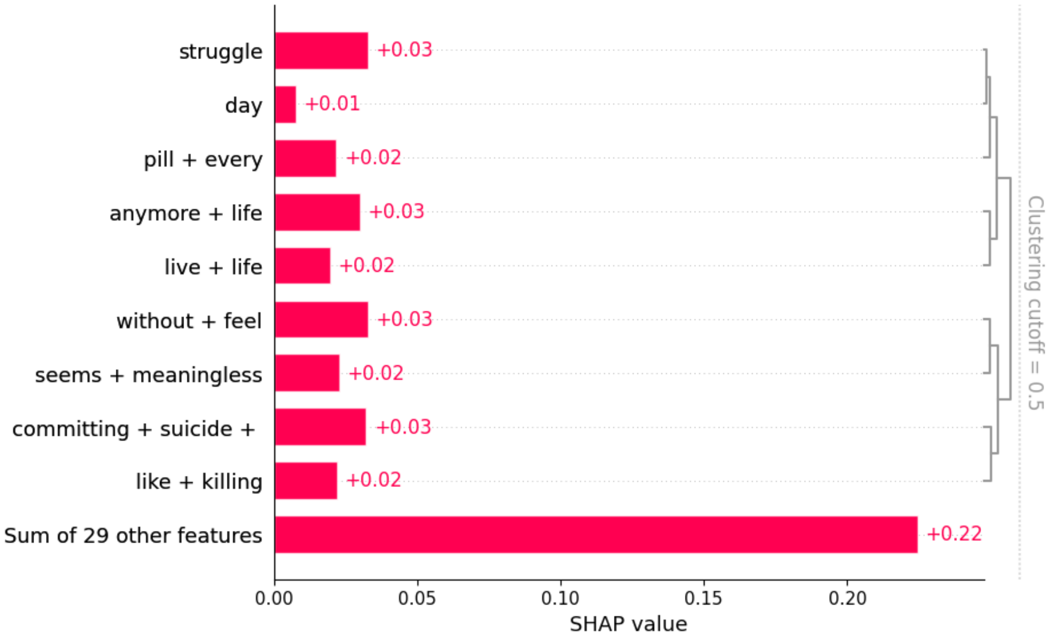


Figure 5.14(a) *Interpreting classification decision of MentalBERT model trained with Dataset 1 for Text Sample #1*

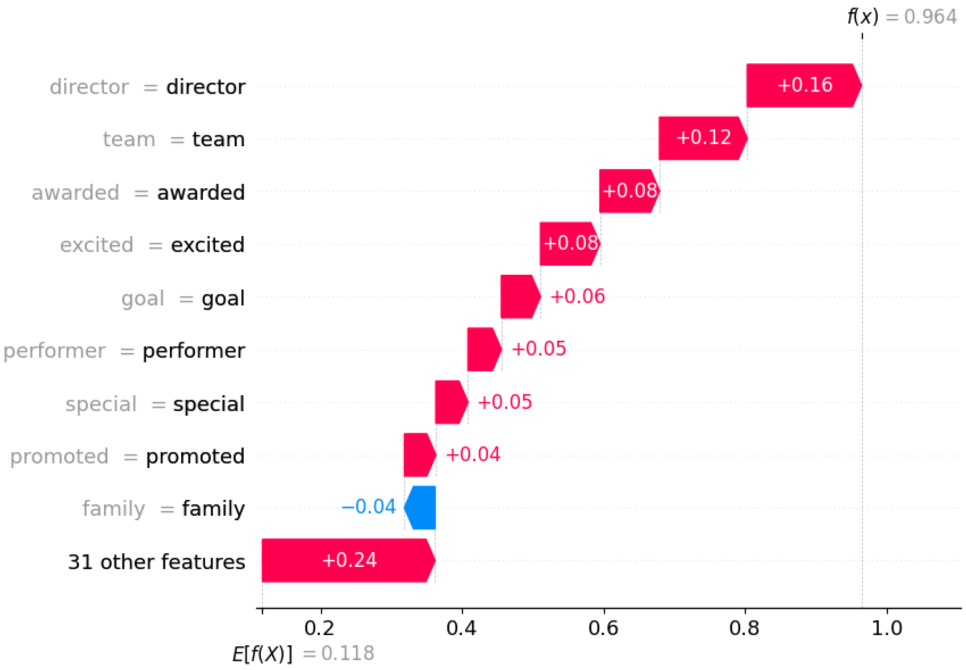


Figure 5.14(b) *Interpreting classification decision of MentalBERT model trained with Dataset 2 for Text Sample #2*

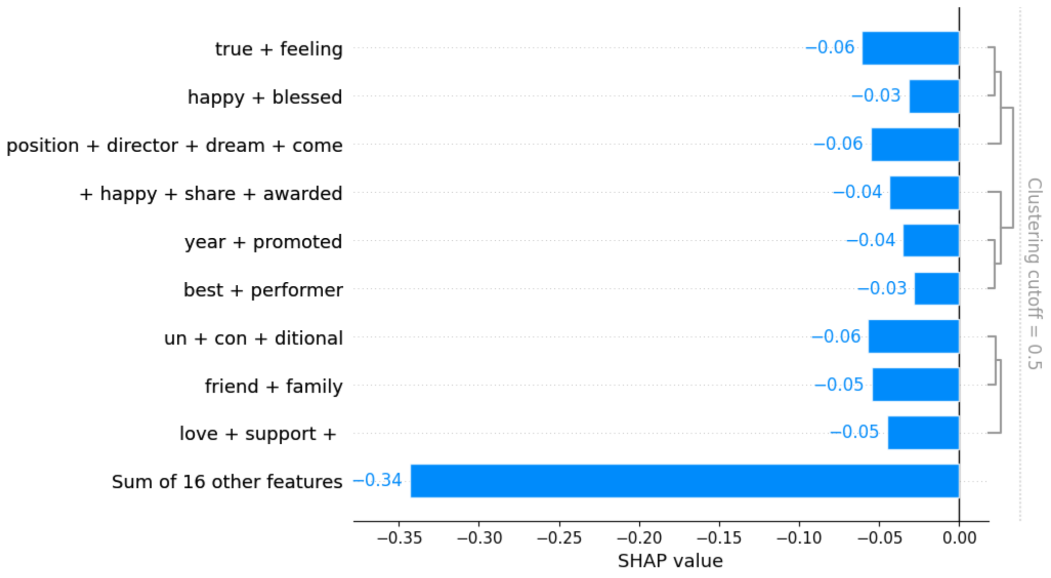


Figure 5.15(a) *Interpreting classification decision of PsychBERT model (trained with Dataset 1) for Text Sample #2 (Label 0). This is a Misclassification / False Positive and was classified as Label 1. The Shapley scores for the words indicate the degree to which these words negatively contributed towards preventing the sample class from being predicted as Label 0*

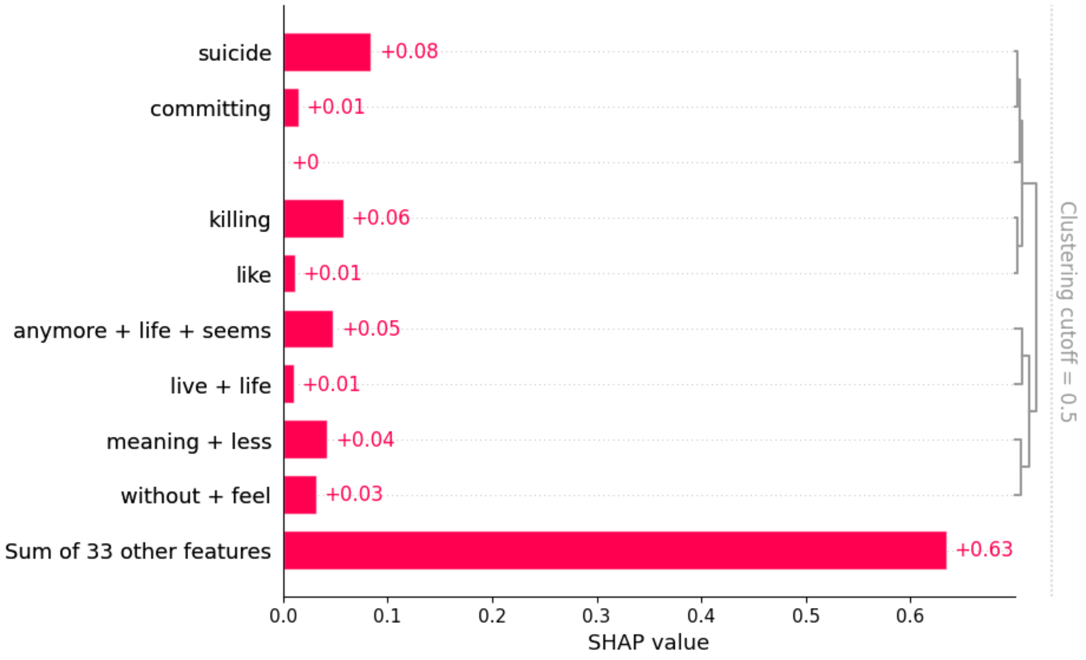


Figure 5.15(b) *Interpreting classification decision of PsychBERT model (trained with Dataset 2) for Text Sample #1*

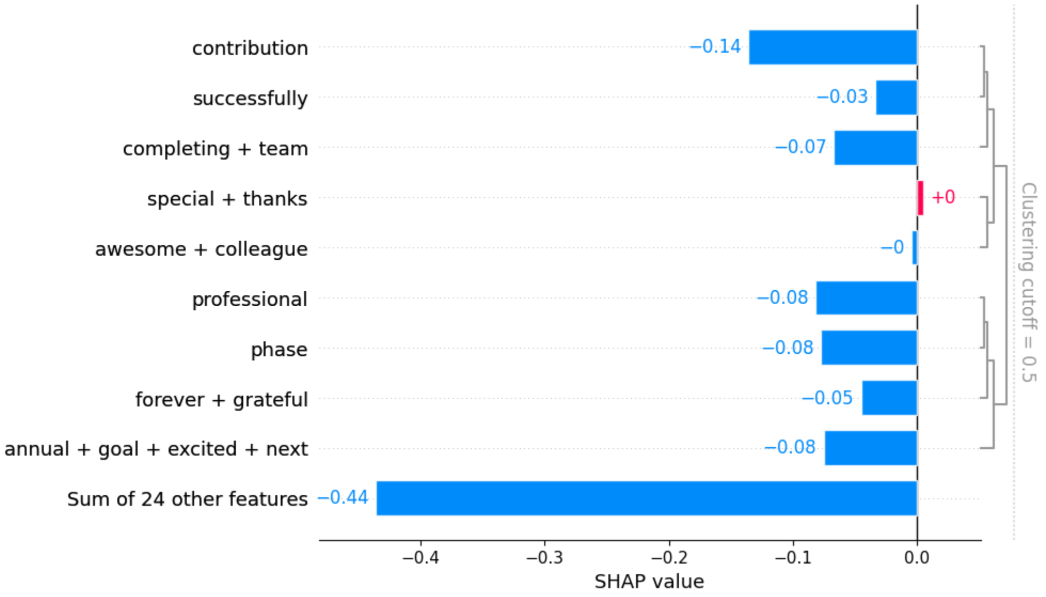


Figure 5.16(a) *Interpreting classification decision of PHSBERT model (trained with Dataset 1) for Text Sample #2 (Label 0). This is a Misclassification / False Positive and was classified as Label 1. The Shapley scores for the words indicate the degree to which these words negatively contributed towards preventing the sample class from being predicted as Label 0*

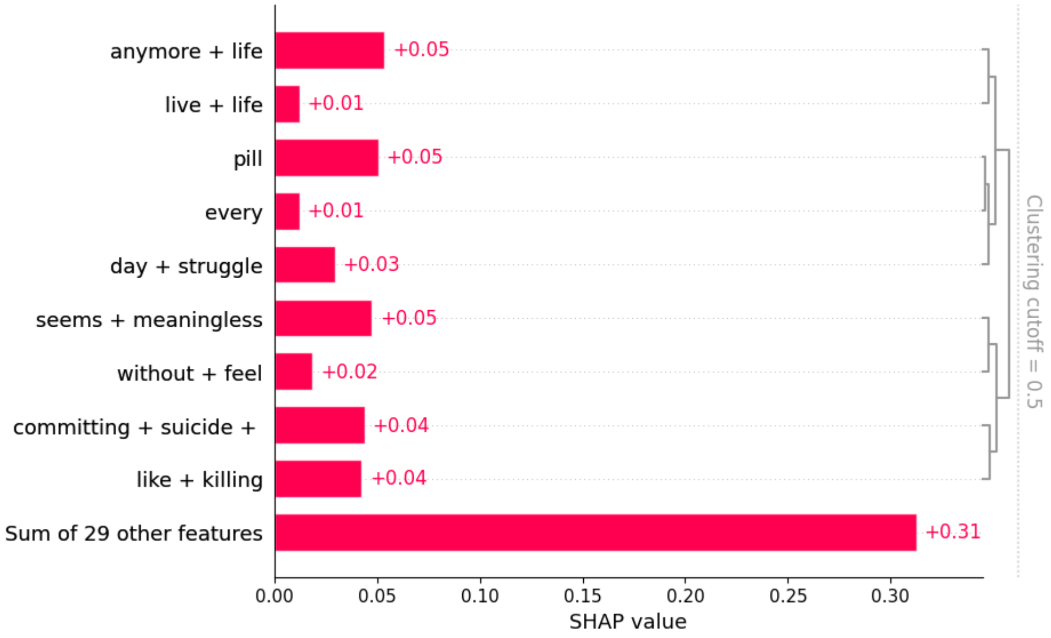


Figure 5.16(b) *Interpreting classification decision of PHSBERT model (trained with Dataset 2) for Text Sample #1.*

5.3.3 Unsupervised BERT Topic Modelling

For topic modeling experiments using BERTopic, we have used Datasets 1, 2, and 3, the results of which are discussed in this section. Figure 5.17 below shows the top fifty topics created from Dataset 1 using BERT Topic Modelling for both classes: Positive (Label 1) and Control (Label 0). It also shows the hierarchical association between these topics. The Hierarchical clustering of topics is obtained by grouping semantically similar topics together based on the (Euclidean) distance between their c-TF-IDF topic representations created by the BERTopic algorithm. BERTopic allows us to compute and compare the intertopic distance (using cosine similarity) and use it to reduce the total number of output topics by hierarchically aggregating topics with lower distances (i.e., more similar) with each other. This can help in summarizing key topics and analyzing their relationships with other topics. Some of the prominent words from the top twelve topics are shown in the following figure. From the topic analysis, it can be inferred that some of the key topics or issues depressed or suicidal users talk about are: family members (who may be possibly suffering), related symptoms (migraine, panic, anxiety, headache, anger issues, stress, alcoholism), possible causes (pregnancy, job, child abuse victim, pandemic, refuge/detention), temporary relief measures (sleep, food, music, meditation, breathing exercise), need for therapy and treatment (antidepressants), possible suicide ideation, social stigma associated with mental health, and negative sentiment words like: hurt, kill, abuse, trauma, illness.

A similar topic analysis was done for Dataset 2 and Dataset 3, the results of which are shown below in Figure 5.18 and Figure 5.19, respectively. In addition to the topics discussed above, the topics discovered from Positive Class (Label 1) posts of this dataset indicate the possible triggers for suicide: mental health issues, divorce, relationship issues, bullying, loneliness, social issues like transgenderism, or struggle with health issues like autism, obesity. Users are seen discussing dangerous methods to harm themselves or take their lives, e.g., jumping from a bridge, consuming sleeping pills, using a gun, tying a rope/belt over the neck, or cutting one's wrist. They are even seen mentioning words like: goodbye and tonight indicative of a possible suicide attempt. Using unsupervised BERT Topic modeling in a public healthcare monitoring system to detect these keywords can serve as an early risk indicator or warning and help prevent suicide attempts. Many users are seen expressing the need for help or the need to talk to someone, which, if detected and addressed promptly through helplines and chatbots, can save precious lives. In contrast, the posts labeled as non-depression / non-suicidal in these Datasets cover varied day-to-day topics like: movies/shows/podcasts, news, politics, shopping/clothes/cosmetics, travel, songs, books, exams, work, science and technology, gadgets, vehicles, weather, animals, astrology and words like: dance, smile, hugs, etc. typically associated with positive sentiment. We can notice from this analysis there is some overlap of topics across the two user groups, e.g., food, music, books, sports, work, and religion. Further research is required to understand the reason for this overlap by possibly analyzing and comparing with other datasets and discovering data quality issues, if any, that may have arisen during data collection or possibly due to a lack of manual validation and annotation (false positives).

