

Detecting Mental Distress: A Comprehensive Analysis of Online Discourses Via ML and NLP

Master's Thesis

in partial fulfillment of the requirements for
the degree of Master of Science (M.Sc.)
in Web and Data Science

submitted by
Bhavya Ashvin Shah
(221202683)

First supervisor: Prof. Dr. Frank Hopfgartner
Institute for Web Science and Technologies

Second supervisor: Dr. Ing. Stefania Zourlidou
Institute for Web Science and Technologies

Koblenz, May 2024

Statement

I hereby certify that this thesis has been composed by me and is based on my own work, that I did not use any further resources than specified – in particular no references unmentioned in the reference section – and that I did not submit this thesis to another examination before. The paper submission is identical to the submitted electronic version.

	Yes	No
I agree to have this thesis published in the library.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
I agree to have this thesis published on the Web.	<input checked="" type="checkbox"/>	<input type="checkbox"/>
The thesis text is available under a Creative Commons License (CC BY-SA 4.0).	<input checked="" type="checkbox"/>	<input type="checkbox"/>
The source code is available under a GNU General Public License (GPLv3).	<input checked="" type="checkbox"/>	<input type="checkbox"/>
The collected data is available under a Creative Commons License (CC BY-SA 4.0).	<input checked="" type="checkbox"/>	<input type="checkbox"/>

Koblenz, May 16, 2024

.....
(Place, Date)



.....
(Bhavya Shah)

Note

- If you would like us to contact you for the graduation ceremony,
please provide your personal E-mail address: bhavya260@gmail.com
- If you would like us to send you an invite to join the WeST Alumni
and Members group on LinkedIn, please provide your LinkedIn ID :
www.linkedin.com/in/bhavya-ashvin-shah

Zusammenfassung

Diese Arbeit erforscht und untersucht die Effektivität und Wirksamkeit traditioneller Modelle des maschinellen Lernens (ML), fortschrittlicher neuronaler Netze (NN) und hochmoderner Modelle des tiefen Lernens (DL) zur Identifizierung von Indikatoren für psychische Probleme in den sozialen Medien Reddit und Twitter, die von Teenagern stark genutzt werden. Verschiedene NLP-Vektorisierungstechniken wie TF-IDF, Word2Vec, GloVe und BERT-Einbettungen werden mit ML-Modellen wie Entscheidungsbaum (DT), Random Forest (RF), logistische Regression (LR) und Support Vector Machine (SVM) eingesetzt, gefolgt von NN-Modellen wie Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) und Long Short-Term Memory (LSTM), um ihre Auswirkungen als Merkmalsdarstellung von Modellen methodisch zu analysieren. DL-Modelle wie BERT, DistilBERT, MentalRoBERTa und MentalBERT sind durchgängig auf die Klassifizierungsaufgabe abgestimmt. In dieser Arbeit werden auch verschiedene Textvorverarbeitungstechniken wie Tokenisierung, Stoppwortentfernung und Lemmatisierung verglichen, um ihre Auswirkungen auf die Modellleistung zu bewerten. Es wurden systematische Experimente mit verschiedenen Konfigurationen von Vektorisierungs- und Vorverarbeitungstechniken in Übereinstimmung mit verschiedenen Modelltypen und -kategorien durchgeführt, um die effektivsten Konfigurationen zu finden und die Stärken, Grenzen und Fähigkeiten zur Erkennung und Interpretation von Indikatoren für psychische Störungen aus dem Text zu ermitteln. Die Analyse der Ergebnisse zeigt, dass das MentalBERT DL-Modell alle anderen Modelltypen und -kategorien signifikant übertrifft, da es durch sein spezifisches Vortraining auf psychische Daten sowie eine rigorose End-to-End-Feinabstimmung einen Vorteil bei der Erkennung von nuancierten linguistischen Indikatoren für psychische Belastung aus dem komplexen kontextuellen Textkorpus hat. Diese Erkenntnisse aus den Ergebnissen bestätigen das hohe Potenzial der ML- und NLP-Technologien für die Entwicklung komplexer KI-Systeme für den Einsatz im Bereich der Analyse der psychischen Gesundheit. Diese Arbeit legt den Grundstein und leitet die künftige Arbeit, die die Notwendigkeit eines kollaborativen Ansatzes verschiedener Domänenexperten zeigt, sowie die Erforschung der nächsten Generation großer Sprachmodelle, um robuste und klinisch bewährte KI-Systeme für psychische Gesundheit zu entwickeln.

Abstract

This thesis explores and examines the effectiveness and efficacy of traditional machine learning (ML), advanced neural networks (NN) and state-of-the-art deep learning (DL) models for identifying mental distress indicators from the social media discourses based on Reddit and Twitter as they are immensely used by teenagers. Different NLP vectorization techniques like TF-IDF, Word2Vec, GloVe, and BERT embeddings are employed with ML models such as Decision Tree (DT), Random Forest (RF), Logistic Regression (LR) and Support Vector Machine (SVM) followed by NN models such as Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM) to methodically analyse their impact as feature representation of models. DL models such as BERT, DistilBERT, MentalRoBERTa and MentalBERT are end-to-end fine tuned for classification task. This thesis also compares different text preprocessing techniques such as tokenization, stopword removal and lemmatization to assess their impact on model performance. Systematic experiments with different configuration of vectorization and preprocessing techniques in accordance with different model types and categories have been implemented to find the most effective configurations and to gauge the strengths, limitations, and capability to detect and interpret the mental distress indicators from the text. The results analysis reveals that MentalBERT DL model significantly outperformed all other model types and categories due to its specific pretraining on mental data as well as rigorous end-to-end fine tuning gave it an edge for detecting nuanced linguistic mental distress indicators from the complex contextual textual corpus. This insights from the results acknowledges the ML and NLP technologies high potential for developing complex AI systems for its intervention in the domain of mental health analysis. This thesis lays the foundation and directs the future work demonstrating the need for collaborative approach of different domain experts as well as to explore next generational large language models to develop robust and clinically approved mental health AI systems.

Acknowledgement

I would like to express my gratitude and kind regards to the Institute of Web Science and Technologies and to its esteemed members as my first supervisors Prof. Dr. Frank Hopfgartner, and my second supervisor, Dr. Ing. Stefania Zourlidou, for constant guidance, feedback and support. Their supervision made a great impact in completing my thesis and shaping it with academic rigor. I am grateful and thankful for their mentorship, patience and insights they shared during this entire research journey. I would also like to thank my Family and Friends for constantly motivating and supporting me during my thesis.

Finally, I would like to express my immense gratitude to the University of Koblenz for providing me with an interesting thesis research direction to work on and to the University's library to provide me the resources and academic environment to complete my thesis successfully.

Contents

1	Introduction	1
2	Literature Review	4
2.1	Related Work	4
2.2	Research Aim and Questions	10
3	Theoretical Background	13
3.1	Machine Learning (ML)	13
3.2	Natural Language Processing (NLP)	13
3.3	Exploratory Data Analysis (EDA)	14
3.4	ML and NLP Pipeline	15
3.5	Data Cleaning and Preprocessing	15
3.6	Vectorization	15
3.7	Text Classification	16
3.8	Hyperparameter Tuning	17
3.9	Modelling	17
3.10	Fine Tuning in NLP	21
3.11	Model Validation and Evaluation Techniques	21
4	Methodology	24
4.1	Research Design and Approach	24
4.2	Environment Stack	25
4.3	Systematic Workflow of the Pipeline	27
4.4	Dataset Overview	29
4.5	Exploratory Data Analysis	30
4.6	Data Cleaning and Pre-processing	31
4.7	Feature Engineering	33
4.8	Modelling	35
4.8.1	Hyperparameter Tuning and Optimization	37
4.8.2	Unified Training Process, Evaluation Metrics, and Model Persistence	38
4.8.3	Traditional Machine Learning Models	40
4.8.4	Advanced Neural Network Models	42
4.8.5	State-of-the-Art Deep Learning Models	45
4.8.6	Model Validation and Evaluation	47
4.9	Model’s Misclassification and Interpretability Analysis	49

5	Results	52
5.1	Exploratory Data Analysis	52
5.2	Insights into Data Cleaning, Pre-processing and Feature Engineering	55
5.3	Model Output Evaluation	57
5.3.1	Traditional Machine Learning Models	57
5.3.2	Advanced Neural Network Models	63
5.3.3	State-of-the-art Transformer based Deep learning Models . . .	67
5.4	Model Category-wise Performance Evaluation	70
5.5	Impact of Different Pre-Processing Techniques	72
5.6	Combined Analysis Engine	73
5.7	Interpretability Analysis	76
5.7.1	Feature Importance and Coefficient weights of ML Models . .	76
5.7.2	Attention Mechanism and SHAP Analysis of DL Model . . .	78
5.8	Misclassification Analysis	83
5.9	Validation Through External Unseen Datasets	85
6	Discussion	87
6.1	Analysis and Integration of Findings	87
6.2	Applications and Implications of Deep Learning Models for Social Media Mental Health	92
6.3	Ethical Considerations	93
6.4	Critical Reflection	95
7	Conclusion	99
7.1	Summary of Findings	99
7.2	Future Work	100
	References	102
	Appendix	113

List of Figures

4.1	Flow Diagram of Pipeline Execution	27
5.1	Dataset information	52
5.2	Labels distribution	52
5.3	Data characteristics	53
5.4	Data samples	53
5.5	Text length distribution	53
5.6	Dataset word cloud	54
5.7	Most frequent words	54
5.8	N-Grams analysis	55
5.9	Comparative Analysis of Vectorization Techniques for ML Models	58
5.10	Different Model's Performance Metrics	58
5.11	DT Classification Report	59
5.12	DT Confusion Matrix	59
5.13	DT AUC-ROC Curve	59
5.14	RF Classification Report	60
5.15	RF Confusion Matrix	60
5.16	RF AUC-ROC Curve	60
5.17	LR Classification Report	61
5.18	LR Confusion Matrix	61
5.19	LR AUC-ROC Curve	61
5.20	SVM Classification Report	62
5.21	SVM Confusion Matrix	62
5.22	SVM AUC-ROC Curve	62
5.23	Comparative Analysis of Vectorization Techniques for NN Models	63
5.24	Different Model's Performance Metrics	63
5.25	CNN Classification Report	64
5.26	CNN Confusion Matrix	64
5.27	CNN AUC-ROC Curve	64
5.28	RNN Classification Report	65
5.29	RNN Confusion Matrix	65
5.30	RNN AUC-ROC Curve	65
5.31	LSTM Classification Report	66
5.32	LSTM Confusion Matrix	66
5.33	LSTM AUC-ROC Curve	66
5.34	Comparative Analysis of DL Model Variants	68

5.35	Different Model's Performance Metrics	68
5.36	Model's Training and Validation Loss	69
5.37	Classification Report	69
5.38	Confusion Matrix	70
5.39	AUC-ROC Curve	70
5.40	Comparative Analysis of All Model Categories	71
5.41	Model Performance Metrics	71
5.42	Impact of Different Pre-processing Techniques on Model's Performance	72
5.43	Preprocessing Techniques Impact on Model's Performance Metrics .	73
5.44	Classification Analysis	74
5.45	Distress Analysis	74
5.46	Sentiment Analysis	74
5.47	Emotion Analysis	75
5.48	DT Feature Importance	76
5.49	RF Feature Importance	76
5.50	LR Label 1 Coeff. Weights	77
5.51	LR Label 0 Coeff. Weights	77
5.52	SVM Label 1 Coeff. Weights	78
5.53	SVM Label 0 Coeff. Weights	78
5.54	Attention Mechanism on MentalBERT with Last Layer	79
5.55	Attention Mechanism on MentalBERT with Aggregated Layers . . .	80
5.56	SHAP Analysis of MentalBERT Model on Distressed Instances	81
5.57	SHAP Analysis of MentalBERT Model on No Distress Instances . . .	82
5.58	Misclassification Instances Evaluation With 2 Pre-Trained Models . .	83
5.59	Mean Scores for Misclassified Distressed Instances into No Distress .	84
5.60	Mean Scores for Misclassified No Distress instances into Distress . .	84
5.61	External Validation Evaluation on Dataset 1 of Depression	85
5.62	External Validation Evaluation on Dataset 2 of Suicide	86

List of Tables

3.1	Confusion Matrix Structure	22
-----	--------------------------------------	----

Abbreviations

- EDA: Exploratory Data Analysis
- TFIDF: Term Frequency-Inverse Document Frequency
- Word2Vec: Word to Vector
- GloVe: Global Vectors for Word Representation
- BERT embeddings: Bidirectional Encoder Representations from Transformers embeddings
- NLP: Natural Language Processing
- ML: Machine Learning
- NN: Neural Network
- DL: Deep Learning
- AI: Artificial Intelligence
- DT: Decision Tree
- RF: Random Forest
- LR: Logistic Regression
- SVM: Support Vector Machine
- CNN: Convolutional Neural Network
- RNN: Recurrent Neural Network
- LSTM: Long Short-Term Memory
- BERT: Bidirectional Encoder Representations from Transformers
- LLM: Large Language Model
- AUC ROC: Area Under the Curve Receiver Operating Characteristic
- SHAP: Shapley Additive exPlanations
- Coeff: Coefficient
- RQ: Research Question

1 Introduction

Ongoing digital transformation is constantly shaping the world and communication. Social media platforms have evolved from time to time. From creating connections with the world to sharing expressions and opinions with the community. These activities generate an enormous amount of data which is directly related to the insights in public psychology and especially trends to their mental state and overall mental well-being. Research reveals that social media channels contribute around 81% of this kind of data [Zhang et al., 2022]. Trends shows that from all the major psychological illnesses, depression and suicide are the major cases with 45% and 20% respectively. To express emotions and thoughts, social media platforms like Reddit and Twitter are used more often. There are Subreddits which are specific to the topics like depression and suicide. These channels make them valuable for mental distress detection and it is possible by exploring and analysing the online discourses textual corpus data form the social media with the combined application of Machine learning (ML) and natural language processing (NLP) [Zhang et al., 2022].

The main motivation that drives this thesis is the increasing problems related to mental healthcare especially linked with anxiety, depression, suicidal ideation, and overall mental distress on the social media channels. The social media usage among teenagers is drastically increasing. In 2018, the usage of social media nearly doubled compared to 2014 – 2015 [Draženović et al., 2023]. Since then, it's continuously on the rise. Another research states that the evolving nature of the interaction between the social media and teenagers with different patterns calls for urgency and highlights the importance of mental health care and its repercussions [Anderson et al., 2023]. This motivates this thesis to study and explore the online discourses further than sentiment analysis to look for the patterns and trends in the complex linguistic corpus.

To detect the mental distress and associated illness from social media discourses is a challenging task, because of its informal and diverse nature of the language. These discourses are always evolving in nature with respect to contextual usage and cultural differences, for instance the use of expression slang's change from generation to generation, culture and geographical regions. This unique challenge demands the creation of such models that can be effective and capable of complex analytics for the distress detection. Let's consider an illustrative example of the available type of data on social media. In a Reddit thread "Daily Life", users express their thoughts and feelings generally via comments, opinions, posts etc. These kinds of statements

express their mental state that can be positive and negative, showcasing distress, anxiety, overwhelming feelings and more. Therefore, this kind of statements or on-line discourses textual data needs to be explored and demands a systematic study. Similar example of a discussion or discourse is followed as:

User 1: "I am just overwhelmed with the ongoing events in my life. It feels like I cannot bear this burden anymore and just want to give up."

User 2: "Damn, lately I feel like I am just existing on the earth to suffer from problems I never asked for."

User 3: "My anxiety has been to the roof and having panic attacks. I hope that I don't harm myself out of frustration."

These texts can be analysed by using ML and NLP to find out the linguistic features and patterns related to mental distress contributing to the domain of mental health analysis.

To tackle this kind of data to identify mental distress indicators, finding patterns, understanding entire context and scenarios, a systematic and comprehensive study will be conducted. This thesis is a mixture of qualitative and quantitative methods of research. Social media data will be explored and analysed with the use of sophisticated Machine Learning (ML) models, Natural Language Processing (NLP) techniques, Neural Network (NN) models and Transformer based Deep Learning (DL) models. This methodology is quite crucial for exploration, experimentation and will be impactful for understanding and interpreting the textual corpus of data. Textual classification models will be built for mental distress detection. The goal of this thesis is to utilize and leverage state-of-the-art technologies for deep exploration and further analysis of the discourses. This benefits early detection of mental health distress that can be attended to improve the mental health trajectory.

This thesis potentially impacts interdisciplinary fields such as psychology professionals, mental health analysis discipline, data science domain, social media, and public policies. By exploring and shedding light on different approaches of mental distress detection on social media platforms and channels, it encourages and promotes the development of digital support systems and mental health intervention systems. This is a step towards promoting mental health care and establishing interventions for supporting the well being of young people's mental health.

This thesis is systematically organized with a structured approach. It starts with an introduction and motivation, the thesis objectives and an overview of the entire thesis. In the second chapter, literature review is stated in depth with the state-of-the-art research on and around the theme followed by the thesis aim and investigative questions acting as a guiding principle of this thesis. In the third chapter, the theoretical background highlighting important concepts, terminologies and techniques used in this thesis will be explained briefly. The next chapter 4, guides through the methodology presenting the design of the thesis, the implementation flow, with

the analytical and explorative experiments performed, explains the choices of techniques and models followed by the rational. The subsequent chapter 5 depicts the results of the modelling experiments and a comparative analysis of the different techniques, models and model categories. The following chapter 6 sheds light on the discussion with respect to the achieved results and their reasoning, as well as application of the models in mental health analysis, ethical considerations followed by critical reflection into this thesis. Last, chapter 7 demonstrates the stage set for the future research perspectives and direction concluding the overall conducted thesis work highlighting the respective key findings and insights.

2 Literature Review

In this chapter, a critical literature review of the most recent scholarly studies is conducted in order to understand the applicability of ML and NLP on social media discourses for detecting mental distress. This literature review based on advanced applied techniques will be focused on and around the theme of how the application of ML and NLP can be applied to the textual data to detect the mental distress. This review will study the articles based on application of ML and NLP on clinical records, non-clinical records, sentiment analysis and surrounding topics. It will showcase the different preprocessing methods, algorithms, validation and evaluation techniques and interpretation approaches applied, different dataset exploration and overall different perspectives and output use cases followed by ethical considerations, biases, limitations, and challenges. This exploration will lay a foundational knowledge understanding that will aid the research in further stages. Followed by the related work, this chapter will also include the research aim and investigative questions acting as the guiding principles of this thesis.

2.1 Related Work

To develop a comprehensive understanding of the applications of studies of ML and NLP in the mental health domain, the review navigates with this insightful article. In the research paper “Machine learning and Natural Language Processing in Mental Health” [Le Glaz et al., 2021], the authors have conducted a systematic review of research papers addressing the medical databases. They included around 58 research papers from a total of 327 research papers for studies to understand the trends and get insights to the applications of ML and NLP with respect to mental health. The authors find out that the applications of ML and NLP are quite versatile and can be employed for many purposes like therapy evaluation, severity classification, symptom extraction and some diagnostic challenges. Most of the studies have used mainly two kinds of data sources such as clinical records and social media data. These two types of data sources are a combination of multiple sources of similar kinds of data in different formats, sizes, complex linguistics, and demographics. The authors also found out that major programming languages behind the model classifiers were based on Python followed by R as good support of standard NLP packages and faster processing of data. Algorithms like DT, RF, SVM were majorly used for most of the research with NLP techniques as feature engineering. Authors

also argue that, instead of proving an existing clinical hypothesis, these methods should be employed to the larger group of audience with different personalities and belongings. Authors argue that the clinical records and data is not publicly available as for the data privacy and patient doctor relationship so the studies cannot be applied to the different groups as there is a chance of personality biasness, ethnicity, cultural and linguistic differences. To completely leverage the applications of ML and NLP to the broader context use, there is a need for complex understanding of the contextual data and for that complex models are required. The risk of biases and limitations are not studied and cannot be comprehended directly from the research. Authors argue that models as well as text preprocessing and vectorization methods selection is based on theoretical knowledge and comprehensive analysis with comparison of such methods and models are missing. Authors suggest that even though there is a potential application of these technologies, Artificial intelligence (AI) based on these models should be only used as a tool with systematic interpretation by the domain health experts for an aid with caution. Ethical considerations with respect to data protection, privacy and exploitation remains a concern for most of the research work [Le Glaz et al., 2021]. The key takeaways from this research paper which studies 58 similar theme research articles is that ML and NLP can be applied but needs further exploration of these techniques with respect to ethical considerations and building complex models which can comprehend semantic meanings and context better. To attend another research gap, a good comparative analysis of different vectorization techniques and multiple training of models should be performed to find the optimal trade off. The foundational insights and trends were quite helpful to set a stage as it discusses multiple applications of ML and NLP with the context of clinical data and social media data. To explore further, transition is made to the following significant article, which discusses the applied techniques and methods in the domain of mental health interventions.

The following article “Natural language processing for mental health interventions: a systematic review and research framework” [Malgaroli et al., 2023] presents the review and analysis of quality medical databases articles closely with application of applied NLP. The scholars have reviewed around 102 articles with respect to investigating their use case and research characteristics from NLP preprocessing methods, algorithms, audio features, complete end to end machine learning pipelines and their outcomes in the conjunction of ground truth of clinical practices, clinical samples as well as overall limitations. The study finds out that there is a noticeable increase in the usage of language models from 2019 with the overgrowing social media data. The authors discuss how different kinds of researchers have used different types of vectorization techniques for training all kinds of models from sophisticated and traditional machine learning models to neural networks and deep learning models. Various techniques like bag of words, Word2Vec, TF-IDF, GloVe, word Embeddings have been used with different text preprocessing methods. This research unveils the trends of common algorithms like DT, RF, LR, SVM, CNN, RNN, LSTM that were used by most of the researchers. Article argues that models trained on

the clinical data or health data with manuscripts from sessions are biased towards certain demographics and the datasets are quite small. According to the study, most of the models (83%) are biased towards the English language and majority of the data used by the researchers belongs to America. Specific country based data as well as language causes another bias of differences in context, linguistic complexity and different cultural views and life. This is the limitation of this kind of models as they cannot be generalized easily towards larger populations geographically. Authors stress here that text preprocessing techniques are not comprehensively analyzed for selection with respect to their effectiveness. Authors argue that most of the researchers have not included code, datasets, and other contextual information for reproducing the models for public evaluation. Researchers discuss that NLP and mental health intervention (MHI) frameworks need to be evolved and simultaneously it should be the work of both clinicians, data scientists and geographical linguistic experts to make the models generalizable over larger context and population. Authors highlight that deployment of such models should make a trade-off between performance and interpretation capability with keeping computational resources and clinical ground truths in mind. Like the previous article, authors here also raise the concern and limitation of using bigger datasets and making sure that data privacy, confidentiality of patient's private manuscripts and personal data remains secure and should be obtained legitimately with consent and follow standard operating procedures. Authors emphasize using large datasets with advanced NLP techniques and deep learning models like BERT and GPT for making generalizable models with better context understanding [Malgaroli et al., 2023].

The essential points to be noted here are that, again it is evident that ML and NLP methods can be strongly leveraged for different kinds of tasks and classification towards mental health data. The concerns of using bigger datasets should be addressed, especially that of publicly available data from social media that contains various lingo's and cultural contexts irrespective of geographical boundaries. Usage of advanced models like BERT in this study can be impactful. To maintain the data privacy and confidentiality of social media data, the masking of usernames or deleting the personal information can be the first step towards ethical considerations. If clinical data is used, then necessary standard operating procedures and permissions should be considered and followed. Another strength of this research was that it reviews most of the latest studies and most of them applied classification approaches and that gives more confidence in this study, the classification application can also be implemented successfully. Overall, the outcome of the models should be evaluated with standard metrics and the models should be reproducible for further research. The gap of comprehensive analysis of text preprocessing needs to be covered to select the most effective technique to make the original content retained after removing noise. Building upon themes and discussions from this article on practical applications of advanced NLP techniques and ethical considerations, the following paper that is more focused on social media data based automated systems will be studied.

Exploration continues further to make the understating better with latest studies and application of advanced techniques. This article “Application of NLP and Machine Learning for Mental Health Improvement ” [Borah et al., 2022] discusses creating an automated smart system based on the social media posts. Pipeline commenced with unsupervised data by scraping it from Reddit and Twitter and then labelling them by employing the Text2Emotion library. Researchers have developed a multiclass classification model that detects the stress and places them in the nearest mental health issues. Similarly, based on the previous articles, here also scholars have again used vectorization techniques like bag of words, TF-IDF, Word2Vec and BERT embeddings. This is quite interesting as BERT embeddings can also be used as an input layer to models when computational resources are limited. Authors also applied NLP preprocessing techniques like tokenization, stop word removal, and handling special characters to treat the noise. These are some of the most important methods that have been used in the previous article research too. Similar to other researches, traditional and sophisticated ML algorithms with NLP techniques applied to dataset is also seen here as authors start training their model with SVM, LR, RF followed by some advanced model like LSTM and LSTM trained with BERT embeddings as an input layer of feature.

The BERT embeddings on the LSTM model have outperformed all the models with overall accuracy of 93%. Researchers state that the limitations of these models or the interpretation of these models can be questionable or can be biased due to manual labelling of the data. To test the model’s evaluation, metrics like confusion matrix, F1 score, and accuracy were used followed by the error rate. Authors argue that there is a gap of models error analysis to study the limitations. [Borah et al., 2022]. Important element identified is to leverage the BERT embeddings with different vectorization techniques on the dataset. My critical analysis here, that makes me argue that these models should be trained on both types of data. This is because, it means that training the different models on aggressively cleaned data as well as basic cleaned data will provide different performance due to the models inherent architecture. Especially if BERT Embeddings or end to end fine tuning of BERT is involved, as these models can learn semantic meanings and overall context of the data more accurately if the raw form of data is preserved. Major gap can be considered in misclassification analysis as it is crucial and needs to be addressed. Insights were gained from this research study highlighting the future use of BERT embeddings and different vectorization techniques. Further exploration of research article focused on specific problem of suicide note classification is explored.

Moving forward with the critical literature review on the similar theme, here is another interesting article which specifically touches the crucial topic of suicide note classification in the article “Suicide Note Classification Using Natural Language Processing: A Content Analysis” [Pestian et al., 2010] where researchers conduct that mental health professionals and machine learning algorithms can classify correctly whether suicide notes were genuine or elicited. Around 66 suicide notes data was

taken into consideration for understanding the features and pattern. Various machine learning algorithms and techniques such as decision trees, lazy learners, meta learners, logistic regression as well as NLP techniques like parts of speech, tagging, classification rules, sentiment polarity were leveraged in combination to create models. Researchers found out that 78% of the time the models were more accurate to classify between genuine and elicited suicide notes compared to health professionals which did 63% of the time. Scholars highlight that features to the models were based on the 19 emotional states derived from the literature review and expert reviews. Their best performing model was Logistic Regression Tree model (LRT), that was highly based on structure of the sentence as well as linguistic aspects of the suicide notes. Authors stress that there are limitations to this kind of research with respect to generalizability because of very small dataset and context. The authors also argue that there is a possibility to use these kinds of sophisticated models in therapy and classification, but the results should be interpreted by health domain experts. Further, quantitative and qualitative grounds should be laid by understanding the models decision making process which is missing in this research. This research [Pestian et al., 2010] makes me question that models are highly overfitting, and the context window is too small. Also, I believe that the models could be biased towards certain personalities due to the dataset limitation. The strength of this article is that it shows a good direction by demonstrating the promising research perspective to develop models on specific kinds of clinical dataset for personality analysis, which can be useful for clinical as well as forensic sciences. Also, authors have stressed that ethical considerations like data privacy and permissions are required to work with this kind of data and their research was granted permission from various governmental organizations in the USA. ML and NLP can also be applied to small datasets but with caution to overfitting. Humans usually focus on the content and machine learning models focus on the overall structure of the data and of course it's clear from the above discussed research article, that need of a domain expert to interpret the results and finalize the outcome is crucial as, it's a matter of sensitive topic or else with the wrong classification or ambiguous results can develop confusion and can manipulate patients thinking. To build upon that, there is a huge requirement to address the gap of the research by conducting rigorous interpretable analysis providing insights into the models features and their decision making process. In this thesis, this concern of interpretable models will be addressed by looking in the model's decision making process to promote transparency. This research article is considered important in regards to this thesis as medium to bigger sized datasets will be focused and aimed to build generalizable models.

Previous research articles have provided understanding on and around this thesis topic, followed by the exploration of the final scholarly article of this literature review, which closely aligns to this thesis. As this thesis is based on the social media data and leveraging advanced ML and NLP techniques, this research article "Machine Learning Driven Mental Stress Detection on Reddit Posts Using Natural Language Processing" [Inamdar et al., 2023] is quite recent and showcases the usage of

social media data as well as advanced ML and NLP techniques. The researchers use the Dreddit dataset which consists of the social media posts and comments. Research scholars employ supervised learning for the task of a classification usecase. Authors use preprocessing techniques such as basic data cleaning of noise, tokenization, stop word removal, stemming. The RAKE algorithm is employed here for keyword extraction. Advanced vectorization techniques are employed for feature engineering such as Bag of words, ELMo and BERT embeddings. Generated feature vectors through vectorization are used as an input layer as feature representation on ML models such as LR, SVM and XGBoost. The result showed around 76% accuracy on their SVM model where ELMo and BERT embeddings were used. Authors argue that the generated results were not that good as expected due to the limitation of small dataset and cannot be generalized as the collection of data is not possible that easily. Authors suggest using state-of-the-art technologies for sentiment analysis and for mental health care applications by leveraging advanced deep learning models, domain specific models and vectorization techniques such as Word2Vec, GloVe and suggest using different and big datasets for multimodal capabilities exploration. Researchers highlight another crucial gap in generalizability of models as the research did not conduct additional external validation on similar to different real world datasets. This research article [Inamdar et al., 2023] made a quite interesting impact to understand the pipeline of applying advanced ML and NLP methods on social media datasets for classification. This research insight is closely related to this thesis, and it makes me criticize certain points here such as, dataset used were too small. Ethical considerations regarding data privacy or anonymity were not quite described in detail or highlighted. I also argue that the model's performance could be improved by applying optimization through optimizers or by hyperparameter tuning. Another important gap will be addressed in this research by leveraging state of the art DL models and domain specific pretrained models. Though the research paper has shown great strength in displaying end to end modelling pipeline and evaluation indexes. This research article has also contributed to overall understanding and providing the foundational knowledge to this thesis.

By reviewing multiple scholarly works, an intersection of interdisciplinary domains of data science and psychology is quite clear, thus advanced data science techniques can be applied to mitigate the problems and can be useful as an aiding tools. Exploring these research articles shed light on different perspectives, usage of different data sources like clinical and non-clinical data, different methodologies ranging from sentiment analysis to complex modelling. Most of the research works used sophisticated ML algorithms and NLP techniques for handling text data. It was seen that most of the research works had limitations with the used dataset as they were either small or manually labelled or they had issues of ethical considerations with respect to data privacy and patient's confidentiality. This created a biasness and hindrances in making the models generalizable. It was clear from these research works that models should not just be technically advanced, but they should also excel at understanding the overall context and semantic meaning of the text to become more

generalizable.

Overall, the literature review provided a foundational knowledge for this thesis. After studying the limitations and weaknesses of these research works, the key takeaways are to use at least medium sized dataset and to leverage the use of pre-trained transformer based deep learning models for our research to address the major gaps. To address another theoretical biased selection gap, multiple models with different text preprocessing and feature engineering techniques of NLP vectorization should be employed and then these need to be compared to find the most effective technique to conclude on the dataset at hand. To address the gap of ethical considerations and maintain the data privacy the dataset will be masked if sensitive information is there to prevent the confidentiality loss. Other ethical considerations regarding the dataset availability, source, usability and scrapping information will be studied. To address the gap of poor to mediocre performing models, hyperparameter tuning will be done in this thesis to find the best set of optimal parameters that improves performance of the model without underfitting and overfitting. To maintain the models integrity and validity, interpretability analysis, misclassification analysis and external dataset validation will be performed to address the major gaps of the current literature work around and on this topic. On top of that to provide additional context which is missing from the current literature, additional two pretrained models will be employed in the combined analysis engine to interpret the sentiment and emotional state of the instance too. This can provide deeper understanding and reasoning to the trained models predictions. Overall, the pipeline is clear from literature review, and it further demands the deeper exploration of advanced techniques with different datasets and methodologies as well as comprehensive analysis of results. With that said, the foundational knowledge base has been laid and this thesis will move ahead for critical exploration addressing the major and crucial gaps from the current research.

2.2 Research Aim and Questions

This section discusses the research aim and respective research questions setting a path for this thesis. As the constant advancements are ongoing digital transformation and the evolving nature of interaction with social media have changed, it has various impacts on human psychology and mental health. This thesis will guide through the investigation of the research aim followed by the research questions that will highlight the importance of linguistic characteristics associated with mental distress.

