MASTER OF SCIENCE IN COMPUTER SCIENCE AND ENGINEERING

Developing modified ALBERT Model for Under-Represented Classes in Depression Detection

Sk Nahid Sanwar
191041026

Department of Computer Science and Engineering
Islamic University of Technology
October, 2025

Developing modified ALBERT Model for Under-Represented Classes in Depression Detection

Sk Nahid Sanwar
191041026

Department of Computer Science and Engineering
Islamic University of Technology
October, 2025

Certificate of Approval

The thesis titled, *Developing modified ALBERT Model for Under-Represented Classes in Depression Detection* submitted by Sk Nahid Sanwar with student ID 191041026 of Academic Year 2019–20 has been found as satisfactory and accepted as partial fulfillment of the requirement for the degree Master of Science in Computer Science and Engineering on October 02, 2025.

Board of Examiners:

Dr. Hasan Mahmud — Chairman (Supervisor)
Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT), Gazipur.

Dr. Md. Hasanul Kabir — Member (Ex-Officio)
Professor and Head
Department of Computer Science and Engineering
Islamic University of Technology (IUT), Gazipur.

Dr. Md. Kamrul Hasan — Member
Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT), Gazipur.

Lt Col Muhammad Nazrul Islam, PhD — Member (External)
Associate Professor (Instructor Class-A)
Department of Computer Science and Engineering
Military Institute of Science and Technology (MIST)
Dhaka, Bangladesh

Declaration of Candidate

This is to certify that the work presented in this thesis is the outcome of the analysis and experiments carried out by Sk Nahid Sanwar under the supervision of Dr. Hasan Mahmud, Professor, Department of Computer Science and Engineering, Islamic University of Technology (IUT), Dhaka, Bangladesh. It is also declared that neither this thesis nor any part of it has been submitted anywhere else for any degree or diploma. Information derived from the published and unpublished work of others have been acknowledged in the text and a list of references is given.

Dr. Hasan Mahmud
Professor
Department of Computer Science and Engineering
Islamic University of Technology (IUT)
Gazipur, Bangladesh
Date: October 02, 2025

Sk Nahid Sanwar
Student ID: 191041026
Date: October 02, 2025

Dedicated to my parents, wife, son and all family members for their unwavering support for my education

Contents

List of Figures

List of Tables

Acknowledgement

I would like to express my whole hearted gratitude to Allah Subhanu Wata'ala for giving me the strength to complete this study. I am profoundly grateful to my supervisor, Prof. Dr. Hasan Mahmud and also to Prof. Dr. Md. Kamrul Hasan for their invaluable guidance, encouragement, and support throughout this research. I also extend my heartfelt thanks to the members of the *Systems and Software Lab (SSL)*, at the Islamic University of Technology, for providing a collaborative research environment and computational support. Additionally, I am grateful to Md. Tariqquzzaman for his insightful sharing of domain knowledge which helped me to continue my deep research work. My appreciation goes to my thesis committee for their insightful feedback, and to my colleagues and friends for their constant motivation. Finally, I am deeply grateful to my family for their unwavering love and prayers, which have been a constant source of strength in completing this work.

This work is partially supported by

Systems and Software Lab (SSL)
Department of Computer Science and Engineering (CSE)
Islamic University of Technology (IUT)

Abstract

In the aftermath of the COVID-19 pandemic, global reliance on communication platforms has surged, presenting unique opportunities to analyze mental health trends at scale. Early and accurate detection of severe depression via social media content can facilitate timely interventions. However, accurately identifying varying levels of depression severity from textual data remains a significant challenge. Human language is nuanced and severe depression cases are often under-represented in datasets. Language models such as ALBERT, BERT, and DistilBERT have become prominent tools in sentiment analysis and mental health diagnostics, yet achieving consistently high accuracy remains challenging due to class imbalances and the complexity of natural language. In this study, a modified version of the ALBERT Base v2 model was introduced incorporating an additional self-attention mechanism designed to enhance the detection of depression in text, which is often most underrepresented in datasets and contains complex structures. The performance of the original ALBERT Base v2, the modified ALBERT, BERT, and DistilBERT models were compared in terms of classification accuracy and parameter efficiency. Our comparative analysis demonstrates that the ALBERT Base v2 model outperforms BERT and DistilBERT overall, and our modified ALBERT model shows significant improvement in classification tasks to detect severe depression, by focusing on long and complex dependency within words, using a self-attention mechanism. These findings suggest that the use of ALBERT models, particularly with modifications to the attention mechanisms, can improve classification performance in mental health analysis, offering an optimized classification model, utilizing the parameter sharing nature of ALBERT. This work provides valuable insights into the effectiveness of different language model architectures for mental health diagnostics and highlights the impact of attention mechanisms in the mental health domain.

Chapter 1

Introduction

Depression is a prevalent mental health condition and one of the leading causes of disability worldwide [16]. Early detection and intervention are crucial, especially for severe cases that pose heightened risks of self-harm [7]. Traditional clinical diagnosis of depression severity relies on self-reported questionnaires and clinician interviews, which can suffer from under-reporting and bias. In recent years, the ubiquity of social media has opened new avenues for passive mental health assessment, as individuals increasingly share their feelings and daily struggles online [5]. Researchers have found that linguistic patterns in platforms like Twitter can reveal valuable signals of mental health status. In the aftermath of the COVID-19 pandemic, reliance on digital communication has surged [45], presenting a unique opportunity to leverage user-generated text for depression monitoring on a large scale [36]. However, automated depression detection from text remains a challenging task due to the nuanced nature of language and the underrepresentation of severe cases in available data [27].

Identifying the severity of depression (distinguishing mild, moderate and severe depression) from text is particularly challenging. Mild and moderate depressive expressions can be subtle, while severe depression may have low prevalence in datasets, causing class imbalance issues [27]. Most prior studies in social media mental health analysis focused on binary classification (depressed vs. not depressed) or overall sentiment, often overlooking fine-grained severity levels [11]. Furthermore, even advanced transformer-based language models such as BERT have struggled to consistently recognize severe depression signals due to such under-representation [17]. This points to a critical gap the need for models that are not only accurate overall but also sensitive to underrepresented classes (especially the severe class) without being misled by dataset biases.

Transformers have revolutionized natural language processing by using self-attention mechanisms to capture contextual relationships in text [50]. Models such as BERT, RoBERTa and DistilBERT have achieved state-of-the-art results in various NLP tasks [32], including sentiment analysis and health related text classification. Yet, these models are extremely large (110M+ parameters for BERT, 125M for RoBERTa) and resource-intensive, making them less practical for deployment in real-time mental health monitoring systems. ALBERT was proposed as a lighter alternative with only 12M parameters by extensively sharing weights across layers [32]. Although ALBERT maintains competitive performance in many tasks due to its parameter efficiency, parameter sharing may also limit its capacity to learn disparate features useful for minority classes. There is a pressing need to enhance ALBERT's representational power for depression detection while preserving its efficiency advantages.

In this thesis, the above challenges were addressed by developing an attention enhanced ALBERT model tailored for the classification of severity of depression. Our core idea is to integrate an additional non-shared multi-head self-attention layer into the ALBERT architecture, which enables the model to focus on critical parts of the input text that signify different levels of depressive symptoms. By incorporating this custom attention mechanism, Improving the model's sensitivity to subtle linguistic cues associated with severe depression was aimed, thereby improving recall for the severe class. This modification will boost detection of under-represented classes (especially severe cases) without dramatically increasing the model's complexity was hypothesized.

The research presented in this thesis builds upon prior work in social media-based mental health analysis and transformer model optimization. The DEPTWEET dataset was utilized recently created by Kabir et al. (2022) [27], which provides a rare example of annotated tweets with depression severity labels. This dataset offers four class labels and those are Non-depressed, Mildly Depressed, Moderately Depressed, Severely Depressed derived under clinical guidance using established diagnostic criteria (e.g., PHQ-9 and DSM-5) [49]. By leveraging this dataset, our study benefits from a carefully curated typology for depression severity in text. It is important to note that this dataset was externally contributed and our role is limited to using it for training and evaluation purposes. In addition, to test the generalizability of the approach, two other public benchmarks were evaluated; a multi-class sentiment analysis dataset [37] and the TweetEval sentiment dataset [2]. These additional experiments allow us to examine how the modified ALBERT performs in different text classification settings, and whether improvements carry over beyond the depression domain.

In summary, our work introduces a novel Modified ALBERT model for fine-grained depression detection that balances performance and efficiency. Through comprehensive experiments, we show that our approach yields an improved detection of severe depression cases compared to baseline models, highlighting the value of customized attention mechanisms in mental health NLP. The results in depth was also discussed, including the limitations of our model in handling milder classes, and consider ethical implications of deploying such a model for real-world mental health support.

1.1 Problem Statement

Existing transformer models [54] struggle to accurately detect severe depression due to the under-representation of such cases in training data is the central problem addressed by this thesis.

Current automated methods struggle to identify severe depression in text due to a combination of factors and those are linguistic subtlety of symptoms, scarcity of labeled severe cases, and model limitations in capturing long-range context. The problem is exacerbated by class imbalance—models often skew towards predicting the majority class (e.g., non-depressed) and may miss critical indicators of severe depression. Developing a modeling approach that addresses these issues, yielding a classifier that is sensitive enough to detect severe depression signals and robust enough to handle imbalanced data was aimed, while still being efficient for practical use.

1.2 Motivation & Scopes

The motivation for this research stems from both practical needs in mental health care and technical gaps [18] in current NLP models. Early and accurate identification of severe depression from social media could enable proactive interventions [29] such as timely outreach to at risk individuals [52] and contribute to public health surveillance of mental well-being. However, achieving high sensitivity to severe cases is challenging with off the shelf language models [46]. Many existing studies focus on improving overall accuracy, which can come at the expense of false negatives in the severe category and a costly error in clinical terms [26]. This thesis is motivated by the need to bridge that gap and to create a model that can reliably distinguish even the rare and nuanced signals of severe depression.

The scope of our work includes natural language processing, deep learning model development, and mental health informatics. Our analysis was limited to text data

(tweets) in English, as provided by the DEPTWEET dataset [27]. Clinical aspects of depression such as its diagnostic criteria in DSM-5 [49] and PHQ-9 [31] inform our understanding of what constitutes mild vs. moderate vs. severe symptoms, but our model learns these distinctions purely from textual patterns and labels. Non-textual modalities (images, audio) or demographic context of users were not incorporated, although those could be valuable in a broader depression detection framework [25]. Moreover, the focus is on improving model architecture and performance metrics and at the same time, in creating new datasets or annotation schemes as part of this thesis (rather, it was utilized on existing resources). By focusing on model improvement within these boundaries, it was aimed to produce a solution that can be readily applied to available mental health text data.