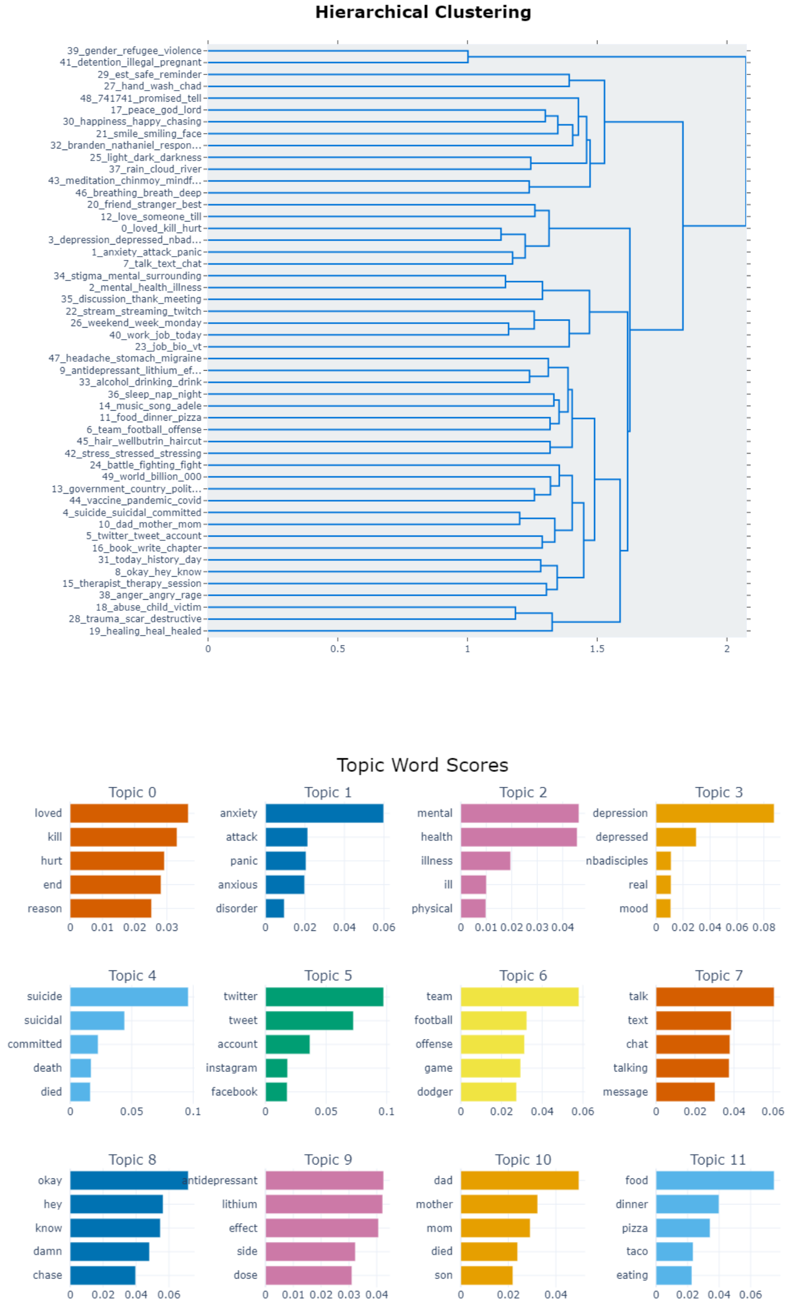


Figure 5.17(a) *BERTopic results for Dataset 1 (Positive Label 1 Class)*

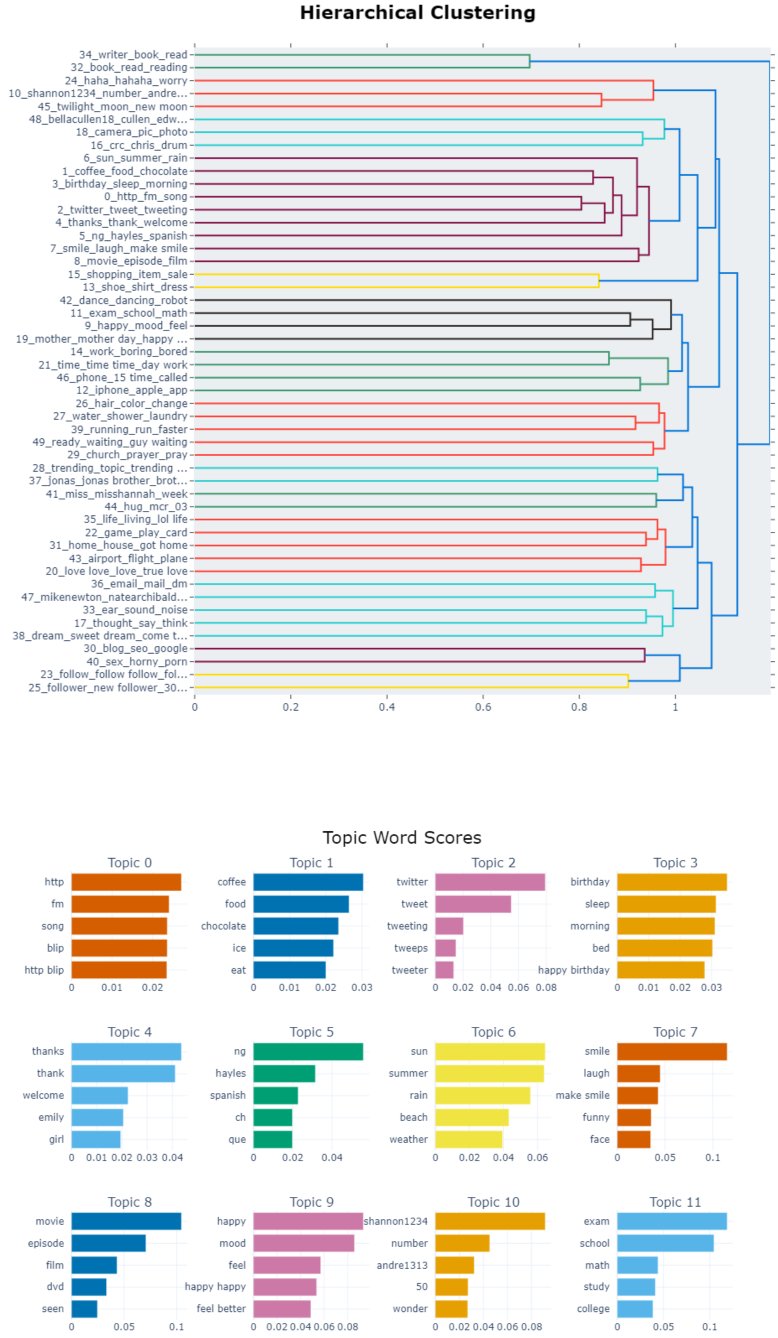


Figure 5.17(b) *BERTopic results for Dataset 1 (Control Label 0 Class)*

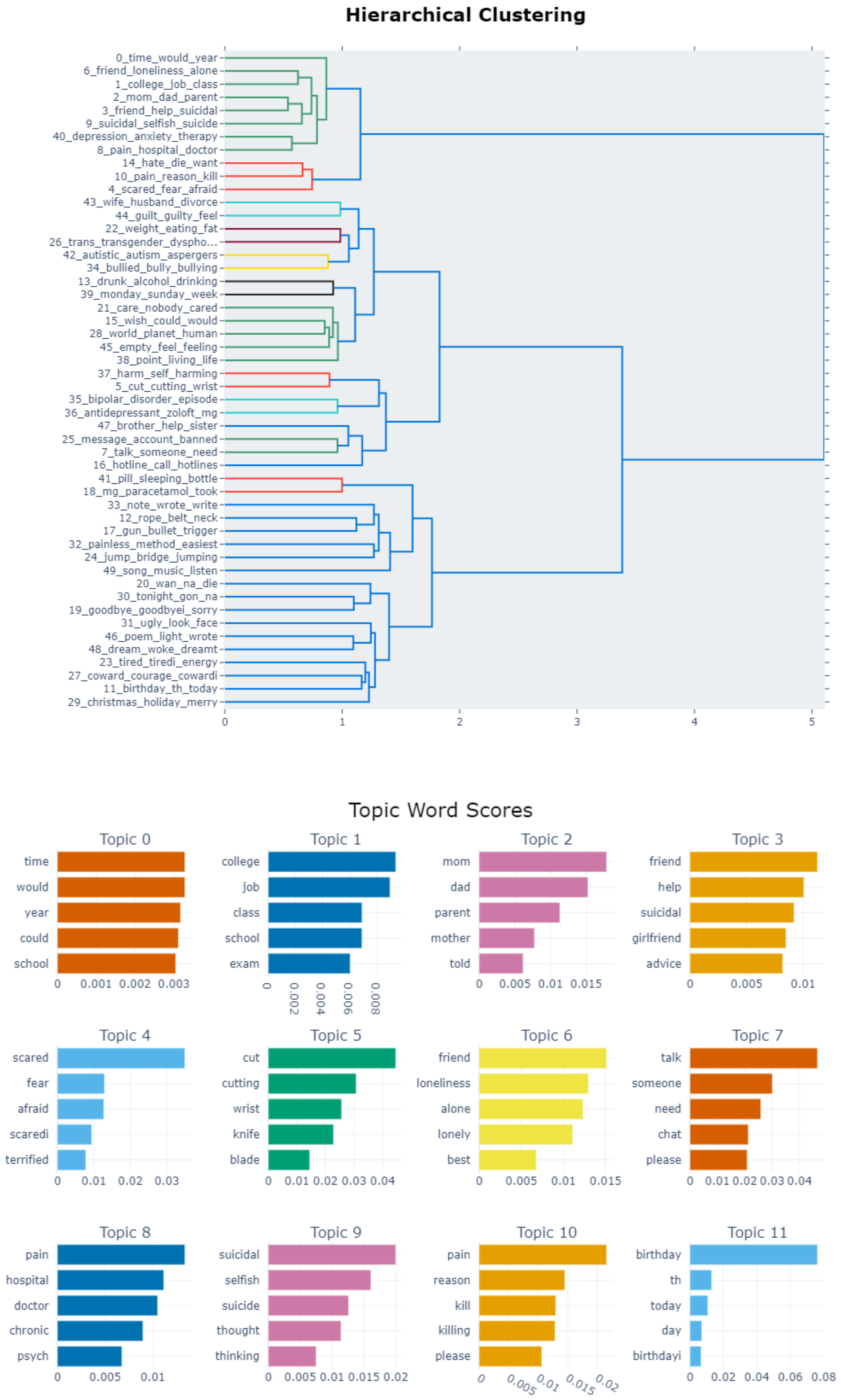


Figure 5.18(a) *BERTopic results for Dataset 2 (Positive Label 1 Class)*

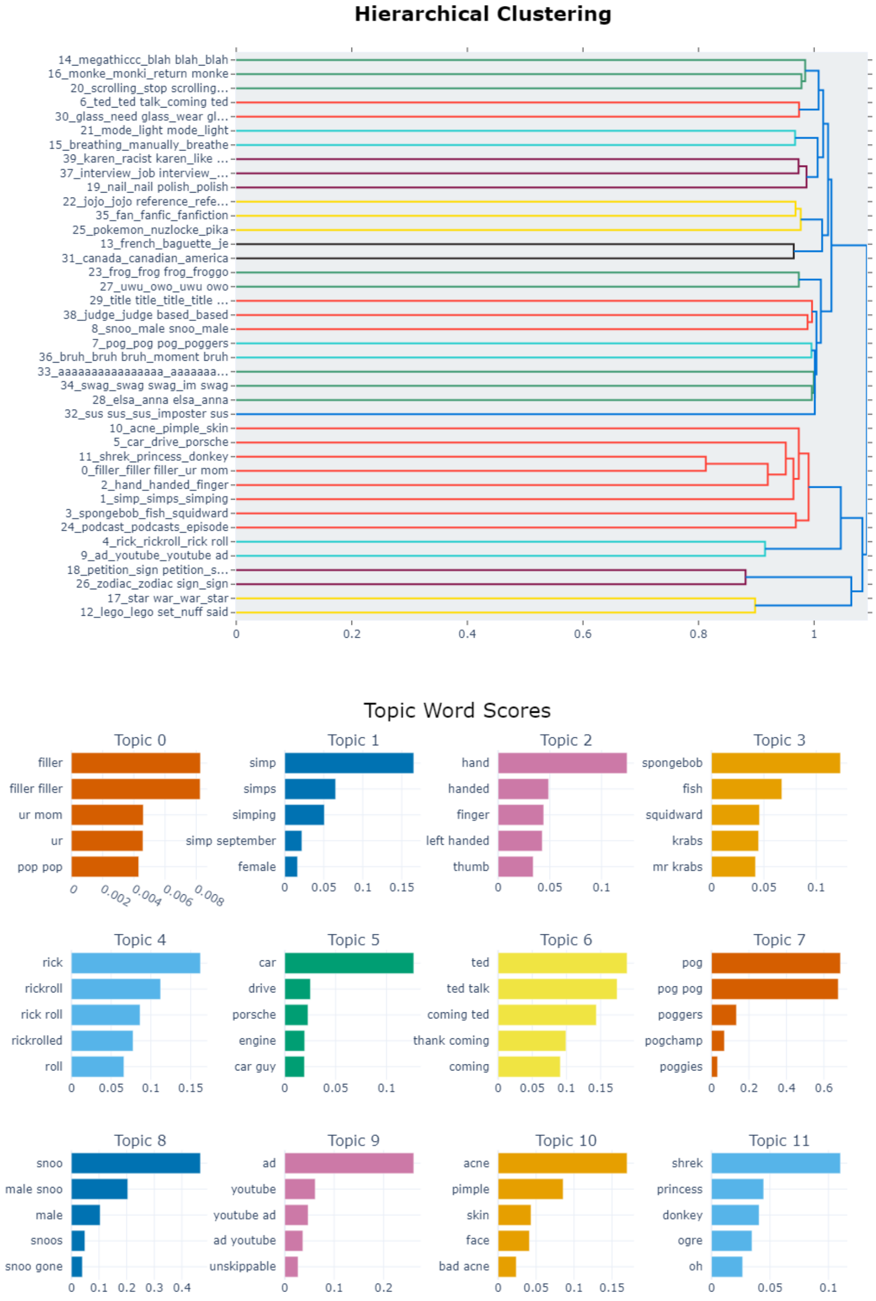


Figure 5.18(b) *BERTopic results for Dataset 2 (Control Label 0 Class)*

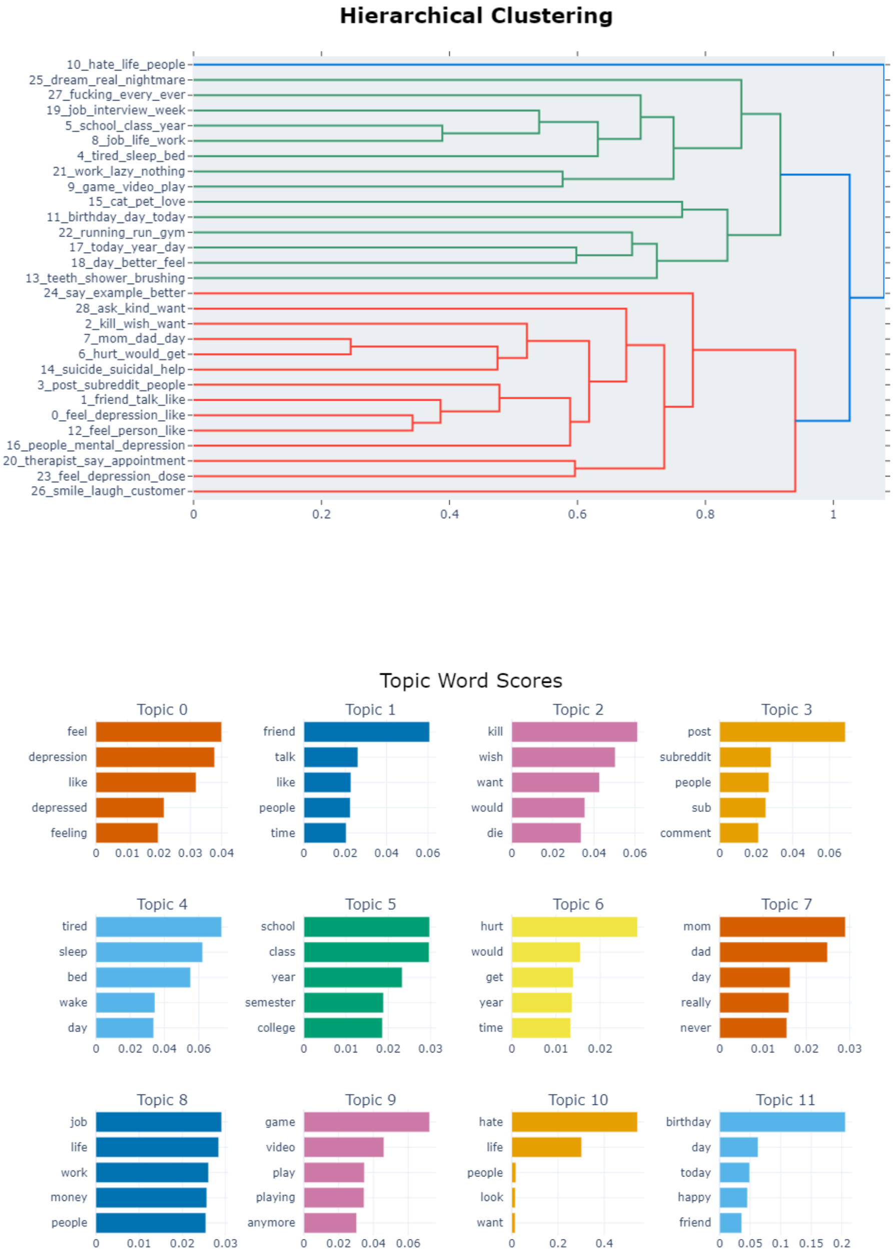


Figure 5.19(a) *BERTopic results for Dataset 3 (Positive Label 1 Class)*

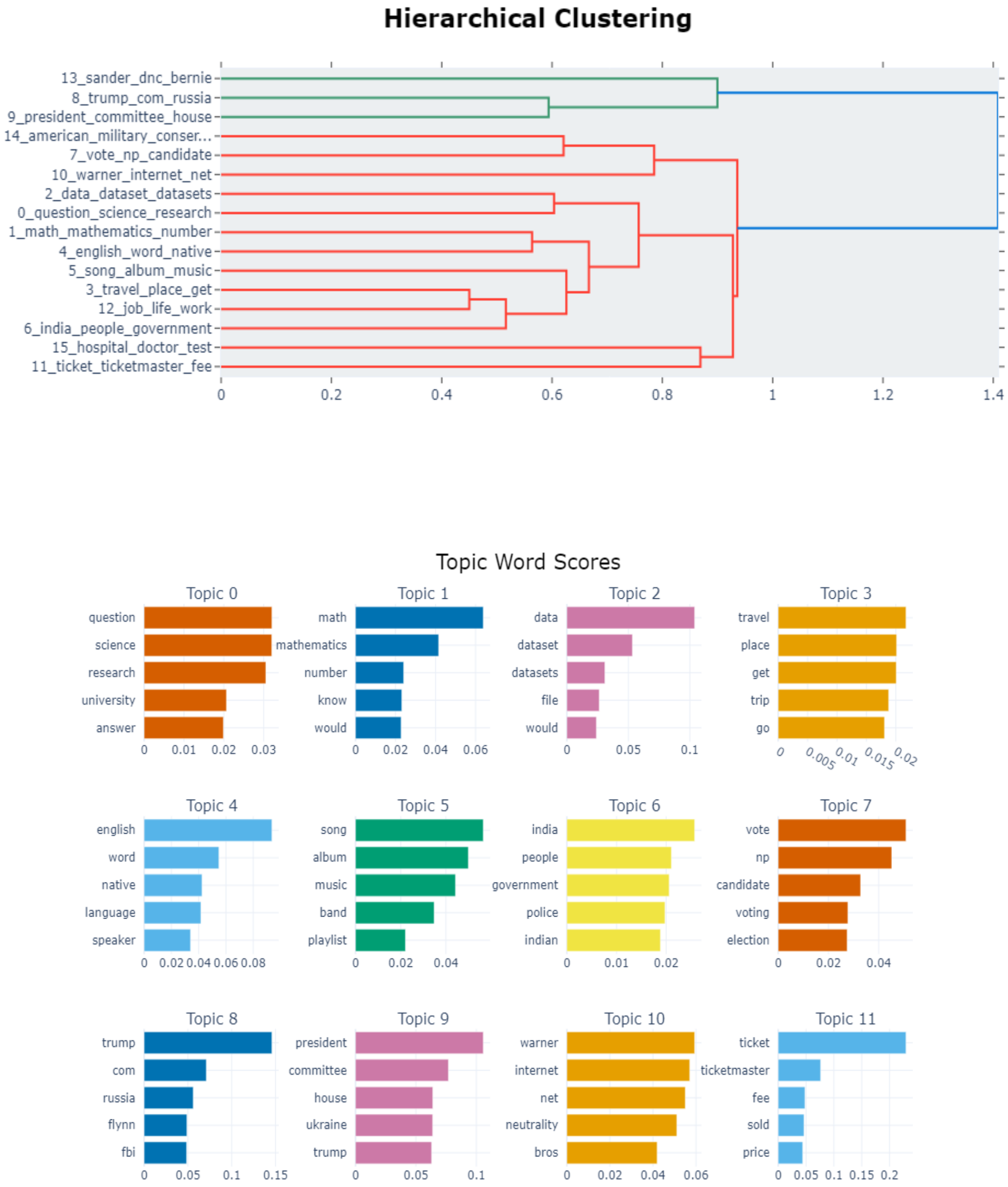


Figure 5.19(b) *BERTopic results for Dataset 3 (Control Label 0 Class)*

5.4 Few Shot Learning Experiments with Transformer Language Models

Traditional machine learning and the recent deep learning algorithms require good quality datasets of significant size to achieve high classification accuracy and generalizability [198]. Online social network datasets to conduct mental health assessment related research are scarce, and only very few good quality datasets are publicly available [146] [199] [200] [201]. For low resource scenarios such as these, Few Shot Learning may prove to be beneficial by training supervised AI algorithms with very few, good quality annotated data samples [198] [202] [203] [204] [205]. Recently, pretrained LLMs have been used for various text classification tasks using Few Shot Learning [22] [205].

In this section, we present the results of two K-Shot Learning experiments we conducted for three pretrained mental health domain adapted LLMs (MentalBERT, PsychBERT, PHSBERT) with Dataset 3 and Dataset 4 (which are relatively very smaller in size as compared to the other two Datasets mentioned in section 5.3.1 above). We use N-way K-shot learning (with K = 5) [198] [202] [203] [204] [206] [207] for two tasks: classifying depression vs. non-depression posts (Dataset 3), and distinguishing between suicide ideation vs depression posts (Dataset 4). We compare their performance with the scenario when they are trained on the full available dataset. In addition, we also compare their classification performance with the other three domain-independent LLMs (BERT, DistilBERT, RoBERTa). Please refer to Table 5.3 and Table 5.4.

Table 5.3 *Few Shot Learning to Distinguish between Depression and Non-Depression related user posts*

LLM	**Dataset 3 (Depression vs Control)**			
	Precision	*Recall*	*F1*	*Accuracy*
Full Dataset				
BERT	0.967	0.984	0.976	0.974
DistilBERT	0.957	0.975	0.966	0.963
RoBERTa	0.962	0.985	0.973	0.97
MentalBERT	0.977	0.974	0.975	0.974
PsychBERT	0.969	0.983	0.976	0.974
PHSBERT	0.981	0.983	0.982	0.98
Few Shot Learning (N-way K-shot Learning, K = 5)				
MentalBERT	0.554	0.54	0.547	0.507
PsychBERT	0.779	0.667	0.718	0.713
PHSBERT	0.515	0.295	0.375	0.451

Table 5.4 *Few Shot Learning to Distinguish between Suicide vs. Depression related user posts*

LLM	**Dataset 4 (Suicide vs Depression)**			
	Precision	*Recall*	*F1*	*Accuracy*
Full Dataset				
BERT	0.591	0.637	0.613	0.583
DistilBERT	0.582	0.22	0.319	0.555
RoBERTa	0.537	0.474	0.504	0.565
MentalBERT	0.678	0.578	0.624	0.643
PsychBERT	0.60	0.35	0.442	0.543
PHSBERT	0.656	0.742	0.697	0.693
Few Shot Learning (N-way K-shot Learning, K = 5)				
MentalBERT	0.509	0.681	0.583	0.530
PsychBERT	0.528	0.07	0.125	0.512
PHSBERT	0.546	0.282	0.372	0.525

5.5 Discussion & Summary

For complex, multi-disciplinary, computational linguistic tasks involving unstructured, multimodal user-generated content from heterogeneous sources on the Internet, deep learning networks have shown better model performance as compared to simpler machine learning models like decision trees, regression, etc. Deep learning networks are opaque, black-box architectures due to their complex internal structure with stacked non-linear transformations, which makes it difficult to understand and explain their decisions. Clearly, there is a trade-off between model performance and model interpretability/explainability. However, model explainability is equally essential as model performance for real-world use cases. Even any exceptionally