Research Aim: The primary objective of this thesis is to explore how Machine Learning and Natural Language Processing methods can detect and analyse mental distress indicators effectively from online textual discussions and discourses. How

the state-of-the-art techniques and concepts can be explored and leveraged to study their effectiveness for text classification. This thesis results will bridge the gap between the latest advancement in ML and NLP technologies and their practical applications in aiding mental health wellbeing, also addressing ethical and practical challenges.

Research Questions: The following research questions will be investigated to address the overarching research aim. This research study follows more of an inductive research approach.

RQ1: Which are the common linguistic patterns and features that are associated with mental distress from the textual corpus of data?

Here, the use of a language with phrases, words and context will be studied with respect to the emotional tone for understanding the overall semantic meaning. This can be achieved by conducting exploratory data analysis with visualizations. This is an important step before modelling to make the algorithms capable of identifying this kind of patterns and indicators. Also, the post analysis of models using interpretability analysis and misclassification analysis will provide the insights into the model's features, influential token weights, patterns and trends directly associated with mental distress indicators. This is important to make the algorithms capable of identifying patterns and indicators as well as to validate their predictions.

RQ2: How various machine learning models can identify mental distress indicators and what metrics can be used to compare a model's accuracy?

Here, different models will be built from traditional and sophisticated ML models such as Decision Tree (DT), Random Forest (RF), Support Vector Machine (SVM), Logistic Regression (LR) followed by advanced neural network models like Convolutional Neural Network (CNN), Recurrent Neural Network (RNN) and Long Short-Term Memory (LSTM). Transform based state-of-the-art deep learning models like Bidirectional Encoder Representations from Transformers (BERT), DistilBERT, MentalRoBERTa and MentalBERT will be also end-to-end fine-tuned on the dataset. All these trained models will undergo a systematic comparative analysis and will be investigated for their effectiveness in identifying mental distress indicators. The evaluation metrics used on these models will be confusion matrix, F1 score, Recall, Accuracy, and precision followed by classification report and AUC-ROC curve. This is important as it will showcase the best performing model as well as the trade-off between model's accuracy and computational resource requirements.

RQ3: How the performance of the model is influenced by the different choice of text preprocessing techniques, feature extraction and hyper parameters?

Here, different kinds of NLP techniques such as tokenization, stop word removal, lemmatization with different combinations will be employed and compared to see the best effective techniques to maintain the overall semantic context in the dataset

at hand. Feature extraction such as n-grams, sentiment analysis and word embeddings will be employed, and the trained models will be optimized by hyperparameter tuning. This is important to make the entire pipeline efficient and accurate.

RQ4: How is the model's efficacy in recognizing mental distress indicators affected by different vectorization techniques such as BERT embeddings, TF-IDF, GloVe and Word2Vec as well as their comparison?

Here, different vectorization techniques will be used as feature engineering on pre-processed dataset to feed the models. Techniques such as TF-IDF, Word2Vec, GloVe and BERT embeddings will be employed. All the ML and NN models will be trained on each of these vectorization techniques and then undergo comparative analysis to find out how the semantic meaning and context are captured for the text classification.

RQ5: What are the possible options to validate and evaluate the model and how it can be used for future work in implication for automated support systems on social media?

The models while training will be validated with validation techniques such as hold out method and k-stratified cross fold. This will make sure that the model is trained equally on different labels and does not overfit or underfit during the process. With performance and computational resources trade off, usage of models in mental health intervention systems and digital support systems on social media for future scope and use will be discussed. External validation with unseen datasets will also be implied for generalizability and adaptability in a broader context.

RQ6: What are the ethical considerations with respect to privacy and security when analysing and interpreting the results of the model?

Here, ethical considerations with respect to data processing, analysing, and interpreting model's results will be discussed. Developed models applications, implications, biases, challenges and potential solutions will be discussed. The guidelines and measures to treat the sensitive data with ethical compliance in this thesis as well as its practical applications as predictive analysis are discussed. Future research and AI applications systems in the mental health domain will be dealt with too.

After setting the guiding principles and objectives of this thesis through thorough recent literature review on and around the topic and outlining the research aim and questions to minimize and attend the gaps, now the next chapter will be theoretical background. This chapter will discuss the techniques and technologies employed in this thesis. It aims to provide the reader a solid understanding of the technical nuances and techniques that this thesis will use for implementing its aim and objectives.

3 Theoretical Background

This chapter includes the theoretical background that acts as a preliminary knowledge base explaining the methods, techniques, algorithms and other important terminologies related directly to this thesis study to provide the technical context of this thesis.

3.1 Machine Learning (ML)

It belongs to a cloud of artificial intelligence, where the machine learns trends and patterns from the data to interpret the similar patterns and nuances from the similar structure of unseen data. It is different from traditional programming and here ML involves the algorithms which can be used for prediction and decision making based on data. Different types of learning include supervised learning, unsupervised learning, and reinforcement learning. This thesis study utilizes supervised learning techniques where the machine learning models are trained on the labelled dataset, so that they can predict the outcomes based on the historical data [Burkart and Huber, 2021]. Further background on these models is provided in Subsection 3.9.

3.2 Natural Language Processing (NLP)

It is another branch of the Artificial Intelligence field that focuses on the interaction between the machine and human language. It is used specifically to deal with large amounts of textual corpus and data for understanding and analysing. Some key techniques that NLP uses are [Khan et al., 2020]:

Tokenization: Breaking sentences into tokens of words and phrases.

Stopword removal: Removal of common words that does not add semantic meanings.

Stemming and Lemmatization: Words are reduced to their base and root form.

Part of Speech Tagging: Identifying the grammatical parts of speech.

Named Entity Recognition: To identify and classify some of the important key figures and information like names, dates, places.

These techniques are employed in this thesis study for making the entire text processing and analysing efficient. This eventually helps to focus on the major and important aspects by eliminating the noise and redundant data to identify the mental distress indicators. NLP helps to reveal the trends, patterns, and meaningful insights from the textual dataset that helps us to understand the mental distress indicators.

3.3 Exploratory Data Analysis (EDA)

Exploratory data analysis is a crucial step before moving ahead with text preprocessing and vectorization. It gives the insights in the dataset directly, such as the structure of the dataset, important variables and fields, rich features, anomalies, outliers, and overview of the entire dataset. This critical analysis is performed here which includes statistical analysis followed by visualizations (graphs, plots, histograms etc.). In this thesis EDA is conducted to explore the dataset closely, finding insights and identifying the initial glimpse of the dataset at hand. It will guide the exploration, strategy building for analysis and modelling, identifying patterns, trends, correlations and underlying insights that are directly related to these thesis research questions. Conducting and interpreting the EDA as the first step informs this thesis and sets a guiding principle for further exploration into advanced analysis and modelling. Techniques that will be employed are the following:

Data Inspection: Checking the number of columns, rows, their data types, and first few entries to see the type of content and overall structure.

Missing value analysis: Checking for the missing or the null values in the entire dataset to see the quality and the completeness of the data.

Class distribution analysis: To understand whether the dataset has balanced or imbalanced class label distribution.

Text length distribution analysis: To understand the variability of the textual data entries in the dataset and visualize it.

Word cloud generation: To study the themes and patterns, most frequent words in the dataset are retrieved and then highlighted in the cloud.

Feature engineering for text: To understand the text-based features by looking at average word counts and word length in the dataset.

N-gram analysis: To identify the most common phrases, patterns and trends, uni-grams, bigrams, and trigrams are extracted from the dataset and then visualized.

Topic modelling: Again, to identify different themes and topics in the dataset.

Sentiment analysis: To study the emotional tone of the textual entries.

Statistical analysis: Performing correlation analysis and T tests, P tests to see the label distribution with sentiment tone analysis for identification of texts and label's relationship.

Outlier detection: To lookout for anomalies and redundant data in the dataset.

By conducting EDA and performing these steps will reveal meaningful insights, trends, patterns, and an overall understanding of the textual dataset characteristics at hand. This will eventually highlight the underlying mental distress indicators and the features that will be helpful for training the advanced models for distress detection.

3.4 ML and NLP Pipeline

This thesis employs the ML and NLP pipeline framework that includes multiple stages and sets of actions. It includes all the steps from data collection to the classification task that passes through stages such as data cleaning, feature engineering (text preprocessing, vectorization), hyperparameter tuning, modelling with validation, optimization, evaluation, and interpretation analysis [Kunft et al., 2019].

3.5 Data Cleaning and Preprocessing

Another important step in the ML and NLP pipeline is the cleaning of the dataset, that includes removing or treating the null and empty strings or values. Removing the html tags and special characters, masking the name, standardizing the text format and other basic cleaning techniques are employed in this thesis. This results in an efficient and accurate dataset to ingest in the pipeline. Also, the text preprocessing techniques mentioned in the above NLP paragraph will be applied to the data. This eventually increases the reliability, validity, noise free data, and emphasizes on the important data space in the entire pipeline for data analysis and exploration. This step sets a foundation for the upcoming actions to be performed on the cleaned and pre-processed dataset [Chapman et al., 2020].

3.6 Vectorization

Vectorization is a technique that is used to convert the textual data to the numerical format data, so that it can become useful for machine learning models. This is a very

important method as machine learning models work better with numerical data as an input [Yang et al., 2022]. The vectorization techniques that will be employed in this thesis are as follows:

TF-IDF (Term frequency-Inverse document frequency) - It is the technique that represents the importance of a word in a document of textual corpus by balancing the word frequency in the particular document towards the other document frequency [Bounabi et al., 2019].

Word2Vec (word to vector) - This technique is advanced than TF-IDF but has a larger feature vector space and is generally used with neural networks to learn associations of words with each other in the larger textual corpus. Here, semantic relationships are captured between words by the representation of words in a continuous vector space done by a group of models [Johnson and Karthik, 2021].

GloVe (Global Vectors for Word Representation) - Just like Word2Vec, GloVe incorporates the combination of the two NLP techniques like local context window methods and matrix factorization. This enhances the word vectorization technique and efficiently represents the words co-occurrences over a textual corpus. [Huong et al., 2022]

BERT (Bidirectional Encoder Representations from Transformers) Embeddings - It's a state-of-the-art methodology and recent development in the domain of NLP. BERT will consider all the occurrences of the word with the context in the textual corpus at hand. This increases the understanding of the use of words in different cases and scenarios [Zhu et al., 2020].

Text preprocessing followed by vectorization in our pipeline is crucial, as it performs feature engineering. By doing this, this thesis can utilize the entire potential of the dataset by deeper understanding of context and semantic meanings. Also, this thesis focuses on the mental health domain, so it's quite obvious that context and semantic meaning are of the utmost importance. The above stated methods have their own advantages and disadvantages. To find the best method for this thesis pipeline, different vectorization techniques will be implemented on ML and NN models with the same dataset for thorough comparison.

3.7 Text Classification

Classification is a fundamental ML task. NLP deals with textual classification. Here the main goal is to assign the predefined labels to the text. Here the classification models are trained to predict the text instance to its respective label. This NLP task is employed in this thesis and is the core of text analytical models. It plays an important role in this thesis, because employing the text classification as the final layer

of analytical models, categorisation of the textual entries between mental distress or not distressed will be detected. [Dogra et al., 2022].

3.8 Hyperparameter Tuning

To initiate a model with the first set of parameters and later improvise the parameters for optimization is called hyperparameter tuning in ML and NLP. It can be conducted in multiple methods such as grid search or random search. Employing them can test a certain range of parameters initially, train the model and again adjust the parameters for optimal performance and resources trade-off. Parameters like number of trees, sequence length, learning rate, regularization variable, number of hidden layers, number of neurons, penalty weight, feature dimensions and more can be tuned according to their suitability with the given dataset at hand. These sets of parameters are important to be tuned because they are directly associated and affect the model's learning and generalization ability. In grid search, the set of parameters are predefined and evaluated and in the random search parameters are selected based on the given range or distribution. These methods eventually sample the best possible parameters and make the model effective for practical applications [Probst et al., 2019].

3.9 Modelling

It's a process of creating mathematical models that are capable of making predictions. These models learn features from the seen data patterns, trends and historical data and use it for prediction on the unseen data of similar structures. From the branch of Artificial Intelligence, Machine Learning models include different modelling techniques from linear regression models to complex statistical models to tackle different types of problems and different structure of data. This thesis specifically deals with the text classification task and sophisticated ML models like Decision Trees, Random Forest, SVM, Logistic Regression are implemented here. Due to the binary label's nature of data, these sets of algorithms will set the foundational stage for baseline modelling [Kang et al., 2020].

Decision Trees: Decision Trees algorithm is a traditional machine learning algorithm. It works like humans in terms of decision making and has a tree-like structure representing decisions and their outcomes. There are multiple nodes and branches in the tree, each node represents a decision point followed by the branches showcasing the outcomes. This helps in easier interpretation of predictability as it increases transparency by demonstrating the reasons behind each decision to support the prediction [Handley et al., 2014].

Random Forest: It is based on the concept of the decision trees itself, but improvising on it by creating an ensemble of trees. During the training, multiple decision trees are trained and produce output of classification or regression as desired for each and every individual tree. RFs are less susceptible to overfitting than decision trees because variance is reduced by the aggregation of multiple tree's prediction results. Also, the accuracy and the robustness of the random forest is more often higher than DT [Cacheda et al., 2019].

SVM: When it comes to the high dimensional vector spaces, SVM is a powerful model. SVM is versatile and also has a higher efficacy for the huge feature vector space. It works very well with the use case of classification. The working of SVM includes the identification of the optimal hyperplane. This hyperplane clearly separates the different labels or classes from the feature space. SVM uses different kernels for solving the nonlinear problems and challenges. To make the data points separable, SVM transforms the original feature space into the higher dimension [Song and Diederich, 2013].

Logistic Regression: To deal with binary classification tasks, logistic regression is another powerful algorithm applying a linear approach for classifying. Even though the name of the model suggests regression, it predicts categorical outcomes between labels. Logistic regression is simple but quite effective with data space that is linearly separable. It predicts the probability of the input to either of the classes. It's a robust classification and regression algorithm [Alishiri et al., 2008].

Neural networks are more complex adaptations of sophisticated machine learning algorithms with human brain-like structure and functionality. Here there are multiple layers, and they are interconnected with each other with neurons. This shows their network of interconnected nodes characterized in multiple layers. In neural network models, all the layers involved have certain responsibilities like transferring the input from layer to another with the abstract composition of the data that makes these models learn patterns, context, trends from the data that is used for predictions or making decisions on the new data based on the historic data learning [Gardner, 1988]. There are multiple neural network models, but in this thesis the following models are employed due to their capabilities and applications related to the current dataset at hand.

Convolutional Neural Network (CNN): This is the commonly used model in the domains of image processing as well as computer vision. CNNs are a special type of model that are built to process the topology-like structure that are usually found in images. Though it was designed for processing images, it has become popular to applications and can also be employed with NLP tasks like text classification. To learn patterns, trends and relationships from the textual data, CNN applies 1 dimension convolution to the textual data and considers the words as special spatial features [Glick and Applbaum, 2010].

Recurrent Neural Networks (RNN): RNN models are capable of handling the sequential data. RNN includes data in the range or format of time series, audio or the textual corpus data. The unique characteristic of RNN model is to have an ability to retain the memory of the previous input. This is achieved by passing the hidden state from the current one step in the sequence to the next step in the sequence. This action is very important when context and the order of certain data in a format have a high feature value. RNNs can be used for language modelling, sequence prediction as well as textual data analysis [Bouarara, 2021].

Long Short-Term Memory Networks (LSTM): In conjunction with RNN, LSTM algorithm is an extended method to mitigate the RNN challenge of long term dependencies. Vanishing gradient was the issue with traditional RNN, as they had difficulty in maintaining the information that had a huge sequence. Here LSTM adapts an additional layer mechanism where there are input, output and forget gates. These gates are designed as filters to regulate the huge sequence for longer periods of information to flow. In this way important data features are preserved then irregular redundant data. This capability makes the LSTM robust when dealing with the extended contextual and semantic informational use cases like speech recognition, language translation, summarization, and sequential text decision making [Singh et al., 2022].

In recent times in the field of NLP, transformer based deep learning models are considered to be the ground breaking development. It handles and performs best with the sequential and textual data mitigating the challenges of previous traditional models. These types of models use the novel approach called the self-attention mechanism. Here the different parts of the textual input data are weighted differently. Due to that, these models can do the parallel processing simultaneously and can handle the long-term dependencies in the textual data efficiently. These models are used for attending the tasks like question answering, sentiment analysis, text classification, summarization, and language translation. Constantly these models are being updated to push the borders in the field of AI for enhancing more nuanced and better language understanding with context and use cases [LeCun et al., 2015]. These state-of-the-art technologies in NLP are getting better and better in understanding the natural human language, can find the patterns, trends in large corpora of the data and provides advanced textual analysis that was previously inaccessible.

BERT (Bidirectional Encoder Representations from Transformers): To understand the words and their contextual relationship in sentence with different use cases and scenarios, BERT models are very effective for modelling as they are state-of-the-art technology in NLP which utilizes bidirectional training approaches. BERT models have two main versions and many different variants that are specifically trained on certain kinds of domain data. BERT base versions have 110 million parameters whereas BERT large has 340 million parameters. BERT base has 12 layers and large has 24 layers. Both have different hidden units like 768 and 1024 units respectively. These models are pre-trained and during pre-training there are certain

methods applied to it like masked language modelling as well as next sentence prediction. These models are very effective with contextual, semantic and language structure understanding. Due to nuanced language understanding, it comes with the cost of high computational resources required to even use pre trained models for end-to-end fine tuning [Devlin et al., 2018].

RoBERTa (Robustly Optimized BERT Approach): This pretrained model is based on the BERT itself. But the training approach of this model is different from the BERT and it leads to improvement in performance on NLP tasks and sets different benchmarks. It includes methods such as dynamic learning rate, mini batches are larger and have more sequence length and it also alters the training variables. Like BERT, it does not perform the pre-training of the next sentence objective. The dataset is more vast here for training and it improves on understanding the language structures and overall context. It outperforms BERT in different tasks due to the different training methods. The computational resources to perform downstream tasks on such a complex architecture model is resource intensive and computationally demanding [Liu et al., 2019].

DistilBERT: As it is seen that these models are computationally and resourcefully intensive and demanding. To deal with that, this model was developed by keeping efficiency as the main motive. The best part of this model is that it successfully retains the BERT's performance by 95% and by keeping the size of the model small by 40%. This model has 66 million parameters. It involves a process called distillation in which during training time it is made to learn from the main BERT model which is full-fledged. Due to this approach, distilBert can maintain all the capabilities of the main model followed by the lower usage of resources and computational power. Certain methods that are employed on it while training are cosine distance loss, language modelling and distillation. It is called a triple loss combination that helps this model to be robust, efficient and maintain the maximum capabilities of the main BERT model making it an overall compact model [Sanh et al., 2019].

Mental BERT: As this thesis study is in the domain of mental health, so this Mental BERT is a specialised model for the applications of mental health domain. This model is closely trained with the datasets that associate the topics of mental health and related context including clinical data, non-clinical data, social media posts and books. Though the parameters and overall architecture is similar to the original BERT model, it is fine-tuned and trained closely with the mental data for in depth nuanced and complex understanding of language related to the indicators of mental health [Ji et al., 2022].

3.10 Fine Tuning in NLP

In the domain of NLP, fine tuning is considered as a technique where a pre-trained model is used for an end-to-end pipeline tuning task. It is called a downstream task and in this thesis, BERT models are used as a pre-trained model and the downstream task here will be the text classification task. This is highly effective on specific tasks like sentiment analysis or language translation because these models are already trained on specific kinds of domain as well as general datasets. To use the pre-trained knowledge, it can be fine-tuned with the dataset at hand. This makes the entire pipeline efficient and has a richer feature base, contextual and semantic understanding. It is an entire pipeline process and the key to achieve maximum performance on these kinds of models is setting the optimal set of parameters that comply with the dataset at hand. This is a leap in the domain of NLP where pre-trained models can be adapted with the dataset at hand and customized as per the desired downstream tasks and requirements [Howard et al., 2018].

3.11 Model Validation and Evaluation Techniques

In the entire pipeline of Machine learning and NLP, the quality of the models should be validated and evaluated because these models should be robust as well as generalizable to the unseen data. So, for validating these models K-Fold cross validation and Hold-out methods are applied due to their effectiveness. In the K-Fold cross validation method, the data is divided into the subsets of the chosen number of K. One of them will be used for testing and the other will be used for training. This method makes sure that the unseen data is performing well on all the kinds of data samples. The hold-out method will split the data into training set, validation set and testing set. It is faster than the K-fold method but not equally robust like it. This is because the K-fold method is usually applied to sophisticated models and deep learning models already have a high layer and performing k-fold on it will make it exhaustive. So, here hold-out methods followed by other metrics are employed for validation of models [Yadav and Shukla, 2016].

Models performance is evaluated by employing different metrics on them to judge them in different ways and perspectives, Evaluation of the models enhances confidence in the practical implications, shows reliability and accuracy of the outcomes [Zhou et al., 2021]. This thesis study focuses on the below evaluation metrics:

Confusion Matrix: It is represented or visualized in table form to study the performance of the classification task executed by the model. It displays the actual labels against the predicted tables. To see the detailed performance of a model, a confusion matrix is used and it further expands to provide insight into the model's outcomes and type of errors associated with it. Confusion matrix comprises of: True Positives

(TP), True Negatives (TN), False Positives (FP) and False Negatives (FN). These metrics are built upon the results of the model's classification prediction vs the actual labels [Deng et al., 2016].

	Predicted Positive	Predicted Negative
Actual Positive	TP	FN
Actual Negative	FP	TN

Table 3.1: Confusion Matrix Structure

Precision: This metrics gives insights into the model's prediction of correctly predicted positive instances towards the total predicted positive instances. This metric is important to study and crucial when the false positives are high. Formula: [Hossin et al., 2015]

$$Precision = \frac{TP}{TP + FP}$$

Here: TP (True Positives) is the count of correct positive predictions. FP (False Positives) is the count of incorrect positive predictions.

Recall: To study the ratio between the correctly predicted instances towards all the actual positive instances. This is another important metric, especially when we are calculating as many positive instances as possible, irrespective of the false positive instances increasing. Formula: [Hossin et al., 2015]

$$Recall = \frac{TP}{TP + FN}$$

Here: TP (True Positives) is the count of correct positive predictions. FN (False Negatives) is the count of incorrect negative predictions.

F1 Score: To study the models weighted average of both precision as well as recall. This metric is also important as it provides overview of both and specially when the dataset's classes are variably distributed. [Hossin et al., 2015] Formula:

$$F1 \text{ Score} = \frac{2 \cdot Precision \cdot Recall}{Precision + Recall}$$

AUC-ROC Curve: This metric provides the insights into the ability of the model in differentiating class labels. So, a good and optimal model should have the AUC-ROC Curve near 1. This demonstrates that it can easily distinguish between classes and is not biased towards an imbalance class [Hossin et al., 2015].

Accuracy: To study the model's overall performance. Accuracy gives the insights of how the ratio of predicted observations towards the total number of observations. But relying completely on accuracy can be misleading due to class imbalance or

different issues. So, other metrics discussed above should also be taken into consideration besides accuracy. [Hossin et al., 2015] Formula:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN}$$

Here: TP (True Positives) is the count of correct positive predictions. TN (True Negatives) is the count of correct negative predictions. FP (False Positives) is the count of incorrect positive predictions. FN (False Negatives) is the count of incorrect negative predictions.

Training and Validation Loss: To study the model's learning capability and its generalization capability. Over the epochs training loss decrease is expected and validation loss should also be decreased over epochs. If initially it decreases and then in the next epoch it increases, then it is a sign that the model is starting to become over-fit [Eelbode et al., 2021].

Validation during the training and evaluation after training the model provides the overall classification report to judge the model's performance from all the aspects. This eventually gives the insights and feedback for fine tuning, optimizing, and selecting the best possible models with optimal trade-offs. Analysing model's performance are the important steps to make the prediction reliable, robust, and generalizable to the unseen data. Hence a strong confident foundation is formed for these models to be used for the future research and deployment.

4 Methodology

This chapter discusses the designed methodology to explore the efficacy of different traditional machine learning models, advanced neural networks, and the state-of-the-art deep learning models with different configurations of preprocessing techniques followed by vectorization techniques in identifying mental distress indicators from the online discourses. This systematic crafted reliable and valid methodology is designed to attend and address the research aim and its following research questions. This designed methodology is robust and ensures this thesis's credibility, practicality, reproducibility, and real-world applicability by producing insights and findings in-turn viable for the future research and exploration.

This methodology provides details to the data collection, exploratory data analysis, data cleaning and preprocessing, feature engineering, model selection, hyperparameter tuning, optimization, evaluation and validation, interpretability analysis and generalizability of the models. These comprehensive details showcases the academic rigor of this thesis followed by strong technicality demonstrating the way of conducting responsible research with regards to the sensitive domain of the mental health analysis. By the implication and the use of cutting-edge data science technology and best practices, this designed robust methodology covers each and every research question responsibly overarching the main research aim that was identified in the outset of this thesis. The results, insights from the analysis and overall findings after applying the methodology creates a clearer path towards the future advancements and makes a contribution for providing the mental health analysis support effectively on the digital platforms and spaces.

4.1 Research Design and Approach

This thesis implies a mixed set of methods including both the experimental as well as correlational designs to see how the context is related, leveraging the strength of quantitative and qualitative research. This multifaceted way of approach and perspective encourages and facilitates deep exploration into the mental health analysis for the identification of mental distress indicators by leveraging ML and NLP.

Research paradigm: Here the integration of quantitative analysis is done with qualitative analysis, implementing multi stage investigative analysis [Mayring et al., 2001].

The quantitative analysis covers the aspect of experimenting with models with different configurations, tests and systematic comparative analysis of around 32 models to quantify the model's efficacy and effectiveness. Parallely, qualitative analysis covers the aspect of the correlational designs for investigating the linguistic features, trends and patterns associated with the nuanced context of mental distress indicators captured, identified, and classified by these developed models.

Research Design for experimental component: It covers the training of different ML, NN and DL models with different sets of configurations based on preprocessing techniques, feature engineering techniques (vectorization), base and domain specific models. Conducting exploratory data analysis, evaluation, validation, misclassification analysis, interpretability analysis and generalizability analysis. This aspect reveals, how effective the models are in identifying distress indicators.

Research Design for Correlational Component: The results, insights and evaluation of these techniques and models reveals the contextual linguistic relationship between features and mental distress indicators. This helps to understand the correlational component where the model's decision making ability is highlighted to show the language and distress indicators relationship. Hence providing qualitative analytical insights into the model's decision making process transparently.

Rationale for the Chosen Approach: The mixed-methods strategy, combining experimental and correlational designs, is selected for its ability to offer both broad and deep insights into the detection of mental distress from text. This approach ensures that, this not only identify the most effective models and techniques quantitatively but also understand the qualitative reasons behind their success, addressing the thesis research questions comprehensively [Mayring et al., 2001].

This kind of mixed approach strategy of merging both experimental as well as correlational designs is chosen for its capability to provide overall insights in the detection of mental distress indicators through the textual analysis. This strategic approach makes sure that identification of the most effective preprocessing and vectorization techniques followed by best models and model categories are done quantitatively followed by the supporting qualitative reasoning behind it. This designed methodology ensures the systematic and reliable research in the domain of mental health analysis with high academic and technical rigor, bridging the gap between the data science applications and mental health analysis.

4.2 Environment Stack

In this thesis, the configuration of the computational environment is most important to support the intensive processing demands of NLP and DL algorithms. Below, is the description of the hardware setup and software tools utilized throughout the

thesis. To execute the methodology and to implement the high intensive processing requirements of the ML and NLP based techniques and models, robust computational resources are required. Below are the details of the hardware and software setup used for this thesis.

Hardware Setup: The computational infrastructure of the methodology is backed by the high performing AMD Ryzen 6900HX processor providing distributed computing with multi core architecture for the requirements of high computational speed for data processing as well as model training. As advance neural network models and state of the art deep learning models are involved, there is high requirement of GPU and the used GPU in this thesis is NVIDIA GeForce RTX 3070 TI 8GB VRAM with 150W of power capacity ensuring high parallel processing and acceleration in operational computation for NN and DL models. The infrastructure also includes the 16 GB of ram with 4800 MHz frequency ensuring the rapid handling of big data sets and its processing. In combination with these specifications, a 1 TB NVME SSD is also used here for the fast data transfer rates for input and outputs as well as storing huge model architectures for inference [ROG, 2022]. Overall, this infrastructure provides a solid hardware configuration for supporting the implementation and execution of complex ML and NLP operations.

Software and Tools: Anaconda distribution is used here as the primary development interface with its superior capability of easy package management and deployment [Anaconda, 2023]. Inside the same ecosystem, Jupyter notebook interface, an open source-based web-based notebook style IDE is used for complete end-to-end coding for building model's pipeline, visualizations and [Jupyter Project, 2023] debugging. To use the parallel computing features and capabilities of [Nvidia, 2023] GPU, Cuda drivers need to be configured to enable the direct programming on the GPU architecture. This makes modelling tasks resource efficient and reduces training and evaluation time of the models. Google Chrome performs a role of the medium for accessing the Jupyter notebook server interface for the codebase setup and version control. Jupyter notebooks' inherent features makes execution of ML and NLP workflows efficient.

Libraries and Frameworks: Python programming language serves as the primary language of the entire codebase [Python, 2023]. Multiple libraries and frameworks are used here for the end-to-end pipeline of building models. Frameworks like PyTorch are used for end-to-end building of neural networks and deep learning models [PyTorch, 2023]. Scikit-learn is used for building machine learning models. NumPy, Pandas, Transformers, Tqdm, NLTK, Spacy, seaborn, matplotlib, safetensors, Autokenizers and other libraries as well as frameworks have been used for data handling as well as model development tasks [HuggingFace, 2023].

With the implication of prominent technical competence followed by the compatible hardware and software configuration makes the designed methodology execution possible to the best of its capability and technical limitations.

4.3 Systematic Workflow of the Pipeline

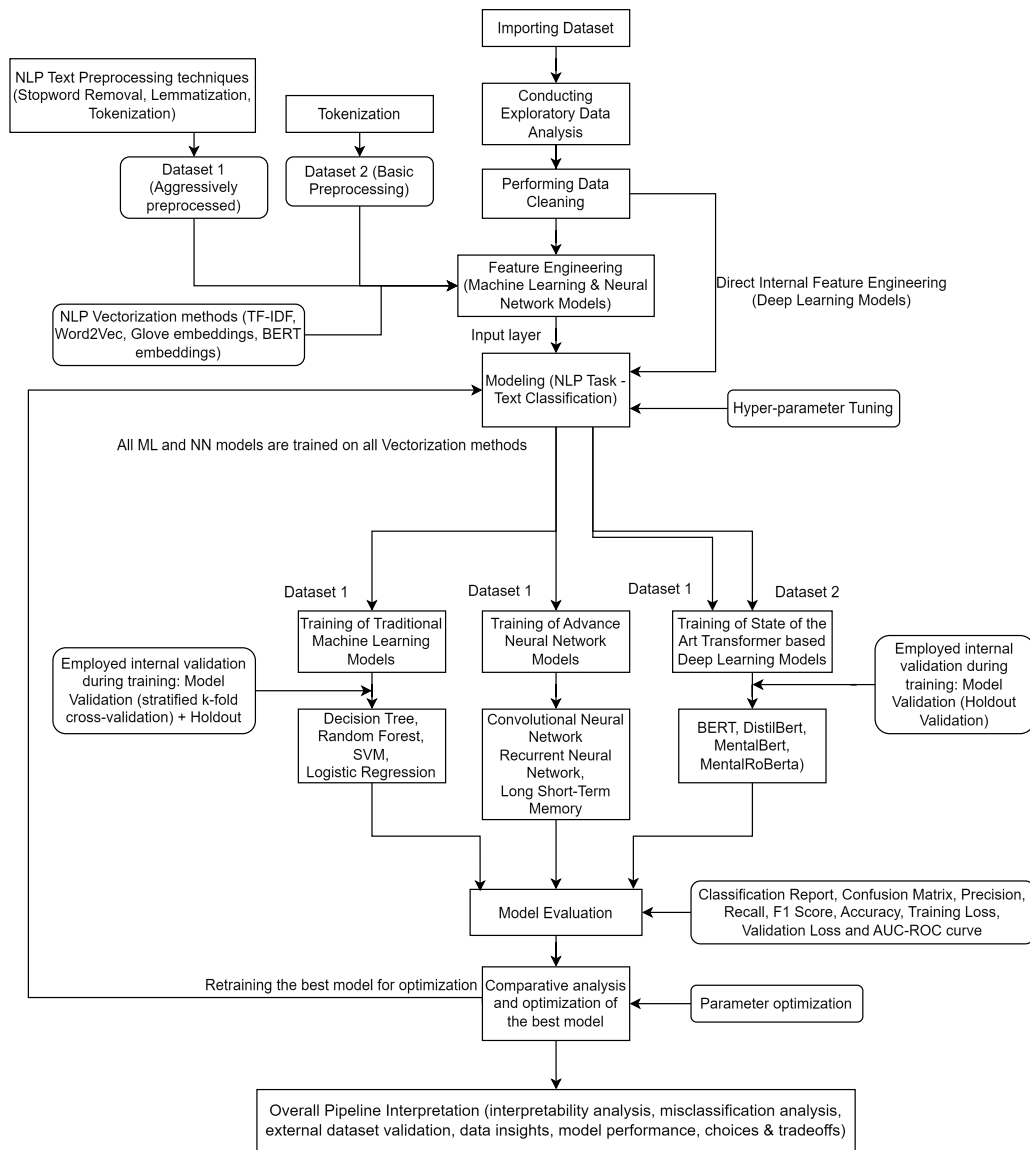


Figure 4.1: Flow Diagram of Pipeline Execution

The above Figure 4.1, represents the flow diagram of pipeline execution that demonstrates the detailed step by step progression of the entire ML and NLP pipeline. This figure provides an overview of the methodological approach from the initial step of dataset preparation till the analytical output from the models.