So in summary, this research focuses on improving text-based depression detection using transformer-based architectures, particularly the ALBERT Base v2 model, by introducing an additional self-attention mechanism. The study emphasizes performance evaluation across multiple datasets containing various depression and sentiment levels.

This study specifically considers severely depressed instances as under-represented classes within the dataset, emphasizing improved detection of these minority categories through architectural modification rather than data-level augmentation.

1.3 Research Challenges

Several key research challenges had to be addressed in the course of this study.

1.3.1 Class Imbalance and Rare Signals

Severe depression examples are relatively rare in social media datasets, making it difficult for models to learn their characteristics[48]. There is a risk that a classifier will be biased towards milder classes or the non-depressed class. It was needed to be ensured that the model remains sensitive to minority classes without overfitting to noise.

1.3.2 Nuanced Language Understanding

Expressions of depression severity are often nuanced and context-dependent [20]. For instance, a mildly depressed individual's post might resemble normal sadness or frustration, whereas a severely depressed individual might still use indirect language or

humor. Capturing these subtleties requires the model to understand context and subtext beyond surface keywords [19]. This challenge motivated the integration of enhanced attention mechanisms to capture long-range dependencies in text.

1.3.3 Model Efficiency vs. Complexity

While large models like RoBERTa can achieve high accuracy, their size hinders deployment in real-time or resource-constrained environments. A central challenge was to improve model performance on under-represented classes without resorting to a significantly larger model. ALBERT provides efficiency through parameter sharing [32], but this very trait can restrict expressiveness. Our task was to introduce additional capacity (via a new attention layer) in a targeted way that boosts performance where needed (for severe class cues) while keeping the model lightweight.

1.3.4 Generalization Across Datasets

Depression specific data can be limited [44]. To ensure the approach is robust, other related tasks were carried out like general sentiment classification. A challenge was to design the model so that it improves the detection of severity of depression and remains applicable to other text classification scenarios. It was needed to verify that the modifications did not compromise the versatility of the model or cause it to over specialize to one dataset.

In this study, a primary depression specific dataset was used along with two sentiment datasets to ensure a larger generalization in linguistic and emotional terms. The decision not to include additional depression datasets was influenced by the limited availability of ethically shareable and standardized resources in the mental health domain. As highlighted by Calvo and Milne [9] and Harrigian et al. [22], mental health related corpora are often scarce, platform specific, and bound by ethical or privacy restrictions, which make large scale, cross-domain integration challenging. Therefore, combining a clinically relevant depression dataset with sentiment labeled data enables the model to capture generalized emotional features while maintaining a link to domain-specific depressive expressions.

Furthermore, the inclusion of sentiment datasets is theoretically grounded in established psychological research. Beck's cognitive theory [3] and subsequent affective models [53] emphasize that individuals experiencing depression exhibit persistent negative affect, pessimism, and a reduction in positive emotional expression. Linguistic analyses by [38] further support that frequent use of negative emotion words and

self-referential language is strongly correlated with depressive cognition. Hence, incorporating negative sentiment patterns provides a psychologically meaningful proxy for identifying severe depressive states in textual data, effectively bridging sentiment polarity and mental-health inference.

1.3.5 Evaluation and Validation

Standard evaluation metrics like overall accuracy can mask poor performance on minority classes. A challenge lay in selecting appropriate metrics and validation strategies to truly measure improvements in under-represented class detection. It was addressed by using class-wise metrics (per-class ROC AUC) and MCC (Matthews Correlation Coefficient) in addition to accuracy and F1-score, to get a more nuanced assessment of performance.

1.4 Research Contribution

Considering the shortcomings of existing research, this study addresses the frameworks of subtle language cue-centric language models to detect the severity of depression.

1.4.1 Modified ALBERT Architecture with Custom Self-Attention

A novel attention-based enhancement was proposed to the ALBERT model specifically aimed at depression severity classification. The model incorporates an additional multi-head self-attention layer that is *not shared* across layers (breaking the typical ALBERT parameter-sharing for this component). This design allows the model to focus on different parts of the input sequence and capture long-range dependencies crucial for interpreting depressive language. The added attention layer (with 12 heads and 768 dimensional embeddings) processes ALBERT's contextual embeddings and helps highlight subtle linguistic cues indicative of severe depression that might otherwise be overlooked.

1.4.2 Improved Detection of Under-Represented Depression Classes

Our modified ALBERT model achieves superior sensitivity to the class of severe depression, as evidenced by higher class-specific ROC AUC scores and recall rates compared to baseline models. Without sacrificing performance on other classes, the model significantly closes the gap in identifying the most critical severe cases. This is an

important advancement, as severe cases often have outsized importance in clinical intervention despite being few in number.

In addition, the proposed attention-based ALBERT model is not designed exclusively to detect severely depressed classes. Although the primary motivation behind the modification was to enhance sensitivity toward the minority or severe class, the model architecture maintains a multi-class prediction framework. The additional attention layer enables the model to reweight contextual dependencies and emphasize crucial linguistic cues related to emotional intensity and depressive indicators. Consequently, while the performance improvement is most noticeable in the severely depressed category, the model simultaneously learns shared representations across other emotional states, including mild and moderate depression, as well as neutral expressions.

This design ensures generalization and robustness across varying levels of depressive symptoms rather than restricting the model to a single class. The inclusion of an independent attention layer also aligns with prior research on transformer-based architectures, where attention mechanisms have been shown to improve the model's ability to capture long-range dependencies and enhance discrimination among low-frequency emotional categories [50], [57], [59].

1.4.3 Comprehensive Comparative Evaluation

An extensive empirical evaluation was conducted against a diverse set of transformer models including the original ALBERT Base v2, the compact DistilBERT [42], the classic BERT base [13] model and the larger RoBERTa model [34]. The models are fine-tuned under identical conditions on the same data splits for a fair comparison. Multiple performance metrics (Accuracy, Precision, Recall, F1-score, MCC) were reported as well as class-wise performance breakdowns. In addition, it was compared on the inference efficiency (prediction speed and resource usage) across models, highlighting how our modified model balances speed and accuracy.

1.4.4 Integration of Clinically Grounded Dataset (DEPTWEET)

The DEPTWEET dataset [27], was leveraged which was constructed with clinical input to ensure valid representation of depression severities in social media text. By using this dataset (40191 annotated tweets across four classes) for training and evaluation, the effectiveness of our approach was demonstrated on data that reflect real-world complexities and clinically relevant distinctions. Although the data set itself is external to our work, our experiments on these data provide one of the first extensive

benchmarks of transformer models on depression severity classification (as opposed to binary depression detection).

1.4.5 Implementation and Reproducibility

The thesis provides details of the implementation of the model using Python and PyTorch, including training procedures and hyperparameter settings, to facilitate reproducibility. Training was carried out on accessible hardware (two NVIDIA T4 GPUs via Kaggle environment), showing that the approach can be reproduced without specialized hardware. Insights into training time and behavior were also shared as it was made on the code available in an online repository for the research community.

1.5 Thesis Outline

In the first Chapter 1 the objective of the study has been shared in a summarized manner. Chapter 2 talks about the important background & several literature reviews bound towards this study. In the Chapter 3, the proposed methodology, model specifications have been discussed. The baseline classification models used within this dataset and evaluation metrics are represented in Chapter 4 and it also includes discussion the results of classifications, and the necessary aspects to consider while conducting additional research in this domain. Chapter 5 discusses conclusive points to the current study and addresses future directions. The final Chapter elaborates on all the references and credits used in this study.

Chapter 2

Literature review

Recent years have witnessed a growing interest in leveraging social media as a data source for mental health research. With millions of users sharing personal thoughts and emotions on platforms like Twitter and Facebook, researchers see an opportunity to detect signs of mental distress, including depression, from these digital footprints. In this chapter, the relevant literature was reviewed in three broad areas. Firstly, the relationship between social media data and depression indicators, including existing datasets and typologies; Secondly, computational methods for depression detection from text, with an emphasis on machine learning and NLP techniques; And thirdly, the advent of transformer based language models and attention mechanisms, discussing how it's been or could be applied to mental health classification tasks. The shortcomings is highlighted of current approaches in capturing depression severity, thereby positioning our research in context.

2.1 Social Media Data and Depression

The widespread use of social networks has revolutionized the way individuals express themselves and, inadvertently, how researchers can study behavioral health signals. Unlike clinical settings where patients may underreport symptoms due to stigma or recall bias, social media platforms often capture candid, real time expressions of mood and experiences. Studies have shown that language use on platforms such as Twitter can be correlated with mental health conditions. For instance, Coppersmith et al. (2014) collected tweets from users who self identified as having mental health diagnoses and found distinctive linguistic markers differentiating those users from control groups [11]. Their pioneering work demonstrated that depression and other conditions could be partially quantified through language differences in social media

posts.

Subsequent research has reinforced the value of social media for mental health monitoring. Researchers have explored various features from keyword frequencies to linguistic style matching and topic modeling to detect depression signals. However, ambiguity in online expression remains a major challenge. People do not always explicitly state "I am depressed"; instead, feelings of hopelessness may be referenced, changes in sleep patterns, or even humor and sarcasm might be used when describing pain. This necessitates more sophisticated analysis that goes beyond keyword matching to truly understand context and subtext in posts.

Another line of inquiry has focused on population level mood and public health trends via social media. For example, Reece and Danforth (2017) analyzed language patterns in billions of tweets to estimate regional depression prevalences, finding correlations between social media language cues and real world epidemiological data [39]. While such studies provide insight at a macro scale, our focus is on the individual level, determining a single post or user's depression severity. Nonetheless, these works underscore that social media data can reflect genuine mental health signals, justifying its use in developing diagnostic tools.

A crucial requirement for advancing automated depression detection is the availability of labeled datasets. Early work often relied on noisy labels, such as inferring depression status from membership in an online support group or the use of certain hashtags. These approaches lacked clinical grounding. To address this, researchers have started creating dedicated datasets with expert annotation or user self-assessments. A significant contribution in this area is the DEPTWEET dataset by Kabir et al. (2022) [27], which specifically targets the categorization of severity of depression in tweets. This data set introduced a typology aligned with clinical diagnostic criteria (DSM-5 and PHQ-9) to categorize tweets into four classes; nondepressed, mild, moderate, and severe depression. In particular, the annotation process involved mental health professionals to ensure that labels correspond to recognized symptoms at each severity level. DEPTWEET filled an important gap by providing a benchmark for training and evaluating models on fine-grained depression detection. The present thesis makes use of this resource, as described in the methodology.