performing model will not have many takers if they find it hard to trust its decisions. XAI is even more crucial for healthcare applications. Hence, we have proposed and demonstrated the application of two most recent XAI techniques, LIME and SHAP, that can be used to explain the black-box classification decisions of LLMs (Transformers) trained for any user generated text categorization task. These model-agnostic post hoc XAI techniques don't require intrinsic changes to the neural network, nor do they require creating elaborate feature representations (text embeddings) for gradient analysis or complex attention weight analysis for model explainability. Due to these reasons, they are easy to use as an out-of-the-box attachment, thereby giving the flexibility to leverage the power of any available open-source pretrained LLM as a black box and yet be able to explain and interpret its output decisions. These post hoc techniques can provide some degree of explainability for LLMs that are not intrinsically interpretable due to their complex internal structure. We experiment with six pre-trained Transformer based language models and four UGC datasets from two commonly used social networks: Twitter (shorter texts due to character limits) and Reddit (usually longer posts and discussions). Results indicate that these techniques can provide reasonable explainability for both short and long user-generated text without the need for intrinsic changes to the network, thereby giving the flexibility to leverage any available, open-source pretrained LLM for transfer learning. The results have also shown that XAI techniques can give insights about data quality and sanity issues in training datasets. For example, in the experiments above, it can be seen that all false positives are mostly on LLMs trained on Dataset 1, whereas the same LLM trained on the other dataset doesn't give false positives. Similar insights are confirmed by the topics created from this dataset using BERTopic. Another interesting observation is around the explainability of false positives. Even though PHSBERT and PsychBERT are mental health domain adapted models, they misclassified samples. SHAP values for these false positives have helped to understand the words or features that have contributed towards this misclassification. Such insights can help improve the training process by ensuring data quality standards of datasets used for pretraining or finetuning LLMs, thereby developing robust and accurate models. Additionally, when labeled data is unavailable, and TLMs cannot be trained using supervised learning, we have shown that unsupervised topic modeling using BERTopic can be leveraged for user generated text categorization problems and for pseudo Interpretability of user posts. At last, we have demonstrated the use of Few Shot Learning paradigm with various domain adapted pretrained LLMs which can be an extremely useful approach when only a few, good quality, expert annotated data samples are available.

CHAPTER 6

PROTOTYPES FOR FUTURE RESEARCH ENHANCEMENTS

This chapter demonstrates the preliminary research work done (prototypes / proofs-of-concept) for extending the scope of the research by exploring the use of recent innovative techniques like: Deep Active Learning, Transfer Learning, and Multimodal Deep Learning for categorizing user generated content on the Internet.

6.1 Deep Active Learning

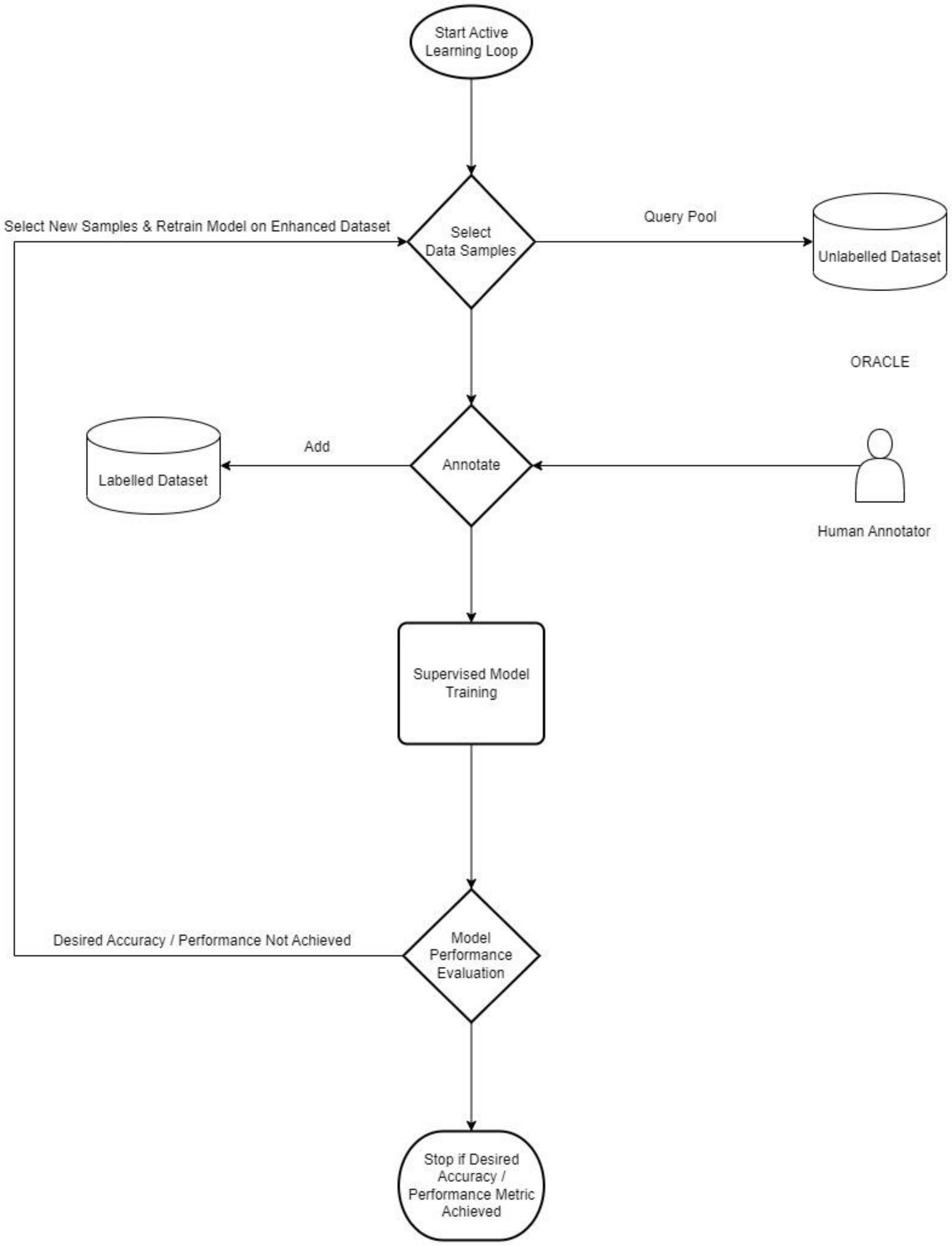
The classification performance of any supervised model in the real world depends heavily on the quality and quantity of annotated datasets used to train it. A huge volume of user generated content is available on the Internet and can be easily collected through the Web APIs to build large datasets for research purposes. There can be numerous real-world applications of user-generated content available on the Internet. However, data annotation of these big datasets remains an arduous task. Data labeling and annotation are costly, time-consuming processes, often requiring extensive effort from domain experts. On the other hand, data labeling done through crowdsourcing can be error-prone.