The first step in the workflow is to acquire a dataset and to explore the dataset by conducting rigorous exploratory data analysis to understand the data at hand

and its detailed characteristics. After performing different analytical functions in EDA, the dataset is understood, and the next stage is initiated focusing on the data cleaning. Here the dataset is cleaned with specific identified noise followed by the standard cleaning and preparation of the dataset for the next stage of data preprocessing. In the stage of data preprocessing, the data is being prepared for feature engineering. To achieve this kind of data preparation, the dataset is pre-processed in two different ways. Dataset 1 is aggressively pre-processed with the NLP techniques like lemmatization, stopword removal, tokenization, and the dataset 2 is just pre-processed with tokenization. Dataset 1 is used for all the modelling types and categories followed by dataset 2 additionally only used on one of the best performing deep learning models to see the impact and setting the stage for comparative analysis.

After the dataset preparation is completed in the stage of data preprocessing, the pipeline moves forward with the feature engineering. Here, in feature engineering to convert the textual data into numerical vectors, vectorization a NLP technique is used. Different types of vectorization techniques are used here like TF-IDF, Word2Vec, Glove and BERT Embeddings on all types and categories of ML and NN models. This further sets a stage for comprehensive comparative analysis and provides insights regarding the most effective type of vectorization and modelling configuration to process nuanced linguistics associated with the mental health analysis. This sets a stage for the NLP task text classification executed and implemented by the modelling stage.

So all the above mentioned stages have prepared the dataset with a strong foundation and its modelling ready. All the feature spaces generated by the vectorization are used as an input of feature vectors for training the ML and NN models. Hyperparameter tuning is done here with the random search technique to identify the best set of optimal parameters for this dataset, vector input and modelling configurations resulting in providing maximum performance, accuracy, and reliability. Different classification techniques like ML models including types like Decision Tree, Random Forest, Logistic Regression, SVM followed by NN models including types like CNN, RNN, LSTM are trained here on the dataset 1 and different vectorization techniques. The DL models are also trained here with dataset 1 including BERT transformer based standard models like BERT and DistilBERT followed by domain specific BERT models like MentalRoBERTA and MentalBERT models. Here, additionally the MentalBERT model will also be trained on dataset 2.

In the modelling stage during the model training, internal validation is done using a two fold approach parallelly. This rigorous strategy includes stratified K-fold cross validation method followed by holdout validation method ensuring that the validation offers a comprehensive assessment on performance as well as there is an unbiased evaluation of the models generalizable capability. All the trained model's performance are evaluated with the metrics like precision, recall, f1-score, accuracy, auc-roc curve, confusion matrix, training loss and validation loss where applicable.

This offers a rich and insightful post model analysis regarding the model's performance, robustness, consistency and reliability.

After modelling and optimization stage, pipeline moves towards the misclassification analysis with two other mental health data pretrained models, interpretability analysis with feature importance, tree, coefficient weights, attention mechanism, SHAP analysis and two external unseen dataset validation to provide overall insights into model performance. This stage is important as it reveals the model's capability to handle the mental health analysis data in the real world and demonstrates the model's adaptability and generalizability towards the reliable mental health analysis on social media discourses.

To summarize the pipeline, it represents strong rigor, replicability as well as ethical approach that includes use of publicly available licensed data, following best data science practices and internal as well as external validation of results followed by interpretability in the decision making process of models. This is performed with regards to the designed methodology that aligns with highest technical precision as well as academic research rigor to conduct this thesis practically and responsibly as per my best knowledge. This pipeline ensures that the foundation is strong to provide solid understanding and insights into the complex task of text classification by reliable identification of mental distress indicators from the nuanced complex text. The measures taken into this pipeline showcases the responsible commitment for developing AI applications to address mental health analysis ethically.

4.4 Dataset Overview

The main dataset used in this thesis is called [Namdari and Gaes, 2022] "Mental health Corpus" centered around the context of depression, anxiety, ADHD and other related mental problems. This is a non-clinical dataset and reflects the online discourses from the Reddit platform. This showcases the richness in the data generated with different spectrum from vivid users representing everyday mental health discussions and conversations. This encourages the systematic and deep exploration of text for the complex nuanced contextual sentiments linked with the mental health distress.

Justification for Selection: The selected dataset is directly related to this thesis. The dataset has 27,972 unique textual entries that ensures the robustness in training, validation, and evaluation of the models. The dataset has two columns, "text" and "labels". It is a supervised dataset and has binary label distribution between "no distress = 0" and "distress = 1" categories. This makes it even better alignment with this thesis for a model to identify potential mental distress indicators from the complex linguistic text with the NLP use case of binary text classification. The dataset is acquired and verified from Kaggle with high usability of 10/10 showcasing the

dataset's quality, ethical authenticity, broader applicability, and collection methods. It is available under the CC By 4.0 license. This makes it sure that it can be further used for academic research as well as educational purposes and application development too. Utilizing this dataset makes this thesis research more transparent and lays the ethical ground for the thesis.

The selected dataset "Mental health Corpus" plays an important role due to its powerful characteristics, variability of data points, supervised labelling and unique entries. This provides a solid base for conducting investigative analysis of the captured, identified and classified complex linguistic features that are directly associated with the mental distress state. Overall, employing this dataset makes the research rigorous, transparent and relevant, eventually resulting in interesting and important insights into the involved subtleties and complexities in the domain of mental health analysis from the online discourses with the application of ML and NLP techniques.

4.5 Exploratory Data Analysis

To understand the dataset's underlying characteristics, exploratory data analysis is a crucial step before moving ahead in the pipeline [Sahoo et al., 2019]. There are multiple analytical steps applied to the dataset for revealing the textual nuances, overall structure, trends and insights. The chosen and applied steps are explained below:

Lowercasing the text: Implied for maintaining similar recognition to words irrespective of their case. This standardizes the dataset for NLP operations. [IŞIK et al., 2020]

Data inspection and info: Performed for verification of the dataset's integrity, null values checks, data type and number of instances checked.

Token length analysis: It reveals the datasets token length size and allocation in quantiles to preprocess the data appropriately in regards with models data processing limits and token limits specifically with BERT models [Koizumi et al., 2012].

Label distribution analysis: It reveals the label distribution ratio. This is a very important aspect for modelling as class imbalance can create biased models. [Yang et al., 2020]

Average word and text length calculation: It provides insights regarding the dataset's internal complexities and revealing results helps to tailor and customize the NLP operations to align with dataset's characteristics [Koizumi et al., 2012].

Unique word count: Performed to analyse the dataset's lexical diversity as it's an important aspect to see the dataset's overall richness and complexity. [Boon-Itt et al., 2020]

Histogram of text lengths: Performed to visualize the distribution of data that provides the insights related to the requirement of normalization steps like padding or truncation of the data entries during preprocessing [Koizumi et al., 2012].

Word clouds: Performed to see the most common words present in each class resulting in deeper understanding of the overall nature of the labels and datasets followed by the identified trends that can be used as a potential features for modelling [Heimerl et al., 2014].

N-gram analysis: Performed to identify most common words and phrases with two words and three words. These insights are again useful to understand the nature, trends and patters in data that could be used as features for modelling [Dey et al., 2018].

Common word frequency analysis: Performed to highlight the most common and dominating words or terms in the entire dataset, providing insights into the data characteristics and useful for feature selection [Boon-Itt et al., 2020].

Rationale for performed analysis: Each specific technique used here is to make sure that the dataset is ready for modelling and the dataset characteristics can be leveraged for feature selection of complex and meaningful patterns. Textual dataset normalization is important for simplification and uniformity. Dataset's structure investigation is important to check the null values and noise that could create bias in the modelling. Token length analysis provides confidence to select BERT's text processing parameters. Label distribution is important to detect imbalance, or the models become sensitive towards either of the classes. Word clouds, N-gram analysis and most frequent words provide the insights into the dataset's trends, patterns, and potential features. To completely understand the nuance and depth of the dataset at hand, visual as well as quantitative assessments are necessary to analyse the complexity, vocabulary, and richness to provide overall data report [Jebb et al., 2017].

Comprehensive exploratory data analysis provides valuable insights and informs the further stages of data preprocessing and feature engineering, eventually aiding in the development of robust ML, NN and DL models.

4.6 Data Cleaning and Pre-processing

In this stage, the dataset is being structured and clean as part of the data preparation. The dataset's structure lies in .csv format. After the identified noise in the dataset from the exploratory data analysis, it is being treated here followed by the

standard cleaning process that suits the further stage of data preprocessing, feature engineering and modelling.

Data Cleaning Rationale: The performed steps in data cleaning are important due to the identified noise and problems in EDA. It includes the presence of noise, outliers, redundant scrapped tags like HTML tags, non-alphabetic characters, irrelevant patters like "br" tags, null entries and others minor problems can create the hurdles and errors in the model training and also affect the model's performance by the presence of wrong features or noise that leads to irregular and inaccurate analysis [Li et al., 2021]. That's the reason cleaning the dataset is important as it prepares the dataset in its most accurate, standard, quality and integrity assured form for further stages.

Removal of HTML tag: To remove the HTML markup and attributes, BeautifulSoup is used here to achieve it [Rao et al., 2022].

Lowercasing: To avoid the duplication and maintain the uniformity in the treating the text and analysing it [IŞIK et al., 2020].

Non-Alphanumeric Filtering: To focus mainly on the linguistic nuance and patterns in the text, symbols and numbers have been filtered out with the use of regex in python [Nguyen-Truong et al., 2022].

Pattern Exclusion: Even after stripping HTML tags and attributes, there were certain patterns like "br" still present. To maintain the data relevance, again regex is employed with multiple conditions to prepare the data appropriately. [Huang et al., 2023].

Null entry and whitespace normalization: Standardization of the sparse null entries and irregular whitespaces to make the dataset streamlined for the further computational processes [Koumarelas et al., 2020].

Data Preprocessing Rationale: After cleaning the dataset, it's ready for the preprocessing stage. Preprocessing is done to prepare the dataset for the further complex NLP tasks and modelling. Here, the textual content is distilled down to its most meaningful elemental form for the easier capturing and identification of mental distress indicators for the models [Sengupta et al., 2020].

Tokenization: Different models need different kinds of tokenization. Here, customized tokenization is implied to process the textual inputs effectively. [Yeskuatov et al., 2022].

Stopword Removal: To make the dataset most effective, stopwords removal is performed by employing the NLTK library. Here, the dataset is narrowed down to words of specific significance that promote the analytical usecases and task. [Yeskuatov et al., 2022].

Lemmatization: By using the NLTK's WordNetLemmatizer, the words in the dataset are converted to its lemmas form. This is done to promote the consistency of vocabulary of the dataset [Yeskuatov et al., 2022].

Both the stages of data cleaning and preprocessing makes the dataset ready for the NLTK tasks in the further stages of feature engineering and modelling. The cleaned and pre-processed data is stored in the new column named "cleaned_text". This cleaned and pre-processed data will be used for feature engineering and modelling to attain the highest accuracy and relevance by reflecting on the real world linguistically context dependent mental distress indicators.

4.7 Feature Engineering

After exploratory data analysis, data cleaning and preprocessing is performed the dataset is well prepared and structured for the crucial stage of feature engineering. Here the raw but cleaned textual data will be transformed into a structured vectored format for the understanding of the machine learning models and to make them capture, learn, and interpret from the textual corpus much more effectively.

Feature Engineering: To move ahead and beyond the manual feature engineering and interpretation of superficial texts to highlight the underlying semantic layers, vectorization a NLP usecase is used to extract the features by transforming the original textual data itself into a format that is machine readable. This is a most important and effective foundational process that boosts the model's performance. In this stage of vectorization, features, trends, patterns are identified that captures the complex contextual, emotional, and semantic depth from the mental health data corpus of online discourses. This extracted vector formatted feature space enhances the model's performance and sensitivity to capture, recognize, interpret and classify highly nuanced linguistic patterns that are directly associated with the mental distress indicators [Mukherjee et al., 2020].

Feature Extraction and Vectorization Techniques: Different vectorization techniques are chosen to evaluate by comparative analysis to see their impact on the models performance. All the ML and NL models are trained on all the below listed traditional to advanced to state-of-the-art vectorization techniques. Different techniques capture the nuances of mental distress indicators.

TF-IDF (Traditional Method): This method quantifies the word's importance based on its relativity to its frequency over the documents, presenting a baseline for the feature representation [Campillo-Ageitos et al., 2021].

Rationale: It is chosen for its high effectiveness and simplicity. It's a traditional method of vectorization and sets a benchmark for comparison too. Highly used

for making the models interpretable to understand the models' decisions making process and its transparency. It provides the core fundamental understanding of particular words with their respective significance and separates the words with high significance [Campillo-Ageitos et al., 2021].

Word2Vec (Advanced Word Embedding): This method vectorizes the words based on their respective contextual similarities that showcases the semantic relationships [Liu et al., 2022].

Rationale: It's an advanced distributed word embedding technique, chosen for its high capability to capture the nuanced semantic relationships. This results in making the analysis richer with deeper word associations that are more dominating and related to the context of mental health discourses [Liu et al., 2022].

GloVe (Advanced Word Embedding): This method generates the word embeddings based on the global word's co-occurrences, eventually providing insights into the word's semantics that are highly based on their collective usage over the entire corpus [Straw and Callison-Burch, 2020].

Rationale: It's an advanced distributed word embedding technique that provides a vivid perspective based on the word's relationships that are focused on the global statistical information. This makes sure that models become capable of highly understanding the context and semantics involved within a linguistically nuanced textual data [Straw and Callison-Burch, 2020].

BERT Embeddings (State-of-the-Art Context-Aware Embeddings): This feature representation is obtained from the pre-trained BERT model. These high dimensional feature spaces are capable and excel at understanding the context of words within the sentences and it also captures the subtleties of the dataset that are lost in the other methods [Zeberga et al., 2022].

Rationale: This is state of the art technique and superior to all the above methods but can't be leveraged and utilized with all the models. This method has high capabilities to leverage context aware nuances to learn and understand the patterns and trends associated with the mental distress from the intricate and linguistically complex languages present in the textual corpus data. This makes the models' understanding better and makes them excel at performance [Zeberga et al., 2022].

By applying selected vectorization techniques to our models, the analysis identifies the best configuration techniques with models as well as most impactful vectorization techniques in creating highly reliable feature vector space. This diversified applied approach ensures that this thesis is comprehensive, and the analysis covers a wide range of applicable techniques from the traditional to advanced and moving to state-of-the-art techniques for feature representation. This feature engineering stage performed by the vectorization results in creating feature vector spaces that will be a guiding input layer in modelling of ML and NN models for classification with respect to mental health analysis.

4.8 Modelling

Now that the dataset and feature vector space are ready, the modelling stage begins here. This phase is systematically structured and it's a comprehensive exploration of different types of models and model categories for identifying the most effective type of models and categories for detecting the mental distress indicators through the comparative analysis. As per the designed methodology, models are categorized into three different groups. They are traditional Machine Learning (ML), advanced Neural Networks (NN) & Deep Learning models (DL)[Iyortsuun et al., 2023]. These categories are chosen from baseline to cutting edge technology models to provide the insights into a broad-spectrum analysis of these model's performance with regards to different computational architectures followed by their inherent learning capabilities.

Traditional ML Models: This category includes the model types like Decision Trees (DT), Random Forest (RF), Logistic Regression (LR), and Support Vector Machines (SVM). These ML models provide a strong foundation in regards to pattern recognition within structured data and each model has different strengths and capabilities with the NLP usecase of text classification. Here, DT and RF models are quite well known for their robust ability to interpret the non-linear data and provides straightforward interpretability into the models. While models like LR and SVM are quite efficient and effective for the NLP usecase of binary classification problems. As well as these models were utilized before and proven capable of handling basic mental health analysis [Mohamed et al., 2023] setting the baseline performances to explore different preprocessing and vectorization techniques.

Advanced Neural Network Models: Moving forward to advance NN models, it includes Neural Networks (CNN), Recurrent Neural Networks (RNN), and Long Short-Term Memory networks (LSTM). CNN model is well known for its solid capability to capture the spatial hierarchies from the data which makes it more suitable for the textual structure sequences. RNN as well as LSTM model are highly capable and excelling to handle sequences with their inherent ability to process and maintain the information over the longer textual passages. This is quite crucial for understanding the nuanced context as well as temporal dependencies present in the language. As discussed in the existing literature review section as well as other literature review [Kour and Gupta, 2022] highlights that these models are capable of capturing more contextual information and offers moderate performances. Their advanced capabilities set the stage for further exploration of different preprocessing and vectorization techniques with larger dataset at hand.

State-of-the-Art Transformer-Based Deep Learning Models: The inclusion of BERT, DistilBERT, MentalBERT, and MentalRoBERTa represents our forte into cutting-edge NLP technologies. These models leverage the transformer architecture, which allows for a deeper understanding of context and semantics in text through attention

mechanisms. Their pre-trained knowledge on vast corpora provides an unparalleled basis for fine-tuning on specific tasks like mental distress detection. Reviewed from the existing literature reviews as well other literatures [Greco et al., 2023], [Bokolo and Liu, 2023] suggests the use of trending transformer based models as well as exploring domain specific pretrained models. To push the edge and explore deeper, both base and domain specific models will be end-to-end fine tuned in this thesis with the dataset at hand to evaluate the efficacy of promising models.

Progression towards more complex and high performing models like BERT, DistilBERT, MentalBERT [Ji et al., 2022], and MentalRoBERTa demonstrates the implementation and investigation into cutting edge NLP technologies. These transformer based large language models (LLM) like BERT family [Luna, 2023] with standard and domain specific variants leverage the transformer architecture for the deeper understanding of the context, semantics and linguistic nuances from the text via attention mechanism. These models excel at NLP tasks because of their pre-trained knowledge on vast and variable corpus of textual data providing an edge for implementing end to end fine tuning on specific datasets for the different tasks like text classification.

Justification for Model Categories and its Types Selection: There is twofold reasoning behind the diversified approach of selecting model types and categories. Firstly, traditional models are selected for baseline benchmarking and its straightforward interpretability [Stiglic et al., 2020]. Whereas advanced neural network models demonstrate complex architectures offering superior [Rustagi et al., 2021] performances. Deep learning models are chosen to showcase its extraordinary capabilities in understanding highly complex, semantic, and nuanced linguistic characteristics from the data, providing insights into the applicability of state-of-the-art NLP technology in the sensitive and complex domain [Martínez-Castaño et al., 2021] of mental health analysis. Secondly, by selecting and implementing the wide range of different model architectures and complexities, the best and the most effective models are identified for capturing, identifying, interpreting, and classifying mental distress indicators from the text. This diversified approach as well as comprehensive analysis is performed to verify the capabilities and limitations of different types of models and categories resulting in the high-level research background and encouraging future research as well as such complex application developments for the mental health analysis through the application of ML and NLP techniques.

Model Training: Model training is the most important phase of the pipeline as well as designed methodology for the identification of the most effective method for detecting mental distress indicators from the textual data [Li et al., 2022]. By the implication and application of the diversified and strategic approach, the result from the analysis highlights the abilities and capabilities of each model type and categories to analyse and interpret the complex nuanced language patterns associated with mental health. The traditional ML models like DT, RF, LR and SVM followed

by advance NN models like CNN, RNN and LSTM are all trained on four vectorization techniques like traditional TF-IDF, advance distributed word embeddings like Word2Vec and Glove followed by the state-of-the-art context aware BERT embeddings. This kind of a diversified modelling approach is chosen for assessing the different techniques of feature representation's impact on the model's ability to perform and process the textual data. These chosen model categories lay a foundation of baseline models as well as advanced models, for analysing their performance and capabilities in handling the textual data providing different analytical perspectives. Additionally, to analyse and shed light on the state of the art NLP technologies, DL models have been end-to-end fine-tuned [Howard et al., 2018] and trained including both standard and domain specific models. These models surpass traditional feature engineering due to their inherent capabilities of generating context aware embeddings and its capability to process and analyse the textual data more efficiently. Aggressively pre-processed, Dataset 1 is used for all the ML, NN and DL models for training followed by dataset 2 with minimal preprocessing is additionally used only on the best performing DL model for comparative analysis assessing the impact of preprocessing levels on the model's performance and understanding.

Rationale and Objectives: This systematic planning of model training and evaluating approach over different types of preprocessing techniques, vectorization techniques, different model variants and different model categories demonstrates rigorous and comprehensive analysis. Evaluation results from the modelling phase creates a solid foundation for comparative analysis of model performance revealing the best preprocessing approach, best vectorization techniques, best model and its configuration followed by the most effective model category in regards of performance and adaptability towards mental health analysis [Dang et al., 2020].

Modelling and optimization phase clearly represents the meticulous approach of training and evaluation with high methodological, technical as well as academic rigor. This approach makes sure that the results are not just scientifically sound but even practically proven and relevant, laying the foundation as well as guiding towards future innovations and research in the domain of mental health analysis via the application of ML and NLP techniques.

4.8.1 Hyperparameter Tuning and Optimization

This thesis methodology executes customized training approaches for each different category of models and model types to maximize their performance in capturing, identification and classification of complex nuanced data. Hence, selection of best optimal parameters for each type of model is important to follow the standard practices of modelling with respect to best practices in data science to achieve the superior acceptable performance [Liao et al., 2022].

Random search method is implemented to explore the hyperparameters for different models from all the categories of ML, NN and DL models. Different parameters like tree depths, split criteria, iterations, kernels, batch size, learning rates, iterations, epochs, dropout rates, hidden layer dimensions, regularization and more parameters were explored in this systematic and important process to find the best suit of optimal architectural parameters [Turner et al., 2021].

By selecting the optimal set of parameters by hyperparameter tuning, models became more optimized resulting in the balanced relationship between model complexity and model's generalization ability. This ensures that developed models are capable enough and demonstrates high performance, robustness, consistency, and reliability in models' prediction revealing no overfitting and adaptability to unseen datasets.

Comprehensive random search approach with evaluation and validation framework reveals the impact of different hyperparameter settings with respect to the model's efficacy and optimization. This iterative as well as data driven process revealed the best set of hyperparameters that enhances the modelling and optimization process and make the models to capture, identify and classify complex nuanced linguistic context from the textual data effectively and to the best of the model's ability. This process sets the stage for modelling and showcases the best practices followed in the domain of data science with respect to sensitive application of mental health analysis.

4.8.2 Unified Training Process, Evaluation Metrics, and Model Persistence

Training Process-

Data Splitting and validation: Used Stratified K-fold cross validation in all ML and NN models with 5 splits. This makes sure that the training and the evaluation is unbiased followed by maintaining the overall class proportionality as well as addresses the bias cases if there is an imbalance pattern in class distribution. [Szeghalmy et al., 2023]

Vectorization Input: Implemented traditional to advance NLP techniques like TF-IDF, Word2Vec, Glove and BERT embeddings to generate the feature space vectors making the modelling process effective by making models learn deep nuanced and linguistic context followed by complex semantic patterns.

Batch Processing (NN models specific process): To maintain the training efficiency and manage computational resource allocation, batch processing is implemented on the advance NN models like CNN, RNN, LSTM [Kandel et al., 2020].

Epochs (NN models specific process): To make the models learn effectively and to optimize the learning process with precautions to under-fitting and overfitting, number of epochs are designed to make the model expose to the training data adequately.

Optimization of Models-

ML Models: Based on cross validation performance, Hyperparameter tuning selected the optimal parameters.

NN Models: To minimize the loss function across different training epochs, Adam optimizer is used because of its adaptive learning rate.

Early stopping mechanisms: Implemented on NN models to avoid the case of overfitting. Over specified number of epochs when validation loss is not improved then the training is halted [Bai et al., 2021].

Evaluation Metrics: Common for both ML and NN models [Hossin et al., 2015], [Deng et al., 2016]

Accuracy: Used here for overall effectiveness of the model

ROC-AUC Score: To assess the model's distinguishing capability across classes.

Precision, Recall and F1-score: Used to assess the model's predictive accuracy as well as reliability.

Confusion Matrix: Used for visualizing and understanding model's performance with respect to true vs. predicted labels and looking at false positives and false negatives.

Model Persistence-

Serialization (ML-specific): ML models are saved with Joblib for easier access, inference and application for future studies [Joblib, 2024].

Model Saving (NN-specific): NN models are saved with compatible framework formats as .pth with PyTorch.

Checkpointing (NN-specific): To prevent the overfitting, models are checkpointed and saved from the state before the early mechanisms stop. This also helps recover and analysis of the model from the specific states.

This unified training framework makes sure that model is going through rigorous training, evaluation as well as persistence with respect to systematic and best modelling practices. Training of neural networks is executed with details and clarity with consideration of batch processing, epochs, optimizers, checkpointing and early stopping mechanisms. This highlights the use of advanced ML and NLP techniques to make the models generalizable, reproducible, and adaptable.

4.8.3 Traditional Machine Learning Models

Traditional Machine learning models trained here are the baseline models which includes Decision Tree, Random Forest, Logistic Regression and Support Vector Machine. All, the four models are trained on all the four chosen vectorization techniques ranging from traditional TF-IDF, advanced distributed word embeddings like Word2Vec and Glove and state-of-the-art context aware BERT embeddings. This modelling process plays an important role in understanding how baseline models perform when configured with different vectorization techniques and see their capability to identify and classify mental distress.

Best Performing Decision Tree Model Configuration (Using Word2Vec)-

Best performing DT model trained with Word2Vec vectorization uses the following key parameters:

Vectorization: Word2Vec feature space used for nuanced understanding of complex linguistically dependent contexts.

criterion: To enhance class separation, Gini is used for optimal split.

Max Depth: To prevent overfitting as well as balance complexity value is 12.

Min Samples Split: To make a meaningful split of internal nodes 25 is the set value.

Min Samples Leaf: To accommodate larger groups adequately, 10 is the set value.

Random State: Reproducibility is important, hence 42.

NLP usecase: Text classification (binary).

This specific set of parameters demonstrates stable performance by leveraging the Word2Vec feature vectors effectively. These parameters make sure that model is sensitive enough to detect mental distress indicators as well as overfitting is prevented on the basis of training data. This best performing DT model with word2vec along other DT variants models is saved for the future research, review, and reproducibility.

Best Performing Random Forest Model Configuration (Using Word2Vec)-

Best performing RF model trained with Word2Vec vectorization uses the following key parameters:

Vectorization: Word2Vec feature space used for nuanced understanding of complex linguistically dependent contexts.

Estimators: To improve the robustness and accuracy of the overall model, 1000 trees is the optimal set value.

Criterion: To enhance class separation, Gini is used for optimal split.

Max Depth: To prevent overfitting as well as balance complexity value is 12.

Min Samples Split: To make a meaningful split of internal nodes 25 is the set value.

Min Samples Leaf: To accommodate larger groups adequately, 10 is the set value.

Random State: Reproducibility is important, hence 42.

N_jobs: -1 to utilize all processors for faster training.

NLP usecase: Text classification (binary).

This specific set of optimal parameters obtained from the hyperparameter tuning and are configured with Word2Vec. These parameters demonstrate stable performance and strike balance between the models capturing and identifying mental distress indicators from complex nuanced text and maintaining generalizability to the unseen data. This best performing RF model with word2vec along other RF variants models is saved for the future research, review, and reproducibility.

Best Performing Logistic Regression Model Configuration (Using Word2Vec)-

Best performing RF model trained with Word2Vec vectorization uses the following key parameters:

Vectorization: Word2Vec feature space used for nuanced understanding of complex linguistically dependent contexts.

NLP usecase: Text classification (binary).

Max Iterations: Value is set to 1000 for allowing notable convergence when dealing with high-dimensional vector spaces.

Random State: Reproducibility is important, hence 42.

This specific set of optimal parameters obtained from the hyperparameter tuning and are configured with Word2Vec. Specifically, when researching in the context of mental health analysis, interpretability of the models is important for validation and logistic regression provides a strong interpretability analysis environment. This best performing LR model with word2vec along other LR variants models is saved for the future research, review, and reproducibility.

Best Performing SVM Model Configuration (Using Word2Vec)-

Best performing SVM model (LinearSVC variant) trained with Word2Vec vectorization uses the following key parameters:

Vectorization: Word2Vec feature space used for nuanced understanding of complex linguistically dependent contexts.

Base Classifier: LinearSVC is used for its effectiveness with high-dimensional space characteristic of Word2Vec.

Calibration: CalibratedClassifierCV is implied with LinearSVC for estimating reliable probability scores, and also for enhancing performance evaluation and tracking metrics like ROC AUC score.

Random State: Reproducibility is important, hence 42.

Max Iterations: Value is set to 1000 for allowing notable convergence when dealing with high-dimensional vector spaces.

NLP usecase: Text classification (binary).

This specific set of optimal parameters obtained from the hyperparameter tuning and are configured with Word2Vec. Here, CalibratedClassifierCV is applied to mitigate the SVM's primary limitations regarding the probability scores. Hence, the LinearSVC variant of SVM is implemented here to not just accurately achieve binary classification but to also have the respective probabilities score. This best performing SVM model with word2vec along other SVM variants models is saved for future research, review, and reproducibility.

4.8.4 Advanced Neural Network Models

Progression and execution towards the implementation of more complex and Advanced Neural Network Models like CNN, RNN and LSTM are also trained on all the four chosen vectorization techniques ranging from traditional TF-IDF, advanced distributed word embedding like Word2Vec and Glove and state of the art context aware BERT embedding. This modelling process plays an important role in understanding how advanced NN models perform when configured with different vectorization techniques and see their capability to identify and classify the complex, contextual and nuanced mental distress indicators, and semantic meaning patterns from the textual corpus.

CNN Model Configuration (Using BERT Embeddings)-

Best performing CNN model trained with BERT embeddings captures more deeper contextualized features from the textual corpus data uses the following key parameters:

Input Dimension: Same as the BERT embeddings vector size of 768 for comprehensive and systematic contextual processing.

Convolutional Layers

Layer 1: Output channels are set to 128, kernel size is set to 5, padding set to 2 for initial feature extraction.

Layer 2: Reduction to 64 channels, similar kernel and padding value for feature optimization.

Batch Normalization: To maintain the output stability, batch normalization is done after each convolution layer.

Pooling: For dimensionality reduction as well as feature extraction, Max pooling is set with kernel size 2.

Dropout: To prevent overfitting by omitting training units randomly by 0.5 value.

Fully Connected Layers: To have a binary output layer, fully connected layers are reduced to 32 dimensions with a pre-final layer.

Activation: ReLU for non-linearity, enhancing pattern recognition.

Batch Size: 32 for efficiency and performance optimization.

Learning Rate: For effective learning it is balanced with 0.0005.

Weight Decay: For regularization it is set with 1e-5 Adam optimizer.

Scheduler: To adjust the learning rate adaptively and dynamically after each epoch, StepLR with step size 5, gamma 0.1 is the set value.

Early Stopping Patience: To prevent overfitting, patience value is set to 7 so adequate learning also carries on.

This specific set of optimal parameters of CNN model are configured with BERT embeddings to provide the maximum performance to seamlessly capture, identify and classify complex and nuanced mental distress indicators from the textual data demonstrating its advance model architecture and application of NLP techniques to address robust and consistent classification with regards to mental health analysis. This BERT embeddings configured CNN model is saved with other CNN variants for the review, future work, and reproducibility.

RNN Model Configuration (Using BERT Embeddings)-

Best performing RNN model trained with BERT embeddings captures more deeper contextualized features from the textual corpus data uses the following key parameters:

Input Dimension: To utilize the full context aware BERT embeddings ,768 is set as the same vector size of BERT embeddings.

RNN Layers

Hidden Dimension: To maintain the complexity and pattern capture balance, 128 is set to the value.

Layers: To maintain the efficiency as well as leverage of sequential data, 1 is set.

Fully Connected Layer: ReLU activation is implied to map the RNN models output to the binary classification from the identification of the mental distress indicators.

Regularization: To prevent overfitting by omitting training units randomly by 0.5 value before the fully connected layer.

Batch Size: 32 for balanced efficiency and performance optimization.

Learning Rate: 0.001 is set for steady as well as non-aggressive convergence.