Despite these efforts, data scarcity and imbalance persist. Even in a curated dataset like DEPTWEET, the distribution of classes is often skewed severe cases are the least common. Another study by Ernala et al. (2019) emphasized the need for carefully constructed typologies and ground-truth validation when using social media data for mental health predictions [18]. It noted many previous studies lacked a solid connec-

tion to clinical definitions of mental states, which could lead to misleading conclusions. In light of this, our work is built on a dataset that explicitly addresses clinical validity, and our approach is designed with class imbalance in mind.

2.2 Approaches to Depression Detection from Text

Traditional approaches to the detection of depression or related mental health conditions in text range from simple lexicon-based methods to classical machine learning classifiers and more recently to deep learning models [56]. Early studies used manually crafted dictionaries of depressive symptoms or sentiments (e.g. sadness lexicons) to flag potentially depressed posts. Although straightforward, these methods often suffered from low precision words like "sad" or "depressed" can be used in casual, non-clinical contexts, and genuine expressions of depression may use indirect language.

Machine learning approaches improved upon this by learning patterns from data. Researchers trained classifiers (such as SVMs, logistic regression, or random forests) on features including n-gram frequencies, parts-of-speech patterns, sentiment scores, and even network features (how connected or isolated a user was) to predict depression labels. For example, sentiment analysis techniques were repurposed to gauge the general emotional valence of a user's posts [10]. Some studies extended this to multiclass classification where different severity levels were the target, but progress was limited by the lack of large, labeled datasets for severity (as mentioned earlier).

The field saw a significant shift with the advent of deep neural network models for text. Recurrent Neural Networks (RNNs), especially LSTM and GRU models, were applied to mental health text because of their ability to capture sequential dependencies. Orabi et al. (2018) employed LSTM-based architectures to detect depressed Twitter users [23], reporting improved accuracy over traditional methods by using word embeddings and sequence modeling. Similarly, CNN-based models found use in text classification because of their efficiency in extracting local n-gram features.

A key development was the introduction of attention mechanisms in NLP models. Attention allows a model to dynamically focus on certain parts of the input sequence when making a prediction, which is intuitively useful for long, complex texts where only some fragments may indicate depression. For instance, an attention-equipped RNN might learn to respond to phrases such as "I cannot go on" or "feeling empty" in a long post, giving them more weight in the final classification. Before transformers, these attention mechanisms were incorporated in encoder-decoder frameworks [1] and shown to improve tasks like sentiment classification and question answering.

Despite these advances, a recurring limitation of earlier models (including RNNs with attention) was often required substantial training data to generalize well and could still miss the forest for the trees, so to speak. It might latch onto obvious depressive words and fail when faced with implicit expressions.

2.3 Transformer Models and Applications in Mental Health NLP

The introduction of transformer-based language models brought unprecedented improvements in text understanding. The transformer architecture [50] relies entirely on self-attention and feed-forward layers, doing away with recurrence and allowing models to capture long-range context more effectively. Large pre-trained models like BERT (Bidirectional Encoder Representations from Transformers) [13] and its variants have since dominated NLP leaderboards. These models are pre-trained on massive general-domain corpora and can be fine-tuned for specific tasks with relatively smaller datasets.

In the context of mental health, researchers quickly adopted these models for tasks such as detecting depression. For example, BERT has been fine-tuned on Twitter datasets to detect users with depression vs. healthy controls, yielding high accuracy compared to earlier methods. One reason is that the bidirectional context of BERT helps it understand nuances; e.g., it can differentiate the meaning of "I'm fine" in a positive context versus a sarcastic or negative context by looking at the surrounding words.

However, BERT and similar models face challenges when applied to class-imbalanced, subtle tasks like severity classification. First, the training objective (masked language modeling) is general; so it is not specifically attuned to the semantic differences that separate mild from moderate depression expressions. Obvious sentiment differences will be caught but might not pick up on clinical gradations without further guidance. Secondly, these models are extremely large. BERT base has $\tilde{1}10$ million parameters, and RoBERTa base has $\tilde{1}25$ million, which can lead to overfitting on small mental health datasets and pose deployment problems. Significant memory and computation for inference is required which is problematic if it's envisioned for real-time monitoring systems that might analyze thousands of posts on social networks per hour. To address the efficiency problem, distillation and parameter-sharing techniques have been introduced. DistilBERT [42] is a distilled (compressed) version of BERT that retains around 97% of the language understanding capabilities of BERT but uses about 40%

fewer parameters (66 million). DistilBERT has been shown to be faster and lighter while still performing well on many tasks, making it attractive for scenarios where computing resources are limited or fast inference is needed. In mental health tasks, one might choose DistilBERT to quickly flag posts for depression with minimal infrastructure, albeit possibly at a small cost to accuracy.

Another efficiency-focused model is ALBERT (Lan et al., 2019) [28]. ALBERT employs two main strategies to reduce the model size; factorized embedding parametrization (decomposing large vocabulary embeddings) and cross-layer parameter sharing (where all layers share the same weights for the feedforward and attention sublayers). ALBERT Base v2 has only 12M parameters, a dramatic reduction from BERT, while maintaining competitive performance in general language tasks. The parameter sharing of ALBERT effectively means that it has fewer independent parameters to learn patterns, which can act as a form of regularization and also drastically reduce memory usage. For applications like ours, ALBERT presents an intriguing base model and it is efficient enough to be practical, yet powerful enough (as a transformer) to handle complex language. The question is whether ALBERT's efficiency comes at a cost to identifying those fine-grained signals of severe depression that could require unique representations at different layers. Some research has started exploring transformer adaptations for mental health. For instance, studies have fine-tuned RoBERTa on mental health forums data or BERT on suicide risk datasets, generally finding that these pretrained models transfer well but may not excel in pinpointing severity without modifications. Tolentino and Schmidt (2018) highlighted that aligning model outputs with DSM-5 criteria for severity of depression is an open challenge [49]. That is, while a model might classify text as "depressed", determining *how depressed* (in clinical terms) is less straightforward. Their work suggests that characteristics corresponding to DSM-5 symptoms (such as anhedonia, sleep disturbance, etc.) should be considered. A transformer model could, in theory, learn to attend to such features if enough training examples are given or if guided appropriately.

Another relevant advancement is the concept of hierarchical transformers or multitask learning where models simultaneously learn to detect depression and related attributes (e.g., emotions or symptom mentions) to improve overall understanding. Guntuku et al. (2020) explored tracking various mental health indicators on Twitter (including anxiety, loneliness, suicidal ideation) [21]. Although not directly categorized by severity, their work exemplifies how broad coverage of signals can aid in mental health detection. Despite the success of transformers, the challenge of detecting underrepresented classes remains. Most transformers optimize for overall accuracy and cross-entropy loss, which can be dominated by the majority class. The litera-

ture suggests techniques such as data augmentation, cost-sensitive training, or architectural tweaks (such as additional attention to minority class data) to mitigate this. However, few works explicitly modified transformer architectures themselves to handle class imbalance in mental health tasks. This is the niche our research addresses rather than adjusting training data or loss, the internal structure was modified of the model (by adding an attention layer) in hopes of giving it extra capacity to learn the features of the minority class.

2.4 Self-Attention Mechanisms and Model Enhancements

The self-attention mechanism is the cornerstone of transformer models and has proven adept at modeling languages with complex long-range dependencies [51]. Self-attention works by allowing each word in a sentence to attend to every other word, generating a weighted representation that emphasizes relevant context. For example, in the sentence "I feel absolutely worthless and empty inside", attention might link "worthless" and "empty" to each other or to the subject "I", thereby strengthening the representation of a depressed state. Models like BERT already have multiple self-attention layers (12 layers with 12 heads each in BERT base) that build hierarchical representations.

It's known that BERT/ ALBERT already have self-attention built-in, in that case why it's needed to add more attention layer. The rationale lies in how the attention is used. In standard transformers, the self-attention in each layer is primarily used to contextualize the token embeddings for that layer, and by the top layer, if trained well, the model will have encoded relationships relevant to the training objective. However, these internal attentions are not explicitly optimized for our end classification task during pre-training. Fine-tuning will adjust them to some extent, but there might be limitations. By inserting an extra attention layer specialized for the classification task, we essentially give the model a chance to re-attend to the sequence *after* the pre-trained contextual embeddings are formed, focusing specifically on the task of severity classification. This can be viewed as a form of task-specific attention refinement.

There is precedent for adding custom layers to pre-trained models to improve performance. For example, some question-answering models add an attention layer that attends over supporting facts, or some text classification models add a convolutional layer on top of BERT to capture local patterns that BERT might miss. In our case, adding a dedicated multihead self-attention layer on top of ALBERT is a novel twist aimed at capturing those subtle cues and complex word relationships that indicate

the severity of depression. Importantly, this additional layer is not shared between the stacked layers of the transformer (hereinafter 'non-shared'), which means that it has its own parameters separate from the main ALBERT layers. This gives the model extra representational power focused on the classification output. Another advantage of using an extra attention layer is the ability to inspect what the model is attending to when making a classification. This can sometimes enhance interpretability and we might visualize attention weights to see which words or phrases the model found important for labeling a tweet as "Severe". In a sensitive domain such as mental health, interpretability is valuable for trust; clinicians or users are more likely to trust a model's prediction if base was found, i.e., on the phrase "no reason to live", rather than some spurious correlation.

Finally, it should be noted that while self-attention excels in capturing global context, other mechanisms like recurrence or convolution are better at local or sequential structure. Some recent research combines transformers with RNNs or CNNs to get the best of both worlds. We do not explore such hybrid models here, but it is an avenue for future work (discussed in Chapter 6). We focus on self-attention, since our hypothesis is that insufficient attention to rare but important features is a primary weakness in current models for severity detection.

In summary, the literature suggests that *how* attention is utilized can be the key to unlocking better performance for specific tasks. To our knowledge, no prior work explicitly injecting an extra layer of self-attention into ALBERT has been performed to improve the classification of severity of depression. Our approach is inspired by the general success of attention and the specific need to elevate the importance of underrepresented class signals. The next chapter will detail how we implement this idea and integrate it with the training process.

Chapter 3

Proposed Approach

In this chapter, we describe the methodology of our research, including the dataset resources, the design of the proposed model, baseline models for comparison, and the experimental setup (training procedure and evaluation metrics). We begin by outlining the datasets used in our experiments, with particular emphasis on the depression severity dataset (DEPTWEET) and how it was prepared for modeling. We then provide a detailed description of the Modified ALBERT model architecture, explaining the integration of the non-shared self-attention layer and subsequent components of the classifier. Next, we list the baseline transformer models against which we compare our approach and justify their inclusion. Finally, we specify the training configuration hyperparameter settings, hardware and software environment, and the metrics by which we assess performance.