This research work demonstrates the idea of finetuning state-of-the-art pre-trained LLMs with minimal annotated data and yet achieving high classification accuracy through the use of an Active Learning loop. Active Learning is a training paradigm of incremental learning where instead of training a supervised model with all the data in one go, it is trained iteratively with incremental data or batches of data sampled from training data. The entire training dataset is not labeled beforehand in one go. Rather, it is divided into two components: labeled and unlabeled pool. Initially, only a sample of data is labeled, and a baseline model is trained. Till the desired classification accuracy is achieved, new batches are sampled from the unlabelled pool, and the human experts or annotators then provide the ground truth labels for this batch. This setup is known as an Information Oracle or a data source that a model can interactively query to fetch new data points along with their ground truth labels. Through our proof of concept, we demonstrate that it is possible to achieve high/comparable accuracy with as few as 10% of samples from the entire dataset by iteratively training a deep learning model using incremental updates of annotated data instead of using the entire dataset for model development. Deep active learning can leverage the high-performing Transformer-based Language Models coupled with Active Learning in order to mitigate the challenges associated with data annotation or in cases when very little labeled data is available.

6.1.1 Background on Active Learning

Active learning is a training paradigm of incremental learning where we attempt to build supervised machine learning models with a minimal amount of labeled data and yet aim to achieve higher or desired classification accuracy. Data labeling and annotation are costly, time-consuming processes requiring extensive effort from domain experts. The classification performance of any supervised model in the real world depends heavily on the quality of annotated datasets used to train it. Yet, most of the Machine Learning and Deep Learning research focuses on algorithmic improvements rather than on the data annotation process. There can be numerous research applications of ML/AI in the real world; however, they fall short of real-world deployment due to the data annotation efforts required. AL can help build real-world applications where a large amount of data is readily available; however, labeling all of that data is not easily feasible, e.g., UGC (text, images) from the Internet/OSNs, images from the WWW, medical domain / clinical images, e.g., X-Rays, IOT camera recording feed, etc.

Active learning is analogous to semi-supervised learning but is less commonly used. In semi-supervised learning, unlabeled data is used to learn feature representation, which is then used to build supervised models with limited labeled data available [208] [209]. In contrast, the fundamental steps in an Active Learning loop are shown in Figure 6.1. The underlying concept behind AL is that instead of training a supervised model with all the data in one go, it is instead trained iteratively with incremental data or batches of data sampled from training data. The entire training dataset is not labeled beforehand in one go. Rather, it is divided into two components: labeled and unlabeled pool. Initially, only a sample of data is labeled, and a baseline model is trained. Till the desired classification accuracy is achieved, new batches are sampled from the unlabelled pool, and the human experts or annotators then provide the ground truth labels for this batch. This setup is known as an Information Oracle or a data source that a model can interactively query to fetch new data points along with their ground truth labels.



Start Active Learning Loop
Select New Samples & Retrain Model on Enhanced Dataset
Select Data Samples
Query Pool
Unlabelled Dataset
ORACLE
Labelled Dataset
Add
Annotate
Human Annotator
Supervised Model Training
Desired Accuracy / Performance Not Achieved
Model Performance Evaluation
Stop if Desired Accuracy / Performance Metric Achieved

Figure 6.1 *Active Learning Flowchart*

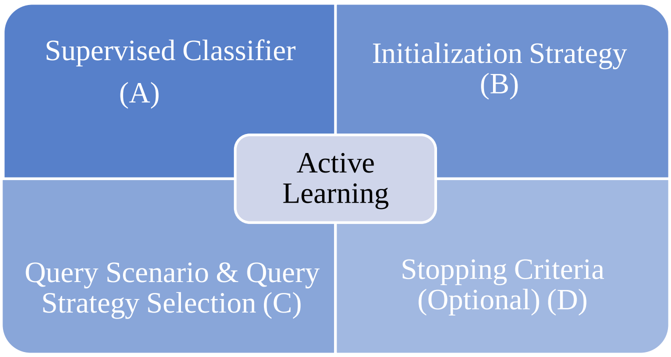

Supervised Classifier (A)
Initialization Strategy (B)
Active Learning
Query Scenario & Query Strategy Selection (C)
Stopping Criteria (Optional) (D)

Figure 6.2 *Active Learning Components*

Active Learning Key Components: To implement an AL process, we typically require the following components [210] (Refer to Figure 6.2):

A. Firstly, we need to decide the supervised machine learning or deep learning model we want to train.

B. Next, we need to train the initial model (baseline, weak learner) with a few randomly sampled and annotated data points (while maintaining the class label's distribution).

C. The most crucial component is deciding on the Query Scenario and Query Strategy [209]. The Query Scenarios determine how the learner and oracle interact and exchange information w.r.t annotated data samples. The three scenarios are: Membership Query Synthesis (learner-generated, i.e., simulated/hypothetical data points), Stream-Based Selective Sampling (learner inspects data samples one by one sequentially and requests labels for the ones it deems informative), and Pool-Based Active Learning. The Pool-based sampling approach is the most commonly used in the real world, where large unlabeled datasets have been collected. The learner requests the human in loop/annotator/oracle to provide ground truth labels for samples selected from this unlabeled pool. The samples to be labeled are selected using different query strategies based on their informativeness (uncertainty sampling) or, representativeness (diversity sampling), or a hybrid measure. Popular query strategies based on informativeness measures are: Least Confidence, Maximum Entropy, and Margin-Based Uncertainty, and the ones based on representativeness are: K-means Clustering and Core-Set Selection. [208] [209] [211].

D. Optional stopping criteria can determine if and when to exit the AL loop early if some desired condition is met [210]

AL training strategy may be combined with any supervised machine learning or deep learning algorithm. Active Learning can especially be beneficial along with current state-of-the-art deep learning algorithms that can learn complex nonlinear patterns from datasets but require a large amount of labeled data for training [212] [213]

6.1.2 Experiments & Results

In this section, we present the empirical evaluation results of our proposed framework and the details of the datasets used. We have used BERT in our experiments since it has become the most frequently used state-of-the-art TLM for the NLP domain; however, other TLMs can also be used in a similar way by using their open-source model checkpoints [194]. We have used Small-Text [210] for our Deep Active Learning experiments. It is a state-of-the-art Python library that provides robust modularized components for initialization strategy, query strategies, and stopping criteria for implementing AL for text classification. All these components can be interchangeably used for AL experiments with various ML and DL classifiers. This library provides integration with GPU-based models such as Transformers and PyTorch. The use of this library has not been explored in the existing literature.

Table 6.1 *Datasets Used*

Dataset	Positive Class	Control Class
Depression Dataset [195]	11466 tweets	12054 tweets
Suicide Dataset [196]	116015 posts	115952 posts

Datasets: We demonstrate and evaluate our proposed Deep Active Learning approach using two publicly available UGC datasets (Table 6.1). The first dataset is a collection of ~ 11K possibly depression related tweets from Twitter, whereas the second dataset consists of 116K posts from Reddits where the users might have expressed suicide ideation in some way. Both these datasets are labeled and have a balanced distribution of positive and control classes.

Experimental Setup: We have used Small-Text integration for the 'bert-base-uncased' (BERT) model [210]. We use balanced random initialization for bootstrapping or initial supervised training of the BERT model. We train this model under Pool-Based Active Learning query scenario and using Entropy-based query strategy. For the Depression dataset, the query sample size was chosen as 100, and for the Suicide dataset, the query sample size was 1000. The results indicate that under ten iterations (i.e., the number of times the unlabeled pool is queried for more samples to be annotated), comparable classification accuracy is achieved, as is equivalent to training on the entire dataset (Refer Figures 6.3, 6.4, 6.5). Though the datasets mentioned above are fully annotated, we do not use the entire datasets for model training. Instead, we sample limited data points from the dataset and request only their labels from the dataset. This setup mimics the human in the loop or Information Oracle of the Active Learning approach in the real world, where the annotator will label only the requested samples from the unlabeled dataset on a need basis.

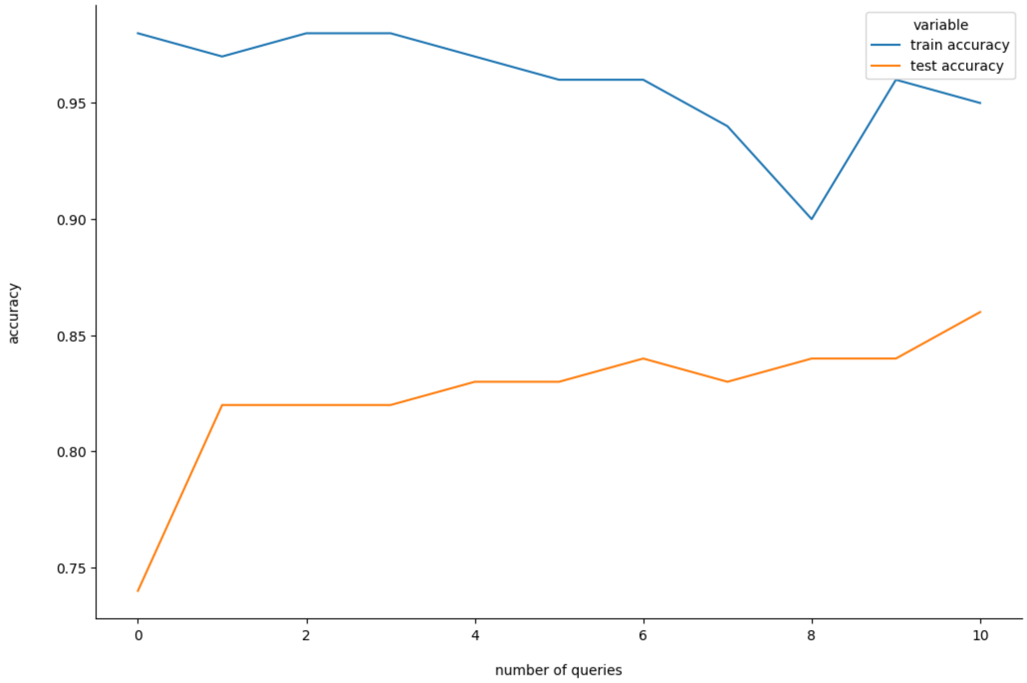


Figure 6.3 *Active Learning Results for Depression Dataset : 0.89 Test Accuracy & 0.93 Train Accuracy is obtained by training on entire labelled dataset, which is equivalent to Accuracy achieved with approx. 10% labelled dataset with Deep Active Learning Loop*

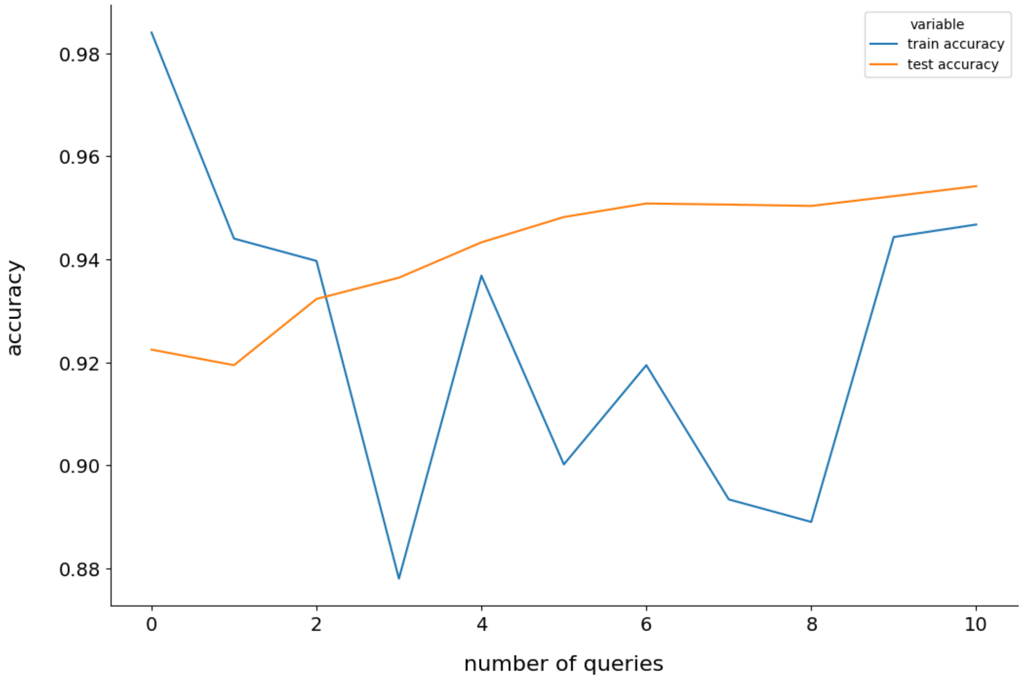
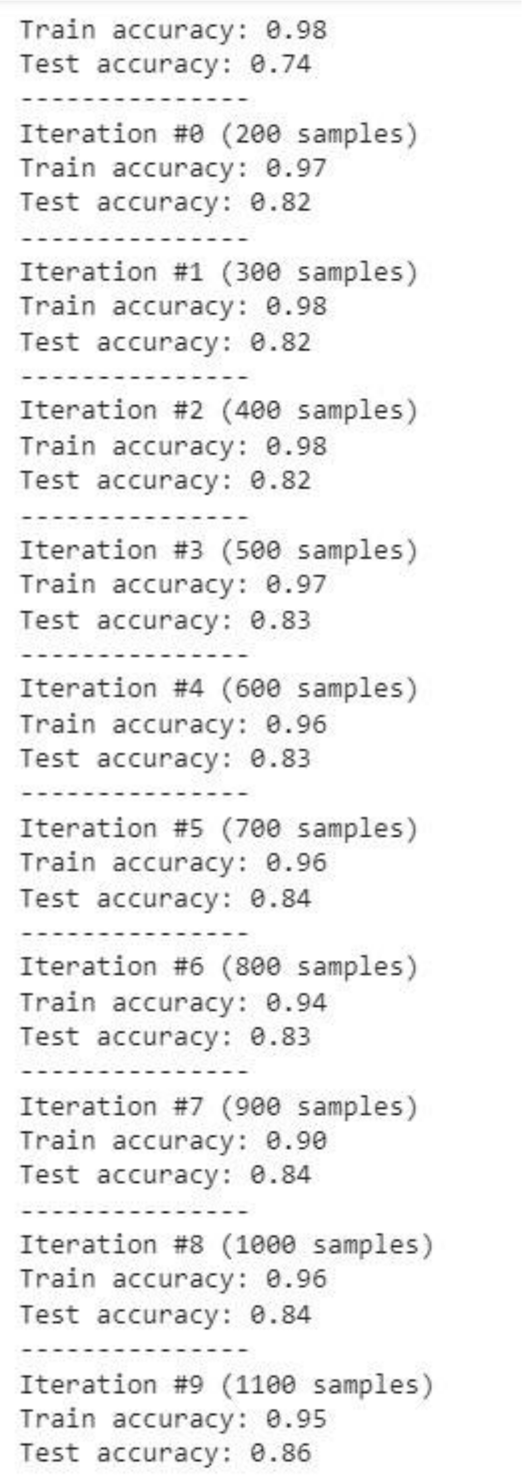
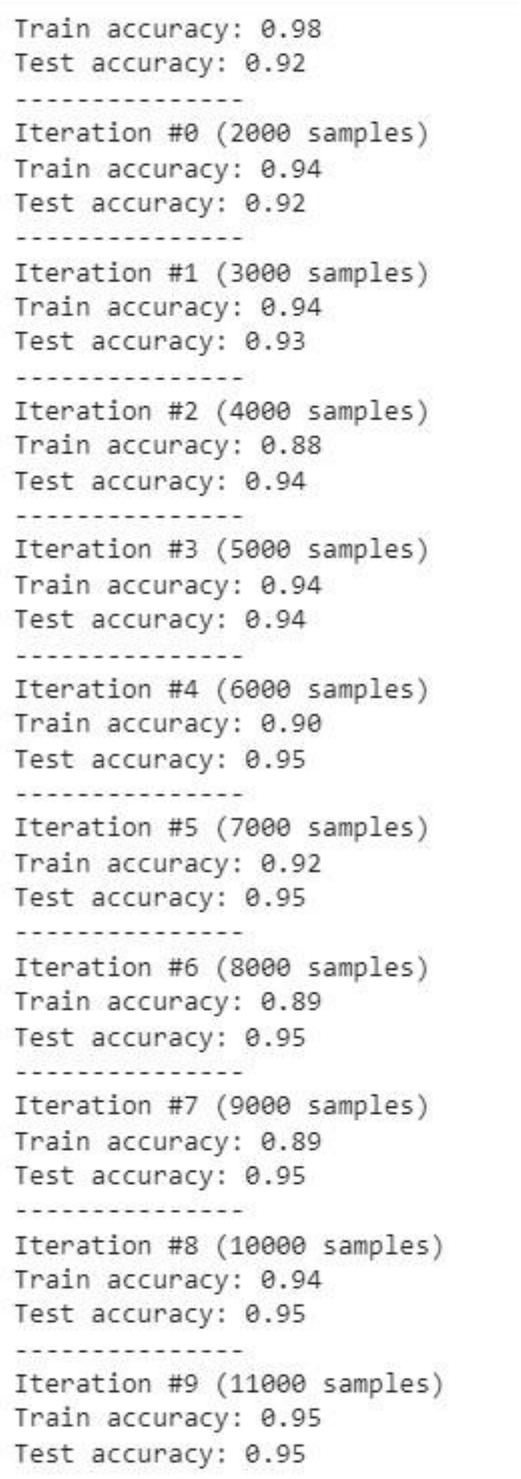