Weight Decay: For regularization it is set with 1e-5 Adam optimizer.

Scheduler: To adjust the learning rate adaptively and dynamically after each epoch, StepLR with step size 5, gamma 0.1 is the set value.

Early Stopping: Delta is set with the value of 0.0001 and to prevent overfitting, patience value is set to 7 so adequate learning also carries on.

This specific set of optimal parameters of the RNN model configured with the BERT embeddings showcases the models effective use of sequential NN architecture that excels to capture, interpret, and classify the complex contextual data. RNN model is effective due to its ability in processing sequences that is directly suitable for the analysis of the textual data where the order and the context of the words are important to get the overall context. This BERT embeddings configured RNN model is saved with other CNN variants for the review, future work, and reproducibility.

LSTM Model Configuration (Using BERT Embeddings)-

Best performing LSTM model trained with BERT embeddings captures more deeper contextualized features from the textual corpus data as well as captures sequential information with long term dependencies uses the following key parameters and architecture:

Input Dimension: To utilize the full context aware BERT embeddings ,768 is set as the same vector size of BERT embeddings.

LSTM Layers

Hidden Dimension: To maintain the complexity and pattern capture balance, 128 is set to the value.

Layers: 1 is set here for balancing the long-term dependency capture as well as maintaining the computational efficiency.

Dropout: To prevent overfitting by omitting training units randomly by 0.5 value for multi-layer configurations.

Fully Connected Layer: To imply the mapping of LSTM output results to the binary classification.

Output Dimension: is set to 2 for customizing binary classification between classes.

Activation Function: To make the model learn complex patterns, ReLU is connected at post-fully layer.

Batch Size: Set to size 16 to align with the LSTM's complexity as well as memory usage for optimal performance.

Learning Rate: To make sure optimal convergence of weights, 0.001 is the set value.

Weight Decay: For regularization it is set with $1e-5$ Adam optimizer.

Scheduler: To adjust the learning rate adaptively and dynamically after each epoch, StepLR with step size 5, gamma 0.1 is the set value.

Early Stopping: To prevent overfitting, patience value is set to 7 so adequate learning also carries on.

This specific set of optimal LSTM models parameters configured with BERT embeddings showcase the advanced and complex architecture of the model and its capabilities to remember information for longer periods results in important understanding of the complex patterns and context from the textual corpus of data. This best performing BERT embeddings-based LSTM' architecture offers robust and consistent nuanced classification and is saved with its other LSTM variants for review, future work, and reproducibility.

4.8.5 State-of-the-Art Deep Learning Models

Training process of these state-of-the-art transformer based deep learning models [Wolf et al., 2020] is different compared to the training of ML and NN models. Here, four transformer based deep learning models are trained from the BERT family including the core BERT models and domain specific BERT models. These four models are BERT, DistilBERT, MentalRoBERTa and MentalBERT. These models do not need traditional vectorization input for feature space. They have intrinsic capabilities and complex architecture to generate their own context aware embedding. All the four DL models are end-to-end fine-tuned with the chosen dataset to make the model robust, reliable, consistent enough to capture, identify and classify the mental distress indicators with high performance and accuracy. Out of all four deep learning models, MentalBERT end-to-end fine-tuned model is the best performing model and its architectural parameters are presented here.

Mental BERT End-to-End Fine-Tuned Model-

The MentalBERT model [Ji et al., 2022] is a pre-trained domain specific variant of the BERT base model which is rigorously fine-tuned with Reddit based mental health data from threads and discussions. This was done to specifically make the model capable of identifying intricate context dependent linguistics related to mental health. This model was trained over four Nvidia Tesla v100 GPUs for around eight days. This model was built for early detection of mental health related problems by analysis online discourses. As its trained on a non-clinical data, it should only be used for the initial identification and supportive insights that too on discourses and not as a substitute or option of medical health professional diagnosis. Ethical considerations like user protection, data privacy, bias and fairness have been taking care of with highest importance during the development of the model showcasing commitment, responsible research and laying foundation for the further research [Ji et al., 2022]. This domain specific pretrained MentalBERT model is further end to end fine-tuned in this thesis with the chosen dataset for the text classification.

Tokenizer: During the model's training, consistency in text preprocessing, tokenization and encoding is important and to achieve that AutoTokenizer is used which is made specifically for the "mental/mental-bert-base-uncased" [Ji et al., 2022].

Input Data Handling

Max Length: Set to 300 for maintaining the balance between the complex contextual details and computational resource efficiency.

DataLoaders: Implied RandomSampler during the training process to introduce some randomness to make the model robust and also implied SequentialSampler for the process of validation for maintaining the consistency in evaluation. [HuggingFace, 2023].

Optimizer: Here Adamw is implied with the learning of $5e-5$ followed by the weight decay of 0.01 to ensure adaptive learning rate and regularization. This results in effective learning by model, optimization and preventing overfitting cases.

Training and Validation

Batch Size: Set to 16 for efficient consumption of GPU memory and also allowing gradient accumulation which is important for the complex and big models like the transformer based BERT family.

Epochs: Limited to one epoch due to exceptional capabilities of the BERT models and epoch 2 starts to overfit.

NLP usecase: Binary text classification.

Evaluation Metrics: Classification report includes accuracy, precision, recall, F1 score, ROC AUC, training loss and validation loss to evaluate and assess the models

performance from different perspectives ensuring model's capability and efficacy in identifying mental health indicators.

Model and Tokenizer Persistence: After training the model, tokenizer and respective model files are saved with `save_pretrained` [HuggingFace, 2023]. This method is effective to load the model for inference, deployment, review, and future work.

Environment and Reproducibility: Fixed seed is set here across torch, random and numpy to make sure the model is reproducible with the similar results.

These sets of specific BERT's architectural parameters demonstrate the state of the art NLP techniques with regards to mental health analysis. Also, the domain specific pretrained model makes a difference and its retained knowledge pushes the boundaries and provides high performance. End-to-end fine tuning the MentalBERT model makes it learn complex contextual patterns and leverages the model's architecture for robust text classification and prediction of classes by identifying nuanced mental distress indicators. MentalBERT models with over BERT model variants are saved for review, future work, and reproducibility.

4.8.6 Model Validation and Evaluation

After the modelling phase, it is important to ensure the robustness of the results. For that, this section will provide the approach for validation and evaluation of ML, NN and DL models.

Validation Techniques for ML and NN models: Implemented stratified K-Fold cross validation with an set value of the 80 to 20 split. This method is chosen for its capability to validate thoroughly as well as computationally efficient. It divides the datasets in the k (5) segments that are distinct. This maintains the class probability as well as makes sure the systematic validation over the entire dataset. This provides the models generalizability as well as performance review [Yadav and Shukla, 2016], [Szeghalmy et al., 2023].

Validation Techniques for DL models: It's a bit different compared to ML models. Here, the hold out method is used of 80 to 20 split ratios. This is chosen due to its computational efficiency and comprehensive validation. Also, DL models are intricate and resource exhaustive, hence computational efficiency is very important. This technique divides the dataset into training as well as validation set for analysing the DL models performance parallelly managing the computational load on the infrastructure. This way essential learning and generalization capabilities of the DL models can be evaluated [Yadav and Shukla, 2016].

Rationale Behind Our Validation Methods: The chosen technique of Stratified K-Fold cross validation followed by the holdout method is because they are reliable, efficient and efficacy of these techniques to conduct a robust validation between

different type of sub datasets. They are computationally adaptable and capable of providing unbiased and rigorous validation for the models.

ML, NN and DL Models Evaluation Metrics: Evaluation techniques for ML, NN and DL models: Precision, recall, F1 score, Accuracy, AUC-ROC curve, and Confusion matrix are the main evaluation techniques used to gauge the performance of the ML models. These techniques in combination provide the overall performance insights of these models from different perspectives. This is crucial to understand the models' capabilities of robust classification, sensitivity and specificity, consistency in prediction, differentiation and identification of distress patterns and reliability in the correct classification. For NN and DL models also these techniques for evaluation remain the same. Training and validation loss are additional two metrics to evaluate the model's learning capabilities and to highlight the overfitting or underfitting scenarios [Zhou et al., 2021], [Deng et al., 2016], [Hossin et al., 2015], [Eelbode et al., 2021].

Rationale Behind Our Evaluation Metrics: These set of metrics are chosen to gauge the models performance, overfitting and underfitting case, class separation capability, misclassification instances, overall accuracy, learning dynamics and validation loss of the models. This kind of holistic approach provides the insights into the models rigorous analysis, demonstration models capability and limitations.

External validation is performed on the MentalBERT DL model which are centered around depression [Kaggle, 2024a] and suicidal ideology [Kaggle, 2024b]. A bit different in the context compared to the training dataset, but it belongs to the same mental health domain with similar labels. External validation with such datasets pushes the MentalBERT model boundaries in classification showcasing the model's real-world adaptability to variable context identification of mental domain as well as proves the model's generalizability and performance. Both external datasets are extensive in the terms of context, nuances and sample size that allows us to gauge and analyse MentalBERT's efficacy to detect different kinds of mental distress indicators related to different problems in the domain of mental health. The dual approach of validation during the training and post training ensures the models robustness, generalizability, and real-world adaptability [Ermer et al., 2020]. Also, it proves the high performance of the models in practicality and not just academic rigor. This makes the model ready for deployment, review, further research or to optimize and evolve the model further with more multilabel training data for different usecases. Such models are a contribution in the domain of mental health analysis and demonstrate real world practicality [Li et al., 2023].

These chosen and applied model validation and evaluation techniques are the best practices in the domain of data science, and they determine the model's capability, limitations and provide an overall robust performance report covering different perspectives and cases.

4.9 Model's Misclassification and Interpretability Analysis

Just developing models, validation and evaluation is not enough, understanding the "why" behind the models classification prediction is equally important especially when dealing with the mental health domain. Interpretability of models helps to make and understand the decision-making process transparent and to generate trust in the system's results. This section explores the interpretability analysis approach chosen and applied on developed models.

Misclassification Analysis: It is performed on the MentalBERT model outputs. Here, the misclassified instances are analysed to gauge the model's limitations and areas for further refinement [Jeatrakul et al., 2010]. To enhance the misclassification analysis, two pretrained domain specific models are employed to improve the effectiveness of the misclassification analysis. The two pretrained models used here are "twitter-roberta-base-sentiment-latest" [Camacho-Collados et al., 2022] and "roberta-base-go_emotions" [Lowe, 2023] are fine-tuned for detecting the sentiment and emotion from the instances and provides multi label classification respectively. Both these models are used to identify sentiments states and emotional states from the misclassified instances for both the labels. Mean scores of sentiments states followed by the mean scores of emotional states provides deeper insights into the misclassification analysis by providing the quantitative grounds to conclude what happened and to gauge the limitations of the model.

Rationale Behind the Analysis: To understand the underlying and predominant sentiments and emotional states of the misclassified instances and to understand the misclassified errors. The calculation of mean scores of these outputs from pretrained models reveals the insights and possible reasons for misclassifications associated with the model's limitations as well as required further refinement like adding more data of similar misclassified instances or needing more labels or needing additional feature engineering and possibly technical limitations. This analysis demonstrates the quantitative ground to provide reasoning for the misclassified instances and limitations are identified for further optimization and laying groundwork for future research and work [Jeatrakul et al., 2010].

To look inside and understand the models' understanding regarding the features, indicators, trends and patterns that associate to mental distress indicators, traditional ML algorithms are straightforward [Stiglic et al., 2020] compared to NN and DL models as they are usually called as black boxes.

Tree and feature importance: To understand the decision tree model, a tree is being printed to understand the root and the split nodes, feature importance is employed to understand the top 50 features that influences the decision tree's model classification outputs. Similarly, to understand the random forest's models understanding, top 50 features analysed via feature importance mechanisms [Kang et al., 2022].

Coefficients weights: To understand the LR and SVM models understanding, coefficient weight mechanisms are applied to reveal the top 50 positive (1) and negative (0) coefficients for models. They represent the words with the most influential lexical feature and their weights for the classification outputs [Kang et al., 2022].

Attention Mechanism: As NN and DL models have a similar and complex architecture, attention mechanisms are used on the DL model that allows us to investigate the influencing tokens or words in the sentence where the model focuses for decision making. This is applied on a single instance and provides visualization with weights and colour coding to understand the influential features that are context dependent. This method is performed with both the sub techniques like on the last layer of the model as well as the aggregated layer of the model [Niu et al., 2021].

SHAP Analysis: To get deeper insights into the complex decision making process and understanding of DL models, SHAP analysis is used on individual instances to provide local explanations showcasing the most influential words identified by the model pushing towards a certain label of classification. It's important to understand that due to the complex architecture and nature of the neural networks and deep learning models, findings for each individual model cannot be tied or generalized to the entire model. This is where it separates from the traditional ML models and showcases supremacy in analysing each individual instances with deeper contextual nuance exploration [Aldughayfiq et al., 2023].

Justification for Interpretability Techniques: Applied different and diversified interpretability techniques showcases the models understanding with the context of mental health analysis based on their architectural capability and textual processing prowess. Implication of different methods provides highlights into the complex human language where context is highly important to understand the meaning and sentiment. Similarly, misclassification analysis demonstrates the limitations, error reasoning and future implications to make the model more robust and optimized. Tree, feature importance and coefficient weights are implied on ML models for their simplicity, straightforwardness, and direct learning into the model's decision-making process. Attention mechanisms and SHAP analysis are applied on NN and DL models for the microscopic view on the difficult black box models by analysis individual instances to understand the decision-making process of this high performing and complex models [Joyce et al., 2023].

The set of chosen misclassification and interpretability analysis techniques is not only to gauge the models predictions but to also understand its adaptability and ethical considerations for real world application. This increases the robustness and stability of the developed models by ensuring the accountability of these model's outputs and demonstrates solid academic and technical rigor of this thesis. This thesis, creates a responsible contribution to the research of Mental health analysis with AI applications [Joyce et al., 2023].

Implying this systematic and structured methodology, this thesis demonstrates a high level of academic, technical, and ethical rigor. Designed methodology offers multifaceted assessments of different preprocessing, vectorization, model types and model categories including their hyperparameter tuning, training, validation, evaluation, misclassification, and interpretability analysis. This methodology not only covers traditional models and techniques but provides practical insights into the latest NLP techniques and deep learning models. All the model configurations from different types and categories are gone through comprehensive comparative analysis and the best preprocessing techniques, vectorization techniques, model types and categories are identified which excels in superior understanding of features that are directly related to mental distress indicators. This ethically grounded and responsible thesis contributes to the research by reducing the gap between practical applications of data science and AI in the domain of mental health analysis.

5 Results

In this result section, the NLP classification pipeline task with its data, models, interpretation, and details with regards to mental health domain textual analysis is explored and examined with multifaceted perspectives. The exploratory data analysis outcomes are presented to check out the dataset characteristics and found insights. Different model outputs are evaluated for performance followed by comparative analysis of model configurations based on different modelling categories, pre-processing techniques, and vectorization techniques to study their impact and overall predictive efficacy. Presentation of combined analysis engine with samples is presented for nuanced interpretability. Model interpretation analysis followed by misclassification analysis to understand the model capability and understanding. External datasets validation with benchmarks is evaluated to make sure the model's capability with real world applicability and generalizability.

5.1 Exploratory Data Analysis

```
Dataset Information:
<class 'pandas.core.frame.DataFrame'>
RangeIndex: 27977 entries, 0 to 27976
Data columns (total 2 columns):
 #   Column  Non-Null Count  Dtype
---  -
 0   text    27977 non-null  object
 1   label   27977 non-null  int64
dtypes: int64(1), object(1)
memory usage: 437.3+ KB

Token indices sequence length is longer than
ll result in indexing errors
count    27977.000000
mean      84.703757
std     118.632974
min         2.000000
25%        22.000000
50%         46.000000
75%        101.000000
90%        194.000000
95%        278.200000
99%        559.240000
max     3118.000000
Name: text, dtype: float64

Basic Statistical Overview for Text:
count          27977
unique         27972
top    real supplieroot hours up day far
freq          3
Name: text, dtype: object

Distribution of labels (0 and 1):
label
0     14139
1     13838
Name: count, dtype: int64
```

Figure 5.1: Dataset information

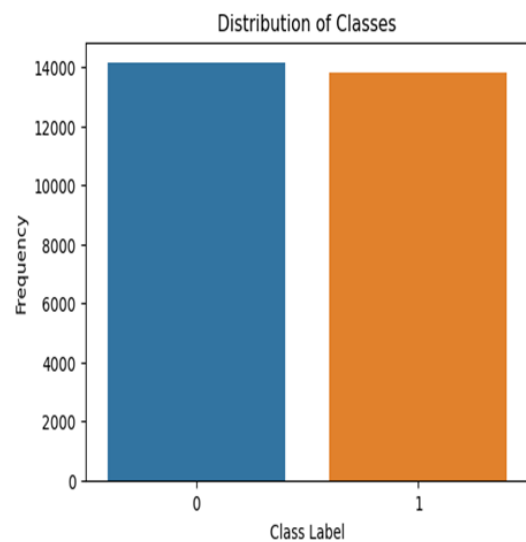


Figure 5.2: Labels distribution

Dataset Overview: Above Figure 5.1 presents an EDA aspect regarding the dataset information. The chosen dataset contains text based unique entries with around 28000 samples that are labelled as no distress (0) and distress (1). The variability and uniqueness in the dataset are very critical for the training of different categories of models to make them find and learn complex and nuanced natural language patterns and indicators that are directly associated with mental health.

Class Label Distribution: To build robust and well performing models over various scenarios with no to minimal bias, class balance is very crucial. The label distribution for the used dataset is relatively balanced with label 0 having 14139 samples and label 1 with 13838 samples. This distribution is visualized with a bar chart and can be seen in Figure 5.2.

```
Missing values in each column:
text      0
label     0
dtype: int64
Average Word Length:
5.8688108054081045
Average Text Length:
71.75940951495872
Total Number of Unique Words in Dataset: 72649
Average Number of Unique Words per Text: 55.55899488865854
```

```
text label
0 dear american teens question dutch person hear... 0
1 nothing look forward lifei dont many reasons k... 1
2 music recommendations im looking expand playli... 0
3 im done trying feel betterthe reason im still ... 1
4 worried year old girl subject domestic physic... 1
```

Figure 5.3: Data characteristics

Figure 5.4: Data samples

Dataset Insights: Variety in the text samples are important to make the models generalized. As seen in Figure 5.3, this dataset has average word length over the dataset is around 5.87 words followed by 71.75 words as overall length of the text. This represents that there is length variability in datasets expression. This dataset also has a strong and rich vocabulary as the total count of unique words here is 72,649. This is beneficial for the analysis of linguistics thoroughly.

Data Samples: Figure 5.4 gives a high level glance into our dataset and labels. These samples demonstrate that language tone is ranging from casual conversation remarks to different and strong emotional state and tone. This gives us the impression of linguistic complexity and varied diversity of this dataset.

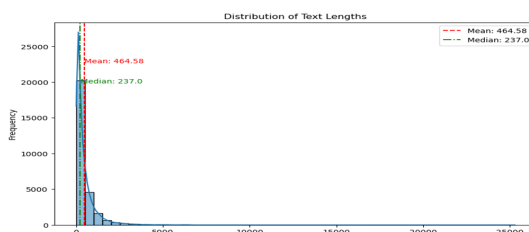


Figure 5.5: Text length distribution


```

Top 10 Unigrams:
[('im', 34720), ('like', 22300), ('want', 17299), ('know', 15475), ('feel', 14508), ('life', 14386), ('ive', 11287), ('people', 11249), ('time', 10829), ('really', 10187), ('think', 8615), ('going', 7959), ('dont', 7829), ('friends', 7071), ('day', 6944), ('good', 6799), ('help', 6534), ('years', 6480), ('make', 5971), ('things', 5936), ('way', 5779), ('got', 5740), ('br', 5610), ('love', 5069), ('anymore', 4972), ('fucking', 4931), ('better', 4812), ('end', 4754), ('family', 4753), ('school', 4713), ('need', 4666), ('live', 4443), ('work', 4351), ('die', 4296), ('talk', 4276), ('right', 4231), ('say', 4188), ('year', 4077), ('kill', 3960), ('bad', 3924), ('ill', 3923), ('film', 3901), ('thing', 3872), ('redflag', 3846), ('movie', 3813), ('hate', 3805), ('shit', 3659), ('told', 3592), ('thought', 3563), ('friend', 3556)]

Top 10 Bi-grams:
[('feel like', 5245), ('filler filler', 2889), ('dont know', 1934), ('im going', 1831), ('want die', 1400), ('dont want', 1347), ('like im', 1342), ('im tired', 1159), ('know im', 1154), ('feels like', 907), ('high school', 795), ('im sure', 778), ('im scared', 774), ('im sorry', 751), ('think im', 737), ('years ago', 723), ('suicidal thoughts', 694), ('best friend', 689), ('im gonna', 677), ('br br', 664), ('need help', 650), ('ive tried', 644), ('long time', 634), ('felt like', 619), ('really want', 610), ('years old', 600), ('year old', 602), ('want kill', 598), ('want live', 593), ('mental health', 569), ('im really', 568), ('want end', 516), ('im fucking', 479), ('im trying', 475), ('life im', 473), ('makes feel', 445), ('im afraid', 445), ('feeling like', 429), ('anymore im', 421), ('end life', 414), ('make feel', 412), ('dont think', 388), ('day day', 383), ('friends family', 381), ('commit redflag', 373), ('months ago', 367), ('want feel', 367), ('im sick', 366), ('really know', 360), ('feel better', 359)]

Top 10 Tri-grams:
[('filler filler filler', 2657), ('feel like im', 730), ('gtpoplt gtpoplt gtpoplt', 326), ('water nowdrink water', 323), ('nowdrink water nowdrink', 287), ('quick brown jumps', 284), ('brown jumps lazy', 284), ('jumps lazy quick', 283), ('lazy quick brown', 283), ('want die want', 219), ('dont know im', 185), ('eve eve eve', 167), ('monitor monitor monitor', 144), ('feel like ive', 143), ('im years old', 140), ('im going kill', 138), ('im pretty sure', 132), ('love love love', 128), ('feel like shit', 126), ('makes feel like', 117), ('life worth living', 112), ('dont know anymore', 107), ('make feel better', 107), ('know im going', 107), ('click click click', 106), ('want live anymore', 105), ('dont want live', 102), ('think im going', 99), ('pee pee po o', 99), ('pee poo poo', 99), ('poo poo check', 99), ('feel like life', 98), ('dont feel like', 98), ('need help need', 98), ('poo check pee', 98), ('check pee pee', 98), ('ampxb ampxb ampxb', 97), ('life feel like', 96), ('help need help', 94), ('help help help', 92), ('long story short', 91), ('want feel like', 90), ('im year old', 89), ('im sorry im', 89), ('day day day', 89), ('feels like im', 88), ('want die im', 86), ('im im im', 85), ('pain sad pain', 85), ('sad pain sad', 84)]

```

Figure 5.8: N-Grams analysis

N-gram Analysis: To find more interesting trends, patterns and insights into dataset, n-gram analysis is conducted here and resulting in unigrams, bigrams and trigrams as seen in the above Figure 5.8. The result from this analysis gives insights into the topics, common phrases and words present in the dataset. This analysis also revealed that dataset contains noise and some bracket fillers that needs to be cleaned before preprocessing. Bigrams like "feel like", "want die", "suicidal thoughts", "need help" and trigrams like "feel like im", "im going kill", "dont know anymore" and "sad pain sad" are highly demonstrative of personal narrative with indicating mental distress patterns and indicators [Huang et al., 2023]. These insights helps to understand the data better before preprocessing and feature engineering as it provides more context into mental health's patterns and features that could improve and refine modelling.

Performing exploratory data analysis and analysing its results, highlights the characteristics of the dataset. This is crucial and the insights gathered here plays an important role for further stages in the pipeline with regards to the NLP usecase of text classification.

5.2 Insights into Data Cleaning, Pre-processing and Feature Engineering

Data Cleaning: Data cleaning resulted in noise free and refined data. The dataset has two columns named "text" and "labels". The cleaned data is appended to the new column named "cleaned_text". The data cleaning preliminary step is significant as it is necessary to remove the HTML tags, specific unwanted patterns of brackets,

lower casing the text, filtering out the non-alphanumeric characters followed by outlier handling. To make sure the next steps and end goal's integrity, this rigorous data cleaning is performed to lay an error free strong foundation. Further null entries and white spaces were handled, removed, and normalized to make the dataset streamlined and prepare the dataset for the next stage of data pre-processing.

Data Pre-Processing: After successful data cleaning, data pre-processing is performed to move ahead further and refine the data with regards to easier NLP adaptability. Mainly, stopword removal, lemmatization and custom tokenization are performed to transform the data to its most meaningful form and elements. As different algorithms and models have vivid requirements, tokenization is customized accordingly to suit them. To maintain and improve the dataset's vocabulary consistency, stopwords were eliminated and the tokens or words were converted to their lemma forms. This pre-processing is performed on the "cleaned_text" column. This data pre-processing step made the dataset noise free and refined to make the modelling process efficient and effective. Data pre-processing resulted in processed data and making two different datasets 1 and 2. Dataset 1 is aggressively pre-processed with stopword removal, lemmatization and tokenization and it is used for training all the model categories and dataset 2 with customized tokenization is additionally used on the best performing deep learning model to see the impact of pre-processing techniques on the model's efficacy. This data pre-processing stage resulted in setting the stage and laying the foundation for the next step of feature engineering.

Feature Engineering: Data cleaning and preprocessing stages laid the foundation and the dataset is ready for feature engineering. Vectorization techniques are employed here for transforming the textual nature data into a structured format and multidimensional feature space for the understanding of ML, NN and DL models. Different types of vectorizations are performed from traditional TF-IDF, advanced distributed word embeddings like Word2Vec and GloVe followed by state of the art context aware BERT embeddings. All these techniques are selected to identify and capture different nuanced facets of linguistic context, semantic meanings, and emotional depth from the online discourses with regards to the mental health domain analysis. The implementation of different vectorization techniques resulted in machine readable feature vectors that allows the model to learn different features, trends, patterns with respect to mental distress indicators and enhances the overall capability of the model to learn from the data corpus. Feature engineering outputs four matrices such as TF-IDF (27975, 1012), GloVe (27975, 300), Word2Vec (27975, 500), and BERT (27975, 768), encapsulating term demonstration, global co-occurrences, semantic relationships, and contextual representations, respectively. These different feature spaces generated from the different vectorization techniques are used as an input layer for the models. All the ML and NN models are trained on all the vectorization feature spaces. This meticulous approach provides insights into comparative analysis of model's efficacy and impact of vectorization techniques in identifying the mental distress indicators. Feature engineering not only makes the

deeper understanding of the textual corpus of data, but it also improves the model performance and optimization for the NLP usecase of text classification. Multiple different types of models from varied categories are trained with different vectorization configurations. The comparative analysis of various model configurations, model types and categories are evaluated in the following section.

5.3 Model Output Evaluation

In this section, output from different categories of models will be evaluated, followed by comparative analysis of different configurations regarding vectorization, pre-processing datasets, and categories of models.

Note: While running the same model pipelines for comparative analysis of models and separate model evaluation with similar codebase architecture and parameters, slight metric value variations can be seen in floating point computation of 0.20% to 0.30% irrespective of seed settings. This is common due to the technical differences and compatibility between AMD processor and Nvidia GPU with Cuda configuration for parallel processing. Rest assured, it does not compromise the model's accuracy, reliability, and validity. All the implemented model pipelines, investigative analysis, datasets used and all the technical implementations performed in this thesis are systematically organized and made easily accessible in the Jupyter notebooks and folders. This ensures complete transparency, ease of access followed by reproducibility of the conducted implementation for the review and future research.

5.3.1 Traditional Machine Learning Models

This section provides the insights into the best traditional machine learning models with respect to NLP usecase of text classification in the domain of mental health analysis.

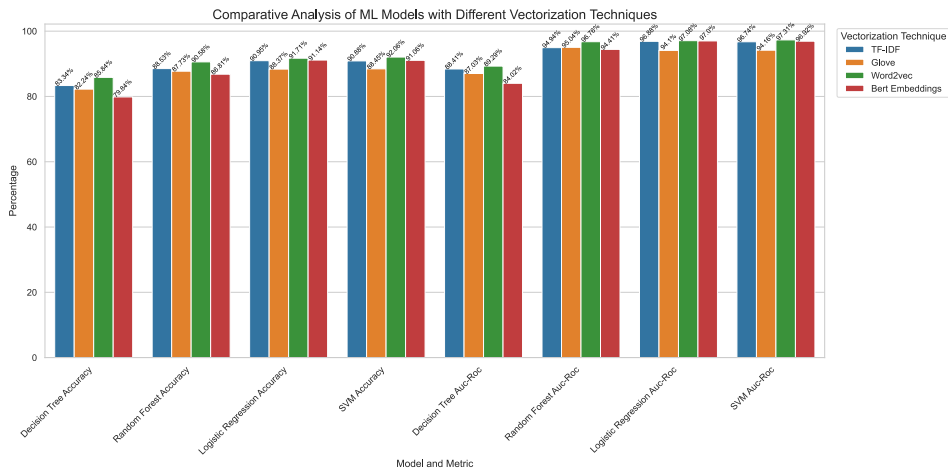


Figure 5.9: Comparative Analysis of Vectorization Techniques for ML Models

Vectorization Technique \ Models	Decision Tree	Random Forest	Logistic Regression	Support Vector Machines
TF-IDF	Accuracy: 83.34 Auc-Roc: 88.41	Accuracy: 88.53 Auc-Roc: 94.94	Accuracy: 90.95 Auc-Roc: 96.88	Accuracy: 90.88 Auc-Roc: 96.74
Glove	Accuracy: 82.24 Auc-Roc: 87.03	Accuracy: 87.73 Auc-Roc: 95.04	Accuracy: 88.37 Auc-Roc: 94.10	Accuracy: 88.45 Auc-Roc: 94.16
Word2vec	Accuracy: 85.84 Auc-Roc: 89.29	Accuracy: 90.58 Auc-Roc: 96.76	Accuracy: 91.71 Auc-Roc: 97.08	Accuracy: 92.06 Auc-Roc: 97.31
BERT Embeddings	Accuracy: 79.84 Auc-Roc: 84.02	Accuracy: 86.81 Auc-Roc: 94.41	Accuracy: 91.14 Auc-Roc: 97.00	Accuracy: 91.06 Auc-Roc: 96.92

Figure 5.10: Different Model's Performance Metrics

The above Figure 5.9 followed by Figure 5.10 represents the comparison of traditional machine learning models based on different vectorization techniques. Each ML model here is trained on 4 different vectorization configurations like TF-IDF, Word2vec, Glove and Bert Embeddings. The comprehensive comparative analysis reveals that the best performing configuration among different vectorization methods on ML models is word2vec technique with highest accuracy followed by Auc-Roc scores for all models and highlighted in dark green color. This showcases that the use of advanced distributed word embeddings in the NLP usecase like text classification is very effective. Analysis also reveals that among all ML models, SVM is the best performing model. Overall, the results of this analysis reveals that the choice of right vectorization method has a great impact over models efficacy in NLP usecase of text classification.

Decision Tree Model:

```
Average Classification Metrics with Standard Deviation (Decision Tree):
Precision: 0.857 ( $\pm 0.004$ )
Recall: 0.857 ( $\pm 0.004$ )
F1-Score: 0.857 ( $\pm 0.001$ )
Accuracy: 0.858 ( $\pm 0.001$ )
Mean AUC-ROC: 0.893 ( $\pm 0.005$ )

Average Confusion Matrix:
[[2432 396]
 [ 396 2371]]
```

Figure 5.11: Classification Report

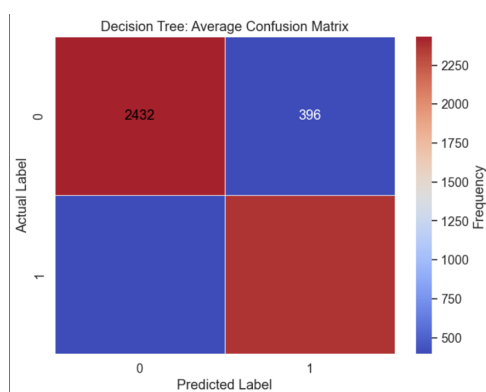


Figure 5.12: Confusion Matrix

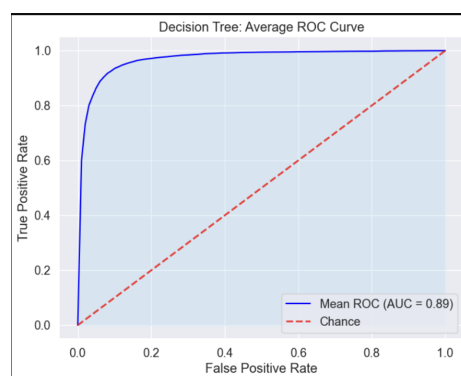


Figure 5.13: AUC-ROC Curve

The above Figure 5.11 is the classification report representing the best configuration decision tree model with word2Vec feature extraction technique for distinguishing between two labels "No distress = 0" and "Distress = 1". The evaluation of the model results in a precision, recall followed by f1-score of 85.7% with consistent standard deviation showcasing constant performance. Model's accuracy resulted in 85.8% followed by the mean Auc-Roc score of 89.3% can also be seen in the figure 5.13 presenting the models reliability. Figure 5.12 showcasing confusion matrix with 2432 true positives followed by 2371 true negatives. Here, false positives and negatives have the same value of 396. Overall, Word2Vec based decision tree model evaluation shows balanced classification report compared to TF-IDF, Glove and Bert Embeddings based decision tree model, hence demonstrating reliable predictions for our NLP text classification task.