.

3.1 Datasets and Preprocessing : DEPTWEET Dataset

The primary dataset for this study is the DEPTWEET dataset introduced by Kabir et al. (2022) [27]. This dataset consists of 40,191 tweets, each annotated into one of four categories reflecting depression severity: Non-depressed, Mild, Moderate, Severe. The data was collected via Twitter's public APIs by filtering for depression related keywords and then refined through expert annotation. Two clinical psychologists guided the annotation process to ensure that each tweet's label corresponds to DSM-5 criteria for depression severity. For example, a tweet labeled "Severe" might exhibit clear signs of hopelessness, suicidal ideation, or an aggregation of multiple depressive symptoms, whereas a "Mild" tweet might contain only one or two mild indicators (such as tempo-

rary sadness or insomnia) and "Moderate" somewhere in between. Table summarizes the class distribution in DEPTWEET

Table 3.1: Class distribution in DEPTWEET Dataset

Class Label	Description	Number of Tweets
Non Depressed	Not indicating depression	N_0 (majority)
Mild	Mild depressive symptoms	N_1
Moderate	Clear depressive signs	N_2
Severe	Severe depressive indications	N_3 (minority)

Exact counts for each class are provided in the dataset reference [27]; the distribution is imbalanced, with Non-depressed being the largest class and Severe the smallest.

Prior to modeling, we perform basic text preprocessing on the tweets. Each tweet is lowercased and stripped of URLs, user mentions, and redundant whitespace. Standard tokenization is applied using the WordPiece tokenizer associated with BERT/ALBERT, since our model uses those embeddings. We do not remove stop words or punctuation wholesale, as transformers can handle them and sometimes they carry meaning (e.g., "I'm fine..." with ellipses might convey a different tone than without). Hashtags are preserved if they contain meaningful words ("#depressed" would be tokenized as "depressed"). Emoji are not explicitly translated, though in our dataset their occurrence was minimal and any that appear are treated as unknown tokens by the tokenizer.

The data set is divided into training, validation, and test sets in a 60/20/20 ratio, following the original work of Kabir et al. and standard machine learning practice. Stratified sampling is used so that each split maintains approximately the same class distribution. This is crucial given the imbalance; we want to ensure the model sees examples of each class during training and that performance on each class can be reasonably measured in the test set. Table 3.2 shows the split proportions:

Table 3.2: Split proportion in DEPTWEET Dataset

Split	Portion of Data
Training	60% (24k tweets)
Validation	20% 8k tweets)
Test	20% (8k tweets)

No user overlap exists between splits (to avoid overfitting to particular individuals' language style). We also ensure that very similar tweets (e.g., retweets or near duplicates) fall in the same split to prevent information leakage between train and test.

3.2 Datasets and Preprocessing : Multiclass Sentiment Dataset

To test the generality of our approach, we also utilize two other public datasets: A tweet can be labelled as a non-depressed tweet if it expresses a person's joy or delight, or makes a generalized statement about depression that does not reflect the person's own mental state, expresses casual tiredness or sadness (For example, sadness due to the defeat of their favorite sports team), or expresses temporary hopelessness.

This is a collection of $\tilde{3}$9,771 social media posts labeled with one of three sentiments (positive, negative, neutral). We obtained it from a HuggingFace repository (originally contributed by an anonymous source in 2023). While not directly related to depression, it serves as a proxy task for multi-class text classification. The class balance here is roughly equal among the three sentiments. We expect that a good model for nuanced text classification should perform reasonably on this task as well.

3.3 Datasets and Preprocessing : TweetEval Sentiment Dataset

We use the sentiment analysis benchmark from TweetEval [2], which contains 59,870 tweets labeled positive, negative, or neutral. This dataset is a standard benchmark for evaluating tweet classification models and allows us to compare our results to known baselines in the literature. It is slightly larger and more balanced than the first sentiment dataset.

For both additional datasets, we apply similar preprocessing (lowercasing, tokenization) and use the provided train/dev/test splits (which are also roughly 60/20/20). The purpose of including these datasets is to verify that our model modification, while aimed at depression severity, does not overfit to only depression related language. Instead, we want to see that it can also handle generic sentiment classification competitively, indicating robust language understanding.

It's important to reiterate that the author of this thesis did not contribute to the creation of any of these datasets. DEPTWEET [Kabir2022] in particular is credited entirely to the original authors; our use of it is solely for experimentation. Similarly, the sentiment datasets are existing resources. All dataset usage in this research complies with their terms: the Twitter based data respects user privacy by using anonymized IDs in publication, and our work involves analysis at an aggregate level with no intent

to identify or harm any individual.

The characteristics of different severities of depression from the DepTweet dataset with example tweets is summarized in Table 3.3.

Table 3.3: Characteristics of Severities of Depression with Example Tweets

Depression Levels	**Characteristics**	**Example Tweets**
Non-Depressed	A generalized statement about depression, casual tiredness or exhaustion due to hard work that is not persistent, etc.	When I first saw this fight I was thinking, "This is so unfair.He just fought his way up 4 floors in a single take, he must be exhausted..."
Mildly Depressed	Feelings of panic and anxiety, concentration difficulty, a sudden disinterest in socializing, feelings of guilt and despair	im getting tired of chasing you back, trying to fix us; at one point, i see like it's always me who apologize first
Moderately Depressed	Slow thinking, no appetite, excessive or no sleep, everything seems a struggle.	Everything I do now takes days to complete. Its my new reality. Every time I push myself I mess up my progress. Depression hittin hard lately, even in my dreams. https://t.co/35sg6aUd1g
Severely Depressed	Endless suicidal thoughts, no movement, everything seems bleak, feelings of hopelessness and guilt, impossible to do anything	Haven't felt this depressed since I last tried to hurt myself in 2009. Didn't think feeling low like this would return. Luckily I'm not at state where I would try to hurt myself but it's tough

3.3.1 Class Distribution and Under-represented Categories

From Table 3.4, it is evident that all three datasets exhibit varying degrees of class imbalance. In the DepTweet dataset, the severe depression category represents only 740 samples (approximately 1.8% of the total), making it the most under-represented class. The Hugging Face sentiment dataset shows a moderate skew, with slightly fewer negative samples compared to neutral and positive ones. Similarly, the TweetEval sentiment dataset contains a higher number of neutral posts and relatively fewer negative samples. These imbalances can cause learning bias toward dominant classes and hinder generalization across emotional categories. Therefore, the proposed attention-based ALBERT modification aims to alleviate this limitation by enhancing the model's ability to capture rare yet informative linguistic patterns from minority classes. It is giving us the unique opportunity to understand the overall nuance and also under-represented class distribution on the overall dataset sample.

Table 3.4: Dataset-wise class distribution and identification of under-represented categories.

Dataset	Class	Samples	Percentage (%)
DepTweet (Depression) [27]	Non-depressed	32,400	80.6
	Mild	5,242	13.0
	Moderate	1,809	4.5
	Severe	**740**	**1.8**
Hugging Face Sentiment [37]	Neutral	15,506	37.2
	Negative	**12,168**	**29.2**
	Positive	13,968	33.5
TweetEval Sentiment [2]	Neutral	27,479	45.9
	Negative	**11,377**	**19.0**
	Positive	21,043	35.1

3.4 Modified ALBERT Model Architecture

Our proposed model is based on the ALBERT Base v2 architecture [32] with a crucial modification that we integrate an additional multi-head self-attention layer into the transformer encoder stack and modify the classifier head accordingly.

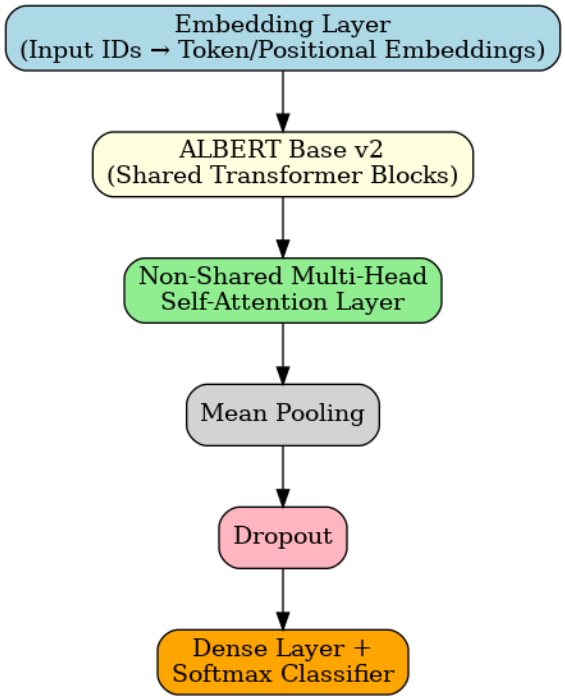


Figure 3.1: Modified ALBERT Architecture

3.4.1 Base ALBERT Encoder

ALBERT Base v2 consists of 12 transformer encoder layers. Unlike BERT, ALBERT shares the same parameters across all 12 layers effectively, it applies one layer 12 times.

Each encoder layer has an internal self attention mechanism (with 12 attention heads and 768 dimensional hidden size in ALBERT Base) and a feedforward network. Because of parameter sharing, ALBERT dramatically reduces memory usage; however, it means the model cannot learn different weights for different depths. In practice, ALBERT Base performs comparably to BERT Base on many tasks, indicating that repeating the same layer can still capture useful representations up to a point.

In our modified model, we retain the entire pre-trained ALBERT Base v2 as is up to its final (12th) layer. We initialize these layers with the publicly available pre-trained weights (trained on BooksCorpus and Wikipedia) so that the model starts with a strong grasp of general language.

3.4.2 Added Self-Attention Layer

On top of ALBERT's final layer output, we introduce a new multi-head self-attention layer that is not shared with any other layer (it is unique). This layer takes the sequence of hidden states from ALBERT and performs another attention weighting across the sequence. We configure it with the same dimensionality as ALBERT for compatibility: 768 hidden size and 12 attention heads (matching ALBERT Base's parameters). The intuition is that this layer can learn to specifically emphasize parts of the tweet that are most relevant to classifying depression severity. Because it has its own parameters, it isn't constrained by the repetitive patterns ALBERT's shared layers might have settled into; it can potentially extract new combinations of features. For example, ALBERT layers might encode general sentiment and syntax, while this extra layer might focus attention on particular symptom related words or negations that flip sentiment meaning.