Figure 6.4 *Active Learning Results for Suicide Dataset: 0.96 Test Accuracy & 0.97 Train Accuracy is obtained by training on entire labelled dataset, which is equivalent to Accuracy achieved with approx. 10% labelled dataset with Deep Active Learning Loop*



```
Train accuracy: 0.98
Test accuracy: 0.74
---------------
Iteration #0 (200 samples)
Train accuracy: 0.97
Test accuracy: 0.82
---------------
Iteration #1 (300 samples)
Train accuracy: 0.98
Test accuracy: 0.82
---------------
Iteration #2 (400 samples)
Train accuracy: 0.98
Test accuracy: 0.82
---------------
Iteration #3 (500 samples)
Train accuracy: 0.97
Test accuracy: 0.83
---------------
Iteration #4 (600 samples)
Train accuracy: 0.96
Test accuracy: 0.83
---------------
Iteration #5 (700 samples)
Train accuracy: 0.96
Test accuracy: 0.84
---------------
Iteration #6 (800 samples)
Train accuracy: 0.94
Test accuracy: 0.83
---------------
Iteration #7 (900 samples)
Train accuracy: 0.90
Test accuracy: 0.84
---------------
Iteration #8 (1000 samples)
Train accuracy: 0.96
Test accuracy: 0.84
---------------
Iteration #9 (1100 samples)
Train accuracy: 0.95
Test accuracy: 0.86
```

(a)



```
Train accuracy: 0.98
Test accuracy: 0.92
---------------
Iteration #0 (2000 samples)
Train accuracy: 0.94
Test accuracy: 0.92
---------------
Iteration #1 (3000 samples)
Train accuracy: 0.94
Test accuracy: 0.93
---------------
Iteration #2 (4000 samples)
Train accuracy: 0.88
Test accuracy: 0.94
---------------
Iteration #3 (5000 samples)
Train accuracy: 0.94
Test accuracy: 0.94
---------------
Iteration #4 (6000 samples)
Train accuracy: 0.90
Test accuracy: 0.95
---------------
Iteration #5 (7000 samples)
Train accuracy: 0.92
Test accuracy: 0.95
---------------
Iteration #6 (8000 samples)
Train accuracy: 0.89
Test accuracy: 0.95
---------------
Iteration #7 (9000 samples)
Train accuracy: 0.89
Test accuracy: 0.95
---------------
Iteration #8 (10000 samples)
Train accuracy: 0.94
Test accuracy: 0.95
---------------
Iteration #9 (11000 samples)
Train accuracy: 0.95
Test accuracy: 0.95
```

(b)

Figure 6.5 *Deep Active Learning Training Iterations for (a) Depression Dataset & (b) Suicide Dataset*

6.2 Multimodal Deep Transfer Learning Architecture

This work demonstrates the preliminary research done for extending the scope to categorizing multimodal user generated content on the Internet by exploring the use of recent innovative advancements in the field of deep learning. We have proposed a deep transfer learning framework for the affective analysis of multimodal user-generated content. In the framework, we have proposed fusing multimodal user-generated content by creating joint representations. Joint representations are created by fusing the individual feature vector representations of multiple UGC modalities: text, image, and videos (Refer to Figure 6.6 below) [217]. These feature vector representations are obtained through state-of-the-art techniques for each modality, e.g., Word2Vec for text, VGG-16 for creating feature embeddings for images, and Faster R-CNN for video frames. These joint representations are then used to compute the weighted average score, which can be used to make the final classification decision using the Softmax prediction layer [217]. The above steps of the proposed architecture for categorizing multimodal user generated content from the Internet are discussed in detail below.

Text Feature Representations: Out of all input modalities, text and images are predominant in user-generated content from the Internet. People post pictures and text more than they post videos. Rich image and textual embeddings can be derived from the user's posts for the downstream task of classification from UGC. The steps and techniques required for UGC text pre-processing have already been discussed in Chapter 2 (section 2.1); hence, we have omitted those details from here to avoid redundancy. Post UGC text cleaning and pre-processing, deep neural embeddings like Word2Vec, GloVE, etc., can be used to create text feature representation vectors. These word feature representation vectors can be updated using optimizers such as SGD to embed the contextual information. From the learned vectorized text embedding, the vector representation is passed as input to a dense Fully Connected layer [217].

Image Feature Representations: As a pre-processing step, the pictures from user posts are downsampled for consistency in upcoming steps. The embeddings are extracted from UGC images using a 16-layer VGG network. VGG-16 [214] is a popular deep learning approach, finding its application from classification to segmentation. The weights of the VGG-16 based classification model are initialized from pre-trained weights of VGG network trained on ImageNet database. It has thirteen convolution layers, followed by three FC layers. The striking feature of a VGG network is its 3 x 3 convolution filters, which drastically reduces the number of trainable weights of the network [217].

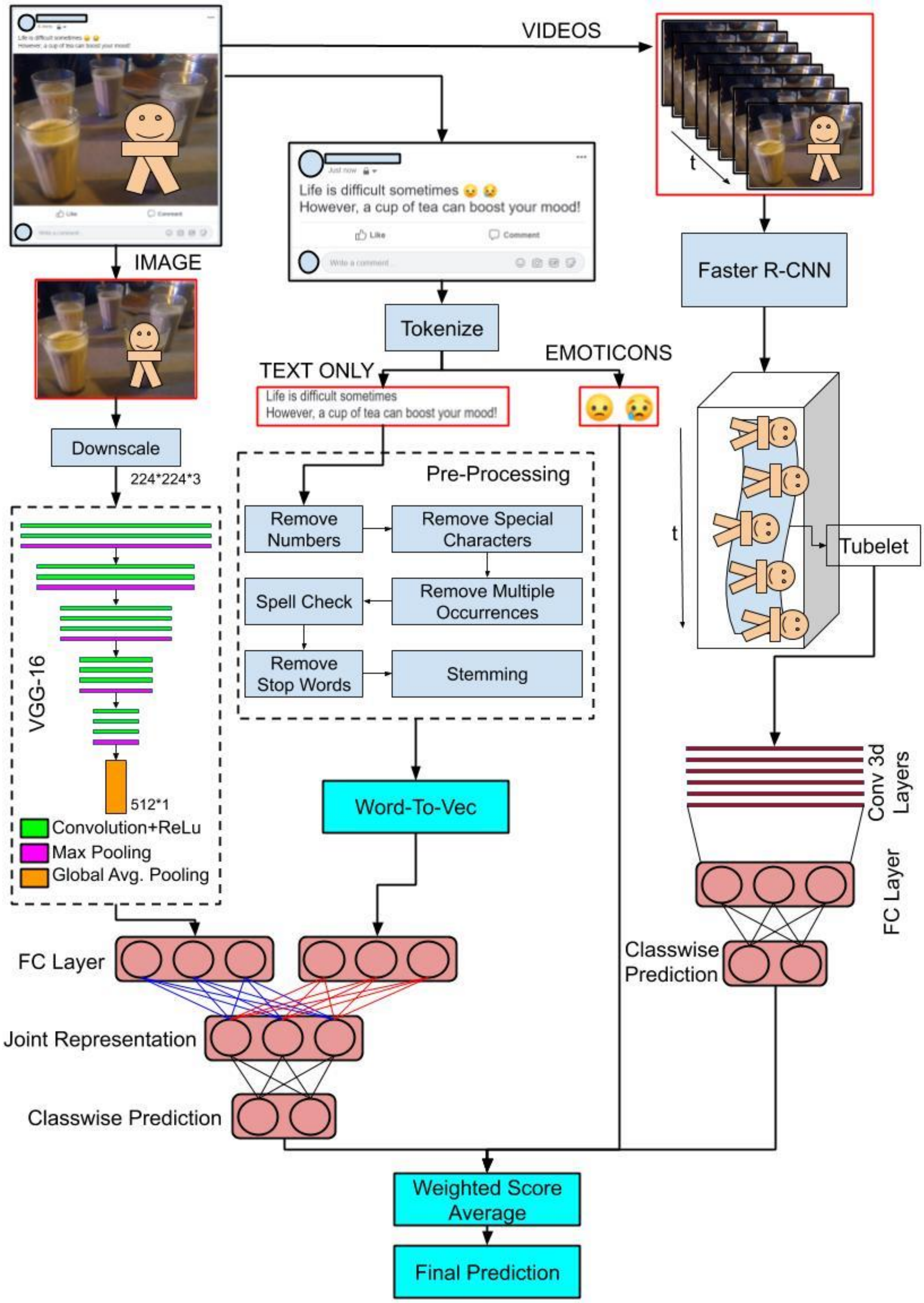


Figure 6.6 *Deep Learning Framework for Categorizing Multimodal UGC*

In the proposed model, we remove the last three dense layers. We do this because they are trained for a classification objective specifically for the ImageNet database, which is not required in our context. Instead, our aim is to extract the embeddings from UGC images for the downstream task of classification. Thus, a global average pooling (GAP) layer is inserted after the 13th convolutional layer, resulting in a representation of dimensionality [512 x 1]. This vector embedding can be utilized for classifying online posted images. Finally, the obtained embedding is fed as input to a dense layer, which is subsequently fused with a similar embedding from text to get a joint feature representation for image and text-based features [217].

Fusing Text and Image Feature Representations: We obtain joint representation by passing text and image embeddings concurrently to the subsequent dense layer. In some scenarios, either image or text based features might be absent as the user just posted either text or image status/post. To handle such a case, during training, we drop a few connections (marked as red and blue colored lines in Fig. 6.6) of either one or both for image or text representation from the previous layer. Even in the absence of one of the modalities, such a mechanism ensures that the forthcoming Softmax classification layer can classify user-posted content. Lastly, the Softmax layer performs binary prediction using joint textual and image embedding [217].

Video Feature Representations: Though text and image modalities dominate in the UGC collected from the Internet, however, videos can also play a vital role in content classification. To classify videos, the architecture should have the capability to: (i) do object detection, (ii) process videos, temporally or at frame level, in real-time, and (iii) classify the video into one of the pre-defined target classes. These three steps for processing videos and creating their feature representation for content classification are described in detail below.

The first step is to detect, localize, and identify the person/object of interest. Firstly, the video is broken down into frames. However, processing each frame is a computationally expensive procedure. Hence, every 20th frame is chosen for object detection and identification. For each chosen frame, Faster R-CNN [215] is used for detection and localization. The CNN in Faster-RCNN produces feature maps for each chosen frame of the video. Internally, a ZF-net model generates feature maps that are subsequently provided as input to the RPN. Here, RPN stands for Region Proposal Network. The objective of RPN is to create region proposals, which are eventually used as candidates for object/person detection. Each region proposal is an input to dense layers, which has a regression task of predicting the bounding box for the person/object in the image, and simultaneously providing class confidence probabilities. Here, a Pascal VOC database pre-trained model is used. Furthermore, the coordinates of the intermediatory tenth frame can be estimated using bi-linear interpolation. Eventually, the video is processed at six FPS, while we perform computation at only three FPS. Here, FPS stands for frame per second. Lastly, the ensemble of the localized person/object through the video sequence is used to form a tubelet, which assists in tracking the activity of the concerned person/object throughout the video [217]. This tubelet subsequently acts as input for the forthcoming 3D convolution layers. Along with spatial convolutions, Conv3d additionally performs convolutions over the additional time dimension. Such across-the-time convolutions encode temporal information, translating to how the person or the object is moving/transitioning across the frames of the video, eventually detecting the activity. The embedding obtained after conv3d is again used for classification using dense layers. In the case of binary classification, the dense layer subsequently has a two-node Softmax layer to classify activity within the video into either of the two target classes [217].

Processing Emoticons: Emoticons express a person's thoughts and feelings. The emoticons pruned out from text can be utilized as a separate modality to encode user's sentiments. The features based on emoticons can be created using their class-wise normalized frequency counts. Similar to the previous three modalities, scores from emoticons are also between [0,1]. A score of 0.5 denotes the absence of emoticons [217].

Fusing Embedding Vectors and Prediction: Lastly, the classification score is calculated by the weighted average over each of the four scores mentioned above. This includes textual, image, video, and emoticon scores. Finally, we would classify the user posted content to the class with the highest score. We plan to implement the proposed multimodal deep transfer learning architecture described above as a part of our future research work [217].

CHAPTER 7

CONCLUSION AND FUTURE WORK

This final chapter discusses the summary of the work done, key takeaways, results, outcomes, limitations, conclusions, and future scope of this research work.

7.1 Summary of Work Done in the Thesis and Key Takeaways

Deep learning techniques are known to improve the classification performance for unstructured data and alleviate the challenges related to handcrafted feature engineering required for training machine learning algorithms. In this research thesis, we have successfully demonstrated the applications of deep learning techniques for categorizing user generated text from the Internet. We have reviewed the current state-of-the-art for this research domain by means of a systematic literature review. We have also empirically evaluated all the popular deep learning techniques for the NLP domain for a real-world application of UGC text categorization using two publicly available datasets. This systematic review has shown that deep learning techniques have wide applications for innovative social computing applications using user generated content from the Internet. Through the literature review, we have focused on understanding the current state-of-the-art, research gaps, open challenges, and future research directions for advancing research applications of deep learning techniques for categorizing user generated content available on the Internet for real-world social computing problems. The survey has helped vastly in learning about the most recent deep learning techniques and model architectures for text categorization and for creating deep neural feature representations/embeddings.

Through the literature review and comparative evaluation of deep learning algorithms, it was identified that clearly, there seems to be a trade-off between model performance and model explainability. Past research has predominantly focused on improving model performance (i.e., accuracy, precision, recall, etc.) by increasing the number of model parameters and stacking more number of hidden and nonlinear layers in neural networks. All of these lead to complex internal structures, which makes it difficult to explain and interpret the model's decisions and decision-making process for humans (users of the system). The currently popular Transformer based Large Language Models have become the state-of-the-art for NLP and NLU domains due to their exceptional prediction correctness and accuracy. However, they are opaque or black box models. Our proposed approach using LIME and SHAP XAI techniques has contributed vastly towards providing explainability and interpretability to these LLM classification decisions. Through multiple experiments, we have demonstrated that our proposed approach can be applied to any open-source pretrained LLM. Results indicate that these model-agnostic techniques can provide reasonable explainability for both short and long user-generated text without the need for intrinsic changes to the network, thereby giving the flexibility to leverage any available, open-source pretrained LLM for transfer learning. We have also demonstrated that unsupervised topic modeling using BERTopic can be leveraged to derive insights from user-generated text and for pseudo-interpretation of user posts.

Additionally, this thesis has also contributed towards addressing challenges with UGC dataset collection and annotation through alternate approaches like Few Shot Learning and Deep Active Learning. We have demonstrated the use of Few Shot Learning paradigm with various domain-adapted pretrained LLMs, which can be an extremely useful approach when only a few, good quality, expert annotated data samples are available. Through another proof of concept for Deep Active Learning, we demonstrate that it is possible to achieve high/comparable accuracy with as few as 10% of samples from the entire dataset by iteratively training an LLM using incremental updates of annotated data instead of using the entire dataset for model development.