Random Forest Model:

```
Average Classification Metrics with Standard Deviation (Random Forest):
Precision: 0.905 (±0.004)
Recall: 0.905 (±0.006)
F1-Score: 0.905 (±0.002)
Accuracy: 0.906 (±0.002)
Mean AUC-ROC: 0.968 (±0.001)

Average Confusion Matrix:
[[2565 263]
 [ 264 2503]]
```

Figure 5.14: Classification Report

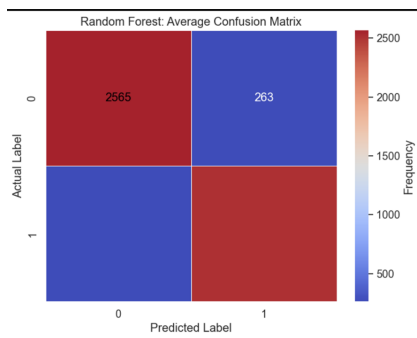


Figure 5.15: Confusion Matrix

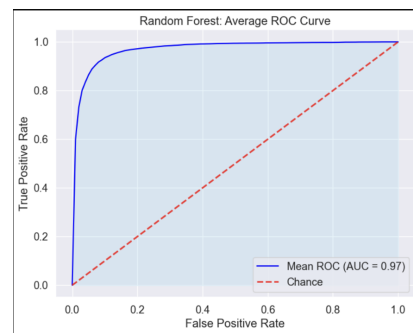


Figure 5.16: AUC-ROC Curve

Like decision tree, here also Word2vec based random forest configuration model is best performing model for our NLP classification usecase in distinguishing precisely between "No distress = 0" and "Distress = 1" labels. Above Figure 5.14 shows the reliable performance of the model's classification report with high precision as well as recall with value of 90.5% with stable standard deviation. The model also achieves a balanced F1-score of 90.5%. The model's evaluation results in high accuracy of 90.6% followed by the AUC-Roc of 96.8% and the same can be seen in the Figure 5.16. The Figure 5.15 demonstrates the confusion matrix of the model with 2565 as true positives followed by 2503 true negatives. The count value of false positives and false negatives is quite low as 263 and 264 respectively. The above model evaluation metrics results highlight the models robustness and outstanding overall performance with strong discriminative capability between labels. This is the best random forest model compared to other Random Forest variants based on TF-IDF, Glove and Bert Embeddings as well as better than the decision tree model.

Logistic Regression Model:

```
Average Classification Metrics with Standard Deviation of Logistic Regression:  
Precision: 0.904 (±0.002)  
Recall: 0.932 (±0.004)  
F1-Score: 0.918 (±0.002)  
Accuracy: 0.917 (±0.002)  
Mean AUC-ROC: 0.971 (±0.001)  
Average Confusion Matrix:  
[[2553 274]  
 [ 189 2578]]
```

Figure 5.17: Classification Report

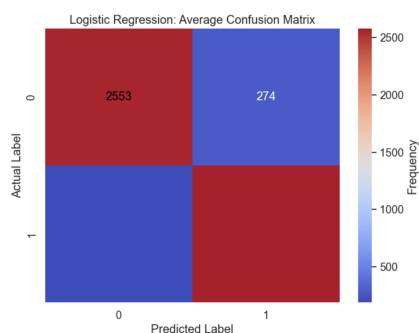


Figure 5.18: Confusion Matrix

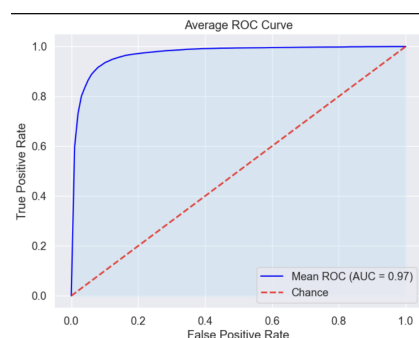


Figure 5.19: AUC-ROC Curve

The logistic regression models evaluation reveals that, word2vec based configuration is best performing model among Logistic Regression variants based on TF-IDF, Glove and Bert Embeddings in differentiating between "No distress = 0" and "Distress = 1" labels. The Figure 5.17 represents the evaluation metrics values showcasing precision with 90.4% and recall with 93.2% followed by balanced F1-Score of 91.8% highlights the stable relationship with no to negligible standard variation across the folds. The model results in accuracy of 91.7% followed by AUC-ROC curve of 97.1% shown in the Figure 5.19 demonstrating the models high predictive capability as well as understanding between labels identification. The above Figure 5.18 gives insights into the confusion matrix with the true positives of 2553, true negatives of 2578 followed by false positives of 274 and false negatives of 189 which are quite low in count. The above evaluation metrics results showcase the models robust capability in identification of labels as well as reliability in prediction. Also, the results reveal that LR model is the successor and better performing than both decision tree and random forest model.

Support Vector Machine:

```
Average Classification Metrics with Standard Deviation (SVM):
Precision: 0.907 (±0.003)
Recall: 0.931 (±0.004)
F1-Score: 0.919 (±0.003)
Accuracy: 0.918 (±0.003)
Mean AUC-ROC: 0.971 (±0.002)

Average Confusion Matrix:
[[2563 265]
 [ 191 2576]]
```

Figure 5.20: Classification Report

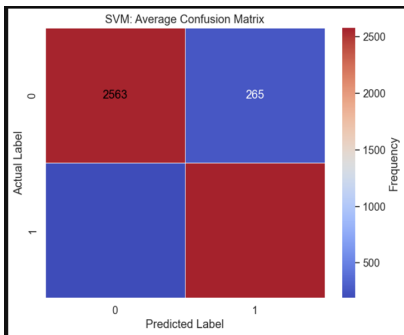


Figure 5.21: Confusion Matrix

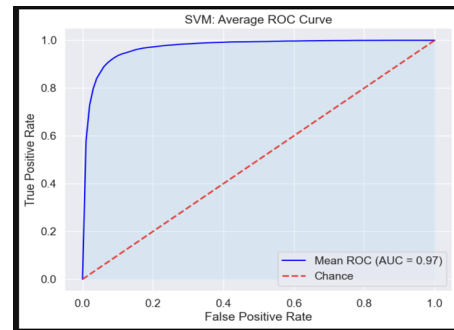


Figure 5.22: AUC-ROC Curve

The Support Vector Machine model evaluation reveals that, word2vec based SVM is the best performing SVM model among other variants based on TF-IDF, Glove and Bert Embeddings in differentiating between "No distress = 0" and "Distress = 1" labels. The classification report in the above Figure 5.20 reveals the precision with 90.7% and recall with 93.1% followed by balanced F1-Score of 91.9% showcasing stability for both classes with no to negligible standard deviation for all the folds. The model performs with accuracy of 91.8% followed by the AUC-ROC curve of 97.1% as shown in Figure 5.22. The Figure 5.21 highlights the confusion matrix with true positives of 2563 and true negatives of 2576 followed by false positives of 265 and false negatives of 191 which are quite low. The above SVM model's evaluation metric results highly demonstrate the model's overall capability in robust prediction and identification of two different labels.

Other than that, this Word2vec configured SVM model is the best performing model among other previously trained traditional Machine Learning models such as Decision Tree, Random Forest, and Logistic Regression because of its kernel based

architecture that uses hyperplane to capture [Song and Diederich, 2013] word2vec’s semantic context [Johnson and Karthik, 2021] effectively . This showcases the practical usability in accordance with the complex and nuanced text classification task.

5.3.2 Advanced Neural Network Models

In the previous section, the best traditional ML models have been evaluated and this following section will give insights into the progression of advanced neural network models with respect to the domain of mental health analysis with the NLP usecase of text classification.

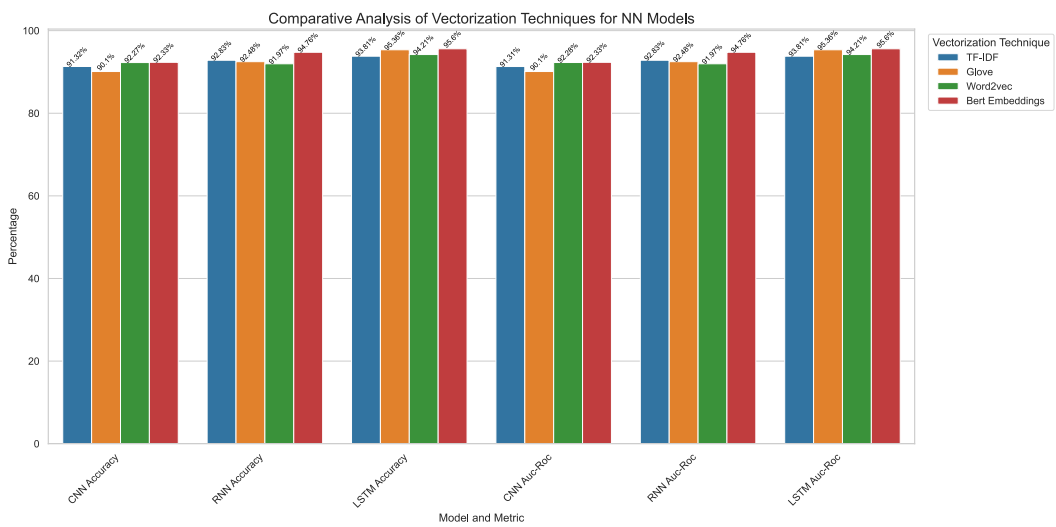


Figure 5.23: Comparative Analysis of Vectorization Techniques for NN Models

Vectorization Technique \ Models	CNN	RNN	LSTM
TF-IDF	Accuracy: 91.32 Auc-Roc: 91.31	Accuracy: 92.83 Auc-Roc: 92.83	Accuracy: 93.81 Auc-Roc: 93.81
Glove	Accuracy: 90.10 Auc-Roc: 90.10	Accuracy: 92.48 Auc-Roc: 92.48	Accuracy: 95.36 Auc-Roc: 95.36
Word2vec	Accuracy: 92.27 Auc-Roc: 92.28	Accuracy: 91.97 Auc-Roc: 91.97	Accuracy: 94.21 Auc-Roc: 94.21
BERT Embeddings	Accuracy: 92.33 Auc-Roc: 92.33	Accuracy: 94.76 Auc-Roc: 94.76	Accuracy: 95.60 Auc-Roc: 95.60

Figure 5.24: Different Model’s Performance Metrics

Advanced neural network models evaluation in the above Figures 5.23 and 5.24 clearly highlights that all CNN, RNN and LSTM models with the configuration of BERT Embeddings as feature engineering technique shows greater performance and is highlighted in dark green colour compared to the TF-IDF, Word2Vec and Glove based feature representation techniques. Among all the three models, the LSTM model has the highest performance with the accuracy and auc-roc both of 95.60%. RNN is the second-best performing model with accuracy and auc-roc both of 94.76%. Also, CNN model shows high performance with accuracy and auc-roc both of 92.33%. It is evident from the results evaluation of advance neural network models that Bert Embeddings configuration is the better choice for these sophisticated models to perform more better when they are leveraged with high dimensional contextual and semantic information relationships for nuanced textual classification usecase in the domain of mental health analysis.

Convolutional Neural Network:

```

Average Precision: 0.9046 ± 0.0266
Average Recall: 0.9014 ± 0.0305
Average F1-Score: 0.9022 ± 0.0076
Average Accuracy: 0.9034 ± 0.0076
Average ROC AUC: 0.9033 ± 0.0075

Average Confusion Matrix:
[[2559 267]
 [ 272 2494]]

```

Figure 5.25: Classification Report

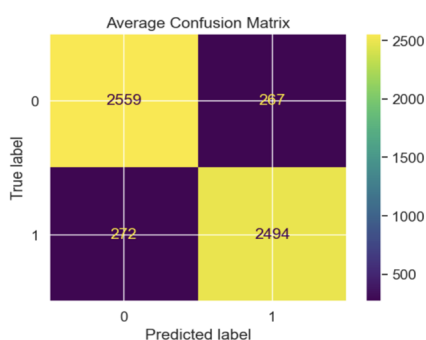


Figure 5.26: Confusion Matrix

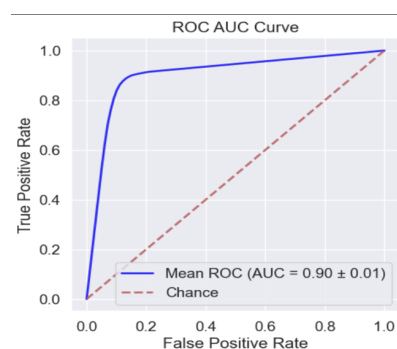


Figure 5.27: AUC-ROC Curve

The CNN model evaluation results that, Bert Embeddings feature engineering based

CNN model is the best performing CNN model among the other variants based on TF-IDF, Word2Vec and Glove techniques in differentiating between "No distress = 0" and "Distress = 1" labels. The Figure 5.25 highlights the consistent precision, recall and F1-score of 90.46%, 90.14%, 90.22% respectively with no to negligible standard deviation showcasing balanced relationship. The models perform with average accuracy of 90.34% and an average Auc-Roc curve of 90.33% seen in the Figure 5.27 demonstrating the consistent predictive performance as well as strong indication of distinguishing capabilities. The Figure 5.26 presents the confusion matrix with true positives of 2559 and true negatives of 2494 followed by the false positives of 267 and false negatives of 272 showcasing strong classification performance with minimal misclassification. Overall, the model shows stable and consistent performance, and it clearly represents that Bert Embeddings makes the CNN's capability more nuanced to handle the textual data quite effectively.

Recurrent Neural Network:

```

Average Precision: 0.9462 ± 0.0377
Average Recall: 0.9447 ± 0.0309
Average F1-Score: 0.9450 ± 0.0285
Average Accuracy: 0.9454 ± 0.0287
Average ROC AUC: 0.9868 ± 0.0107
Average Confusion Matrix:
[[2675 152]
 [ 153 2614]]

```

Figure 5.28: Classification Report

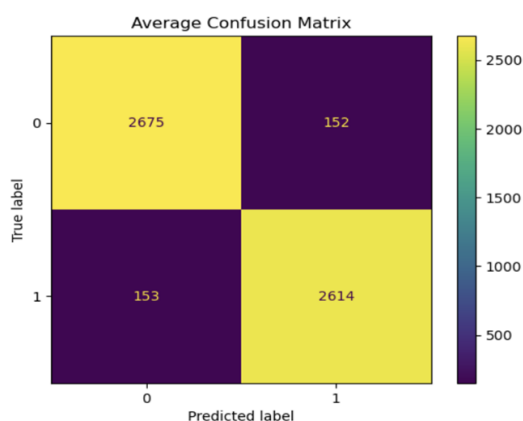


Figure 5.29: Confusion Matrix

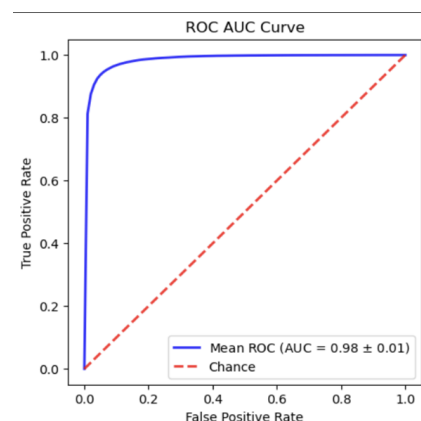


Figure 5.30: AUC-ROC Curve

The RNN model's evaluation results that, BERT Embeddings based RNN model is the best RNN configuration performing model among the other variants based on TF-IDF, Word2Vec and Glove techniques in differentiating between "No distress = 0" and "Distress = 1" labels. The Figure 5.28 represents the classification results with high precision of 94.62%, recall with 94.47% and a balanced f1-score of 94.50%. The model performs with high accuracy of 94.54% and auc-roc curve as seen in the Figure 5.30 of 98.68%. All the evaluation metrics show constant performance over all the folds with no to negligible standard deviation. The figure 5.29 highlights the confusion matrix of the model with true positives of 2675, true negatives of 2614 followed by false positives of 152 and false negatives of 153 showcasing the strong predictive stability of the model. Overall, RNN model exhibits strong capability in robust prediction and adaptability in identification of different labels demonstrating the models understanding of nuanced and complex mental health domain language patterns and is better performing than CNN model as well as traditional ML models.

Long Short-Term Memory:

```

Average Precision: 0.9577 ± 0.0298
Average Recall: 0.9543 ± 0.0411
Average F1-Score: 0.9557 ± 0.0328
Average Accuracy: 0.9565 ± 0.0319
Average ROC AUC: 0.9897 ± 0.0111
Average Confusion Matrix:
[[2710  116]
 [ 126 2640]]

```

Figure 5.31: Classification Report

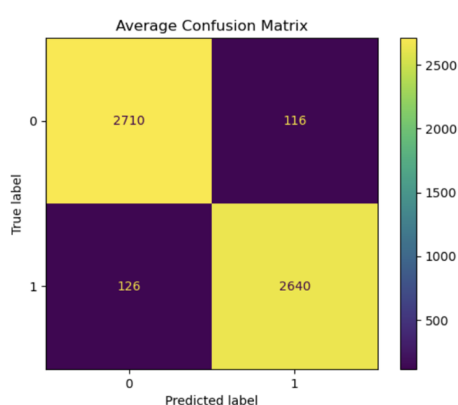


Figure 5.32: Confusion Matrix

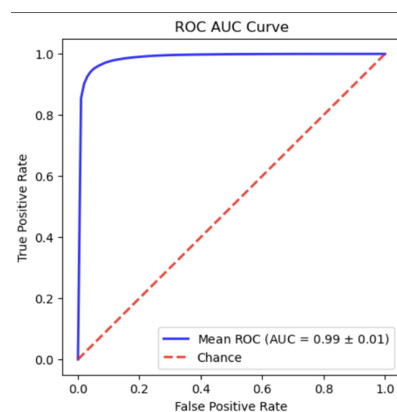


Figure 5.33: AUC-ROC Curve

The LSTM models evaluation results that, Bert Embeddings based LSTM model configuration outperforms all other LSTM variants based on TF-IDF, Word2Vec and Glove techniques in differentiating between "No distress = 0" and "Distress = 1" labels. The Figure 5.31 represents the classification report of the LSTM model with high precision and recall of 95.77% and 95.43% respectively followed by f1-score of 95.57% showcasing a highly balanced relationship. The model performs with high accuracy of 95.65% followed by the AUC-ROC curve of 98.97% and the same can be seen in Figure 5.33. All the evaluation metrics show exceptional and consistent performance across all the folds with no to negligible standard deviation. The Figure 5.32 highlights the average confusion matrix of the model with true positives of 2710 and the true negatives of 2640 followed by the relatively quite low false positives of 116 and false negatives of 126.

Overall, the LSTM model demonstrates high performance showcasing its reliability and robustness in predictions and capability in identifying different labels effectively. Among other advanced neural network models discussed before like CNN and RNN, LSTM outperforms both of them and traditional ML models such as DT, RF, LR and SVM too. This reveals its superiority to capture the long-term dependencies from the contextual semantic textual information and its advanced capability to deal and process the sequential data [Singh et al., 2022]. Of course, BERT Embeddings feature representation is the better choice here with these sophisticated advanced neural network models for processing highly nuanced contextual patterns and information from the data [Zhu et al., 2020].

5.3.3 State-of-the-art Transformer based Deep learning Models

This section highlights the ultimate progression of the models and evaluation compared to the previously discussed traditional ML models and advanced neural network models evaluation. It provides insights into the state-of-the-art transformer based deep learning models like BERT and domain specific BERT variants with respect to the usecase of NLP text classification in the domain of mental health analysis.

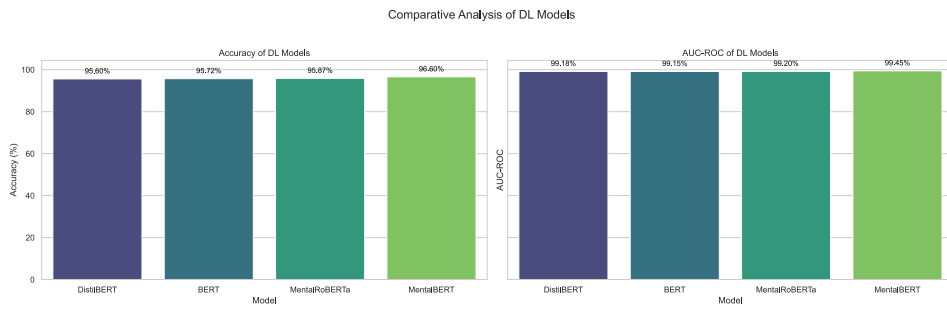


Figure 5.34: Comparative Analysis of DL Model Variants

Metrics\Models	DistilBERT	BERT	MentalRoBERTa	MentalBERT
Accuracy	95.60 %	95.72 %	95.87 %	96.60 %
AUC-ROC	99.18 %	99.15 %	99.20 %	99.45 %
Training Loss	0.16	0.17	0.16	0.14
Validation Loss	0.11	0.11	0.12	0.09

Figure 5.35: Different Model’s Performance Metrics

The above Figures 5.34 and 5.35 provides the insights into the state-of-the-art transformer based deep learning models like BERT and its variants including domain specific variants. Here all the 4 models like DistilBERT, BERT, MentalRoBERTa and MentalBERT have been end-to-end fine-tuned for the NLP text classification use-case. The model evaluation results highlight that all the deep learning models show exceptional performance and have surpassed traditional ML models as well as advanced NN models. Out of the 4 DL models, MentalBert highlighted in dark green colour shows the extraordinary performance with the highest accuracy of 96.60% followed by the Auc-roc of 99.45% which is outstanding.

Overall, performance of DL models showcases the high capability and reliability to understand and classify the highly contextual and complex semantical nuanced language patterns from the textual data corpus [LeCun et al., 2015]. Also, the highest performance of MentalBERT DL model indicates that due the specific mental domain variant and pre-training on mental data, it has maximum edge to understand the complex linguistic patterns and the main BERT architecture behind it, is developed to analyse and handle complex textual data [Ji et al., 2022]. This analysis highlights that for the usecase of NLP text classification in the mental health analysis domain, these state-of-the-art transformer based BERT deep learning models [Wolf et al., 2020] are highly adaptable, reliable and robust providing maximum performance.

MentalBERT Model:

To analyse the MentalBERT model closely, the below Figure 5.36 represents the models training and validation loss of 0.14 and 0.09 respectively which is lowest compared to other DL models indicating promising performance as well as it's effective learning and generalization capabilities.

```
Epoch 1 - Training Loss: 0.1412643752587737, Validation Loss: 0.09272549198541258
```

Figure 5.36: Model's Training and Validation Loss

```
Classification Report:
      precision    recall  f1-score   support

     0       0.96     0.97     0.97     2857
     1       0.97     0.96     0.97     2738

 accuracy          0.97     5595
 macro avg       0.97     0.97     0.97     5595
 weighted avg   0.97     0.97     0.97     5595

Confusion Matrix:
[[2785  72]
 [ 118 2620]]

Epoch 1
Accuracy: 0.966041108132261
Precision: 0.9732540861812778
Recall: 0.9569028487947406
F1: 0.9650092081031306
ROC AUC: 0.9945085092092443
```

Figure 5.37: Classification Report

The above Figure 5.37 represents the classification report of the MentalBERT end-to-end fine-tuned model. It has been trained for an epoch. The results are exceptional and impressive with the precision, recall and f1 score all at around 97% showcasing balanced relationship. The MentalBERT model performs with the accuracy of 96.60% and AUC-ROC curve of 99.45% can be seen in the below Figure of 5.39.

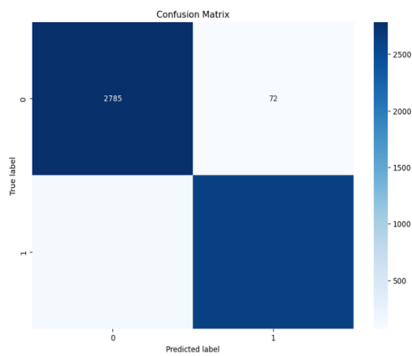


Figure 5.38: Confusion Matrix

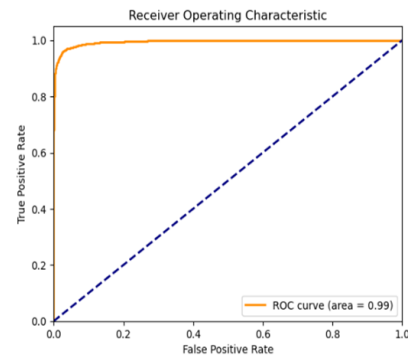


Figure 5.39: AUC-ROC Curve

The above Figure 5.38 represents the confusion matrix of the model with true positives of 2785, true negatives of 2620 followed by relatively very low mis-classification instances of false positives with 72 and false negatives with 118.

Overall, after analysing the models evaluation results, it is clear that the MentalBERT deep learning model outperforms all the other deep learning model variants followed by all the traditional ML as well as advance NN models trained so far in this thesis and has proven exceptional performance showcasing the models capability in robust identification of different labels between "no distress = 0" and "distress = 1" followed by the models consistent reliability in correct prediction of instances [Bokolo and Liu, 2023]. All deep learning models are high performing models, but this specific domain variant MentalBERT model has the edge because of its pre-training of the similar mental health data from the various platforms like Twitter and Reddit [Ji et al., 2022]. It is understood that deep learning models followed by domain specific pre-trained models are highly adaptable for understanding complex and nuanced linguistic patterns from the textual data and are proven to be superior for the usecase of NLP text classification [Devlin et al., 2018]. These types of models can be confidently leveraged in the domain of mental health analysis [Greco et al., 2023].

5.4 Model Category-wise Performance Evaluation

After analysing all the best configuration models for each model type from all the model categories, now this section will provide grand comparative analysis of the best models from all the categories.

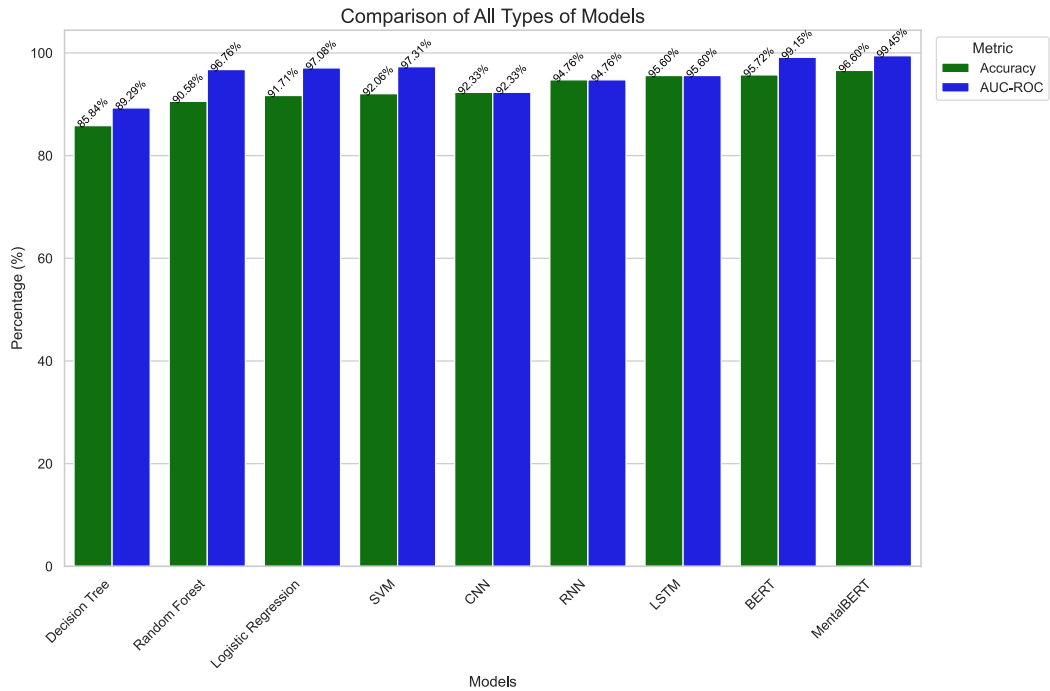


Figure 5.40: Comparative Analysis of All Model Categories

Metrics\ Models	Decision Tree	Random Forest	Logistic Regression	SVM	CNN	RNN	LSTM	BERT	MentalBERT
Accuracy	85.84	90.58	91.71	92.06	92.33	94.76	95.60	95.72	96.60
AUC-ROC	89.29	96.76	97.08	97.31	92.33	94.76	95.60	99.15	99.45

Figure 5.41: Model Performance Metrics

After modelling and analysing around 32 models which includes traditional ML models, advanced NN models and state-of-the-art DL models, the above Figure 5.40 highlights the best performing models from each type and category. Here, the progression trend in performance is seen when moving from left to right, that is from ML to NN to DL models.

The Figure 5.41 highlights the same left to right progression with metrics of each model from each category with accuracy and auc roc curve and has been marked in different colours showcasing the best model in that category. The comparative analysis revealed that among all the categories of ML, NN and DL models, ML is the least performing category at the 3rd position with the highest performing model as SVM with an accuracy of 92.06% followed by an AUC-ROC of 97.31% highlighted

in yellow colour. NN models show promising results compared to ML models and stand at 2nd position with LSTM as best performing model in this category with an accuracy of 95.60% followed by a matching AUC-ROC curve highlighted with green colour. DL models show exceptional results surpassing both ML and NN models with MentalBERT as the best performing model in this category with an accuracy of 96.60% followed by an AUC-ROC of 99.45% highlighted in dark green colour.

The detailed comparative analysis results proves that the progressive trends from traditional machine learning models to advance neural network models to the state of the art and domain specific deep learning models is the key to achieve significantly enhanced performance in the NLP usecase of text classification. It's evident from the above analysis that, as the complexity of the model's architecture evolves and increases the model's interpretation and understanding of nuanced linguistic patterns, trends from textual corpus also improves and categorization ability becomes better.

5.5 Impact of Different Pre-Processing Techniques

This section will provide insights into the evaluation of implying different pre-processing techniques on the best performing deep learning MentalBERT model.

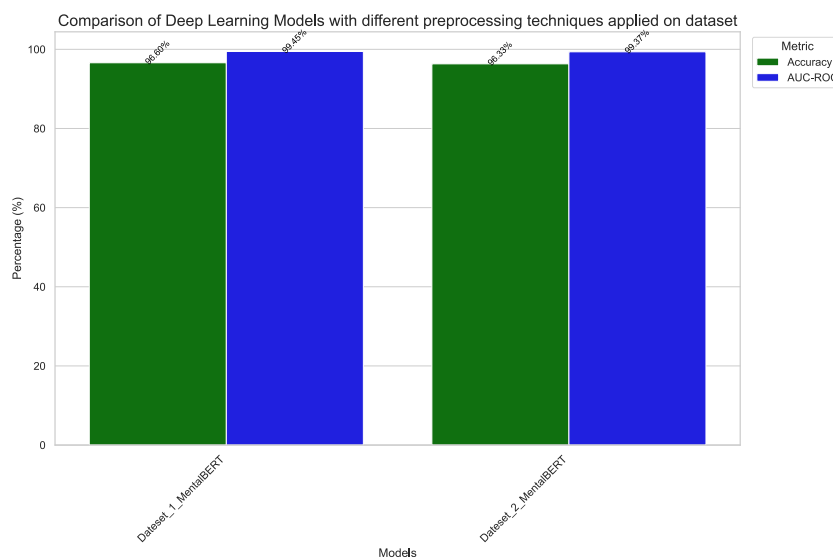


Figure 5.42: Impact of Different Pre-processing Techniques on Model's Performance

Models Metrics	MentalBERT (Dataset 1 - Tokenization, stopwords removal & lemmatization)	MentalBERT (Dataset 2 - Tokenization)
Accuracy	96.60	96.33
AUC-ROC curve	99.45	99.37

Figure 5.43: Preprocessing Techniques Impact on Model's Performance Metrics

The above Figure 5.42 and 5.43 shed light on the impact of different preprocessing techniques on deep learning MentalBERT model. Dataset 1 is aggressively pre-processed with stopwords removal, lemmatization, and tokenization whereas the Dataset 2 is only pre-processed with tokenization. Though the results with chosen dataset for this thesis shows very low difference and impact in overall performance. The metrics with Dataset 1 is 96.60% and an AUC-ROC of 99.45% highlighted in green colour is compared to Dataset 2 metrics 96.33% and an AUC-ROC of 99.37%. The difference between the performance metrics is relatively low, yet aggressive preprocessing for the Dataset 1 enhances the model performance by some extent [Abbe et al., 2016]. This can differ from model to model and dataset to dataset, but in this case the results are closely similar with negligible impact because of the choice of transformer based models. These types of models have advanced capabilities to tackle the extensive need for the preprocessing of datasets before training. This is likely due to the complex and robust internal understanding of the linguistically context related text and nuanced semantic language meanings during its pre-training on data or domain specific data [Chai, 2023].