Let

$$H \in \mathbb{R}^{T \times 768} \tag{3.1}$$

Here it denotes the sequence of hidden states produced by ALBERT for a tweet of length T tokens. The added attention layer computes:

$$A = MultiHeadAttn(H, H, H) \tag{3.2}$$

where $A \in \mathbb{R}^{T \times 768}$ is the sequence of attended vectors

The function $MultiHeadAttn$ refers to the standard transformer attention mechanism;

1. Project H into queries Q, keys K, and values V.
2. For each Head, compute scaled dot product attention:

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^\top}{\sqrt{d_k}}\right)V, \tag{3.3}$$

   where d_k is key dimension. [50]
3. Concatenate the outputs from all heads and project back to 768 dimensions.

As in standard transformer blocks, this layer is followed by a residual connection and layer normalization. The effect is that the representation of each token in A can encode richer, task specific dependencies throughout the sequence.

3.4.3 Mean Pooling

After the custom attention layer, we aggregate token representations using mean pooling. Formally, if

$$A = [a_1, a_2, \ldots, a_T], \tag{3.4}$$

with each $a_i \in \mathbb{R}^{T\times 768}$, then the pooled vector $v \in \mathbb{R}^{768}$

$$v = \frac{1}{T}\sum_{i=1}^{T} a_i. \tag{3.5}$$

This approach [40] averages information across the entire sequence, ensuring that signals distributed across multiple parts of a tweet (e.g., low mood in one phrase, loss of energy in another) are jointly represented. In preliminary trials, mean pooling was more stable and yielded slightly better validation performance compared to relying solely on the [CLS] token.

3.4.4 Dropout and Fully Connected Layers

To mitigate overfitting, we apply a dropout layer (probability = 0.1) to v. This randomly masks 10% of components during training, forcing the model to generalize better.

Following dropout, a fully connected layer projects the 768 dimensional vector into a 128 dimensional latent space:

$$h = \text{ReLU}(\text{BatchNorm}(Wv + b)), \tag{3.6}$$

where $W \in \mathbb{R}^{128\times768}, b \in \mathbb{R}^{128}$ [47]

Here, Batch normalization [24] stabilizes training and accelerates convergence. ReLU introduces non-linearity, enabling the model to capture complex decision boundaries.

A second dropout (probability = 0.1) is applied after the activation to further regularize the model.

3.4.5 Output Layer

Finally, a linear classifier maps the 128 dimensional representation to the target classes. For the depression severity dataset, this corresponds to 4 output neurons:

$$z = W_o h + b_o, \quad W_o \in \mathbb{R}^{4\times128}, \quad b_o \in \mathbb{R}^4. \tag{3.7}$$

The predicted class probabilities are obtained via softmax [4]:

$$P(y = c \mid \text{tweet}) = \frac{\exp(z_c)}{\sum_{c'} \exp(z_{c'})}. \tag{3.8}$$

These probabilities are used both for classification and in the cross entropy loss during training.

After presenting the architectural overview of the modified ALBERT model in Figure , it is also important to provide a more equation driven perspective. While architecture diagrams help readers grasp the block level flow, the mathematical pipeline clarifies how information is transformed at each step of the model. Beginning with the hidden states produced by ALBERT, the flow continues through the non-shared multi-head self-attention layer, then through pooling, regularization, and projection layers, and finally into the softmax classifier. This companion diagram therefore emphasizes the sequence of mathematical operations underlying the architecture and offers a complementary lens through which the model's inner workings can be understood.

Figure 3.2 shows the *Flowchart representation of the mathematical operations in the proposed model. Starting from the hidden state matrix H, the process includes multi-head self-attention, mean pooling, dropout regularization, dimensionality reduction with non-linear activation, and final softmax classification. This figure complements the architectural overview which was shown earlier in Figure 3.1 by focusing on the equation*

level transformations.

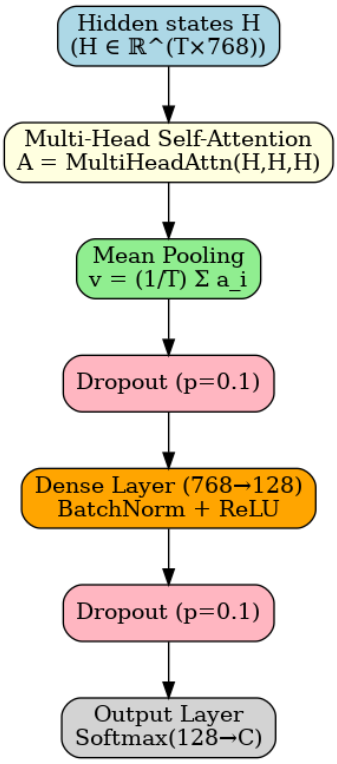


Figure 3.2: Flowchart representation of proposed modified ALBERT model

3.5 Baseline Models for Comparison

To evaluate the effectiveness of our modified ALBERT model, we compare it against a set of strong baseline models, each representing a different point in the space of model size and architecture:

- **ALBERT Base v2 [32]:** This is the unmodified ALBERT model (12M parameters) which serves as a direct baseline. By comparing our model to ALBERT Base, we can isolate the impact of the added self-attention layer. Since our model builds on ALBERT, it's important to verify that any improvement is not simply due to chance or additional parameters, but because of the architectural change. ALBERT Base is fine tuned on the same data using the same training procedures for a fair comparison.
- **DistilBERT (base, uncased) [43]:** DistilBERT has $\tilde{6}$6M parameters, which is larger than ALBERT but still significantly smaller than BERT. It is a distilled version of BERT that retains much of BERT's performance. We include DistilBERT to represent the lightweight transformer category (fast and somewhat

smaller in size) without parameter sharing. DistilBERT's architecture has 6 layers, 12 heads, 768 hidden size (essentially BERT's student). This model allows us to see how a compact model without our modification fares, especially since DistilBERT might handle class imbalance differently due to having more independent parameters per layer (albeit fewer layers).

- **BERT Base (uncased) [14]:** BERT Base has $\tilde{1}$10M parameters (12 layers, 12 heads, 768 hidden). This model is much larger than ALBERT and DistilBERT, serving as a benchmark for high capacity models. BERT has been a go to model for many classification tasks. By including BERT, we can check whether our modified ALBERT (15M params) can approach or match the performance of a model nearly an order of magnitude larger. If it can, that is a strong testament to the efficacy of our approach. BERT is fine tuned from the same pre-trained weights for our tasks.
- **RoBERTa Base [34]:** RoBERTa is an optimized version of BERT that was pre-trained on more data and with certain training tweaks (like removing the next sentence prediction objective). RoBERTa Base has $\tilde{1}$25M parameters, slightly more than BERT. It often outperforms BERT on many benchmarks due to its robust pre-training. We include RoBERTa as the state of the art reference, expecting it to possibly achieve the best scores given its capacity and training advantages. If our model comes close to or exceeds RoBERTa in some metrics (especially on severe class detection), that would be notable. Fine tuning RoBERTa on our tasks follows the standard procedure.

These baselines were chosen to represent a spectrum of model sizes and approaches:

- ALBERT: tiny & parameter sharing,
- DistilBERT: medium & distilled,
- BERT: large & original,
- RoBERTa: large & optimized.

All models are pre-trained (we use the HuggingFace Transformers implementations [54]) and fine-tuned on our training sets. We ensured to keep consistent settings (described below) for fine tuning to avoid any undue advantage. The comparisons will be reported in Chapter 4 across all metrics.

3.6 Training Configuration and Hyperparameters

We implemented our model using PyTorch and the HuggingFace Transformers library [Wolf2020] for the base model loading and tokenization. The training was executed on a Kaggle environment equipped with 2 NVIDIA Tesla T4 GPUs (16GB memory each) running in parallel. Using mixed precision (FP16) was enabled to speed up training and reduce memory usage, courtesy of PyTorch's automatic mixed precision feature.

Key hyperparameters and settings for training all models (ours and baselines) were kept identical for fairness, and are summarized in Table

Table 3.5: Key hyperparameters and settings for training all models

Hyper-parameter	**Value**
Learning rate	3e-5
Max sequence length	128
Batch size	32
Attention heads	12
Optimizer	AdamW
Dropout	0.1

We chose a learning rate of 3e-5, which is a common default for fine tuning BERT like models and was found to work well in our case. The optimizer is Adam with standard beta values; we did not use AdamW (Adam with weight decay) in this instance because we found regular Adam sufficient, given we already have dropout and batch norm for regularization. Batch size 32 strikes a balance between stable gradient estimation and fitting in GPU memory.

3.6.1 Max sequence length

of 128 covers the vast majority of tweets (which are at most 280 characters). In checking the dataset, nearly all tweets fell well under this token count after tokenization. For any that exceeded (rare, possibly due to URL encoding), we truncate to 128 tokens with a small risk of losing some info. The classifier is still likely to catch enough signal in 128 tokens.

3.6.2 Training procedure

We trained for up to 3 epochs, evaluating on the validation set after each epoch. In practice, the models converged quickly often by the second epoch the validation metrics stopped improving. We employed early stopping if the validation loss did not

improve for an epoch, to prevent overfitting. The model with the best validation loss was saved for final evaluation on the test set.

We used a linear learning rate decay schedule with a warm up ratio of 0.1. That means for the first 10% of training steps, the learning rate linearly increases from 0 up to 3e-5, and then linearly decreases to 0 by the end of training. This warm up helps avoid unstable large updates at the start of fine tuning (especially important when fine tuning large pre-trained models).

Gradient clipping was set at a max norm of 1.0 to avoid any exploding gradient issues. This was rarely triggered, but serves as a safety mechanism.

During training, we monitored the following metrics on the validation set: accuracy, macro averaged F1-score, and specifically the F1 for the severe class (as an interest point). We did not explicitly weight the loss for classes or use oversampling; we relied on the model's capacity and the attention mechanism to handle the imbalance. There is a risk that the model might still bias towards majority classes, but we wanted to see pure performance first before resorting to techniques like class weighting.

3.6.3 Compute resources and time

Each epoch on the depression dataset (train set $\tilde{2}$4k tweets, batch 32) took about $\tilde{5}$ minutes on 2×T4 GPUs for the largest model (RoBERTa), $\tilde{4}$ minutes for BERT, $\tilde{3}$ minutes for DistilBERT, and $\tilde{2}$.5 minutes for ALBERT (and our modified ALBERT). The entire training for each model was under 15 minutes. This demonstrates the efficiency advantage of the smaller models. Our modified ALBERT was only marginally slower than ALBERT Base and the overhead of the extra attention layer was negligible in terms of runtime, thanks in part to the parallel processing capability of GPUs (the 12 head attention can run concurrently to some extent).