The review, analysis, empirical evaluations, and experimental results demonstrate the applications of proposed explainable deep learning techniques for social computing applications using text from the Internet. Additionally, we have done preliminary research to extend our work for multimodal UGC categorization. We have proposed a deep transfer learning framework for the affective analysis of multimodal user generated content from the Internet. We plan to implement the proposed multimodal UGC categorization framework in our future work. The other possible future research enhancements of our research work are discussed in the next subsection. This thesis successfully helps in advancing the research related to the applications of deep learning techniques for categorizing user generated content from the Internet.

7.2 Future Work

Through the systematic literature review, we identified some additional potential research gaps that we plan to address in our future research work. We would like to extend our research in the future to address and cover the following aspects related to the categorization of UGC from the Internet.

1. To study the applications of Deep Learning techniques for Multimodal and Multilingual / Code-mixed user-generated content from the Internet.

2. To research and implement a Multitask Learning framework using various LLMs for multiple UGC text categorization tasks and purposes.

3. To understand temporal and ordinal classification research problems related to user generated content from the Internet and propose solutions for them using deep learning techniques.

4. To understand and address aspects related to Imbalance, Bias, Fairness, Ethics w.r.t. user-generated text collected from the Internet.

5. To understand and address the issues related to context and sarcasm and their impact on UGC categorization tasks.

6. To analyze UGC from other Internet sources and cover new use cases and real-world applications.

7. We also plan to explore and evaluate other surrogate XAI techniques and benchmark their computational performance, quality, and type of explanations generated.

LIST OF PUBLICATIONS FROM THE THESIS

Papers Published in International Journals:

- Malhotra, A., & Jindal, R. (2022). Deep learning techniques for suicide and depression detection from online social media: A scoping review. *Applied Soft Computing*, *130*, 109713. (SCIE, IF: 8.7)
- Malhotra, A., & Jindal, R. (2024). Xai transformer based approach for interpreting depressed and suicidal user behavior on online social networks. *Cognitive Systems Research*, *84*, 101186. (SCIE, IF: 3.9)
- Malhotra, A., & Jindal, R. (2020). Multimodal deep learning based framework for detecting depression and suicidal behaviour by affective analysis of social media posts. *EAI Endorsed Transactions on Pervasive Health and Technology*, *6*(21). (Scopus)

Papers Published in International Conferences:

- Malhotra, A., & Jindal, R. (2021). Multimodal deep learning architecture for identifying victims of online death games. In *Data Analytics and Management: Proceedings of ICDAM* (pp. 827-841). Springer Singapore. (ICDAM 2020 organized virtually by Jan Wyzykowski University Poland and B.M. Institute of Engineering and Technology, India on 18th June 2020) (Scopus)
- Jindal, R., & Malhotra, A. (2022). Efficacious Governance During Pandemics Like Covid-19 Using Intelligent Decision Support Framework for User Generated Content. In *Proceedings of Second Doctoral Symposium on Computational Intelligence: DoSCI 2021* (pp. 435-448). Springer Singapore. (DoSCI 2021, International Conference organized virtually by Institute of Engineering and Technology, Dr APJ Abdul Kalam Technical University, Lucknow, India on 06th March 2021) (Scopus)

Papers Accepted/Presented/In-Press:

- Malhotra, A., & Jindal, R. (Accepted, In-Press). Social media analytics using deep neural networks for mental healthcare applications. In A. Khamparia & D. Gupta (Eds.), *Recent Advances in Computational Intelligence Applications for Biometrics and Biomedical Devices*. Elsevier. (Book Chapter). (Scopus)
- Jindal, R., & Malhotra, A. (Accepted & Presented, In-Press). Leveraging Deep Active Learning and Large Language Models for Cost Efficient Categorization of User Generated Content. In *Proceedings of Fifth International Conference on Data Analytics and Management 2024.* Springer. (ICDAM 2024 organized jointly by London Metropolitian University, London, UK on 14th -15th June 2024) (Scopus)

REFERENCES

1. Paul, M., & Dredze, M. (2011). You are what you tweet: Analyzing twitter for public health. In *Proceedings of the international AAAI conference on web and social media* (Vol. 5, No. 1, pp. 265-272).

2. Paul, M. J., & Dredze, M. (2014). Discovering health topics in social media using topic models. *PloS one*, *9*(8), e103408.

3. Mowery, D., Smith, H. A., Cheney, T., Bryan, C., & Conway, M. (2016). Identifying depression-related tweets from Twitter for public health monitoring. *Online Journal of Public Health Informatics*, 8(1).

4. Carbonell, J. G., Michalski, R. S., & Mitchell, T. M. (1983). An overview of machine learning. *Machine learning*, 3-23.

5. Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction*. MIT press.

6. LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *nature*, *521*(7553), 436-444.

7. Jain, A. K., Mao, J., & Mohiuddin, K. M. (1996). Artificial neural networks: A tutorial. *Computer*, *29*(3), 31-44.

8. Yosinski, J., Clune, J., Bengio, Y., & Lipson, H. (2014). How transferable are features in deep neural networks?. *Advances in neural information processing systems*, *27*.

9. Mind. (2023). Retrieved from https://www.mind.org.uk/ (last accessed November 2023).

10. De Choudhury, M., Kiciman, E., Dredze, M., Coppersmith, G., & Kumar, M. (2016, May). Discovering shifts to suicidal ideation from mental health content in social media. In *Proceedings of the 2016 CHI conference on human factors in computing systems* (pp. 2098-2110).

11. Suicide Awareness Voices of Education (SAVE). (2023). Suicide Statistics Report for USA 2020. https://save.org/about-suicide/suicide-statistics/ (last accessed June 2022).

12. World Health Organization. (2023). Detailed fact sheet on suicide. Retrieved from https://www.who.int/news-room/fact-sheets/detail/suicide (last accessed Feb 2023).

13. World Health Organization. (2023). Mental Health and Substance Use research data. https://www.who.int/teams/mental-health-and-substanceuse/data-research/suicide-data (last accessed Feb 2023).

14. Coppersmith, G., Dredze, M., Harman, C., & Hollingshead, K. (2015). From ADHD to SAD: Analyzing the language of mental health on Twitter through self-reported diagnoses. In *Proceedings of the 2nd workshop on computational linguistics and clinical psychology: from linguistic signal to clinical reality* (pp. 1-10).

15. De Choudhury, M., Gamon, M., Counts, S., & Horvitz, E. (2013). Predicting depression via social media. In *Proceedings of the international AAAI conference on web and social media* (Vol. 7, No. 1, pp. 128-137).

16. Safa, R., Bayat, P., & Moghtader, L. (2022). Automatic detection of depression symptoms in twitter using multimodal analysis. *The Journal of Supercomputing*, *78*(4), 4709-4744.

17. Manning, C. D., Raghavan, P., & Schütze, H. (2008)., *Introduction to Information Retrieval*, Cambridge University Press.

18. Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of machine Learning research*, *3*(Jan), 993-1022.

19. Yin, J., & Wang, J. (2014, August). A dirichlet multinomial mixture model-based approach for short text clustering. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining* (pp. 233-242).

20. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., & Polosukhin, I. (2017). Attention is all you need. *Advances in neural information processing systems*, *30*.

21. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)* (pp. 4171-4186).

22. Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J. D., Dhariwal, P., ... & Amodei, D. (2020). Language models are few-shot learners. *Advances in neural information processing systems*, *33*, 1877-1901.

23. Yang, Z., Dai, Z., Yang, Y., Carbonell, J., Salakhutdinov, R. R., & Le, Q. V. (2019). Xlnet: Generalized autoregressive pretraining for language understanding. *Advances in neural information processing systems*, *32*.

24. LeCun, Y., Boser, B., Denker, J., Henderson, D., Howard, R., Hubbard, W., & Jackel, L. (1989). Handwritten digit recognition with a back-propagation network. *Advances in neural information processing systems*, *2*.

25. Cho, K., Merrienboer, B., Gulcehre, C., Bougares, F., Schwenk, H., & Bengio, Y. (2014). Learning Phrase Representations using RNN Encoder-Decoder for Statistical Machine Translation. In *EMNLP*.

26. Greff, K., Srivastava, R. K., Koutník, J., Steunebrink, B. R., & Schmidhuber, J. (2016). LSTM: A search space odyssey. *IEEE transactions on neural networks and learning systems*, *28*(10), 2222-2232.

27. Chung, J., Gulcehre, C., Cho, K., & Bengio, Y. (2014). Empirical evaluation of gated recurrent neural networks on sequence modeling. In *NIPS 2014 Workshop on Deep Learning, December 2014*.

28. Graves, A., & Schmidhuber, J. (2005). Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural networks*, *18*(5-6), 602-610.

29. Goodfellow, I., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., ... & Bengio, Y. (2014). Generative adversarial nets. *Advances in neural information processing systems*, *27*.

30. Hinton, G. E. (2012). A practical guide to training restricted Boltzmann machines. In *Neural Networks: Tricks of the Trade: Second Edition* (pp. 599-619). Berlin, Heidelberg: Springer Berlin Heidelberg.

31. Hinton, G. E., Osindero, S., & Teh, Y. W. (2006). A fast learning algorithm for deep belief nets. *Neural computation*, *18*(7), 1527-1554.

32. Salakhutdinov, R., & Larochelle, H. (2010, March). Efficient learning of deep Boltzmann machines. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics* (pp. 693-700). JMLR Workshop and Conference Proceedings.

33. Bengio, Y., Lamblin, P., Popovici, D., & Larochelle, H. (2006). Greedy layer-wise training of deep networks. *Advances in neural information processing systems*, *19*.

34. Vincent, P., Larochelle, H., Lajoie, I., Bengio, Y., Manzagol, P. A., & Bottou, L. (2010). Stacked denoising autoencoders: Learning useful representations in a deep network with a local denoising criterion. *Journal of machine learning research*, *11*(12).

35. Ng, A. (2011). Sparse autoencoder. *CS294A Lecture notes*, *72*(2011), 1-19.

36. Vincent, P., Larochelle, H., Bengio, Y., & Manzagol, P. A. (2008, July). Extracting and composing robust features with denoising autoencoders. In *Proceedings of the 25th international conference on Machine learning* (pp. 1096-1103).

37. Kingma, D. P., & Welling, M. (2013). Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*.

38. Forsyth, D. A., Mundy, J. L., di Gesú, V., Cipolla, R., LeCun, Y., Haffner, P., ... & Bengio, Y. (1999). Object recognition with gradient-based learning. *Shape, contour and grouping in computer vision*, 319-345.

39. Bai, S., Kolter, J. Z., & Koltun, V. (2018). An empirical evaluation of generic convolutional and recurrent networks for sequence modeling. *arXiv preprint arXiv:1803.01271*.

40. He, K., Zhang, X., Ren, S., & Sun, J. (2016). Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 770-778).

41. Page, M. J., McKenzie, J. E., Bossuyt, P. M., Boutron, I., Hoffmann, T. C., Mulrow, C. D., ... & Moher, D. (2021). The PRISMA 2020 statement: an updated guideline for reporting systematic reviews. *International journal of surgery*, *88*, 105906.

42. Mowery, D. L., Park, Y. A., Bryan, C., & Conway, M. (2016, December). Towards automatically classifying depressive symptoms from Twitter data for population health. In *Proceedings of the workshop on computational modeling of people's opinions, personality, and emotions in social media (PEOPLES)* (pp. 182-191).

43. Benton, A., Mitchell, M., & Hovy, D. (2017). Multitask learning for mental health conditions with limited social media data. In *15th Conference of the European Chapter of the Association for Computational Linguistics, EACL 2017. Proceedings of Conference*. Association for Computational Linguistics.

44. Gkotsis, G., Oellrich, A., Velupillai, S., Liakata, M., Hubbard, T. J., Dobson, R. J., & Dutta, R. (2017). Characterisation of mental health conditions in social media using Informed Deep Learning. *Scientific reports*, *7*(1), 45141.

45. Yates, A., Cohan, A., & Goharian, N. (2017, September). Depression and Self-Harm Risk Assessment in Online Forums. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing* (pp. 2968-2978).

46. Halder, K., Poddar, L., & Kan, M. Y. (2017, September). Modeling temporal progression of emotional status in mental health forum: A recurrent neural net approach. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis* (pp. 127-135).

47. Trotzek, M., Koitka, S., & Friedrich, C. M. (2017, September). Linguistic Metadata Augmented Classifiers at the CLEF 2017 Task for Early Detection of Depression. In *CLEF (working notes)* (pp. 1-17).

48. Sadeque, F., Xu, D., & Bethard, S. (2017, September). Uarizona at the clef erisk 2017 pilot task: linear and recurrent models for early depression detection. In *CEUR workshop proceedings* (Vol. 1866). NIH Public Access.

49. Sadeque, F., Xu, D., & Bethard, S. (2018, February). Measuring the latency of depression detection in social media. In *Proceedings of the Eleventh ACM International Conference on Web Search and Data Mining* (pp. 495-503).

50. Maupomé, D., & Meurs, M. J. (2018). Using Topic Extraction on Social Media Content for the Early Detection of Depression. *CLEF (working notes)*, *2125*.

51. Wang, Y. T., Huang, H. H., & Chen, H. H. (2018, September). A Neural Network Approach to Early Risk Detection of Depression and Anorexia on Social Media Text. In *CLEF (Working Notes).*

52. Paul, S., Jandhyala, S. K., & Basu, T. (2018, August). Early Detection of Signs of Anorexia and Depression Over Social Media using Effective Machine Learning Frameworks. In *CLEF (Working notes)*.

53. Trotzek, M., Koitka, S., & Friedrich, C. M. (2018, September). Word embeddings and linguistic metadata at the CLEF 2018 tasks for early detection of depression and anorexia. In *CLEF (working notes)*.

54. Liu, N., Zhou, Z., Xin, K., & Ren, F. (2018, September). TUA1 at eRisk 2018. In *CLEF (Working Notes)*.

55. Trotzek, M., Koitka, S., & Friedrich, C. M. (2018). Utilizing neural networks and linguistic metadata for early detection of depression indications in text sequences. *IEEE Transactions on Knowledge and Data Engineering*, *32*(3), 588-601.

56. Orabi, A. H., Buddhitha, P., Orabi, M. H., & Inkpen, D. Deep Learning for Depression Detection of Twitter Users. *NAACL HLT 2018*, 88.

57. Shing, H. C., Nair, S., Zirikly, A., Friedenberg, M., Hal Daumé, I. I. I., & Resnik, P. Expert, Crowdsourced, and Machine Assessment of Suicide Risk via Online Postings. *NAACL HLT 2018*, 25.

58. Wu, M. Y., Shen, C. Y., Wang, E. T., & Chen, A. L. (2020). A deep architecture for depression detection using posting, behavior, and living environment data. *Journal of Intelligent Information Systems*, *54*, 225-244.

59. Cohan, A., Desmet, B., Yates, A., Soldaini, L., MacAvaney, S., & Goharian, N. (2018). SMHD: a Large-Scale Resource for Exploring Online Language Usage for Multiple Mental Health Conditions. In *27th International Conference on Computational Linguistics* (pp. 1485-1497). ACL.

60. Sawhney, R., Manchanda, P., Mathur, P., Shah, R., & Singh, R. (2018, October). Exploring and learning suicidal ideation connotations on social media with deep learning. In *Proceedings of the 9th workshop on computational approaches to subjectivity, sentiment and social media analysis* (pp. 167-175).

61. Du, J., Zhang, Y., Luo, J., Jia, Y., Wei, Q., Tao, C., & Xu, H. (2018). Extracting psychiatric stressors for suicide from social media using deep learning. *BMC medical informatics and decision making*, *18*, 77-87.

62. Coppersmith, G., Leary, R., Crutchley, P., & Fine, A. (2018). Natural language processing of social media as screening for suicide risk. *Biomedical informatics insights*, *10*, 1178222618792860.

63. Ji, S., Yu, C. P., Fung, S. F., Pan, S., & Long, G. (2018). Supervised learning for suicidal ideation detection in online user content. *Complexity*, *2018*.

64. Cong, Q., Feng, Z., Li, F., Xiang, Y., Rao, G., & Tao, C. (2018, December). XA-BiLSTM: a deep learning approach for depression detection in imbalanced data. In *2018 IEEE international conference on bioinformatics and biomedicine (BIBM)* (pp. 1624-1627). IEEE.

65. Shen, T., Jia, J., Shen, G., Feng, F., He, X., Luan, H., ... & Hall, W. (2018, July). Cross-domain depression detection via harvesting social media. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence* (pp. 1611-1617).