5.6 Combined Analysis Engine

For classifying each text instance with the similar domain and sentence structure like the original dataset, the combined analysis engine is developed here. It loads MentalBERT model, which is the best performing model in all the aspects. The textual instance will be passed here to the end points of the trained model for prediction and probability to analyse the classification into "no distress" or "distress" labels. To get the more interpretable understanding and additional context to support the decisions, two additional pretrained models have been deployed here which are pre-trained on mental health data platforms like Reddit and Twitter to analyze the sentiment and emotional state from the text instances. By leveraging these two pretrained models with our end-to-end fine-tuned MentalBERT model, this combined analysis engine will not only classify the text between labels but also provides the additional insights into the overall sentiment state of that instance [Camacho-Collados et al., 2022] followed by the emotional states expressed in that instance [Lowe, 2023] with classification and determined probabilities to interpret

the main model's results better. This kind of combined analysis engine is more useful when highly mixed to very complex emotions are involved. For instance, a user is trying to provide a positive review of the film and also talks about how they relate to some negative character in the movie and respective incidents. In such cases model might be confused and may classify either of distress labels, but with analysis of additional pretrained models, overall sentiment and emotional state of the textual instance can be gauged to understand the mental state of the user between not distressed or distressed.

```

Enter your text for analysis: Struggling with self-worth today. Feels like I'm dragging everyone down with me. Can't escape this suffocating loneliness.
Just want to vanish.

Analysis Results:

Distress Detection (MentalBERT Fine-Tuned Model): Result: Distress Detected,
Prediction Probabilities: No Distress Detected: 0.007307864725589752, Distress Detected: 0.9926921725273132,

Sentiment Analysis (Pretrained Model): Negative: 0.9274, Neutral: 0.0653, Positive: 0.0073,

Emotional State (Pretrained Model): Sadness: 0.5589, Disappointment: 0.4219, Annoyance: 0.0942, Neutral: 0.0612, Desire: 0.0596, Disapproval: 0.027, Nervousness: 0.0148, Realization: 0.0127, Anger: 0.0117, Disgust: 0.0098, Grief: 0.0078, Approval: 0.0074, Caring: 0.0069, Curiosity: 0.0057, Remorse: 0.0051, Confusion: 0.0047, Joy: 0.0044, Optimism: 0.0043, Love: 0.0043, Embarrassment: 0.0039, Fear: 0.0039, Excitement: 0.0033, Surprise: 0.003, Admiration: 0.0022, Relief: 0.0021, Amusement: 0.0018, Gratitude: 0.0007, Pride: 0.0006,

```

Figure 5.44: Classification Analysis

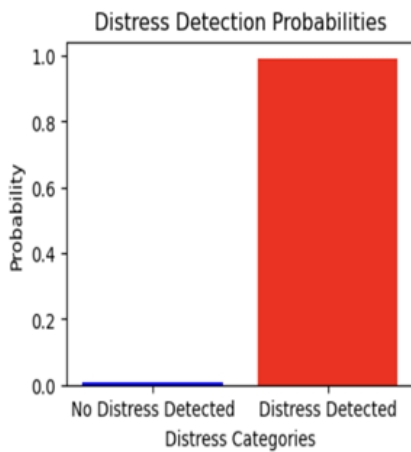


Figure 5.45: Distress Analysis

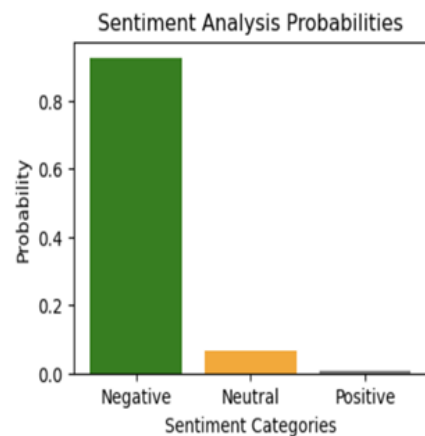


Figure 5.46: Sentiment Analysis

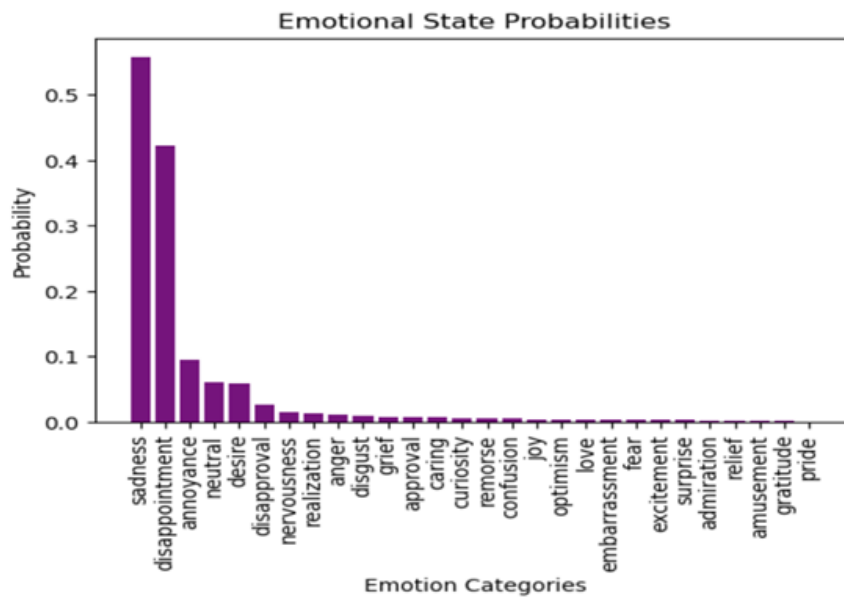


Figure 5.47: Emotion Analysis

The above Figure 5.44 represents the results of the combined analysis engine, where a text instance is passed through the end to end fine-tuned MentalBERT model followed by the other two pre-trained models analysing overall sentiment tone and the emotional states of the textual instance. For a better interpretability and visualization of the same results, Figure 5.45 shows the bar chart from the MentalBERT model, Figure 5.46 shows the sentiment analysis predicted by the pre-trained model and plotted the same with bar chart followed by the Figure 5.47 representing different emotional states identified by the pre-trained model and plotted the distribution with bar chart.

The instance analysed here can be seen in the Figure 5.44 with its results and visualizations are presented in the Figure 5.45, 5.46 and 5.47. The text sample evaluated is "Struggling with self-worth today. Feels like I'm dragging everyone down with me. Can't escape this suffocating loneliness. Just want to vanish". The end-to-end fine-tuned MentalBERT model predicted the sample as a distress sample with high confidence. The sentiment pre-trained model also classified the instance with high predominance in negative sentiment. The emotional state pre-trained model identified "sadness" as the most prevailing emotional state followed by the "disappointment" and "annoyance" present too.

This detailed interpretation provides nuanced understanding behind the models decision by correlating the identified distress or no distress with respect to overall sentiment tone and emotional state providing more context and understanding with regards to mental health analysis. This detailed analysis with visualizations

is highly valuable when considered for real world applications to understand the complete spectrum of the mental health state.

5.7 Interpretability Analysis

This section highlights the mental distress indicators identified by models and highlights insights into the model's decision-making characteristics.

5.7.1 Feature Importance and Coefficient weights of ML Models

As traditional machine learning models are straightforward and reliable for interpretation [Stiglic et al., 2020], feature importance of decision tree and random forest models are explored. Coefficient weights analysis of logistic regression and SVM model are explored. Also, the decision tree is visualized as pdf for the decision tree model and can be found in the appendix due to its size. By interpreting the DT and RF model's decision-making process, they becomes transparent [Kang et al., 2022].

```

Decision Tree Feature Importance:
feature      life      want  redflag      film      feel      kill  \
importance  0.288273  0.138473  0.09529  0.075252  0.066918  0.066038

feature      anymore  die  suicidal  movie      man      dont  \
importance  0.047139  0.040421  0.035087  0.034167  0.022205  0.018999

feature      end  im going  one      live      say  want end  \
importance  0.014493  0.013972  0.01227  0.012192  0.011469  0.011434

feature      know  think      ive  stuff  story  im going  \
importance  0.011355  0.010743  0.010334  0.010208  0.01008  0.00984

feature      one      live      say  want end  know  think  \
importance  0.009441  0.009242  0.009084  0.008835  0.00864  0.008542

feature      maybe  cant  time  try  pill  make  \
importance  0.008457  0.008376  0.008314  0.008261  0.008173  0.007927

feature      want die      go  care  going  lost  need  \
importance  0.007882  0.007786  0.007786  0.007659  0.007573  0.007468

feature      right  lonely  im  life  day  night  \
importance  0.007356  0.007295  0.007174  0.007092  0.00709  0.007081

feature      want  feel
importance  0.006978  0.00697
    
```

Figure 5.48: DT Feature Importance

```

Random Forest Feature Importance:
feature      life      feel  redflag      want      kill      die  \
importance  0.034695  0.028438  0.027596  0.027255  0.026976  0.022185

feature      anymore  im  cant  suicidal  feel like  film  \
importance  0.020957  0.014833  0.014246  0.012465  0.012218  0.012101

feature      going  movie  know  depression  and  ive  \
importance  0.011555  0.011116  0.011042  0.01005  0.009669  0.009536

feature      live  help  pain  everything  family  even  \
importance  0.009235  0.008116  0.007862  0.00682  0.006749  0.00662

feature      nothing  get  like  thought  alone  year  \
importance  0.006452  0.006014  0.006009  0.005897  0.005769  0.005755

feature      tired  tried  feeling  cannot  never  living  \
importance  0.005715  0.005613  0.005431  0.005388  0.005308  0.005181

feature      character  thing  worse  depressed  would  way  \
importance  0.004752  0.004729  0.004485  0.004476  0.004466  0.004466

feature      go  killing  want die  care  much  pill  \
importance  0.004462  0.004436  0.004343  0.004337  0.004133  0.004081

feature      time  think
importance  0.003941  0.003935
    
```

Figure 5.49: RF Feature Importance

The above Figures 5.48 and 5.49 represents the top 50 feature importance analysis and provides insights into the decision tree and random forest models. By this analysis, key tokens are identified that have different degrees of influence on the model's classification process. From the above Figure 5.48, it is seen that the word "life" has

the significant and highest importance followed by other words like "want", "red-flag", "kill", "im going", "die" and other words highlighting their impact and influence in classification of mental distress. Similarly, from the above Figure 5.49, it is seen that across the wide set of words random forest assigns close to relatively equal weights to words like "life", "feel", "feel like", "pain", "redflag", "depressed", "want die", "suicidal" and other words highlighting that random forest assess more wider distributed approach for decision making with larger context.

From the feature importance results analysis, it is clear that both the DT and RF models have nuanced understanding of factors and are able to identify indicators to classify the instance between "distress" and "no distress" state. Due to its architecture, the decision tree prioritizes a certain set of tokens whereas the random forest model focuses on more of a holistic set of tokens that are very crucial to capture, understand and identify the linguistic complexity of natural language associated with the domain of mental health analysis.

```

Top positively influencing features:
Feature  redflag  kill  suicidal  die  pill  killing \
Coefficient 8.749411 7.066664 5.603119 5.09553 4.381883 4.372365

Feature  cannot  life  alive  anymore  end  depression \
Coefficient 4.207644 4.182972 4.079781 3.714624 3.69641 3.555305

Feature  feel  job  living  pain  tried  worse \
Coefficient 3.337042 3.320602 3.218523 3.137681 3.098469 3.05313

Feature  tired  tonight  live  point  alone  depressed \
Coefficient 2.955997 2.952683 2.946192 2.941628 2.912419 2.847161

Feature  therapy  death  feel like  want  method  hospital \
Coefficient 2.834109 2.781972 2.752403 2.742322 2.666716 2.646395

Feature  jump  ending  suffering  hang  goodbye  medication \
Coefficient 2.641971 2.548891 2.539981 2.498539 2.470838 2.434057

Feature  way  gun  seems  family  done  thought \
Coefficient 2.430002 2.429935 2.419315 2.389002 2.388162 2.386261

Feature  going  deserve  get better  help  nothing  painful \
Coefficient 2.349875 2.335692 2.284894 2.264464 2.262537 2.24928

Feature  im done  last
Coefficient 2.228517 2.20297
    
```

Figure 5.50: LR Label 1 Coeff. Weights

```

Top negatively influencing features:
Feature  theyre  white  good  rn  bit  example \
Coefficient -1.527446 -1.536593 -1.53797 -1.548281 -1.558335 -1.562457

Feature  watched  performance  yes  play  war  american \
Coefficient -1.577364 -1.59605 -1.643254 -1.649621 -1.651514 -1.652055

Feature  lol  join  plot  power  great  teen \
Coefficient -1.689366 -1.706167 -1.740067 -1.749801 -1.751718 -1.754671

Feature  first  show  actor  man  played  fact \
Coefficient -1.759952 -1.774936 -1.782945 -1.789588 -1.82382 -1.835585

Feature  also  story  playing  scene  test  gay \
Coefficient -1.878129 -1.910818 -1.930436 -1.932111 -1.997648 -2.014636

Feature  fun  song  teacher  class  bored  teenager \
Coefficient -2.081136 -2.302038 -2.384193 -2.398869 -2.41751 -2.419561

Feature  dm  cool  seen  like  guy  boy \
Coefficient -2.46745 -2.498213 -2.514246 -2.558355 -2.60204 -2.730322

Feature  girl  school  yell  kinda  character  crush \
Coefficient -2.884231 -2.938181 -2.963543 -3.005981 -3.356922 -3.464025

Feature  movie  film
Coefficient -5.124507 -6.713608
    
```

Figure 5.51: LR Label 0 Coeff. Weights

To understand the logistic regression model's decision making process, refer the image above Figure 5.50 and Figure 5.51 where the top 50 coefficient weights analysis is presented. Here, it's evident that most influencing features for the positive class that is distress label have several tokens with weight distribution. Among them "redflag", "kill", "sucidial", "die", "killing", "depression", "suffering", "im done", "feel like" and more words are strong indicators associated with presence of mental distress. Whereas the Figure 5.51 represents the negative class with negative weights showcasing the indicators that are less likely or not associated with mental distress. Such tokens are "good", "school", "performance", "lol", "actor", "film", "fun", "play"

and other words. This kind of coefficient analysis provides the transparent decision-making process and indicates that what kinds of words or tokens are acting as indicators and influence the model's decisions to classify the text between distress and no distress.

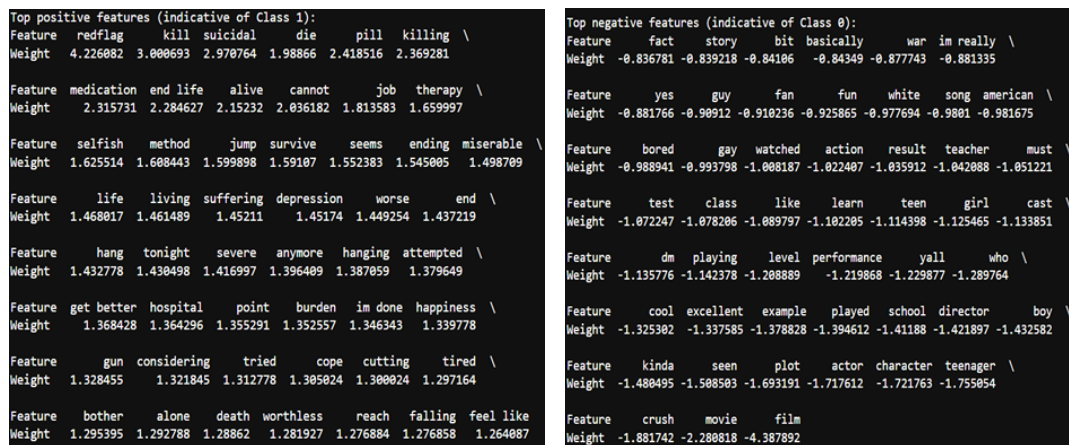


Figure 5.52: SVM Label 1 Coeff. Weights Figure 5.53: SVM Label 0 Coeff. Weights

Similarly to understand the decision making process of the SVM model, refer the above Figures 5.52 and 5.53 representing top 50 coefficient weight analysis for both the labels. For the positive class that is label 1 distress, top positive features are "end life", "suicidal", "medication", "attempted", "die", "killing", "alone", "worthless", "hang", "jump" and other words are strongly influencing the model to classify them with mental distress acting as indicators directly associated with mental distress. Whereas the above Figure 5.53 highlights the negative features that are less likely or not associated with mental distress. Such top features are "film", "story", "fact", "im really", "watched", "cool", "excellent", "played", "song" and other words. This coefficient weights positive and negative analysis is very important and crucial for the understanding of a model's decision-making process and looking at the models own understanding of the indicators, trends and patterns that are associated with mental distress which plays an important role in influencing the model's decisions between classifications.

5.7.2 Attention Mechanism and SHAP Analysis of DL Model

Interpretation and decision making process of traditional ML models is straightforward compared to advance NN and state-of-the-art DL models. They are usually seen as black boxes due to its internal advance and complex [Rudin, 2019] architecture. As our advanced NN models were trained with the BERT embeddings fea-

ture extraction technique, it is more logical to directly interpret and understand the decision-making process of our best and sophisticated DL model MentalBERT. Due to the complexity and technical limitations, attention weight [Niu et al., 2021] mechanism and SHAP analysis [Aldughayfiq et al., 2023] are explored here to study the decision-making process and models understanding in the domain of mental health analysis. These methods give us a way to look inside the black box models and see the classification process and models prediction to a certain extent.

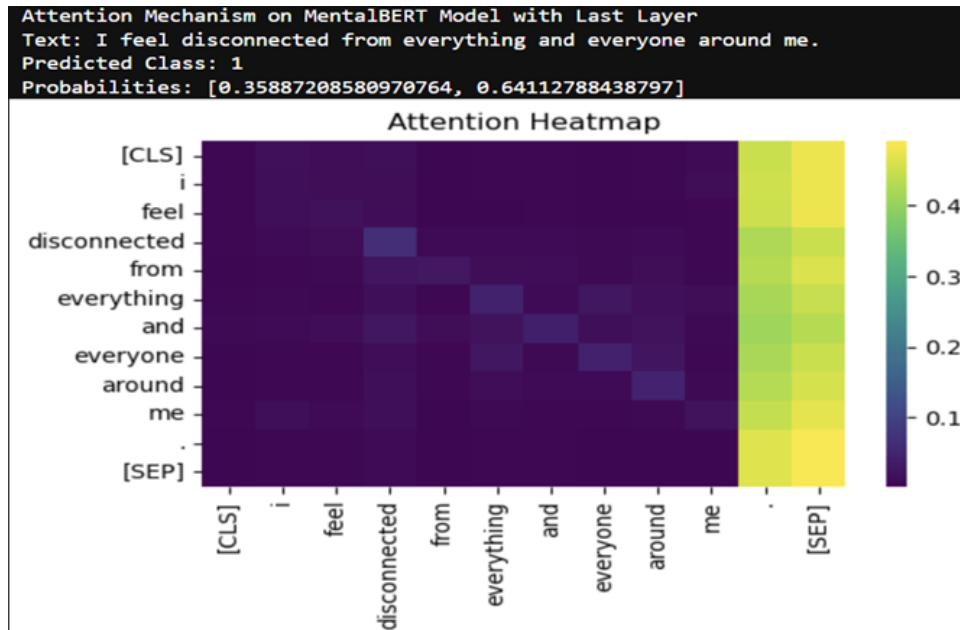


Figure 5.54: Attention Mechanism on MentalBERT with Last Layer

The above Figure 5.54 is a heatmap that represents the attention mechanism analysis of the MentalBERT model that focuses on the weights of the model from the last layer. Heatmap demonstrates the importance of each, and every token or words processed by the model. Here the instance is "I feel disconnected from everything and everyone around me". For instance, heatmap shows that higher attention levels are indicated with the darker shades where the tokens or words like "disconnected", "everyone", as well as "around" receive most of the focus. It highlights their strong contextual relationship to the instance's sentiment. Here, the model predicts this instance to class 1 with notable confidence as seen from the probabilities and that label is a distress label. This kind of attention weight mechanism with heatmap visualization provides the insights into this kind of sophisticated black box models interpretability and highlights the words or the key phrases which influences the decision of labels prediction.

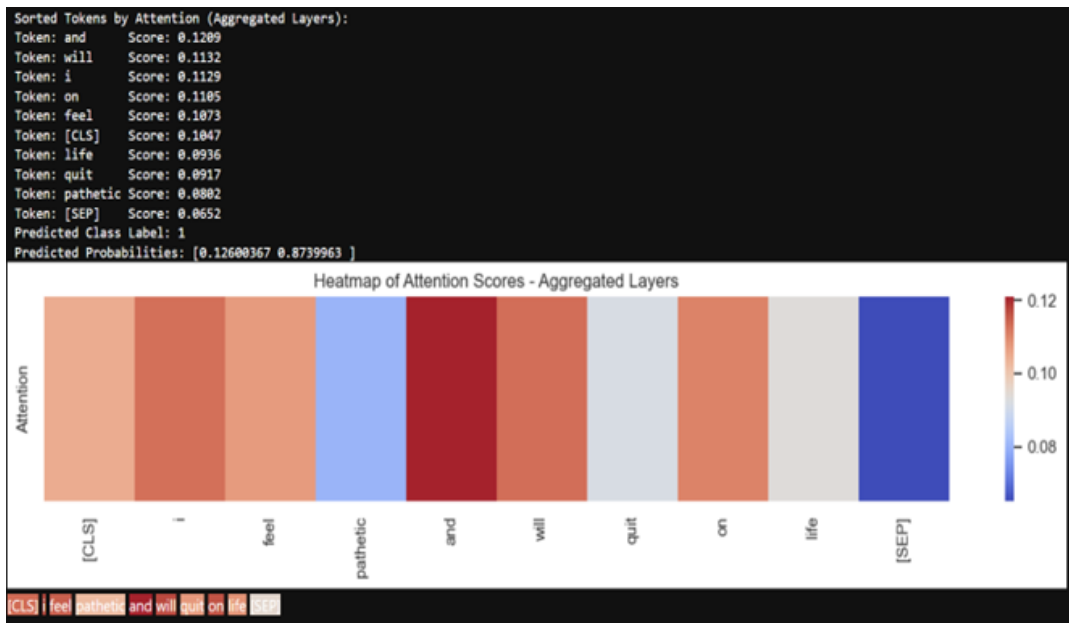


Figure 5.55: Attention Mechanism on MentalBERT with Aggregated Layers

The above Figure 5.55 represents the attention mechanisms results analysis from the aggregated layers of the MentalBERT model providing insights of how different tokens or words are influencing and contributing in the decision making process for the final prediction. From the heatmap, it is clear that tokens or words like "feel", "pathetic", and "quit" have higher attention scores followed by the colour grading of instance in the end highlighting that "pathetic", "quit" and "life" is indicating a contextual relationship and overall model predicts this instance into the distress label 1 with high confidence as seen from the probability distribution between labels. Other colour intensity in the heatmap highlights the context relevant tokens or model views it as similar weights tokens or influencing words. This aggregated layer attention mechanism analysis is highly informative and provides a transparent decision-making process for instances and provides a high level of interpretability.

Another interpretability technique to understand the decision making process and most influencing tokens or words on model's prediction are analysed by the SHAP technique. The below Figure 5.56 is the SHAP analysis visualization with Colour coding to weight the scale accordingly followed by the classification output probabilities and the distribution of the highly influencing words placed on the scale. The tone of red colour indicates weights pointing towards distress label and blue tone indicators represent the opposite no distress label.

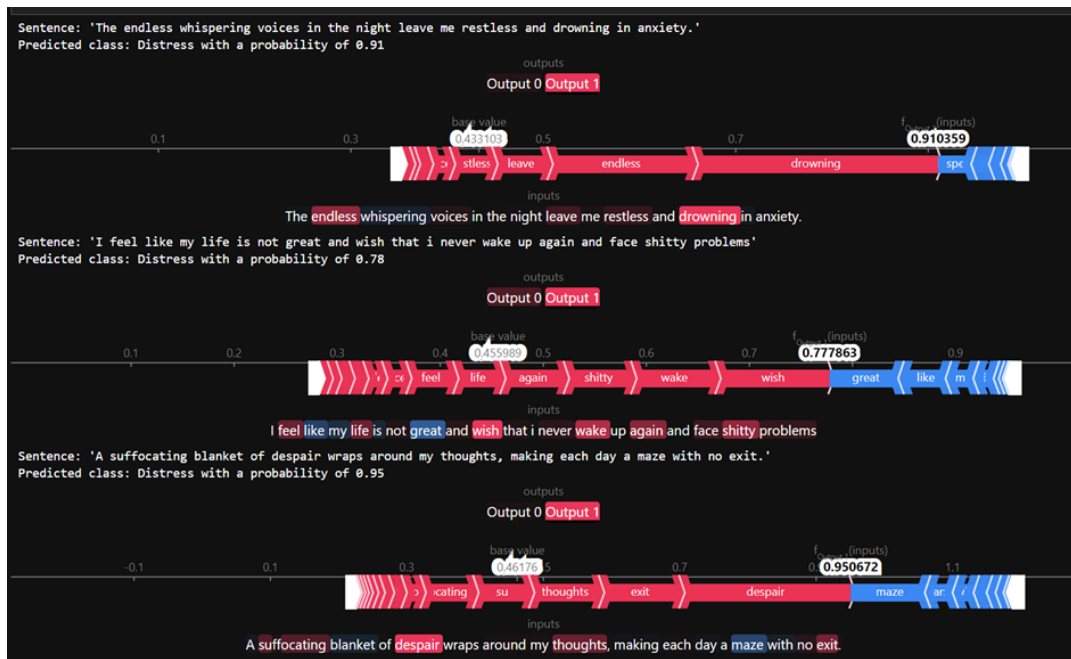


Figure 5.56: SHAP Analysis of MentalBERT Model on Distressed Instances

In the above Figure 5.56, three complex text instances are analysed with the MentalBERT model with SHAP interpretability analysis. The visualization provides insights on each word that influences the model's prediction. The first instance "The endless whispering voices in the night leave me restless and drowning in anxiety." is predicted to the label 1 of distress. The model's distress result and the expected result distress is achieved with the high confidence as probabilities are 0.91%. SHAP visualization reveals that words heavily weighted by the model are "drowning", "restless", "leave", "voices" followed by other tokens. The heavily weighted words clearly associate with a psychological context related to distress exhibiting language.

Instance 2, text is "I feel like my life is not great and wish that i never wake up again and face shitty problems". The visualization provides the insights into highly weighted words by the model are "wish", "wake", "shitty", "again", "feel", "life" and others. The overall prediction is towards the distress label with notable confidence in probability of 0.78%. It can be clearly seen that the model is identifying contextual relationships and indicates that the model is capable of identifying mental distress indicators from the nuanced and complex linguistic sentences.

In instance 3, text is "A suffocating blanket of despair wraps around my thoughts making each day a maze with no exit." The visualization provides insights into highly weighted words by the model are "despair", "suffocating", "thoughts", "exit"

and others. The overall prediction is towards a distress label with high confidence of 0.95%. Model is capable of detecting high likelihood of distress by identifying contextual relationship and semantic meanings.

The results and analysis of above 3 text instances were correct predictions made by the model showcasing its nuanced linguistic capability to detect the mental distress indicators and evaluating its expressions with high accuracy. Practical value of the SHAP analysis is also visualized here to interpret the complex and black box deep learning models. This analysis highlights better interpretation as well as insights to keep the model's decision making transparent, reliable for accurate validation and providing rationale behind the prediction due to the sensitive domain.



Figure 5.57: SHAP Analysis of MentalBERT Model on No Distress Instances

The above Figure 5.57 represents the SHAP analysis of MentalBERT model on the No Distress Instances. The investigative analysis reveals that all 3 sentences are classified towards no distress label as expected. In the first sentence "Had a chill night with friends, feeling pretty good about it", words like chill, pretty, good, night, friends have the highest influence on the model to classify it to be no distress. For the second sentence "Woke up on time today, coffee's hitting just right and I'm actually looking forward to our meeting", words like coffee, actually, our, meeting, today are highly influencing the model to classify this instance to the no distress label. The third sentence "Dude, finally after ages I picked up that book which I've been ignor-

ing for a pretty long time", words like dude, book, picked, pretty are top influencing tokens that makes the model to predict the output towards no distress class. From the SHAP result's analysis, it is clearly seen that highly influencing words are related towards neutral to positive state of mind, showcasing that MentalBERT model is capable of identifying robustly between different linguistic context based mental states indicators. Model classified all the three instances to no distress class with high confidence probabilities ranging from 0.93%, 0.94% and 0.96% respectively.

This interpretability analysis shows the reliable, accurate and robust performance of the MentalBERT model. It demonstrates that transformer based deep learning models are highly capable to identify, capture, differentiate and understand complex and nuanced linguistic patterns and semantic meanings with regards to mental health domain [Joyce et al., 2023].

5.8 Misclassification Analysis

To understand the model's capabilities, limitations and shortcomings, misclassification analysis is explored here [Jeatrakul et al., 2010]. Misclassified instances for both the labels that are diverged from the original labels are analysed to identify and understand the complex and intricate dynamics of the interlinked sentiment, emotional and our classification mental health label states. These instances are analysed by the use of two pre-trained models based on mental data from the Reddit and Twitter platform. The results provide insights into the models capability to navigate these kinds of complexities and evaluate the model from a different scientific perspective.

	text	label	predicted_label	sentiment_analysis	emotion_detection
0	feel like someone need hear tonight feeling ri...	0	1	negative: 0.365,neutral: 0.494,positive: 0.141	optimism: 0.417,caring: 0.257,admiration: 0.24...
1	tried put sugar coffee back spoon happy monday...	1	0	negative: 0.009,neutral: 0.114,positive: 0.877	joy: 0.685,neutral: 0.104,excitement: 0.089,de...
2	hello anyone feeling suicidal hear people lyn...	0	1	negative: 0.867,neutral: 0.125,positive: 0.008	neutral: 0.614,sadness: 0.215,disappointment: ...
3	think need actual help okay depress lately eve...	0	1	negative: 0.913,neutral: 0.076,positive: 0.011	fear: 0.757,nervousness: 0.379,sadness: 0.096,...
4	want talk support group wife cant handle much ...	1	0	negative: 0.630,neutral: 0.353,positive: 0.016	optimism: 0.426,neutral: 0.312,desire: 0.161,d...

Figure 5.58: Misclassification Instances Evaluation With 2 Pre-Trained Models

The above Figure 5.58 represents a glance from a dataset that includes the prediction from the other two pre-trained models that are used to understand and analyse the misclassification instances by the MentalBERT model. The first pre-trained model does the sentiment analysis on the instances and classifies them between positive, negative, and neutral states [Camacho-Collados et al., 2022]. The second pre-trained

model represents the presence of different emotional states from a wide set of emotions [Lowe, 2023]. All the predictions are followed by their probabilities and this analysis provides insights into nuanced contextual linguistics and identification of trends and patterns associated with mental health.

```
Mean Scores for Misclassified Distressed Instances into no distress:
label 1.0, misclassified 1.0, negative 0.53801, neutral 0.338426, positive 0.163249, annoyance 0.086117, sadness 0.085426, gratitude 0.073086, disappoint
ment 0.062695, anger 0.055041, amusement 0.052416, approval 0.036381, fear 0.032188, admiration 0.031995, disapproval 0.031726, optimism 0.031178, love
0.027777, joy 0.027655, confusion 0.026919, caring 0.025701, disgust 0.021609, desire 0.017183, remorse 0.016482, nervousness 0.015472, realization 0.015
041, curiosity 0.012609, excitement 0.011863, surprise 0.010614, embarrassment 0.005782, relief 0.00302, grief 0.002528, pride 0.001142, predicted_label
0.0
```

Figure 5.59: Mean Scores for Misclassified Distressed Instances into No Distress

The above Figure 5.59 represents the mean scores of misclassified distressed instances into no distress labels. Analysis of these instances reveals that there is a negative sentiment state present with the probability of 0.538 showcasing that there is a negative sentiment present but not enough to predict it to the distress labels. From the analysis of results, it can be seen that, there are mixed sentiments present here with probability of 0.338 for neutral followed by the positive sentiment probability of 0.163. It is understood from here that the model might struggle when the probabilities are the border line between these sentiments revealing that there is a need for a neutral label to classify highly mixed contextual sentiments to be predicted in neutral state. The emotional state analysis revealed that, there is a prevailing presence of annoyance, sadness and disappointment linking to negative or distressful states followed by some neutral and positive states found like gratitude and amusement. So, both the pre-trained model's analysis suggests that neutral labels will tackle the negligible problem of misclassification.

```
Mean Scores for Misclassified No Distress Instances into Distress:
misclassified 1.0, predicted_label 1.0, negative 0.670133, neutral 0.241437, positive 0.096656, sadness 0.188367, annoyance 0.098576, disappointment 0.09
6888, anger 0.075375, fear 0.066156, joy 0.034742, caring 0.033984, disapproval 0.033531, gratitude 0.030375, love 0.029018, nervousness 0.025633, optimi
sm 0.025464, approval 0.025419, disgust 0.024185, realization 0.023818, desire 0.023638, remorse 0.021448, admiration 0.018659, confusion 0.017669, amuse
ment 0.011531, embarrassment 0.009096, excitement 0.008422, curiosity 0.007768, grief 0.005359, relief 0.004685, surprise 0.004516, pride 0.001169, label
0.0
```

Figure 5.60: Mean Scores for Misclassified No Distress instances into Distress

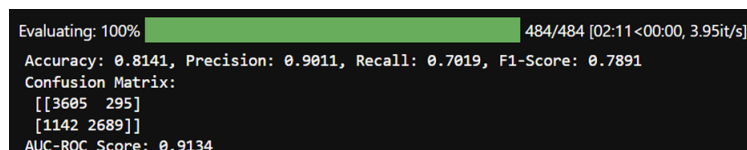
The above Figure 5.60 represents the mean scores for misclassified no distress instances into distress labels. Analysis of these misclassified instances reveals that there is a high negative sentiment score with 0.670 found. When compared with the opposite group discussed above, neutral, and positive states are low here. The emotional state analyses reveals that these emotions found like sadness, annoyance, disappointment, fear and anger are dominating emotional states. These analyses of both the results demonstrates that there is a combination of negative sentiment states followed by the distressful emotional states associated with mental distress.