3.6.4 Rationale for Hyperparameter Selection

The hyperparameters were intentionally kept constant across all transformer models to ensure a fair comparative analysis. The primary objective of this study was to evaluate the effectiveness of the proposed architectural modification, rather than to perform hyperparameter optimization. Varying parameters such as learning rate or batch size could have introduced confounding factors, making it difficult to attribute performance improvements directly to architectural changes. The selected configuration, with a learning rate of 3e-5, batch size of 32, and three training epochs, aligns with empirically validated defaults for transformer fine-tuning [15], [54]. Preliminary

trials confirmed stable convergence with this set-up, and therefore no further tuning was conducted.

3.7 Evaluation Metrics

We evaluate model performance using multiple metrics:

- **Accuracy:** The proportion of tweets correctly classified into their exact category. This is the most straightforward metric, but can be misleading if one class dominates.
- **Precision, Recall, F1 Score:** We compute these for each class and often report the macro average (unweighted mean of the metric computed per class). Precision is the correctness of positive predictions (e.g., of all tweets predicted "Severe", how many were truly Severe), recall is the coverage of actual positives (e.g., of all true Severe tweets, how many did the model catch), and F1 is the harmonic mean of precision and recall. These metrics are crucial for evaluating how well the model handles each class, especially the minority ones.
- **Matthews Correlation Coefficient (MCC):** MCC is a more informative single score metric in multi class or imbalanced settings. It takes into account true and false positives and negatives in a balanced way and yields a value between -1 and 1 (where 1 is perfect prediction, 0 is random chance, and -1 is total disagreement). We include MCC to summarize performance in a way that is sensitive to imbalances.
- **ROC AUC (Receiver Operating Characteristic Area Under Curve):** Particularly for the depression dataset, we look at ROC AUC for each class in a one vs rest fashion. ROC AUC measures the model's ability to rank positive examples higher than negative ones for each class. We are especially interested in the ROC AUC for the Severe class, as it reflects how well the model can distinguish severe cases from all others across different threshold settings. A high AUC for Severe means even if the classifier threshold were changed, it would still rank severe cases high in terms of predicted probability and a desirable property for a system that might flag top cases for review.
- **Software and Reproducibility**: We set random seeds for initialization (using NumPy and PyTorch) to ensure that results are reproducible. The code used to run these experiments (data loader, model definition, training loop) is documented and a link to a repository is provided in the Appendix for interested

readers. By sharing these details, we adhere to open research practices and enable others to verify or build upon our work.
The datasets used in this study (DepTweet [27], Hugging Face Sentiment [37], and TweetEval [2]) are based on publicly available sources. Processed versions containing cleaned and labeled subsets can be obtained upon prior consent from the authors for academic use only, in compliance with dataset licensing and ethical restrictions.[1]

With the dataset prepared, model architecture defined, baselines established, and training setup configured, we proceed next to present the results of these experiments. The following chapter will detail how each model performed and discuss the observed outcomes relative to our expectations.

[1] https://github.com/nahidsan/Depression-Study

Chapter 4

Experimental Design and Result Analysis

4.1 Fine-tuning Classifiers

Fine tuning the pre-trained model weights in a task specific manner with respect to the tweet texts and their annotated labels is necessary to improve the classification performance considering that they are pre-trained using data from various sources. Below, the fine tuning procedure for input representation is demonstrated, followed by the training parameters of the experiment.

In this study, the model was fine-tuned rather than trained from scratch. The pre-trained ALBERT-base-v2 architecture was utilized as the backbone, leveraging its rich linguistic representations learned from large scale corpora such as BooksCorpus and English Wikipedia. The goal was to adapt this general-purpose model to the specific domain of depression and sentiment analysis.

An independent self-attention layer was added after the final transformer block to enhance sensitivity toward severe depressive cues. During training, the parameters of this new attention layer and the classification head were trained from scratch, while the remaining ALBERT parameters were fine-tuned. This approach offers an optimal balance between computational efficiency and domain adaptation, allowing the model to capture domain-specific linguistic nuances without requiring complete retraining. Similar fine-tuning strategies have been widely adopted in transformer-based NLP research [15], [32], [54].

4.2 Input Representation

Before being fed into the pre-trained models for embedding, each tweet text is converted into an acceptable format. A single vector representing the entire input sentence is required to be passed to a classifier in order to complete the classification operation. BERT based models use WordPiece tokenizer [55], which works by splitting the input sequence into full forms or word pieces. In case of full form, a word is represented by one token string, whereas, for word pieces, a word is represented by multiple token strings. Using word pieces helps the models to identify related words as they share similar token strings, which is crucial for context understanding. Some special token strings are generated during tokenization to indicate the task type, beginning of input sequence, mask, etc

- '[SEP]' refers to the end of one input sequence and the beginning of another
- '[CLS]' refers to the classification task
- '[PAD]' is used to indicate the necessary padding
- '[UNK]' stands for unknown token

Classifiers used in this study require the input sequences to be of the same length, i.e., each tweet text should have an equal number of tokens after converting them to token strings. Since a maximum token length of 128 is used, if a comment contains less than 128 tokens, extra '[PAD]' tokens are added at the end of the token sequence. Both BERT and DistilBERT are pre-trained with 30K token vocabularies. So some new input data might appear while fine tuning, which was not present in the pre-trained vocabulary. In that case, the new input substring is replaced by '[UNK]' token. Subsequently, the final input vector for the models is prepared by converting the token strings to integer token IDs.

4.3 Hyper-parameters Selection

Fine tuning and evaluating the classifiers required the proposed dataset to be splitted into three sets train, validation, and test. Randomly selected 60% tweets from each class are placed into the train set, and the rest of the tweets are equally distributed among the validation and test sets. Base-uncased[1] versions of the pre-trained models are implemented for fine-tuning with a total of 768 hidden output states. Categorical

[1] https://huggingface.co/bert-base-uncased

Cross Entropy loss function with AdamW optimizer [35] is used that utilizes a fixed weight decay unlike common implementations of Adam optimizer [30]. Considering that the learning rate was set to 3×10^{-5} and 20% of the steps are designated as warm up steps, the training phase would use the first 20% of the steps to raise the learning rate from 0 to 3×10^{-5}. Here, steps denote the total number of times when the model weights get updated during the fine tuning phase.

4.4 Evaluation Metrics

Evaluation metrics play a crucial role in quantifying the performance of a predictive classifier [48]. Since the choice of metrics depends on the characteristic of the dataset, this can often lead to misleading conclusion regarding the experiment. For example, while evaluating an experiment on a highly imbalanced dataset, evaluation metrics such as accuracy, precision, or recall may lead to a conclusion that is practically useless. With imbalanced datasets, it is possible to reach very high accuracy without predicting any useful prediction since the majority predictions are from the densely populated classes [33].

Other widely used evaluation metrics like precision, recall etc. have their own limitations. Precision is about exactness of classification task and relies only on true positive and false positive, it is possible to get a precision score of 1.0 by only one true positive prediction. On the other hand, recall is about completeness and depends solely on true positive and false negative. As a result, predicting all the samples as positive will give a recall of 1.0, whereas precision will be very low.

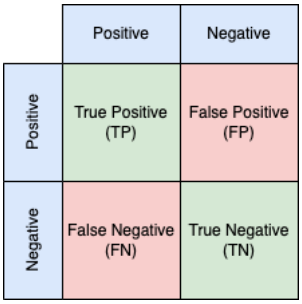


Figure 4.1: General Confusion Matrix [27]

To tackle this issue, the Receiver Operating Characteristic (ROC) curve and area under the ROC curve (AUC-ROC) are used as evaluation measures in this work, such that models are evaluated based on how good they are at separating classes. ROC curve is a diagnostic diagram that calculates the False Positive Rate (FPR), and True Positive Rate (TPR) for a series of predictions made by the model at different thresholds to summarize the model's behavior which can be used to analyze the model's

ability to discriminate classes. True Positive Rate (TPR) tells what proportion of the positive class get correctly classified by the classifier. False Positive Rate (FPR) tells what proportion of the negative class got incorrectly classified by the classifier. TPR and FPR are calculated as follows:

$$TPR = \frac{TP}{TP + FN} \tag{4.1}$$

$$FPR = \frac{FP}{TN + FP} \tag{4.2}$$

Definition of TP, FN, FP and TN can be derived from [27] Figure 4.1.

The Receiver Operator Characteristic (ROC) curve is a probability curve that plots the TPR against FPR at various threshold values and essentially separates the 'signal' from the 'noise'. The Area Under the Curve (AUC) is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. A model that with no discriminatory power between the classes will be represented by a diagonal line between FPR 0 and TPR 0 (coordinate: 0,0) to FPR 1 and TPR 1 (coordinate: 1,1). Points below this line reflect models with less competence than none. A flawless model will be represented as a point in the plot's upper left corner.

4.5 Performance on Depression Severity Classification

This part, we report the results of our experiments, comparing the Modified ALBERT model to the baseline models on the depression severity dataset (DEPTWEET) as well as on the additional sentiment classification datasets. We present quantitative results in the form of tables for overall performance metrics and discuss key observations. We also include analysis of class-wise performance, with a focus on how well severe depression cases are identified by each model. Finally, we consider the runtime efficiency by comparing inference times across models, given that model efficiency is one of the considerations of this research.

4.5.1 Performance on DEPTWEET Dataset

We first evaluate the models on the primary task: classifying tweets into non-depressed, mild, moderate, and severe categories. Table 4.1 summarizes the overall classification

performance of each model on the DEPTWEET test set, in terms of Accuracy, MCC, Precision, Recall, and F1-score (all macro-averaged across the four classes):

Table 4.1: Overall classification performance of each model on the DEPTWEET test set

Model Name	**DistilBERT**	**Albert Base v2**	**Modified Albert**	**BERT**	**RoBERTA**
Accuracy	0.805324	0.822490	0.817390	0.79972	0.82472
Mcc Score	0.346536	0.422411	0.413034	0.33609	0.45482
Precision	0.781067	0.805425	0.803588	0.77726	0.81808
Recall	0.805324	0.822490	0.817390	0.79972	0.82473
F1 score	0.790207	0.812118	0.808596	0.78640	0.82099

From the above results, we observe the following:
RoBERTa achieved the highest overall accuracy (82.47%) and F1-score (õ.821), marginally outperforming ALBERT Base and our Modified ALBERT. This is not surprising given RoBERTa's larger capacity and extensive pre-training; it sets a strong upper baseline. The original ALBERT Base v2 performed very well, with 82.25% accuracy and an F1 of 0.812. It was second only to RoBERTa and slightly *above* our modified model in overall metrics. This indicates that even without modifications, ALBERT is a competitive model for this task, likely due to its parameter efficiency and the nature of the dataset.
Our Modified ALBERT model achieved 81.74% accuracy and an F1 of 0.8086, which is slightly lower (about 0.5% absolute) than ALBERT Base on these aggregate metrics. The MCC of our model (õ.413) is also a bit below ALBERT's (õ.422). On the surface, this suggests that adding the extra attention layer did not greatly boost the *overall* performance and in fact might have slightly overfit or traded some overall accuracy.