66. Song, H. Y., You, J., Chung, J. W., & Park, J. C. (2018). Feature Attention Network: Interpretable Depression Detection from Social Media. In *32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 32)*. Pacific Asia Conference on Language, Information and Computation (PACLIC 32).

67. Naderi, N., Gobeill, J., Teodoro, D., Pasche, E., & Ruch, P. (2019). A baseline approach for early detection of signs of anorexia and self-harm in reddit posts. In *Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2019 Working Notes*. 9-12 September 2019.

68. Ragheb, W., Azé, J., Bringay, S., & Servajean, M. (2019, September). Attentive multi-stage learning for early risk detection of signs of anorexia and self-harm on social media. In *CLEF 2019-Conference and Labs of the Evaluation Forum* (Vol. 2380, No. 126).

69. Allen, K., Bagroy, S., Davis, A., & Krishnamurti, T. (2019). ConvSent at CLPsych 2019 Task A: Using Post-level Sentiment Features for Suicide Risk Prediction on Reddit. *NAACL HLT 2019*, 182.

70. Morales, M., Belitz, D., Chernova, N., Dey, P., & Theisen, T. (2019). An Investigation of Deep Learning Systems for Suicide Risk Assessment. *NAACL HLT 2019*, 177.

71. Mohammadi, E., Amini, H., & Kosseim, L. (2019). CLaC at CLPsych 2019: Fusion of Neural Features and Predicted Class Probabilities for Suicide Risk Assessment Based on Online Posts. *NAACL HLT 2019,* 34.

72. Ambalavanan, A. K., Jagtap, P. D., Adhya, S., & Devarakonda, M. (2019). Using Contextual Representations for Suicide Risk Assessment from Internet Forums. *NAACL HLT 2019*, 172.

73. Matero, M., Idnani, A., Son, Y., Giorgi, S., Vu, H., Zamani, M., ... & Schwartz, H. A. (2019). Suicide Risk Assessment with Multi-level Dual-Context Language and BERT. *NAACL HLT 2019,* 39.

74. Stankevich, M., Smirnov, I., Kiselnikova, N., & Ushakova, A. (2020). Depression detection from social media profiles. In *Data Analytics and Management in Data Intensive Domains: 21st International Conference, DAMDID/RCDL 2019, Kazan, Russia, October 15–18, 2019, Revised Selected Papers 21* (pp. 181-194). Springer International Publishing.

75. Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of depression-related posts in reddit social media forum. *Ieee Access, 7*, 44883-44893.

76. Maupomé, D., Queudot, M., & Meurs, M. J. (2019). Inter and intra document attention for depression risk assessment. In *Advances in Artificial Intelligence: 32nd Canadian Conference on Artificial Intelligence, Canadian AI 2019, Kingston, ON, Canada, May 28–31, 2019, Proceedings 32* (pp. 333-341). Springer International Publishing.

77. Buddhitha, P., & Inkpen, D. (2019). Multi-Task, Multi-Channel, Multi-Input Learning for Mental Illness Detection using Social Media Text. *EMNLP-IJCNLP 2019,* 54.

78. Gaur, M., Alambo, A., Sain, J. P., Kursuncu, U., Thirunarayan, K., Kavuluru, R., ... & Pathak, J. (2019, May). Knowledge-aware assessment of severity of suicide risk for early intervention. In *The world wide web conference* (pp. 514-525).

79. Sinha, P. P., Mishra, R., Sawhney, R., Mahata, D., Shah, R. R., & Liu, H. (2019, November). # suicidal-A multipronged approach to identify and explore suicidal ideation in twitter. In *Proceedings of the 28th ACM international conference on information and knowledge management* (pp. 941-950).

80. Mishra, R., Sinha, P. P., Sawhney, R., Mahata, D., Mathur, P., MIDAS, I., & Shah, R. R. (2019). SNAP-BATNET: Cascading Author Profiling and Social Network Graphs for Suicide Ideation Detection on Social Media. *NAACL HLT 2019*, 147.

81. Cao, L., Zhang, H., Feng, L., Wei, Z., Wang, X., Li, N., & He, X. (2019, November). Latent Suicide Risk Detection on Microblog via Suicide-Oriented Word Embeddings and Layered Attention. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 1718-1728).

82. Gui, T., Zhang, Q., Zhu, L., Zhou, X., Peng, M., & Huang, X. (2019). Depression detection on social media with reinforcement learning. In *Chinese Computational Linguistics: 18th China National Conference, CCL 2019, Kunming, China, October 18–20, 2019, Proceedings 18* (pp. 613-624). Springer International Publishing.

83. Gui, T., Zhu, L., Zhang, Q., Peng, M., Zhou, X., Ding, K., & Chen, Z. (2019, July). Cooperative multimodal approach to depression detection in twitter. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 33, No. 01, pp. 110-117).

84. Wang, X., Chen, S., Li, T., Li, W., Zhou, Y., Zheng, J., ... & Tang, B. (2019, June). Assessing depression risk in Chinese microblogs: a corpus and machine learning methods. In *2019 IEEE International Conference on Healthcare Informatics (ICHI)* (pp. 1-5). IEEE.

85. Wang, X., Chen, S., Li, T., Li, W., Zhou, Y., Zheng, J., ... & Tang, B. (2020). Depression risk prediction for chinese microblogs via deep-learning methods: Content analysis. *JMIR medical informatics*, *8*(7), e17958.

86. An, M., Wang, J., Li, S., & Zhou, G. (2020, December). Multimodal topic-enriched auxiliary learning for depression detection. In *proceedings of the 28th international conference on computational linguistics* (pp. 1078-1089).

87. Ophir, Y., Tikochinski, R., Asterhan, C. S., Sisso, I., & Reichart, R. (2020). Deep neural networks detect suicide risk from textual facebook posts. *Scientific reports*, *10*(1), 16685.

88. Sawhney, R., Joshi, H., Gandhi, S., & Shah, R. (2020, November). A time-aware transformer based model for suicide ideation detection on social media. In *Proceedings of the 2020 conference on empirical methods in natural language processing (EMNLP)* (pp. 7685-7697).

89. Lee, D., Park, S., Kang, J., Choi, D., & Han, J. (2020, November). Cross-lingual suicidal-oriented word embedding toward suicide prevention. In *Findings of the Association for Computational Linguistics: EMNLP 2020* (pp. 2208-2217).

90. Kim, J., Lee, J., Park, E., & Han, J. (2020). A deep learning model for detecting mental illness from user content on social media. *Scientific reports*, *10*(1), 11846.

91. Alabdulkreem, E. (2021). Prediction of depressed Arab women using their tweets. *Journal of Decision Systems*, *30*(2-3), 102-117.

92. de Carvalho, V. F., Giacon, B., Nascimento, C., & Nogueira, B. M. (2020, October). Machine learning for suicidal ideation identification on Twitter for the portuguese language. In *Brazilian Conference on Intelligent Systems* (pp. 536-550). Cham: Springer International Publishing.

93. Tadesse, M. M., Lin, H., Xu, B., & Yang, L. (2019). Detection of suicide ideation in social media forums using deep learning. *Algorithms*, *13*(1), 7.

94. Yao, H., Rashidian, S., Dong, X., Duanmu, H., Rosenthal, R. N., & Wang, F. (2020). Detection of suicidality among opioid users on reddit: machine learning–based approach. *Journal of medical internet research*, *22*(11), e15293.

95. Rao, G., Peng, C., Zhang, L., Wang, X., & Feng, Z. (2020, August). A knowledge enhanced ensemble learning model for mental disorder detection on social media. In *International Conference on Knowledge Science, Engineering and Management* (pp. 181-192). Cham: Springer International Publishing.

96. Sekulić, I., & Strube, M. (2019, November). Adapting Deep Learning Methods for Mental Health Prediction on Social Media. In *Proceedings of the 5th Workshop on Noisy User-generated Text (W-NUT 2019)* (pp. 322-327).

97. Jiang, Z. P., Levitan, S. I., Zomick, J., & Hirschberg, J. (2020, November). Detection of mental health from reddit via deep contextualized representations. In *Proceedings of the 11th international workshop on health text mining and information analysis* (pp. 147-156).

98. Rao, G., Zhang, Y., Zhang, L., Cong, Q., & Feng, Z. (2020). MGL-CNN: a hierarchical posts representations model for identifying depressed individuals in online forums. *IEEE Access*, *8*, 32395-32403.

99. Ji, S., Li, X., Huang, Z., & Cambria, E. (2022). Suicidal ideation and mental disorder detection with attentive relation networks. *Neural Computing and Applications*, *34*(13), 10309-10319.

100. Liu, D., Fu, Q., Wan, C., Liu, X., Jiang, T., Liao, G., ... & Liu, R. (2020). Suicidal ideation cause extraction from social texts. *Ieee Access*, *8*, 169333-169351.

101. Mann, P., Paes, A., & Matsushima, E. H. (2020, May). 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* (Vol. 14, pp. 440-451).

102. Lin, C., Hu, P., Su, H., Li, S., Mei, J., Zhou, J., & Leung, H. (2020, June). Sensemood: depression detection on social media. In *Proceedings of the 2020 international conference on multimedia retrieval* (pp. 407-411).

103. Ramírez-Cifuentes, D., Freire, A., Baeza-Yates, R., Puntí, J., Medina-Bravo, P., Velazquez, D. A., ... & Gonzàlez, J. (2020). Detection of suicidal ideation on social media: multimodal, relational, and behavioral analysis. *Journal of medical internet research*, *22*(7), e17758.

104. Cao, L., Zhang, H., & Feng, L. (2020). Building and using personal knowledge graph to improve suicidal ideation detection on social media. *IEEE Transactions on Multimedia*, *24*, 87-102.

105. Chiu, C. Y., Lane, H. Y., Koh, J. L., & Chen, A. L. (2021). Multimodal depression detection on instagram considering time interval of posts. *Journal of Intelligent Information Systems*, *56*, 25-47.

106. Bagherzadeh, H., Fazl-Ersi, E., & Vahedian, A. (2020). Detection of early sign of self-harm on Reddit using multi-level machine.

107. Achilles, L., Kisselew, M., Schäfer, J., & Koelle, R. (2020, September). Using Surface and Semantic Features for Detecting Early Signs of Self-Harm in Social Media Postings. In *CLEF (Working Notes)*.

108. Uban, A. S., & Rosso, P. (2020). Deep learning architectures and strategies for early detection of self-harm and depression level prediction. In *CEUR workshop proceedings* (Vol. 2696, pp. 1-12). Sun SITE Central Europe.

109. Madani, A., Boumahdi, F., Boukenaoui, A., Kritli, M. C., & Hentabli, H. (2020, September). USDB at eRisk 2020: Deep Learning Models to Measure the Severity of the Signs of Depression using Reddit Posts. In *CLEF (Working Notes)*.

110. Martınez-Castano, R., Htait, A., Azzopardi, L., & Moshfeghi, Y. (2020). Early risk detection of self-harm and depression severity using BERT-based transformers. *Working Notes of CLEF*, 16.

111. Maupomé, D., Armstrong, M. D., Belbahar, R. M., Alezot, J., Balassiano, R., Queudot, M., ... & Meurs, M. J. (2020, September). Early Mental Health Risk Assessment through Writing Styles, Topics and Neural Models. In *CLEF (Working Notes)*.

112. Maupomé, D., Armstrong, M. D., Rancourt, F., & Meurs, M. J. (2021). Leveraging Textual Similarity to Predict Beck Depression Inventory Answers. In *Canadian Conference on AI*.

113. Sawhney, R., Joshi, H., Gandhi, S., & Shah, R. R. (2021, March). Towards ordinal suicide ideation detection on social media. In *Proceedings of the 14th ACM international conference on web search and data mining* (pp. 22-30).

114. Sawhney, R., Joshi, H., Shah, R., & Flek, L. (2021, June). Suicide ideation detection via social and temporal user representations using hyperbolic learning. In *Proceedings of the 2021 conference of the North American Chapter of the Association for Computational Linguistics: human language technologies* (pp. 2176-2190).

115. Uban, A. S., Chulvi, B., & Rosso, P. (2021). An emotion and cognitive based analysis of mental health disorders from social media data. *Future Generation Computer Systems*, *124*, 480-494.

116. Ren, L., Lin, H., Xu, B., Zhang, S., Yang, L., & Sun, S. (2021). Depression detection on reddit with an emotion-based attention network: algorithm development and validation. *JMIR medical informatics*, *9*(7), e28754.

117. Ragheb, W., Aze, J., Bringay, S., & Servajean, M. (2021). Negatively Correlated Noisy Learners for At-Risk User Detection on Social Networks: A Study on Depression, Anorexia, Self-Harm, and Suicide. *IEEE Transactions on Knowledge and Data Engineering*, *35*(1), 770-783.

118. Hamad, Z., Imran, R., Jameel, M. S., & Guandong, X. (2021, July). DepressionNet: A Novel Summarization Boosted Deep Framework for Depression Detection on Social Media. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 133-142). ACM (Association for Computing Machinery).

119. Murarka, A., Radhakrishnan, B., & Ravichandran, S. (2021, April). Classification of mental illnesses on social media using RoBERTa. In *Proceedings of the 12th international workshop on health text mining and information analysis* (pp. 59-68).

120. Gollapalli, S. D., Zagatti, G. A., & Ng, S. K. (2021). Suicide Risk Prediction by Tracking Self-Harm Aspects in Tweets: NUS-IDS at the CLPsych 2021 Shared Task. *NAACL HLT 2021*, 93.

121. Morales, M., Dey, P., & Kohli, K. (2021). A Comparison of Simple vs. Complex Models for Suicide Risk Assessment. *NAACL HLT 2021*, 99.

122. Wang, N., Luo, F., Shivtare, Y., Badal, V., Subbalakshmi, K. P., Chandramouli, R., & Lee, E. (2021). Learning Models for Suicide Prediction from Social Media Posts. *NAACL HLT 2021*, 87.

123. Bayram, U., & Benhiba, L. (2021). Determining a Person's Suicide Risk by Voting on the Short-Term History of Tweets for the CLPsych 2021 Shared Task. *NAACL HLT 2021*, 81.

124. Renjith, S., Abraham, A., Jyothi, S. B., Chandran, L., & Thomson, J. (2022). An ensemble deep learning technique for detecting suicidal ideation from posts in social media platforms. *Journal of King Saud University-Computer and Information Sciences*, *34*(10), 9564-9575.

125. Basile, A., Chinea-Rios, M., Uban, A. S., Müller, T., Rössler, L., Yenikent, S., ... & Franco-Salvador, M. (2021, September). Upv-symanto at erisk 2021: Mental health author profiling for early risk prediction on the internet. In *Proceedings of the Working Notes of CLEF 2021, Conference and Labs of the Evaluation Forum, Bucharest, Romania, September 21st to 24th, 2021* (pp. 908-927). CEUR.

126. Inkpen, D., Skaik, R., Buddhitha, P., Angelov, D., & Fredenburgh, M. T. (2021). uOttawa at eRisk 2021: Automatic Filling of the Beck's Depression Inventory Questionnaire using Deep Learning. In *CLEF (Working Notes)* (pp. 966-980).

127. Lopes, R. P. (2021). CeDRI at eRisk 2021: A naive approach to early detection of psychological disorders in social media. In *CEUR Workshop Proceedings* (pp. 981-991). CEUR Workshop Proceedings.

128. Ghosh, S., Ekbal, A., & Bhattacharyya, P. (2021). What does your bio say? inferring twitter users' depression status from multimodal profile information using deep learning. *IEEE transactions on computational social systems*, *9*(5), 1484-1494.

129. Uban, A. S., Chulvi, B., & Rosso, P. (2021, June). On the explainability of automatic predictions of mental disorders from social media data. In *International Conference on Applications of Natural Language to Information Systems* (pp. 301-314). Cham: Springer International Publishing.

130. Farruque, N., Goebel, R., Zaïane, O. R., & Sivapalan, S. (2021, December). Explainable zero-shot modelling of clinical depression symptoms from text. In *2021 20th IEEE International Conference on Machine Learning and Applications (ICMLA)* (pp. 1472-1477). IEEE.

131. Zogan, H., Razzak, I., Wang, X., Jameel, S., & Xu, G. (2022). Explainable depression detection with multi-aspect features using a hybrid deep learning model on social media. *World Wide Web*, *25*(1), 281-304.

132. Kour, H., & Gupta, M. K. (2022). An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM. *Multimedia Tools and Applications*, *81*(17), 23649-23685.