From the above misclassification results analysis, it is understood that models rely primarily on the distressful indicators for the prediction. Analyses, also reveals that there is a need for a neutral label in classification as there can be cases where the sentiment is overall negative just because it's a bad review of any product, movies or negative opinion does not always mean distress. Analyses highlights that the model is sophisticated and capable of robust performance, due to which in the case where no distress were classified as distressed are initially wrong because of original labels. But with these analyses it is found that there are dominating negative sentiment, and emotions present in these instances. From this trend and qualitative analysis suggests that, there can be some mislabelled data in the original dataset. Due to the domain specific pre-training of the MentalBERT model, it is capable of detecting and predicting the instances into the correct labels demonstrating its extensive learning of complex and nuanced mental distress indicators surpassing the expectations.

Overall, the investigative analysis also highlights that to improve the classification and to excel ahead with the performance and nuanced understanding of the model, there is a need for a neutral label which acts as a bridge for those instances who are around the borderline between no distress and distress.

5.9 Validation Through External Unseen Datasets

To validate the MentalBERT model's real world adaptability and generalizability, external mental domain datasets are used to perform validation and this section will provide the highlights of the results.



```
Evaluating: 100% ██████████ 484/484 [02:11<00:00, 3.95it/s]
Accuracy: 0.8141, Precision: 0.9011, Recall: 0.7019, F1-Score: 0.7891
Confusion Matrix:
[[3605 295]
 [1142 2689]]
AUC-ROC Score: 0.9134
```

Figure 5.61: External Validation Evaluation on Dataset 1 of Depression

In the above Figure 5.61, it represents the evaluation results from the first external dataset validation. This dataset is centered around depression, a domain of mental health [Kaggle, 2024a]. The results reveal that the trained MentalBERT model is robust and proficient to identify, detect and classify the complex and nuanced instances. Total 7731 unique instances were present in this dataset and evaluation results in accuracy of 81.41% followed by Auc-Roc curve of 91.34%. The F1 score is 78.91%. Overall, these metrics highlight that MentalBERT's capability to understand and classify the instances of different context than that of training data shows high reliability and further shows its capability to detect depression from complex instances.

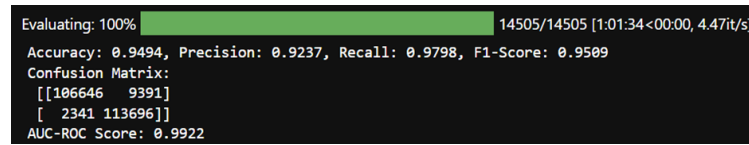


Figure 5.62: External Validation Evaluation on Dataset 2 of Suicide

To explore further, another challenging dataset was chosen centering suicidal ideology [Kaggle, 2024b]. This dataset has 232,074 unique instances that are complex and will definitely push the boundaries of MentalBERT models analytical capabilities. The above Figure 5.62 represents the validation results on this dataset is exceptional and surpassed the expectations with achieving an accuracy of 94.94% with precision of 92.37% and recall of 97.98% followed by F1-Score of 95.09% showcasing balanced relationship and consistent performance. The model also achieved an impressive AUC-ROC score of 99.22% demonstrating its robustness identification between labels. This validation results on this dataset reveals that MentalBERT is not only capable of detecting distress but is also capable of detecting suicidal ideology which has quite intricate, nuanced and complicated linguistic context interlinked with the mental health domain.

The process of these external validations of the MentalBERT model were conducted on the datasets which were distinct, unique and different in context compared to the model's training data demonstrates the versatility and effectiveness of the model's performance with regards to the mental health domain. The training data fed to the model was centered around the labels of distress and no distress laying the foundational understanding of the mental health domain. Model learned the nuanced complex patterns, trends and understood the semantic context from the corpus and became capable to push the boundaries further and extrapolate different scenarios in the domain of mental health related depression and suicide. This validation evaluation shed light upon the model's intricate understanding of the subtleties with respect to the mental health domain.

This remarkable external validation results reveals model's real-world generalizability and adaptability in the domain of mental health analytics for potential to significantly contribute in diagnostics for wide spectrum of mental health problems [Zhu et al., 2020], [Ermeres et al., 2020].

6 Discussion

This chapter provides the insights into the discussion of the obtained results, sheds the light upon applications and implications of deep learning models for mental health analysis on social media discourses, ethical considerations, and finally critical reflection on the thesis's research.

6.1 Analysis and Integration of Findings

Comprehensive analysis conducted on the traditional machine learning models, advanced neural network models and state-of-the-art deep learning models to identify most effective models and their efficacy in identifying distress indicators from the text that are associated with mental distress state. Comparative analysis of ML models such as DT, RF, LR, SVM followed by NN models such as CNN, RNN, LSTM and ultimate progression to DL models such as BERT, DistilBERT, MentalRoBERTa and MentalBERT was performed. ML and NN models were trained on different vectorization techniques such as traditional TF-IDF, advance distributed word embeddings such as Word2Vec and Glove followed by state-of-the-art context aware BERT Embeddings configurations. Also, different preprocessing techniques were implemented to find the most effective set of techniques influencing better model performance. In total, 32 models were trained with different types and configurations and 2 other pretrained models were used for the effective misclassification analysis. This thesis research not only identify the most effective techniques and best performing model with different computational and architectural configurations but also proves that ML and NLP techniques are capable of capturing, identifying, interpreting, and successfully classifying complex nuanced linguistically context dependent semantic meanings from the text corpus by learning features, trends, patterns, and indicators. This thesis lays a foundational ground especially by providing insights into the state-of-the-art technologies that can be very relevant for the future work due to the ongoing revolution in the field of AI as well as how further it can be leveraged for the practical applications in the domain of mental health analysis.

Exploratory data analysis performed in the initial stage proved to be an important step, as it revealed the datasets characteristics, label distribution, text length, word distribution and insights into the containing noise and irregular patterns in the dataset. This guided the data cleaning to tackle and remove the noise, unwanted

patterns of HTML tags as well as standard cleaning techniques were applied. This further prepared the dataset for preprocessing. Due to insightful EDA and data cleaning, the preprocessing stage comprised of stopword removal, tokenization, and lemmatization on the dataset 1 for transforming it into model analysable form. To study the impact of different preprocessing techniques in preserving the essential linguistic context and elements associated with mental distress, dataset 2 was also created with applying tokenization.

Different visualizations created in the EDA phase were quite insightful like word cloud of the entire dataset, N-gram, Bi-gram and Tri-gram analysis highlighting words, expressions, patterns associated with mental health discussions guiding the further stages like cleaning, preprocessing as well as feature engineering to ensure that important features are retained that are truly associated with the mental health context. This strategic approach and sequence of stages enhanced the modelling and classification of different types of models and categories achieving high performance. This demonstrates that different NLP techniques and models with different configurations provide different variation in efficacy and effectiveness in identifying significant mental distress indicators by processing the text.

Comprehensive and comparative analysis among ML with different configurations reveals that, all the ML models like DT, RF, LR and SVM trained with the configuration of Word2Vec have outperformed all other model configurations trained on TF-IDF, Glove and BERT Embeddings. This is due to Word2Vec's characteristics of capturing the contextual relationships between words. Here, TF-IDF is ineffective due to its simple characteristics of just counting word frequencies without understanding and capturing context related to complex semantic meanings. Glove is ineffective due its characteristics of static representation of word relationships which does not have the dynamics to process the contextual sensitivity required to understand the complexities. BERT Embeddings are not effective rather they are not utilized completely by the ML models due to its architectural complexity and context rich representations [Birunda et al., 2021]. Out of all 4 models and their configurations, SVM is the top performing model with the accuracy 92.06%. This is due to the SVM's inherent robust capability to handle and process the high dimensional data as well as non-separable linear data. This is quite the challenge when dealing with the textual nature of the data, so DT, RF and LR did not perform well in comparison to SVM. Usually, DT and RF tend to overfit on the textual data whereas LR model struggles to handle and process the textual data that is high dimensional and sparse in nature [Jain et al., 2021]. Hence, SVM configured with Word2Vec performed well and is reliable for the classification by detecting and interpreting complex linguistics associated with mental health discourses.

Moving towards NN models analysis, it reveals that models such as CNN, RNN and LSTM performed best with the BERT embeddings. This is because of the BERT embeddings effective and deep contextual understanding of the mental distress indicators from the data as well as they were maximally utilized by the sequential

processing capability of the advanced neural networks. TF-IDF is way too simple for neural networks as it does not hold contextual nuances. Word2Vec and Glove were able to capture the semantic relationship and offer performance, yet it could not match up with dynamic context specific embeddings offered by BERT embeddings [Zeberga et al., 2022]. Out of all NN models, LSTM model excels at performance with the accuracy 95.60% due its inherent capability of long-term textual dependencies that is quite crucial to interpret and understand the complex context of indicators associated with mental distress. CNN and RNN models offer high performance, yet they have architectural limitations and struggle with the long term dependencies making them less effective compared to LSTM. Overall, LSTM configured with BERT embeddings feature space offers outstanding performance and surpasses all the traditional ML models and other advanced NN models. Also, it stands out as a great model of choice where the resources are limited or there is a high computational cost involved [Kour and Gupta, 2022]. So just by extracting BERT embeddings from the sentence layer and passing it as an input feature to the LSTM model can result in trade off balance between computational resource consumption and performance. But, as it's a matter related to the mental health domain and overarching medical domain, each percentage of increase in performance is crucial and top performing models need to be used.

Deep Learning models comprehensive and comparative analysis reveals that, MentalBERT model is the best performing DL model with the accuracy 96.60% among other DL models like standard BERT, DistilBERT and MentalRoBERTa. The superior performance can be understandable due its pretraining on the mental data from the similar social media platforms like Twitter and Reddit. MentalBERT's pretraining retained context and fine tuning on top of that with our dataset have made this model robust, consistent, and reliable in capturing, identifying, interpreting indicators, patterns, features, and trends that are directly associated with the mental distress [Ji et al., 2022]. These models are high level models due to its inherent state of the art transformer architecture and are capable of understanding simple to complex human-like language and linguistic nuances. It is the entry level large language model [Wolf et al., 2020], [Devlin et al., 2018]. As, it is seen that all the models including MentalBERT were trained on dataset 1 with its preprocessing done with tokenization, stopword removal and lemmatization, MentalBERT was trained additionally and separately again on the similar dataset 2 that just underwent tokenization. The performance gap is around 1% and modest. This is because lemmatization breaks the word down to its lemma form and reduces the linguistic variability and stopword removal eliminates common but like the most uninformative words from the text to make it more refined, removes noise and model can focus its attention on the important context. This prepares the dataset, with directly having core content that the model can focus and interpret directly without breaking it down further. Though, dataset 2 did not went under all three preprocessing techniques, but it performed well and better than ML and NN models because of its inherent capabilities of the transformer architecture and specialized training of textual data provided the

edge and demonstrates that even without extensive preprocessing of the dataset these models are highly capable and efficient of detecting and interpreting nuanced linguistic context and complex variations from the text. But, this can be subjective as per the dataset at hand and selected models. So, always understand the data in EDA and select the techniques as well as experiment accordingly [Abbe et al., 2016], [Chai, 2023], [Khan et al., 2020].

All the trained models including the best DL model MentalBERT are designed for binary classification into "Distress" and "No Distress" due to the dataset characteristics. To provide more nuanced insights, reasoning, and detailed context, two more pretrained models have been used in the combined analysis engine. It includes the best performing MentalBERT model followed by sentiment state pretrained model [Camacho-Collados et al., 2022] and emotional state pretrained model [Lowe, 2023]. So, for instance, a text is passed through all the three models, and it puts the classification labels with their respective probabilities and bar chart visualization. This collaborative approach provides detailed analysis of the text from different perspectives with regards to overall mental health. This collaborative approach of combined analysis engine is quite effective in combining the best trained model with two other pretrained models because of their retained knowledge of mental health data due to its extensive pretraining on millions of Tweets and Reddits. This also provides reliable interpretability and validity of the results.

To enhance the execution of misclassification analysis on the MentalBERT's misclassified instances, two pretrained models from the mental domain were employed to assess the different sentiment and emotional states present in these misclassified instances to understand the reasons behind it. Along with this quantitative analysis, qualitative analysis was also employed in the misclassification analysis. The results reveal that there is a need for a "neutral" label as most of the errors were borderline instances. Also, some of the instances where it was supposed to be "No Distress" according to the original labels and MentalBERT classified them as "Distress" was due to its dominating negative sentiments with around 67% followed by sadness, annoyance, disappointment, and depressing emotions present. This showcases that, the end to end fine-tuned pretrained MentalBERT model has surpassed the expectations and can actually pickup adeptly "Distress" or "No Distress" showcasing that, there might be labelling errors in the original dataset.

Also, interpretability analysis was performed on different models with their possible techniques to understand the model's own understanding and decision making process behind the classification outputs. For DT and RF model, top features were analysed by the feature importance technique and for DT, tree was printed additionally. For LG and SVM, coefficient weights technique was employed to understand the top influential features for both labels identified by these models. As NN and DL models are not straightforward like ML models to interpret [Stiglic et al., 2020] and usually called black boxes [Rudin, 2019], yet there were certain techniques employed. Here, attention mechanisms for the last layer as well as aggregated layers

were employed to understand the individual instance classification via model attention weights to the tokens. To look deeper and understand the decision-making process of DL models, SHAP analysis was applied to the MentalBERT model with 6 different instances from both the labels showcasing local explanations and context dependent influential words pushing the model towards the output label. This proves that the MentalBERT DL model could robustly detect and separate between both the labels and correctly classify most of the time. Misclassification and interpretability analysis provides the trust and reliable foundation of this complex high performing state-of-the-art DL models by providing insights into their decision-making process.

To provide this thesis more strong and foundational ground, external datasets validation was employed to test the models robustness, adaptability and generalizability with similar to relevant data from the mental health domain. This was performed by using two external publicly available high rated and ethically backed datasets centered around depression [Kaggle, 2024a] and suicidal ideology [Kaggle, 2024b]. As these datasets are a bit different from the original dataset centered around mental distress, but the format, labels and domain are similar. So, it's definitely quite challenging for the MentalBERT model to push its boundaries. The results were astonishing and exceptional with high accuracy of 81.41% and 94.94% for both the datasets respectively. This proves the MentalBERT end-to-end fine-tuned DL model's superiority to interpret and classify the real world like data from the overall mental health domain. This demonstrates the model's capability to understand the textual data that contains different context and nuances which pushes its practical applicability into the diverse domain of mental health analysis applications [Ji et al., 2022].

To reflect back on this comprehensive exploration and comparative analysis of different preprocessing techniques, vectorization techniques, model types and categories was a multistage selection approach. It revealed that state-of-the-art DL models with a MentalBERT specific model is the best performing model compared to all other model types and categories of ML and NN. Though in advanced NN models, LSTM is the best performing model followed by SVM model as the best performing model in the traditional ML category. Advanced distributed word embeddings Word2Vec configured with ML models and state-of-the-art context aware BERT Embeddings configured with NN models provided the best performance metrics compared to the traditional TF-IDF and advanced Glove techniques. Moving towards discussing the preprocessing techniques, the dataset aggressively processed with tokenization, stopword removal and lemmatization provided superior performance compared to the dataset with only tokenization applied. Though the difference between the performance is relatively low but preprocessing stage needs to be customized as per the matter at hand. It's evident that deep learning models are the cutting-edge NLP technology and definitely the future because of its strong and nuanced interplay in capturing and interpreting human-like language, like it did here to capture and interpret complicated mental health intricacies [Bokolo and Liu, 2023] and underlying subtleties.

The diversified exploration and comprehensive analysis highlights the need for selection of preprocessing techniques, vectorization techniques followed by model type and category selection is strategic and meticulous in regards with the NLP use case, dataset at hand and its direct application to the domain. By rigorous research, exploration, implementation and comprehensive detailed analysis of different ML and NLP technologies, significant and interesting insights are gained with respect to developing complex NLP systems for mental health analysis on social media platforms. This thesis not only provides insights and advances in the domain of data science and AI with its applicability to interpret and support mental health domain but also towards how these technologies should be considered, selected, used with high level and attention to technical rigor, ethical considerations and real world adaptability by making this research an intra domain research where the different subject matter and domain experts come together to build robust and reliable NLP systems to reduce the clinically interpretations gap.

This thesis proves that deep learning models and that to specifically domain specific models when end- to-end fine-tuned with specific interest dataset provides exceptional high performance results. The diversified methodological approach of this research demonstrates with a strong quantitative and qualitative result proven grounds that these models are highly capable of capturing, detecting, interpreting, and classifying the instances based on mental distress indicators into different labels consistently, accurately and are robust enough to understand the difference and separate them between labels. These models are highly generalizable and can be applied to real world applications with different adaptability requirements in mental health analysis, of course with professional medical experts involved to continuously review and interpret the model's output and use these systems as aid to provide support on the digital spaces. This thesis research study findings integrated with its existing literature ensures that AI backed with ML and NLP have strong practical implications for building applications that can perform mental health analysis by analysing online discourses on the social media platforms. This will open new avenues for exploration and collaboration of different subject and domain matter experts, AI engineers, medical professionals, commercial as well as public organizations to come together and build clinically approved robust AI systems by leveraging cutting edge technologies and benefit from the ongoing AI revolution.

6.2 Applications and Implications of Deep Learning Models for Social Media Mental Health

To address the different kinds of mental health issues on social media platforms, integration of these deep learning models into these platforms can drastically change and shift the scenario. By using these models and its robust and consistent mental distress indicators detecting capabilities from the textual content, real time support

to the distress users can be provided. The support to the distressed users can be provided in real time [Kalyan et al., 2021].

Early detection and timely support: Deployment of DL models on social media for analysis of the users social media content in real time offers early detection of mental distress followed by providing interventions and timely support. [Iyortsuun et al., 2023].

Professional interventions: Automated support systems can be capable enough to identify mental distress indicators especially extreme or severe cases of distress leading to major and critical cases where immediate escalation is required. Here, direct human moderators or mental domain professionals like psychologists can be informed and notified for the intervention to prevent the critical scenarios and ensure users mental state well-being [Zhou et al., 2022].

Constant analysis and adaptability: By constantly analysing these models keeps on evolving and adapting new cases, scenarios and situations and develops a sense of belonging which results in better distress detection capabilities. Continuous adaptation and self-evolution to newer use of social media trends and linguistic context improves the detection robustness and consistent prediction capabilities of model [Garcia-Ceja et al., 2018].

Ethical considerations and implications: Integration of such deep learning models in social media support systems is beneficial, yet it has certain challenges related to ethical considerations and computational resources. Data handling and privacy of user's data should be the top priority and user's consent for processing and analysis of the data should be respected to maintain the trust and autonomy. Also, such systems should be adhering and compatible with the geographical boundary operation laws like GDPR in Europe [Ienca and Malgieri, 2022].

6.3 Ethical Considerations

To research and work with the domain of mental health analysis in accordance with ML, NLP and AI, there is a strict need for adherence to ethical considerations and standards followed by the geographical boundary laws. Here, the data's privacy, security, usability, processing, and analysis should be transparent as well as reviewed constantly to stay within the legal boundary [Thieme et al., 2020].

In this thesis, all the 3 datasets used for training and validation have been chosen with utmost responsibility and considering ethical considerations. They have been selected from the Kaggle datasets platform with a near to perfect usability scores of 10/10 given by the Kaggle organization, a subsidiary of Google itself. The scores evaluation was based on the dataset's completeness, credibility, and compatibility

metrics. All the three datasets are evaluated through the thorough sourcing and the regular updates made available from the public notebooks on the platform. All the three datasets have proper publicly available licensing versions of (CC BY 4.0, CC0 1.0, CC BY-SA 4.0) showcasing the usage rights of the data as well giving respect to attribution, sharing, and providing credits. The original source of these datasets are Reddit and Twitter platforms. The data is scrapped from this social media platform ethically with the usage of pushshift api, beautiful soup, and web scrapers. The datasets are clean from personal identification data and reflect the ethical standard in the process of data gathering and usage [Namdari and Gaes, 2022], [Camacho-Collados et al., 2022], [Lowe, 2023].

The following points highlight the important aspects and key practices that need to be followed for the responsible use of data and its life cycle given the sensitive domain of mental health.

Privacy protection and security: When handling the data sources from clinical or non-clinical records, social media platforms and online forums, data fields like name, phone numbers, email-ids, usernames, address, dates and places needs to be anonymized and masked. Also, aggregate findings and summary analysis can also be a protective step in securing individual entities [Ngiam and Khor, 2019].

Data security: To protect the data storage and server processing of the data, strict industry standard protocols and advanced encryption needs to be in place to safeguard the data from the data breaches, unauthorized access and cyber-attacks. [Ngiam and Khor, 2019].

Ethical consideration in data usage and transparency: Making sure that the data usage is ethically correct and transparent is the foundation of the research. Especially handling and processing clinical data, consent is the most important requirement. Data scraped or made available from the online forums, websites and social media platforms needs to be tagged accordingly as the source of the original data and context is important and meets the ethical and legal standards for the usage with respect to privacy, security and data protection. Data should be handled, processed, and stored with utmost care, responsibly and transparency into the entire data life cycle should be maintained. Geographical boundary laws are important, and research needs to be compliant with that. Especially when handling the clinical data like patients data or session scripts to protect the patients individual characteristics and privacy. Gaining approval from the governmental, medical, and responsible organizations and administrative bodies is required followed by strict compliance to data protection laws like GDPR and HIPAA [Ajmani et al., 2023], [Glaz et al., 2019].

The above aspects highlight the importance of ethical practice and legal permissions for the fair use of data handling and processing practices in the research related to the sensitive domain of mental health analysis.

6.4 Critical Reflection

To reflect this thesis's research of ML and NLP applications in the analysis of online discourses related to the domain of mental health, there is a relationship between practicality and responsibility. There are biases, limitations, challenges, and potential solutions encountered in this thesis and critical reflection on that is explored here.

Biases: As every research, this thesis research also has its own biases. Like selecting the datasets, NLP techniques and model selections based on the available resources. For instance, the MentalBERT model is quite powerful and effective, but this thesis may only reflect its most powerful aspects and lesser limitations. Also, the datasets are non-clinical as they are taken from social media due to its easier availability and ethical ground, so the research application findings may be more suitable for the people who uses social media for venting, sharing opinions or as a daily communication channel and it may be not suitable to apply on everyone. Technical capabilities of these models to understand certain types of language or cultural linguistics is another hurdle as this thesis focuses on English as the base language of the discourses.

Limitations: One of the significant identified limitations of this thesis research is its binary classification between "distress" and "no distress" labels. The misclassification analysis followed by the interpretability analysis finds that there should be a third label called "neutral" to address the problem of borderline classification cases that does not fit well into the framework of binary classification. There is a need for a mental health professional and domain experts to be in a loop for training as well as reviewing and validating the model's output to make it clinically sound and practically applicable. Another significant limitation is that there is a need for multilingual models that can also capture, identify, interpret, and classify the mental distress labels from different languages by having the nuanced capability to understand different demographics, culture, linguistics, and geographical border differences. Addressing these limitations makes these models excel in their robustness and generalizability.

Challenges: There are certain challenges that came across during this thesis research. Like, the high computational cost and resource utilization for training and validation the deep learning models. Due to the high processing requirement of processors and GPU's, high electricity consumption showcases that training, validating, optimizing and deployment of such state-of-the-art and high performing model's comes with a high cost. The non-clinical datasets are scrapped data from online social media portals. Such data instances have high noise, irregular patterns, personal information, and unstructured data requires rigorous data cleaning and preprocessing with multiple cycles of validation. For instance, in this thesis there were certain patterns of HTML tags and brackets patterns like "br" was not removed

irrespective of data cleaning. Later, models learned it as a feature which is erroneous and impacts models performance. Heavy RegEx, a python package with multiple conditions was applied to clean the data of such patterns and noises to make it error free. Other challenges such as getting access to gated models from HuggingFace for fine tuning were difficult to access and required formal applications to use it for academic research. Technical challenges like setting up CUDA base to harness the maximum power of GPU's for parallel processing with CPU is a complicated task due their different variants, versions, libraries and required compatible frameworks like PyTorch. To use next generation models like GPT, Llama and Mixtral requires heavy computational resources which is out of the scope and league of the current infrastructure used in this thesis research.

Potential Solutions: To excel and move ahead of limitations and challenges requires a multi staged approach. Like, to pair with government bodies, clinical organizations, and mental health professionals to gain the access of clinical data as well as validation and review of the model outputs [Miner et al., 2019]. Followed by that, there is a rigorous need of data cleaning in different cycles to make it perfect and error free for effective modelling performance. Building multilingual models for broader understanding of cultural context [D'Alfonso, 2020]. Real life validation with intra disciplinary domain experts is required to reduce the gap between the application of AI in mental health analysis and clinical interpretations [Soenksen et al., 2022]. Another important thing is to have high computational resources at hand to experiment and develop next generational models to make these thesis research generalizable to a wider extent and possibly global. Cloud computing can be used to certain extent to reduce the computational cost, but it comes with a monetary cost implication. Partnering up with tech giants or research organizations can make the funding available for the resources, research and open the doors for further innovation [Sharma et al., 2022].

This thesis acknowledges the involved biases, limitations, challenges and potential solutions for using AI implications on mental health analysis domain showcases the ethical grounds for innovation. By representing the critical reflection on this thesis demonstrates the responsible approach taken so far. This not only highlights the current stage and understanding of the research but also provides a vision and lays a foundation for the future research to become more transformative as well as conscientious creating a way for the continuous improvement and development towards individual mental health well-being in the digital space of this modern digital age.

The designed methodology and detailed comprehensive and comparative analysis directly contributes and reflects to the thesis research aim and its following research questions defined in the outset of this thesis. The first research question is addressed by performing EDA to understand the data and the patterns that can be understood by the machine. To clearly expand upon that, applying tree printing on DT model, studying top 50 features of DT and RF by feature importance techniques, analysing top 50 positive feature and negative features by applying coefficient weight analysis

on LG and SVM models. To analyse complex NN and DL models, attention mechanisms on both last layer as well as integrated layers have been implemented on instances to understand the model's attention on influencing words. SHAP analysis has also been employed on both label instances to see how the model understands the indicators and classify them differently. So, it's evident from the results of analysis that ML and NLP techniques are effective to capture and detect the mental distress indicators from the text and the RQ1 has been addressed completely.

Second research question is addressed by end-to-end training of traditional ML models such as DT, RF, SVM, and LG then progressing to advanced NN models such as CNN, RNN and LSTM and finally moving towards the end-to-end fine tuning of DL models like BERT standard, DistilBERT, MentalRoBERTa and MentalBERT. These models have been completely trained with sophisticated and systemic NLP usecase text classification pipelines ensuring best data science practices have been followed. This set of 32 trained models covers a wide range of different complexity and capability based technological and architectural models demonstrating different effectiveness to deal with mental distress indicators. DL models were the best performing models among all model types and categories. To assess the model's performance, a wide range of evaluation metrics have been used to capture different perspectives and aspects of trained models. Evaluation Metrics used were precision, recall, f1-score, accuracy, auc-roc curve, confusion matrix, classification report, training, and validation loss. In this way, RQ2 has been attended rigorously. To address the third research question, different text preprocessing techniques have been implemented. Two datasets were created. Dataset 1 underwent extensive preprocessing with tokenization, lemmatization, stopword removal whereas dataset 2 underwent only tokenization. Though the impact was relatively low and depends on model to model with dataset and domain at hand. Further, hyperparameter tuning has been employed with random search techniques to find the best and optimal set of hyperparameters for the training of models. This is how the models behaved to showcase optimal to high performance. These techniques ensure completeness in responding to RQ3.

To deal with the fourth research question, different vectorization techniques were explored and implemented to study their efficacy on model performance. This includes traditional TF-IDF, advanced distributed word embeddings such as Word2Vec and GloVe followed by state-of-the-art BERT Embeddings. Most effective feature representation technique was Word2Vec configured with ML models followed by BERT embeddings configured with NN models. DL does not require traditional feature engineering due its advanced inherent architectural capabilities. In this way, all the ML and NN models were trained on all the four methods and compared to analyse their efficacy on model's performance and RQ4 is completely confronted. To attend the research question five, two internal validation techniques namely stratified k-fold cross validation followed by holdout method to make the model training without biases and to prevent underfitting and overfitting were used. To gauge

the model's generalizability, external validation with two datasets have been implemented resulting in exceptional high performance. Misclassification analysis and interpretability analysis with different techniques and pretrained models have been implemented to evaluate the models with real world adaptability. To further address the RQ5 completely, implications and applications of these models with respect to automated support digital systems have been explained briefly as well as the possible future research directions and perspectives have been explained in detail too. To address the research question six RQ6, detailed description regarding the ethical considerations with respect to privacy, security of the user's data as well as the interpretability of the results have been explained briefly. By addressing and dealing with each research questions in depth, the overarching main research aim is successfully accomplished and completely approached as this conducted thesis research directly answers the ultimate research aim of this thesis "How ML and NLP methods can detect and analyse the mental distress indicators effectively from the online textual discourses".

This thesis sets a foundational stage for further exploration and research. This thesis offers a more systematic, nuanced, and responsible approach to develop AI technologies for mental health analysis that are scientifically stable, robust, ethically sound, and applicable by providing insights into state-of-the-art technologies. This thesis advances and reduces the gaps from the existing literature by providing insights into state-of-the-art DL models and domain specific DL models followed by impact of different preprocessing, vectorization, model types and categories. The results are also justified with robust misclassification, interpretability and external validation analysis surpassing the existing research ground truth. This thesis is significant in reducing the gaps between theoretical AI and Mental health domain analysis by providing insights into the transformative application of AI technologies. This thesis has been conducted responsibly and ethically with high academic and technical rigor to explore the capability of various underlying AI technologies for detection of mental distress indicators from the online textual discourses. The results of this thesis have addressed multiple research gaps with respect to existing literature studies and provide a solid background to build advanced AI systems and direction for future research with evolving AI technologies. This thesis definitely offers a solution towards the social challenge where, ML and NLP branches of AI can be leveraged and experts from interdisciplinary backgrounds can come together to conduct further research and build advanced real world resilient AI systems.

7 Conclusion

This final chapter of the thesis highlights overall findings and discusses possible future research perspectives.

7.1 Summary of Findings

This thesis has rigorously explored and implemented the designed methodology to evaluate the capabilities of ML and NLP technologies for their efficacy and effectiveness in detecting mental distress indicators from social media textual discourses. By conducting a comparative analysis of traditional ML models, it revealed that Word2Vec is the best feature representation technique compared to TF-IDF, GloVe, BERT Embeddings. SVM is the best performing model with accuracy of 92.06% among other models such as DT, RF, LG. This is due to the SVM's and Word2Vec combined capability to capture and identify contextual relationships between words. Among advanced NN models, BERT Embeddings is identified as the most effective feature representation technique due to its superior understanding of context aware embeddings compared to TF-IDF, Word2Vec and GloVe. Among all NN models, LSTM configured with BERT Embeddings outperformed other models such as CNN and RNN followed by all the traditional ML models with accuracy of 95.60%. This is due to LSTM inherent architecture of remembering and processing long term sequential dependencies, also benefited from BERT Embeddings resulting in deeper understanding of mental distress indicators from the text.