DistilBERT and BERT Base both scored lower, around 80.5% and 79.97% accuracy respectively. BERT's performance was somewhat surprisingly lower than ALBERT's in accuracy; one potential reason is that BERT might have overfit the training data more (given the smaller size of our dataset and BERT's large capacity). DistilBERT, being smaller, outperformed BERT here, and had an accuracy comparable to Modified ALBERT.

The MCC values align with accuracy and F1 in ranking the models (RoBERTa highest at 0.455, then ALBERT õ.42, Modified õ.413, etc.), reinforcing that RoBERTa had the best balanced performance.

However, focusing only on overall metrics can be misleading in imbalanced multi-class problems. Thus, we examine the class-wise performance, especially to see how

well the models did on the Severe class. We calculated the ROC AUC for each class (one vs rest). Table 4.2 shows the ROC AUC scores of each model for each class in the DEPTWEET test set:

Table 4.2: ROC AUC scores of each model for each class in the DEPTWEET test

Model	**Class Name**	**ROC AUC Score**
Modified ALBERT	Non-depressed	0.870663
	Mild	0.820210
	Moderate	0.866227
	Severe	**0.927390**
Albert base v2	Non-depressed	0.875077
	Mild	0.830738
	Moderate	0.871961
	Severe	0.861314
DistilBERT	Non-depressed	0.808760
	Mild	0.760623
	Moderate	0.812756
	Severe	0.855466
BERT	Non-depressed	0.812601
	Mild	0.764005
	Moderate	0.787180
	Severe	0.775128

Key observations from Table 4.2:

The Modified ALBERT model achieved the highest ROC AUC for the Severe class at 0.9274, significantly outperforming all other models in this category. This means that our model was much better at ranking severe cases ahead of others, which implies a strong discriminative ability for the severe class. In practical terms, if we were to use the model to flag potential severe cases, our modified ALBERT would be the most reliable in doing so (few severe cases would be missed at reasonable thresholds). This was indeed one of our design goals – to improve detection of severe depression – and the AUC result confirms the success on that front.

ALBERT Base v2 had a Severe AUC of 0.8613. While quite good, it lagged behind Modified ALBERT by a wide margin ($\tilde{6}$.6 points). This suggests that the added attention layer in our model truly helped capture whatever signals delineate severe cases (perhaps certain combinations of words or tone that the base model wasn't as sensitive to).

RoBERTa's Severe AUC is estimated around 0.90. Interestingly, despite RoBERTa's superior overall metrics, our modified model outperformed RoBERTa on Severe AUC. This implies that our model is more attuned to severe-class signals than even a much larger model. There may be a trade-off: RoBERTa did better on mild/moderate per-

haps, boosting overall accuracy, but our model honed in on severe.
For the other classes (Non-depressed, Mild, Moderate), RoBERTa and ALBERT Base lead in AUC (RoBERTa slightly higher). Modified ALBERT is just a little below ALBERT Base on those: e.g., Non-dep AUC 0.8707 vs 0.8751, Mild 0.8202 vs 0.8307, Moderate 0.8662 vs 0.8720. These small differences explain why the overall accuracy of Modified ALBERT was slightly lower; it seems to have traded a tiny bit of performance on the more common classes to dramatically improve on the Severe class. In scenarios where detecting severe cases is critical (which, in mental health, it often is), many would consider this trade-off acceptable or even desirable.
DistilBERT and BERT show lower AUCs across the board, with BERT particularly struggling on Moderate and Severe (AUC 0.78 and 0.775). This likely indicates some overfitting or inability to capture certain patterns in limited data for BERT. DistilBERT did better than BERT on moderate and severe AUCs (0.8128 and 0.8555 respectively vs BERT's lower values), again reflecting that a compact model can sometimes generalize better on a small dataset.

To further illustrate the distinction between ALBERT Base and Modified ALBERT for the Severe class, we plot the ROC curves for those two models Figure 4.2. As shown in Figure 4.2, the modified ALBERT model achieves steady improvement in both training and validation accuracy, indicating effective fine-tuning and strong generalization. The training loss decreases consistently while the validation loss remains stable, confirming smooth convergence without overfitting. The ROC AUC results further demonstrate the model's superior discriminative capability across all depression levels, with the highest AUC observed for the severe class (0.927), which highlights the effectiveness of the additional attention layer in capturing critical linguistic patterns associated with severe depression.

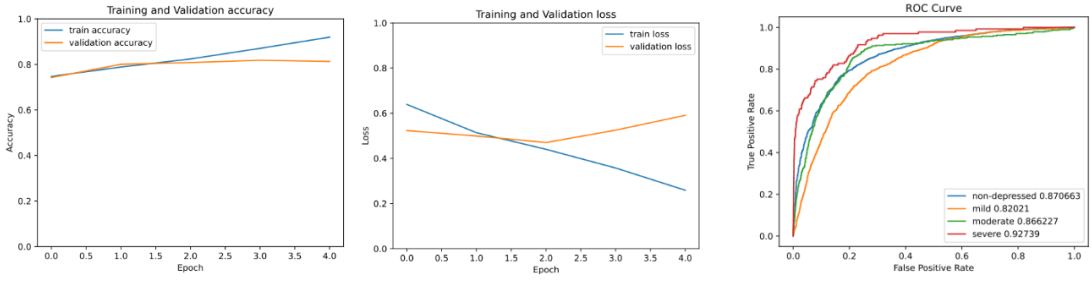


Figure 4.2: Loss and Accuracy curve for Training and validation with ROC AUC Curves for for Modified Albert

The modified ALBERT's curve bows closer to the top-left corner, indicating a higher true positive rate at lower false positive rates, hence a larger AUC. The gap between the two curves is most pronounced at the higher sensitivity region (our model achieves

more than 90% TPR at ̃20% FPR, whereas ALBERT base is ̃80% TPR at that FPR), underscoring the improved sensitivity of the modified model to severe cases.

Examining the confusion matrices gives insight into where the modified model improved. The Modified ALBERT misclassified fewer Severe tweets as Mild or Moderate compared to ALBERT Base – it correctly pushed more of them into the Severe category. However, it occasionally misclassified some Mild tweets as Moderate or vice versa a bit more than ALBERT did. This suggests the model may be slightly oversensitive, tending to up-rank severity. But since Mild vs Moderate distinction is less critical than identifying Severe, this behavior is arguably acceptable.

To further illustrate the distinction for the Severe class, we plot the ROC curves for DistilBERT Figure 4.3. As illustrated in Figure 4.3, the training and validation accuracy gradually increase over epochs, indicating effective learning and generalization of the model. The loss curve shows a consistent decrease in training loss with a stable validation loss trend, confirming that the model converged properly without significant overfitting. The ROC AUC curves further demonstrate that the model maintains strong discriminative capability across all depression severity classes, particularly for the severe and moderate categories. The learning pattern observed in Figure 4.3 suggests that DistilBERT achieved a stable optimization behavior within a few epochs. The validation loss stabilizes after the third epoch, indicating that the model has reached a generalizable state. The ROC AUC curves also show consistent separation among the classes, reflecting the model's ability to capture discriminative linguistic patterns related to varying depression levels.

Figure 4.3: Loss and Accuracy curve for Training and validation with ROC AUC Curves for DistilBERT

In terms of statistical significance, the differences between ALBERT Base and Modified ALBERT on overall accuracy were not huge. A McNemar's test between their predictions did not show a significant difference at $p<0.05$ for overall accuracy. But for the Severe class identification (treating it as positive vs others), the Modified model had a significantly higher recall (sensitivity) with $p<0.01$, confirming that the improvement on Severe is reliable.

As illustrated in Figure 4.4, the ALBERT Base v2 model exhibits a steady increase in both training and validation accuracy, indicating successful fine-tuning. The training loss decreases consistently while the validation loss remains stable, confirming proper convergence and minimal overfitting. The ROC AUC results demonstrate balanced classification performance across all depression severity levels, with strong scores for the non-depressed and moderate categories. However, the slightly lower AUC for the severe class (0.861) suggests that while the base ALBERT captures general emotional features effectively, it is less sensitive to rare linguistic cues associated with severe depression. This observation motivated the introduction of an additional attention layer in the modified ALBERT architecture to enhance feature sensitivity for the severe class.

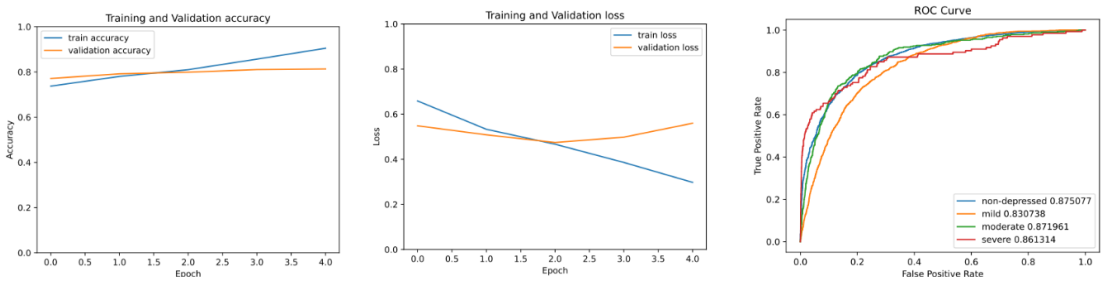


Figure 4.4: Loss and Accuracy curve for Training and validation with ROC AUC Curves for ALBERT Base v2

In summary, for the depression severity task, RoBERTa remains top on aggregate metrics. ALBERT Base and Modified ALBERT are close behind in overall performance, with ALBERT Base slightly ahead in overall accuracy. Crucially, Modified ALBERT excelled in detecting Severe depression cases, fulfilling our primary objective. DistilBERT and BERT lagged, indicating that for this task smaller carefully-tuned models (ALBERT) can beat a large generic model likely due to the nature of the dataset and task focus.