133. Ahmed, U., Srivastava, G., Yun, U., & Lin, J. C. W. (2022). EANDC: An explainable attention network based deep adaptive clustering model for mental health treatment. *Future Generation Computer Systems*, *130*, 106-113.

134. Naseem, U., Dunn, A. G., Kim, J., & Khushi, M. (2022, April). Early identification of depression severity levels on reddit using ordinal classification. In *Proceedings of the ACM Web Conference 2022* (pp. 2563-2572).

135. Ansari, L., Ji, S., Chen, Q., & Cambria, E. (2022). Ensemble hybrid learning methods for automated depression detection. *IEEE transactions on computational social systems*, *10*(1), 211-219.

136. Cheng, J. C., & Chen, A. L. (2022). Multimodal time-aware attention networks for depression detection. *Journal of Intelligent Information Systems*, *59*(2), 319-339.

137. MacAvaney, S., Desmet, B., Cohan, A., Soldaini, L., Yates, A., Zirikly, A., & Goharian, N. (2018, June). RSDD-Time: Temporal Annotation of Self-Reported Mental Health Diagnoses. In *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic* (pp. 168-173).

138. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. (2015). CLPsych 2015 Shared Task: Depression and PTSD on Twitter. *NAACL HLT 2015*, 31.

139. Mowery, D., Smith, H., Cheney, T., Stoddard, G., Coppersmith, G., Bryan, C., & Conway, M. (2017). Understanding depressive symptoms and psychosocial stressors on Twitter: a corpus-based study. *Journal of medical Internet research*, *19*(2), e6895.

140. Losada, D. E., & Crestani, F. (2016, August). A test collection for research on depression and language use. In *International conference of the cross-language evaluation forum for European languages* (pp. 28-39). Cham: Springer International Publishing.

141. Losada, D. E., Crestani, F., & Parapar, J. (2017). eRISK 2017: CLEF lab on early risk prediction on the internet: experimental foundations. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 8th International Conference of the CLEF Association, CLEF 2017, Dublin, Ireland, September 11–14, 2017, Proceedings 8* (pp. 346-360). Springer International Publishing.

142. Losada, D. E., Crestani, F., & Parapar, J. (2018). Overview of eRisk: early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 9th International Conference of the CLEF Association, CLEF 2018, Avignon, France, September 10-14, 2018, Proceedings 9* (pp. 343-361). Springer International Publishing.

143. Losada, D. E., Crestani, F., & Parapar, J. (2020). Overview of eRisk at CLEF 2020: Early Risk Prediction on the Internet (Extended Overview). *CLEF (Working Notes)*.

144. Milne, D. N., Pink, G., Hachey, B., & Calvo, R. A. (2016, June). Clpsych 2016 shared task: Triaging content in online peer-support forums. In *Proceedings of the third workshop on computational linguistics and clinical psychology* (pp. 118-127).

145. Zirikly, A., Resnik, P., Uzuner, O., & Hollingshead, K. (2019). CLPsych 2019 Shared Task: Predicting the Degree of Suicide Risk in Reddit Posts. *NAACL HLT 2019*, 24.

146. MacAvaney, S., Mittu, A., Coppersmith, G., Leintz, J., & Resnik, P. (2021). Community-level Research on Suicidality Prediction in a Secure Environment: Overview of the CLPsych 2021 Shared Task. *NAACL HLT 2021*, 70.

147. Coppersmith, G., Leary, R., Whyne, E., & Wood, T. (2015, August). Quantifying suicidal ideation via language usage on social media. In *Joint statistics meetings proceedings, statistical computing section, JSM* (Vol. 110).

148. Jamil, Z., Inkpen, D., Buddhitha, P., & White, K. (2017, August). Monitoring Tweets for Depression to Detect At-risk Users. In *Proceedings of the Fourth Workshop on Computational Linguistics and Clinical Psychology—From Linguistic Signal to Clinical Reality* (pp. 32-40).

149. Shen, G., Jia, J., Nie, L., Feng, F., Zhang, C., Hu, T., ... & Zhu, W. (2017, August). Depression detection via harvesting social media: A multimodal dictionary learning solution. In *IJCAI* (pp. 3838-3844).

150. Parapar, J., Martín-Rodilla, P., Losada, D. E., & Crestani, F. (2021). Overview of eRisk at CLEF 2021: Early Risk Prediction on the Internet (Extended Overview). *CLEF (Working Notes)*, 864-887.

151. Pirina, I., & Çöltekin, Ç. (2018, October). Identifying depression on reddit: The effect of training data. In *Proceedings of the 2018 EMNLP workshop SMM4H: the 3rd social media mining for health applications workshop & shared task* (pp. 9-12).

152. Losada, D. E., Crestani, F., & Parapar, J. (2019). Overview of erisk 2019 early risk prediction on the internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 10th International Conference of the CLEF Association, CLEF 2019, Lugano, Switzerland, September 9–12, 2019, Proceedings 10* (pp. 340-357). Springer International Publishing.

153. Zeberga, K., Attique, M., Shah, B., Ali, F., Jembre, Y. Z., & Chung, T. S. (2022). A novel text mining approach for mental health prediction using Bi-LSTM and BERT model. *Computational Intelligence and Neuroscience, 2022.*

154. Coppersmith, G., Dredze, M., & Harman, C. (2014, June). Quantifying mental health signals in Twitter. In *Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality* (pp. 51-60).

155. Coppersmith, G., Ngo, K., Leary, R., & Wood, A. (2016, June). Exploratory analysis of social media prior to a suicide attempt. In *Proceedings of the third workshop on computational linguistics and clinical psychology* (pp. 106-117).

156. Aggarwal, C. C., & Zhai, C. (2012). *Mining text data*. Boston, MA: Springer US. https://link.springer.com/book/10.1007/978-1-4614-3223-4

157. Scikit-learn module documentation for text pre-processing using OneHotEncoder. Retrieved from: https://scikit-learn.org/stable/modules/generated/sklearn.preprocessing.OneHotEncoder.html (last accessed June 2022)

158. Song, X. (2021). A Fast WordPiece Tokenization System. *Google Research Blog*. Retrieved from https://ai.googleblog.com/2021/12/a-fast-wordpiece-tokenization-system.html (last accessed June 2022).

159. TensorFlow, Resource Guide for Text Word Embeddings. https://www.tensorflow.org/text/guide/word_embeddings (last accessed June 2022).

160. TensorFlow. (2023). Tutorial on Word Embeddings. Retrieved from https://www.tensorflow.org/tutorials/text/word2vec (last accessed June 2022).

161. Pennington, J., Socher, R., & Manning, C. D. (2014, October). Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)* (pp. 1532-1543).

162. Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching word vectors with subword information. *Transactions of the association for computational linguistics*, *5*, 135-146.

163. Devlin, J. (2018 November 02). Open Sourcing BERT: State-of-the-Art Pre-training for Natural Language Processing. *Google Research Blog.* Retrieved from https://ai.googleblog.com/2018/11/open-sourcing-bert-state-of-art-pre.html (last accessed June 2022).

164. Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2020 March 11). BERT. *Google Research GitHub Repository*. Retrieved from https://github.com/google-research/bert (last accessed November 2023).

165. Muller B. (2022 March 2). BERT 101 State of The Art NLP Model Explained. *Hugging Face*. Retrieved from https://huggingface.co/blog/bert-101(last accessed November 2023).

166. Molnar, C. (2022). *Interpretable Machine Learning: A Guide for Making Black Box Models Explainable (2nd ed.)*. Retrieved from christophm.github.io/interpretable-ml-book/ (last accessed November 2023).

167. Reddy, S. (2022). Explainability and artificial intelligence in medicine. *The Lancet Digital Health*, *4*(4), e214-e215.

168. Amann, J., Blasimme, A., Vayena, E., Frey, D., & Madai, V. I. (2020). Explainability for artificial intelligence in healthcare: a multidisciplinary perspective. *BMC medical informatics and decision making*, *20*(1), 1-9.

169. Das, A., & Rad, P. (2020). Opportunities and challenges in explainable artificial intelligence (xai): A survey. *arXiv preprint arXiv:2006.11371*.

170. Cloudera Fast Forward Lab. (2017). *Interpretability*. Retrieved from https://ff06-2020.fastforwardlabs.com/ (last accessed November 2023).

171. Balasubramaniam, N., Kauppinen, M., Rannisto, A., Hiekkanen, K., & Kujala, S. (2023). Transparency and explainability of AI systems: From ethical guidelines to requirements. *Information and Software Technology*, *159*, 107197.

172. Caruana, R., Lou, Y., Gehrke, J., Koch, P., Sturm, M., & Elhadad, N. (2015, August). Intelligible models for healthcare: Predicting pneumonia risk and hospital 30-day readmission. In *Proceedings of the 21th ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1721-1730).

173. Kim, B., Khanna, R., & Koyejo, O. O. (2016). Examples are not enough, learn to criticize! criticism for interpretability. *Advances in neural information processing systems*, *29*.

174. Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence*, *267*, 1-38.

175. Ribeiro, M. T., Singh, S., & Guestrin, C. (2016, August). " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135-1144).

176. Lundberg, S. M., & Lee, S. I. (2017, December). A unified approach to interpreting model predictions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems* (pp. 4768-4777).

177. Grootendorst, M. (2022). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

178. Speer R. (2019 March 12). ftfy (Version 5.5). *Zenodo*. Retrieved from https://zenodo.org/record/2591652 (last accessed November 2023).

179. Hugging Face. (2016). Retrieved from https://huggingface.co/ (last accessed November 2023).

180. Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., & Cambria, E. (2022, June). MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 7184-7190).

181. Vajre, V., Naylor, M., Kamath, U., & Shehu, A. (2021, December). PsychBERT: a mental health language model for social media mental health behavioral analysis. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)* (pp. 1077-1082). IEEE.

182. Naseem, U., Lee, B. C., Khushi, M., Kim, J., & Dunn, A. G. (2022). Benchmarking for Public Health Surveillance tasks on Social Media with a Domain-Specific Pretrained Language Model. *NLP-Power 2022*, 22.

183. Lundberg, S. M., & Lee, S. I. (2017). *SHAP GitHub Repository*. Retrieved from https://github.com/shap/shap (last accessed November 2023).

184. Nayak, A. (2019). Idea behind LIME and SHAP. *Towards Data Science*. Retrieved from https://towardsdatascience.com/idea-behind-lime-and-shap-b603d35d34eb (last accessed November 2023).

185. SHAPLEY, L. (1953). A value for n-person games. *Contributions to the Theory of Games*, 307-317.

186. Duda, R. O., Hart, P. E., & Stork, D. G. (2000, November). *Pattern Classification*. (2nd ed.) John Wiley & Sons.

187. Hoffman, T. (1990). Probabilistic latent semantic indexing. In *Proceedings of the 22nd International Conference on Research and Development in Information Retrieval, 1990* (pp. 50-57).

188. Févotte, C., & Idier, J. (2011). Algorithms for nonnegative matrix factorization with the β-divergence. *Neural computation*, *23*(9), 2421-2456.

189. Angelov, D. (2020). Top2vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*.

190. McInnes, L., Healy, J., & Melville, J. (2018). Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

191. McInnes, L., Healy, J., Saul, N., & Großberger, L. (2018). UMAP: Uniform Manifold Approximation and Projection. *Journal of Open Source Software*, *3*(29), 861.

192. McInnes, L., & Healy, J. (2017, November). Accelerated hierarchical density based clustering. In *2017 IEEE International Conference on Data Mining Workshops (ICDMW)* (pp. 33-42). IEEE.

193. McInnes, L., Healy, J., & Astels, S. (2017). hdbscan: Hierarchical density based clustering. *J. Open Source Softw.*, *2*(11), 205.

194. Hugging Face. (2023). *Transformers*. Retrieved from https://huggingface.co/docs/transformers/index (last accessed November 2023).

195. Rezazadeh M. (2021 November 27). *Twitter Depression Detection GitHub Repository*. Retrieved from https://github.com/miladrezazadeh/twitter_depression_detection. (last accessed November 2023).

196. Komati N. (2021 January). Suicide and Depression Detection. *Kaggle Datasets*. Retrieved from https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch. (last accessed November 2023).

197. Haque, A., Reddi, V., & Giallanza, T. (2021). Deep learning for suicide and depression identification with unsupervised label correction. In *Artificial Neural Networks and Machine*

Learning–ICANN 2021: 30th International Conference on Artificial Neural Networks, Bratislava, Slovakia, September 14–17, 2021, Proceedings, Part V 30 (pp. 436-447). Springer International Publishing.

198. Sung, F., Yang, Y., Zhang, L., Xiang, T., Torr, P. H., & Hospedales, T. M. (2018). Learning to compare: Relation network for few-shot learning. In *Proceedings of the IEEE conference on computer vision and pattern recognition* (pp. 1199-1208).

199. Ríssola, E. A., Losada, D. E., & Crestani, F. (2021). A survey of computational methods for online mental state assessment on social media. *ACM Transactions on Computing for Healthcare*, *2*(2), 1-31.

200. Uban, A. S., Chulvi, B., & Rosso, P. (2022, June). Multi-aspect transfer learning for detecting low resource mental disorders on social media. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference* (pp. 3202-3219).

201. Mikal, J., Hurst, S., & Conway, M. (2016). Ethical issues in using Twitter for population-level depression monitoring: a qualitative study. *BMC medical ethics*, *17*(1), 1-11.

202. Fei-Fei, L., Fergus, R., & Perona, P. (2006). One-shot learning of object categories. *IEEE transactions on pattern analysis and machine intelligence*, *28*(4), 594-611.

203. Lake, B., Salakhutdinov, R., Gross, J., & Tenenbaum, J. (2011). One shot learning of simple visual concepts. In *Proceedings of the annual meeting of the cognitive science society* (Vol. 33, No. 33).

204. Lampert, C. H., Nickisch, H., & Harmeling, S. (2013). Attribute-based classification for zero-shot visual object categorization. *IEEE transactions on pattern analysis and machine intelligence*, *36*(3), 453-465.

205. Alex, N., Lifland, E., Tunstall, L., Thakur, A., Maham, P., Riedel, C. J., ... & Stuhlmüller, A. (2021, August). RAFT: A Real-World Few-Shot Text Classification Benchmark. In *Thirty-fifth Conference on Neural Information Processing Systems Datasets and Benchmarks Track (Round 2)*.

206. Kang, D., & Cho, M. (2022). Integrative few-shot learning for classification and segmentation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 9979-9990).

207. Ling, J., Liao, L., Yang, M., & Shuai, J. (2022). Semi-supervised few-shot learning via multi-factor clustering. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 14564-14573).

208. Gildenblat, J. Overview of Active Learning for Deep Learning. Retrieved from https://jacobgil.github.io/deeplearning/activelearning (last accessed November 2023).

209. Settles, B. (2009). *Active learning literature survey*. Computer Sciences Technical Report 1648. University of Wisconsin–Madison.

210. Dd Schröder, C., Müller, L., Niekler, A., & Potthast, M. (2023, May). Small-Text: Active Learning for Text Classification in Python. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics: System Demonstrations* (pp. 84-95).

211. Eversberg, L. (2023 April 26). How to Train Neural Networks with Fewer Data Using Active Learning. *Towards AI.* Retrieved from https://pub.towardsai.net/how-to-train-neural-networks-with-fewer-data-using-active-learning-445154c30ddf (last accessed November 2023).

212. Ren, P., Xiao, Y., Chang, X., Huang, P. Y., Li, Z., Gupta, B. B., ... & Wang, X. (2021). A survey of deep active learning. *ACM computing surveys (CSUR)*, *54*(9), 1-40.

213. Gal, Y., Islam, R., & Ghahramani, Z. (2017, July). Deep bayesian active learning with image data. In *International conference on machine learning* (pp. 1183-1192). PMLR.

214. Simonyan, K., & Zisserman, A. (2014). Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*.

215. Ren, S., He, K., Girshick, R., & Sun, J. (2015). Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, *28*.

216. Malhotra, A., & Jindal, R. (2022). Deep learning techniques for suicide and depression detection from online social media: A scoping review. *Applied Soft Computing*, 109713.

217. Malhotra, A., & Jindal, R. (2020). Multimodal deep learning based framework for detecting depression and suicidal behaviour by affective analysis of social media posts. *EAI Endorsed Transactions on Pervasive Health and Technology*, *6*(21).