As ML and NN models have set the benchmark, the thesis was navigated to the ultimate progression towards the state-of-the-art DL models. Here, four different transformer based DL models were trained, namely the standard BERT, DistilBERT, MentalRoBERTa and MentalBERT. The domain specific MentalBERT model provided the highest performance in all the evaluation metrics with a remarkable accuracy of 96.60% surpassing all other DL, NN and ML models. In comparison, all DL models outperformed ML and NN models. This demonstrates that in this new dawn, DL models have evolved with complex and advanced inherent technical architectures that can understand and interpret the human level language surpassing predecessor models from ML and NN domain. DL models should be leveraged when dealing with NLP task and textual data, as it has proven in this thesis by showcasing the exceptional performance with capturing, identifying, interpreting, and classifying

normal and mental distress indicators from the nuanced complex linguistic textual corpus. These cutting-edge models can be further pretrained on the domain specific data, like in this thesis case MentalBERT model is pretrained on mental data based on Reddit and Twitter platform and then underwent end-to-end fine tuning with the selected dataset for this thesis resulting in high performance, robustness, reliability, consistency followed by real world generalizability and adaptability.

This thesis also shed light on the impact of text preprocessing. MentalBERT being the best DL model was trained twice with the same datasets, but different text preprocessing techniques resulted in different performance. The difference between extensively pre-processed dataset with tokenization, stopword removal and lemmatization performed better with around 0.30% accuracy higher than dataset that was only pre-processed with tokenization. Though the impact of extensive preprocessing in performance is modest, it's because of the DL model's inherent capabilities. Yet, it is important to preprocess the dataset for ML and NN models, DL models too as it depends on the dataset, pretraining of the model and domain. This thesis has been conducted by following the best data science practices to the best of my knowledge to select and compare different techniques, models, and their categories to maintain the integrity, reliability, and validity of this thesis findings. This thesis directs the future research to explore the next generation LLM models like GPT, Gemini, Claude, Llama, Mixtral, and Bloom as they offer more advancements and are highly capable to understand human level language with intelligence quotient that enhances NLP capabilities. This requires non exhaustive computational resources followed by multiple data sources including clinical data. This exploration will navigate research further and make the models more robust and reliable to capture different cultures, multilingual nuances, geographical boundaries, various demographics and can advance at multi-label modelling.

Overall to conclude, this thesis emphasizes on the application of sophisticated ML and state-of-the-art NLP transformative technologies for their intervention in the mental health analysis. This thesis has laid solid foundation for future exploration and navigation, it further calls for a collaborative approach by promoting interdisciplinary domain experts to build refined, robust, ethically sound as well clinically approved AI systems to solve the social challenge of mental health well-being on social media. Overall, this thesis research contributes its findings and implications by laying the groundwork, to the best of my knowledge, in accordance with academic, technical, and ethical rigor for the upcoming future advancements in the AI evolution with its applicability in the domain of mental health analysis.

7.2 Future Work

Large language models are on the rise and are continuously evolving and transforming into different domains and segments. State-of-the-art deep learning models

like BERT and its effective domain specific model MentalBERT was end-to-end fine-tuned and analysed in this thesis. Transformer based BERT models are the standing pillars in the domain of LLM with regards to NLP tasks and have shown potential to generate and understand text like humans. The results of this thesis demonstrates the need to push the boundaries, explore and develop efficient, robust, stable, consistent, and adaptable systems to deploy them on large scale social media platforms. To achieve this, there is the need to leverage the latest models. There are several secure and powerful paid models like GPT, Gemini and Claude followed by open-source models like Llama, Mixtral and Bloom showing next gen performance with regards to NLP tasks. These models will provide better performance by detecting complex linguistic patterns related to different mental health diseases and also excel at multilingual text processing. Exploring these new generational models will not only show algorithmic improvements but also supports faster and efficient data processing and robust infrastructure for pipelining and inference [Minaee et al., 2024].

For training and end-to-end fine tuning these advanced and next generation deep learning models, intensive resources are required. Costs with respect to powerful processors, GPU's, memory storage, electricity and domain experts are incurred. To enhance these model's overall understanding of the human language diversity with regards to different demographics, context, language, and cultures, there is a need for enormous training data. The source of data can be clinical records, non-clinical records and social media platform data. Intra domain and disciplinary experts are required to coordinate together and review the building of such powerful systems by leveraging artificial intelligence for mental health analysis and pushing the boundaries to solve real-world problems [Yang et al., 2023].

To improve the overall experience and deal with the diagnosis at initial stage, in addition to mental health state detection, chat-bots can be implemented for the realistic, nuanced and reflective conversations showcasing empathy, options to treatments, further actions to be taken and offer first support to the users. In further research and future work of building this kind of sophisticated, user-centric digital support systems, there will be an important need to address the trade-offs between ethical considerations, advanced technologies, and the [Cabrera et al., 2023] computational resource.

Continuous responsible research and exploration into future work will lead towards constant and evolving innovation in the domain of mental health analysis powered by artificial intelligence. This is important as the digital systems are at a constant rise which requires an immensely calm, stable and relaxed mental well-being state. The ongoing evolution and transformation in the AI technologies, showcases the potential to create proactive digital support systems and provide immediate support to the distressed mental health state users worldwide and possibly could set a new standard for the digital well-being [Graham et al., 2019], [Inkster et al., 2018].

References

- [Abbe et al., 2016] Abbe, A., Grouin, C., Zweigenbaum, P., and Falissard, B. (2016). Text mining applications in psychiatry: a systematic literature review. *International Journal of Methods in Psychiatric Research*, 25(2):86–100.
- [Ajmani et al., 2023] Ajmani, L., Chancellor, S., Mehta, B., Fiesler, C., Zimmer, M., and de Choudhury, M. (2023). A Systematic Review of Ethics Disclosures in Predictive Mental Health Research. *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*.
- [Aldughayfiq et al., 2023] Aldughayfiq, B., Ashfaq, F., Jhanjhi, N., and Humayun, M. (2023). Explainable ai for retinoblastoma diagnosis: interpreting deep learning models with lime and shap. *Diagnostics*, 13(11):1932.
- [Alishiri et al., 2008] Alishiri, G. H., Bayat, N., Fathi Ashtiani, A., Tavallaii, S. A., Assari, S., and Moharamzad, Y. (2008). Logistic regression models for predicting physical and mental health-related quality of life in rheumatoid arthritis patients. *Modern Rheumatology*, 18(6):601–608.
- [Anaconda, 2023] Anaconda, I. (2023). Anaconda. <https://www.anaconda.com/>. Accessed: April 23, 2024.
- [Anderson et al., 2023] Anderson, M., Faverio, M., and Gottfried, J. (2023). Teens, social media and technology 2023.
- [Bai et al., 2021] Bai, Y., Yang, E., Han, B., Yang, Y., Li, J., Mao, Y., Niu, G., and Liu, T. (2021). Understanding and improving early stopping for learning with noisy labels. In Ranzato, M., Beygelzimer, A., Dauphin, Y., Liang, P., and Vaughan, J. W., editors, *Advances in Neural Information Processing Systems*, volume 34, pages 24392–24403. Curran Associates, Inc.
- [Birunda et al., 2021] Birunda, S. et al. (2021). A review on word embedding techniques for text classification. *Innovative Data Communication Technologies and Application: Proceedings of ICIDCA 2020*, pages 267–281.
- [Bokolo and Liu, 2023] Bokolo, B. G. and Liu, Q. (2023). Deep learning-based depression detection from social media: Comparative evaluation of ml and transformer techniques. *Electronics*, 12(21).

- [Boon-Itt et al., 2020] Boon-Itt et al. (2020). Public perception of the COVID-19 pandemic on Twitter: sentiment analysis and topic modeling study. *JMIR Public Health and Surveillance*, 6(4):e21978.
- [Borah et al., 2022] Borah, Trinayan, and Kumar, G. (2022). Application of NLP and Machine Learning for Mental Health Improvement. In *International Conference on Innovative Computing and Communications: Proceedings of ICICC 2022, Volume 3*, pages 219–228. Springer.
- [Bouarara, 2021] Bouarara, H. A. (2021). Recurrent neural network (rnn) to analyse mental behaviour in social media. *International Journal of Software Science and Computational Intelligence (IJSSCI)*, 13(3):1–11.
- [Bounabi et al., 2019] Bounabi, M., Moutaouakil, K. E., and Satori, K. (2019). Text Classification using Fuzzy TF-IDF and Machine Learning Models. *Proceedings of the 4th International Conference on Big Data and Internet of Things*.
- [Burkart and Huber, 2021] Burkart, N. and Huber, M. F. (2021). A survey on the explainability of supervised machine learning. *Journal of Artificial Intelligence Research*, 70:245–317.
- [Cabrera et al., 2023] Cabrera, J., Loyola, M. S., Magaña, I., and Rojas, R. (2023). Ethical dilemmas, mental health, artificial intelligence, and llm-based chatbots. In *International Work-Conference on Bioinformatics and Biomedical Engineering*, pages 313–326. Springer.
- [Cacheda et al., 2019] Cacheda, F., Fernandez, D., Novoa, F. J., and Carneiro, V. (2019). Early detection of depression: social network analysis and random forest techniques. *Journal of Medical Internet Research*, 21(6):e12554.
- [Camacho-Collados et al., 2022] Camacho-Collados, J., Rezaee, K., Riahi, T., Ushio, A., Loureiro, D., Antypas, D., Boisson, J., Espinosa Anke, L., Liu, F., and Martínez Cámara, E. (2022). TweetNLP: Cutting-edge natural language processing for social media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–49, Abu Dhabi, UAE. Association for Computational Linguistics.
- [Campillo-Ageitos et al., 2021] Campillo-Ageitos, E., Fabregat, H., Araujo, L., and Martinez-Romo, J. (2021). NLP-UNED at eRisk 2021: self-harm early risk detection with TF-IDF and linguistic features. *Working notes of CLEF*, pages 21–24.
- [Chai, 2023] Chai, C. P. (2023). Comparison of text preprocessing methods. *Natural Language Engineering*, 29(3):509–553.
- [Chapman et al., 2020] Chapman, A. P., Missier, P., Simonelli, G., and Torlone, R. (2020). Capturing and querying fine-grained provenance of preprocessing pipelines in data science. *Proc. VLDB Endow.*, 14:507–520.

- [Dang et al., 2020] Dang, N. C., Moreno-García, M. N., and De la Prieta, F. (2020). Sentiment analysis based on deep learning: A comparative study. *Electronics*, 9(3).
- [Deng et al., 2016] Deng, X., Liu, Q., Deng, Y., and Mahadevan, S. (2016). An improved method to construct basic probability assignment based on the confusion matrix for classification problem. *Information Sciences*, 340:250–261.
- [Devlin et al., 2018] Devlin, J., Chang, M., Lee, K., and Toutanova, K. (2018). BERT: pre-training of deep bidirectional transformers for language understanding. *CoRR*, abs/1810.04805.
- [Dey et al., 2018] Dey, A. et al. (2018). Senti-n-gram: An n-gram lexicon for sentiment analysis. *Expert Systems with Applications*, 103:92–105.
- [Dogra et al., 2022] Dogra, V., Verma, S., Chatterjee, P., Shafi, J., Choi, J., Ijaz, M. F., et al. (2022). A complete process of text classification system using state-of-the-art nlp models. *Computational Intelligence and Neuroscience*, 2022.
- [Draženović et al., 2023] Draženović, M., Vukušić Rukavina, T., and Machala Poplašen, L. (2023). Impact of social media use on mental health within adolescent and student populations during covid-19 pandemic. *International Journal of Environmental Research and Public Health*, 20(4):3392.
- [D’Alfonso, 2020] D’Alfonso, S. (2020). AI in mental health. *Current Opinion in Psychology*, 36:112–117.
- [Eelbode et al., 2021] Eelbode, T., Sinonquel, P., Maes, F., and Bisschops, R. (2021). Pitfalls in training and validation of deep learning systems. *Best Practice & Research Clinical Gastroenterology*, 52:101712.
- [Ermers et al., 2020] Ermers, N. J., Hagoort, K., and Scheepers, F. E. (2020). The predictive validity of machine learning models in the classification and treatment of major depressive disorder: State of the art and future directions. *Frontiers in Psychiatry*, 11.
- [Garcia-Ceja et al., 2018] Garcia-Ceja, E., Riegler, M., Nordgreen, T., Jakobsen, P., Oedegaard, K. J., and Tørresen, J. (2018). Mental health monitoring with multi-modal sensing and machine learning: A survey. *Pervasive and Mobile Computing*, 51:1–26.
- [Gardner, 1988] Gardner, E. (1988). The space of interactions in neural network models. *Journal of Physics A: Mathematical and General*, 21(1):257.
- [Glaz et al., 2019] Glaz, A. L., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Billot, R., Ryan, T. C., Marsh, J. J., Devylder, J., Walter, M., Berrouiguet, S., and Lemey, C. (2019). Machine learning and natural language processing in mental health: Systematic review. *Journal of Medical Internet Research*, 23.

- [Glick and Applbaum, 2010] Glick, D. and Applbaum, K. (2010). Dangerous non-compliance: a narrative analysis of a cnn special investigation of mental illness. *Anthropology & Medicine*, 17(2):229–244.
- [Graham et al., 2019] Graham, S., Depp, C., Lee, E. E., Nebeker, C., Tu, X., Kim, H.-C., and Jeste, D. V. (2019). Artificial intelligence for mental health and mental illnesses: an overview. *Current psychiatry reports*, 21:1–18.
- [Greco et al., 2023] Greco, C. M., Simeri, A., Tagarelli, A., and Zumpano, E. (2023). Transformer-based language models for mental health issues: A survey. *Pattern Recognition Letters*, 167:204–211.
- [Handley et al., 2014] Handley, T. E., Hiles, S. A., Inder, K. J., Kay-Lambkin, F. J., Kelly, B. J., Lewin, T. J., McEvoy, M., Peel, R., and Attia, J. R. (2014). Predictors of suicidal ideation in older people: a decision tree analysis. *The American Journal of Geriatric Psychiatry*, 22(11):1325–1335.
- [Heimerl et al., 2014] Heimerl, F. et al. (2014). Word cloud explorer: Text analytics based on word clouds. In *2014 47th Hawaii international conference on system sciences*, pages 1833–1842. IEEE.
- [Hossin et al., 2015] Hossin et al. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, 5(2):1.
- [Howard et al., 2018] Howard et al. (2018). Universal language model fine-tuning for text classification. *arXiv preprint arXiv:1801.06146*.
- [Huang et al., 2023] Huang, Y., Li, Y., Wu, W., Zhang, J., and Lyu, M. R. (2023). Do not give away my secrets: Uncovering the privacy issue of neural code completion tools. *arXiv preprint arXiv:2309.07639*.
- [HuggingFace, 2023] HuggingFace (2023). Hugging face. <https://huggingface.co/>. Accessed: April 23, 2024.
- [Huong et al., 2022] Huong, T. H., Tran-Trung, K., Lai, D., and Hoang, V. T. (2022). Sentiment analysis based on word vector representation for short comments in vietnamese language. *2022 9th NAFOSTED Conference on Information and Computer Science (NICS)*, pages 165–169.
- [Ienca and Malgieri, 2022] Ienca, M. and Malgieri, G. (2022). Mental data protection and the gdpr. *Journal of Law and the Biosciences*, 9(1):lsac006.
- [Inamdar et al., 2023] Inamdar, S., Chapekar, R., Gite, S., and Pradhan, B. (2023). Machine learning driven mental stress detection on reddit posts using natural language processing. *Human-Centric Intelligent Systems*, 3(2):80–91.

- [Inkster et al., 2018] Inkster, B., Sarda, S., Subramanian, V., et al. (2018). An empathy-driven, conversational artificial intelligence agent (wysa) for digital mental well-being: real-world data evaluation mixed-methods study. *JMIR mHealth and uHealth*, 6(11):e12106.
- [IŞIK et al., 2020] IŞIK et al. (2020). The impact of text preprocessing on the prediction of review ratings. *Turkish Journal of Electrical Engineering and Computer Sciences*, 28(3):1405–1421.
- [Iyortsuun et al., 2023] Iyortsuun, N. K., Kim, S.-H., Jhon, M., Yang, H.-J., and Pant, S. (2023). A review of machine learning and deep learning approaches on mental health diagnosis. In *Healthcare*, volume 11, page 285. MDPI.
- [Jain et al., 2021] Jain, T., Jain, A., Hada, P. S., Kumar, H., Verma, V. K., and Patni, A. (2021). Machine learning techniques for prediction of mental health. In *2021 Third International Conference on Inventive Research in Computing Applications (ICIRCA)*, pages 1606–1613. IEEE.
- [Jeatrakul et al., 2010] Jeatrakul, P., Wong, K. W., Fung, C. C., and Takama, Y. (2010). Misclassification analysis for the class imbalance problem. In *2010 World Automation Congress*, pages 1–6. IEEE.
- [Jebb et al., 2017] Jebb, A. T. et al. (2017). Exploratory data analysis as a foundation of inductive research. *Human Resource Management Review*, 27(2):265–276.
- [Ji et al., 2022] Ji, S., Zhang, T., Ansari, L., Fu, J., Tiwari, P., and Cambria, E. (2022). MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. In *Proceedings of LREC*.
- [Joblib, 2024] Joblib, D. T. (Accessed 2024). Joblib Documentation. <https://joblib.readthedocs.io/en/stable/>.
- [Johnson and Karthik, 2021] Johnson, A. and Karthik, R. (2021). Performance evaluation of word embeddings for sarcasm detection- a deep learning approach.
- [Joyce et al., 2023] Joyce, D. W., Kormilitzin, A., Smith, K. A., and Cipriani, A. (2023). Explainable artificial intelligence for mental health through transparency and interpretability for understandability. *NPJ Digital Medicine*, 6(1):6.
- [Jupyter Project, 2023] Jupyter Project, C. (2023). Project jupyter. <https://jupyter.org/>. Accessed: April 23, 2024.
- [Kaggle, 2024a] Kaggle, I. C. (Accessed 2024a). Depression Reddit. <https://www.kaggle.com/datasets/infamouscoder/depression-reddit-cleaned>.
- [Kaggle, 2024b] Kaggle, N. K. (Accessed 2024b). Suicide Watch. <https://www.kaggle.com/datasets/nikhileswarkomati/suicide-watch>.

- [Kalyan et al., 2021] Kalyan, S., Ravishankar, H., and Arunkumar, C. (2021). Distress-level detection using deep learning and transfer learning methods. In *Smart Computing Techniques and Applications: Proceedings of the Fourth International Conference on Smart Computing and Informatics, Volume 1*, pages 407–414. Springer.
- [Kandel et al., 2020] Kandel, I. et al. (2020). The effect of batch size on the generalizability of the convolutional neural networks on a histopathology dataset. *ICT Express*, 6(4):312–315.
- [Kang et al., 2020] Kang, D., Raghavan, D., Bailis, P., and Zaharia, M. (2020). Model assertions for monitoring and improving ml models. *Proceedings of Machine Learning and Systems*, 2:481–496.
- [Kang et al., 2022] Kang, K.-S., Koo, C., and Ryu, H.-G. (2022). An interpretable machine learning approach for evaluating the feature importance affecting lost workdays at construction sites. *Journal of Building Engineering*, 53:104534.
- [Khan et al., 2020] Khan, H., Srivastav, A., and Mishra, A. K. (2020). Use of classification algorithms in health care. In *Big data analytics and intelligence: A perspective for health care*, pages 31–54. Emerald Publishing Limited.
- [Koizumi et al., 2012] Koizumi et al. (2012). Effects of text length on lexical diversity measures: Using short texts with less than 200 tokens. *System*, 40(4):554–564.
- [Koumarelas et al., 2020] Koumarelas, I., Jiang, L., and Naumann, F. (2020). Data preparation for duplicate detection. *Journal of Data and Information Quality (JDIQ)*, 12(3):1–24.
- [Kour and Gupta, 2022] Kour, H. and Gupta, M. (2022). An hybrid deep learning approach for depression prediction from user tweets using feature-rich CNN and bi-directional LSTM. *Multimedia Tools and Applications*, 81:23649–23685.
- [Kunft et al., 2019] Kunft, A., Katsifodimos, A., Schelter, S., Breß, S., Rabl, T., and Markl, V. (2019). An intermediate representation for optimizing machine learning pipelines. *Proc. VLDB Endow.*, 12:1553–1567.
- [Le Glaz et al., 2021] Le Glaz, A., Haralambous, Y., Kim-Dufor, D.-H., Lenca, P., Bilot, R., Ryan, T. C., Marsh, J., Devylder, J., Walter, M., Berrouiguet, S., et al. (2021). Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*, 23(5):e15708.
- [LeCun et al., 2015] LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *nature*, 521(7553):436–444.
- [Li et al., 2023] Li, M., Dahmani, L., Hubbard, C., and et al. (2023). Individualized functional connectome identified generalizable biomarkers for psychiatric symptoms in transdiagnostic patients. *Neuropsychopharmacology*, 48:633–641.

- [Li et al., 2021] Li, P., Rao, X., Blase, J., Zhang, Y., Chu, X., and Zhang, C. (2021). Cleanml: A study for evaluating the impact of data cleaning on ml classification tasks. In *2021 IEEE 37th International Conference on Data Engineering (ICDE)*, pages 13–24. IEEE.
- [Li et al., 2022] Li, Q., Peng, H., Li, J., Xia, C., Yang, R., Sun, L., Yu, P. S., and He, L. (2022). A survey on text classification: From traditional to deep learning. *ACM Trans. Intell. Syst. Technol.*, 13(2).
- [Liao et al., 2022] Liao, L., Li, H., Shang, W., and Ma, L. (2022). An empirical study of the impact of hyperparameter tuning and model optimization on the performance properties of deep neural networks. *ACM Trans. Softw. Eng. Methodol.*, 31(3).
- [Liu et al., 2019] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692.
- [Liu et al., 2022] Liu, Y., Xu, C., Kuai, X., Deng, H., Wang, K., and Luo, Q. (2022). Analysis of the causes of inferiority feelings based on social media data with word2vec. *Scientific Reports*, 12(1):5218.
- [Lowe, 2023] Lowe, S. (2023). Model trained from roberta-base on the go_emotions dataset for multi-label classification.
- [Luna, 2023] Luna, J. C. (2023). What is bert? an intro to bert models. DataCamp Blog. Available online: <https://www.datacamp.com/blog/what-is-bert-an-intro-to-bert-models>.
- [Malgaroli et al., 2023] Malgaroli, M., Hull, T. D., Zech, J. M., and Althoff, T. (2023). Natural language processing for mental health interventions: a systematic review and research framework. *Translational Psychiatry*, 13(1):309.
- [Martínez-Castaño et al., 2021] Martínez-Castaño, R., Htait, A., Azzopardi, L., and Moshfeghi, Y. (2021). Bert-based transformers for early detection of mental health illnesses. In Candan, K. S., Ionescu, B., Goeuriot, L., Larsen, B., Müller, H., Joly, A., Maistro, M., Piroi, F., Faggioli, G., and Ferro, N., editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 189–200, Cham. Springer International Publishing.
- [Mayring et al., 2001] Mayring, P. et al. (2001). Combination and integration of qualitative and quantitative analysis. In *Forum Qualitative Sozialforschung/Forum: Qualitative Social Research*, volume 2.
- [Minaee et al., 2024] Minaee, S., Mikolov, T., Nikzad, N., Chenaghlu, M., Socher, R., Amatriain, X., and Gao, J. (2024). Large language models: A survey. *arXiv preprint arXiv:2402.06196*.

- [Miner et al., 2019] Miner, A. S., Shah, N., Bullock, K., Arnow, B., Bailenson, J., and Hancock, J. (2019). Key considerations for incorporating conversational ai in psychotherapy. *Frontiers in Psychiatry*, 10.
- [Mohamed et al., 2023] Mohamed, E. S., Naqishbandi, T. A., Bukhari, S. A. C., Rauf, I., Sawrikar, V., and Hussain, A. (2023). A hybrid mental health prediction model using support vector machine, multilayer perceptron, and random forest algorithms. *Healthcare Analytics*, 3:100185.
- [Mukherjee et al., 2020] Mukherjee, S. S., Yu, J., Won, Y., McClay, M. J., Wang, L., Rush, A. J., and Sarkar, J. (2020). Natural language processing-based quantification of the mental state of psychiatric patients. *Computational Psychiatry*, 4.
- [Namdari and Gaes, 2022] Namdari, R. and Gaes, J. (2022). Mental health corpus. <https://www.kaggle.com/datasets/reihanenamdari/mental-health-corpus/data> and <https://huggingface.co/datasets/joangaes/depression>. Accessed: April 23, 2024.
- [Ngiam and Khor, 2019] Ngiam, K. and Khor, I. W. (2019). Big data and machine learning algorithms for health-care delivery. *The Lancet. Oncology*, 20 5:e262–e273.
- [Nguyen-Truong et al., 2022] Nguyen-Truong, G., Kang, H. J., Lo, D., Sharma, A., Santosa, A. E., Sharma, A., and Ang, M. Y. (2022). Hermes: Using commit-issue linking to detect vulnerability-fixing commits. In *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 51–62. IEEE.
- [Niu et al., 2021] Niu, Z., Zhong, G., and Yu, H. (2021). A review on the attention mechanism of deep learning. *Neurocomputing*, 452:48–62.
- [Nvidia, 2023] Nvidia (2023). Nvidia cuda toolkit. <https://developer.nvidia.com/cuda-toolkit>. Accessed: April 23, 2024.
- [Pestian et al., 2010] Pestian, J., Nasrallah, H., Matykiewicz, P., Bennett, A., and Leenaars, A. (2010). Suicide note classification using natural language processing: A content analysis. *Biomedical informatics insights*, 3:BII–S4706.
- [Probst et al., 2019] Probst, P., Boulesteix, A.-L., and Bischl, B. (2019). Tunability: Importance of hyperparameters of machine learning algorithms. *Journal of Machine Learning Research*, 20(53):1–32.
- [Python, 2023] Python, S. F. (2023). Python. <https://www.python.org/>. Accessed: April 23, 2024.
- [PyTorch, 2023] PyTorch, C. (2023). Pytorch. <https://pytorch.org/>. Accessed: April 23, 2024.
- [Rao et al., 2022] Rao, N. K., Naseeba, B., Challa, N. P., and Chakrvarthi, S. (2022). Web scraping (imdb) using python. *Telematique*, 21(1):235–247.

- [ROG, 2022] ROG, A. (2022). Rog strix g15 (2022) series. <https://rog.asus.com/de/laptops/rog-strix/rog-strix-g15-2022-series/wtb/>. Accessed: April 23, 2024.
- [Rudin, 2019] Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence*, 1(5):206–215.
- [Rustagi et al., 2021] Rustagi, A., Manchanda, C., Sharma, N., and Kaushik, I. (2021). Depression anatomy using combinational deep neural network. In Gupta, D., Khanna, A., Bhattacharyya, S., Hassanien, A. E., Anand, S., and Jaiswal, A., editors, *International Conference on Innovative Computing and Communications*, pages 19–33, Singapore. Springer Singapore.
- [Sahoo et al., 2019] Sahoo et al. (2019). Exploratory data analysis using python. *International Journal of Innovative Technology and Exploring Engineering*, 8(12):4727–4735.
- [Sanh et al., 2019] Sanh, V., Debut, L., Chaumond, J., and Wolf, T. (2019). Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. *ArXiv*, abs/1910.01108.
- [Sengupta et al., 2020] Sengupta, S., Mugde, S., and Sharma, G. (2020). An exploration of impact of covid 19 on mental health-analysis of tweets using natural language processing techniques. *medRxiv*, pages 2020–07.
- [Sharma et al., 2022] Sharma, A., Lin, I. W., Miner, A. S., Atkins, D. C., and Althoff, T. (2022). Human-ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5:46–57.
- [Singh et al., 2022] Singh, J., Wazid, M., Singh, D., and Pundir, S. (2022). An embedded lstm based scheme for depression detection and analysis. *Procedia Computer Science*, 215:166–175.
- [Soenksen et al., 2022] Soenksen, L., Ma, Y., Zeng, C., Boussioux, L., Villalobos Carballo, K., Na, L., Wiberg, H., Li, M. L., Fuentes, I., and Bertsimas, D. (2022). Integrated multimodal artificial intelligence framework for healthcare applications. *NPJ Digital Medicine*, 5.
- [Song and Diederich, 2013] Song, I. and Diederich, J. (2013). Speech analysis for mental health assessment using support vector machines. In *Mental Health Informatics*, pages 79–105. Springer.
- [Stiglic et al., 2020] Stiglic, G., Kocbek, P., Fijacko, N., Zitnik, M., Verbert, K., and Cilar, L. (2020). Interpretability of machine learning-based prediction models in healthcare. *WIREs Data Mining and Knowledge Discovery*, 10(5):e1379.

- [Straw and Callison-Burch, 2020] Straw, I. and Callison-Burch, C. (2020). Artificial intelligence in mental health and the biases of language based models. *PloS one*, 15(12):e0240376.
- [Szeghalmy et al., 2023] Szeghalmy et al. (2023). A comparative study of the use of stratified cross-validation and distribution-balanced stratified cross-validation in imbalanced learning. *Sensors*, 23(4).
- [Thieme et al., 2020] Thieme, A., Belgrave, D., and Doherty, G. (2020). Machine learning in mental health. *ACM Transactions on Computer-Human Interaction (TOCHI)*, 27:1 – 53.
- [Turner et al., 2021] Turner, R., Eriksson, D., McCourt, M., Kiili, J., Laaksonen, E., Xu, Z., and Guyon, I. (2021). Bayesian optimization is superior to random search for machine learning hyperparameter tuning: Analysis of the black-box optimization challenge 2020. In Escalante, H. J. and Hofmann, K., editors, *Proceedings of the NeurIPS 2020 Competition and Demonstration Track*, volume 133 of *Proceedings of Machine Learning Research*, pages 3–26. PMLR.
- [Wolf et al., 2020] Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Davison, J., Shleifer, S., von Platen, P., Ma, C., Jernite, Y., Plu, J., Xu, C., Le Scao, T., Gugger, S., Drame, M., Lhoest, Q., and Rush, A. (2020). Transformers: State-of-the-art natural language processing. In Liu, Q. and Schlangen, D., editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.
- [Yadav and Shukla, 2016] Yadav, S. and Shukla, S. (2016). Analysis of k-fold cross-validation over hold-out validation on colossal datasets for quality classification. In *2016 IEEE 6th International conference on advanced computing (IACC)*, pages 78–83. IEEE.
- [Yang et al., 2020] Yang et al. (2020). Rethinking the value of labels for improving class-imbalanced learning. *Advances in neural information processing systems*, 33:19290–19301.
- [Yang et al., 2023] Yang, R., Tan, T. F., Lu, W., Thirunavukarasu, A. J., Ting, D. S. W., and Liu, N. (2023). Large language models in health care: Development, applications, and challenges. *Health Care Science*, 2(4):255–263.
- [Yang et al., 2022] Yang, X., Yang, K., Cui, T., Chen, M., and He, L. (2022). A study of text vectorization method combining topic model and transfer learning. *Processes*, 10(2):350.
- [Yeskuatov et al., 2022] Yeskuatov, E., Chua, S.-L., and Foo, L. K. (2022). Leveraging reddit for suicidal ideation detection: A review of machine learning and natural language processing techniques. *International Journal of Environmental Research and Public Health*, 19(16):10347.

- [Zeberga et al., 2022] Zeberga, K., Attique, M., Shah, B., Ali, F., Jembre, Y. Z., and Chung, T.-S. (2022). A novel text mining approach for mental health prediction using Bi-LSTM and BERT model. *Computational Intelligence and Neuroscience*, 2022.
- [Zhang et al., 2022] Zhang, T., Schoene, A. M., Ji, S., and Ananiadou, S. (2022). Natural language processing applied to mental illness detection: a narrative review. *NPJ digital medicine*, 5(1):46.
- [Zhou et al., 2021] Zhou, J., Gandomi, A. H., Chen, F., and Holzinger, A. (2021). Evaluating the quality of machine learning explanations: A survey on methods and metrics. *Electronics*, 10(5):593.
- [Zhou et al., 2022] Zhou, S., Zhao, J., and Zhang, L. (2022). Application of artificial intelligence on psychological interventions and diagnosis: an overview. *Frontiers in Psychiatry*, 13:811665.
- [Zhu et al., 2020] Zhu, J., Xia, Y., Wu, L., He, D., Qin, T., gang Zhou, W., Li, H., and Liu, T.-Y. (2020). Incorporating BERT into Neural Machine Translation. *ArXiv*, abs/2002.06823.

Appendix

Complete Codebase, Datasets, Model files, Artifacts and related files are made available for internal University access and reference at the following link:

Click here for navigation to the University SVN link to access my Master Thesis Implementation and Documentation files

or for direct reference:

<https://svn.uni-koblenz.de/westteaching/theses/master/bhavyashah>