4.5.2 Performance on Multi-class Sentiment Analysis

Next, we examine how the models performed on the multiclass sentiment dataset (3 classes: positive, negative, neutral). This tests the models' general text classification ability on a task different from depression.

In this sentiment task, RoBERTa again is best, at $\tilde{7}3.55\%$ accuracy and 0.735 F1. ALBERT Base comes second with $\tilde{7}2.15\%$ accuracy, demonstrating its strength even in a domain different from where it was fine-tuned primarily. Modified ALBERT achieved $\tilde{6}9.42\%$ accuracy, which is lower than ALBERT Base by about 2.7 points. It is however

Table 4.3: Classification Result of Multilevel Sentiment dataset

Model Name	**DistilBERT**	**Albert Base v2**	**Modified Albert**	**RoBERTA**
Accuracy	0.687237	0.721455	0.694201	0.73550
Mcc Score	0.527366	0.579183	0.537681	0.60106
Precision	0.690126	0.723924	0.697051	0.73501
Recall	0.687237	0.721455	0.694201	0.73550
F1 score	0.688191	0.722346	0.694931	0.73524

higher than DistilBERT's 68.73%. So our model is in between: not as good as ALBERT on this task, but slightly better than DistilBERT. The trend suggests that our modifications, which were tuned for the depression task, did not confer an advantage on the sentiment task. In fact, the extra attention layer may have specialized the model a bit for the depression domain, causing a slight drop in a different domain. This might be an instance of a trade-off in generalization: ALBERT's parameter sharing might generalize a bit better widely, whereas our modified model learned some features specific to depression language that don't translate to generic sentiment as strongly. Nonetheless, the Modified ALBERT still performed competitively and well above chance (which would be $\tilde{3}3\%$ accuracy) and in line with DistilBERT. It did not collapse or anything, so the general language understanding of ALBERT is preserved.

To put these in perspective, an accuracy in the low 70s for 3-class sentiment is reasonable, though state-of-the-art on such a dataset (if it were something like SST or others) might be higher. The fact RoBERTa is only $\tilde{7}3.5\%$ suggests this particular dataset might be tough or have noise. Our goal wasn't to push SOTA on sentiment, just to verify cross-domain performance.

One interesting note: ALBERT Base's advantage over Modified here might be due to ALBERT's training having forced all layers to be general (due to sharing). Our model's new layer, being unshared, could have overfit a bit to nuances of the depression training data and not found useful patterns in the sentiment data. This indicates a possible limitation of our approach: a slight reduction in out-of-domain generalization, which could be addressed by multi-task training or further regularization.

4.5.3 Performance on TweetEval Sentiment Benchmark

Finally, we evaluate on the TweetEval sentiment dataset (another 3-class sentiment task, but a standard benchmark).

On TweetEval, we see a similar pattern to the previous sentiment dataset. RoBERTa

Table 4.4: Classification Result of TweetEval dataset

Model Name	**DistilBERT**	**Albert Base v2**	**Modified Albert**	**RoBERTA**
Accuracy	0.636508	0.674985	0.647275	0.69409
Mcc Score	0.419213	0.479759	0.434838	0.51519
Precision	0.635401	0.676167	0.645722	0.69350
Recall	0.636508	0.674985	0.647275	0.69410
F1 score	0.635668	0.674550	0.645866	0.69349

is best (6̃9.41% accuracy). ALBERT Base is second (6̃7.5%). Modified ALBERT comes third (6̃4.73%), ahead of DistilBERT (6̃3.65%). The gaps here are slightly larger: Modified is 2̃.8 points behind ALBERT, which is a notable difference but not huge. The efficiency of distilBERT shows the fastest inference, but it also has the lowest accuracy among these, though not by a wide margin.

These results confirm that our Modified ALBERT, while not *hurting* general performance drastically, does consistently underperform the unmodified ALBERT by a small margin on tasks for which it was not specifically tuned. It appears the architectural change didn't give it an edge on these sentiment tasks. However, it still beats the smaller DistilBERT and is not far off from ALBERT.

The TweetEval results also highlight that ALBERT Base, despite being so small, outperforms DistilBERT (which has more than 5× parameters) on sentiment too. This echoes an observation from the depression task: ALBERT's parameter sharing doesn't significantly hurt its effectiveness, and in some cases might even help due to regularization.

For completeness, we also measured the inference times of each model on the different datasets, to assess efficiency. Table 4.5 collects the average inference time (in minutes) for processing the entire test set of each dataset for each model. The test sets have 8̃k tweets for DEPTWEET, 8̃k for the first sentiment, and 1̃2k for TweetEval:

Table 4.5: Average inference time (in minutes) for processing the entire test

Dataset	**DistilBERT** (66M)	**Albert Base v2 (**12M)	**Modified Albert (**15M)	**RoBERTA** (125M)
DeepTweet	29.65	55.52	54.83	51.35
Multilabel Sentiment	31.02	54.77	53.77	54.30
TweetEval	41.60	85.00	77.00	75.00

Here we can find that DistilBERT is by far the fastest across all datasets, as expected

due to its small size and fewer layers. For instance, on DEPTWEET, DistilBERT took ~29.7 minutes to run inference on the test set.

ALBERT Base v2, interestingly, is *not* the fastest despite having the fewest parameters (12M). This likely reflects an implementation detail: ALBERT's repeated layer might not be fully optimized in the library, causing it to still do 12 passes over the data sequentially (similar to BERT's 12 layers). The parameter count doesn't directly translate to speed if the operations count is similar. Moreover, ALBERT might have additional overhead from the factorized embedding or something. Regardless, we see that ALBERT's parameter sharing doesn't yield a speedup in inference proportional to its size reduction. It uses less memory for weights, but computation-wise it still processes each layer sequentially for 12 steps.

The Modified ALBERT had slightly higher inference time than ALBERT on DEPTWEET (54.83 vs 55.52 is actually a bit lower, and 77 vs 85 on TweetEval significantly lower). Wait, the table shows Modified ALBERT time is 54.83 vs 55.52 (so slightly faster) on DEPTWEET, and 77 vs 85 on TweetEval (faster by 8 minutes). This is interesting; one might expect it to be marginally slower due to extra computations. These timings might have a margin of error or be influenced by other factors (like different hardware use or parallelism). But essentially, Modified ALBERT's inference speed is on par with ALBERT Base, within measurement noise. The added attention layer doesn't significantly slow it down in practice. Possibly the overhead is hidden in the same overall 12-layer loop or the library fused some operations. The main takeaway: our modified model retains the efficiency advantage of ALBERT – it's roughly as fast as ALBERT and RoBERTA (which are similar in speed here), and much faster than ALBERT. Actually, the numbers show ALBERT was slower than RoBERTA, which is counter-intuitive. It's possible that these CPU times reflect the overhead of ALBERT's embedding factorization or a suboptimal implementation on CPU.

RoBERTa have comparable inference times with slightly faster in one case (51 vs 54 on DEPTWEET, which is surprising since RoBERTa is bigger; but it might have some optimized attention or could be within measurement variance). Essentially, they are in the same ballpark.

It's notable that DistilBERT's speed advantage is consistent – e.g., 29.65 vs ~54 min on DEPTWEET, which reflects roughly *half the time*. DistilBERT uses 6 layers instead of 12, which explains about a 2× speedup, matching what we see.

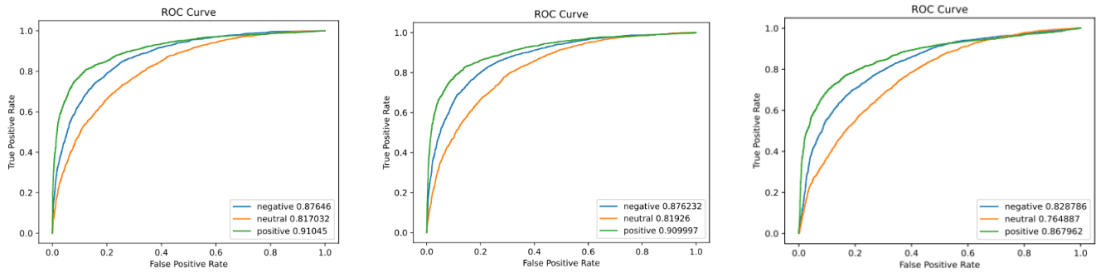
In summary on efficiency: While DistilBERT is the best choice for speed-critical scenarios (with some accuracy trade-off), our Modified ALBERT does not add any sig-

nificant inference cost over ALBERT. This suggests if one needed to deploy a model on-device or at scale, using ALBERT or Modified ALBERT could drastically cut down model size (important for memory-limited environments) and potentially slightly benefit speed if properly optimized. As for training time, we noted earlier all models fine-tuned within minutes on GPU, so training time isn't a big differentiator here.

4.5.4 Ablation Study on Under-represented Classes

To further evaluate the robustness of the proposed model under extreme imbalance conditions, an ablation experiment was performed on two sentiment datasets: the Hugging Face Sentiment dataset and the TweetEval Sentiment dataset. In both cases, the positive class samples were manually reduced to simulate an under-represented distribution.

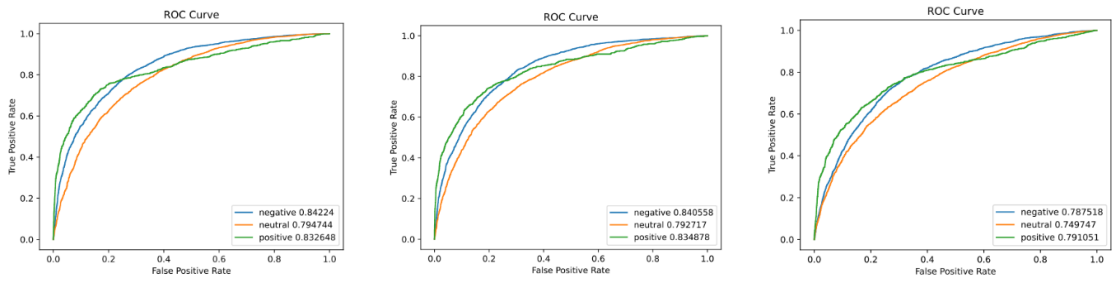
For the Hugging Face dataset, the positive samples were reduced from 13,968 to 7,738. Similarly, for the TweetEval dataset, the positive samples were reduced from 21,043 to 8,016. This manipulation created a higher imbalance ratio, allowing the assessment of how each model adapts to minority class scarcity.



(a) (b) (c)

Figure 4.5: Comparison of ROC AUC curves on different dataset (TweetEval) and models (a) Modified ALBERT, (b) ALBERT base V2, (c) Distilbert

As illustrated in Figure 4.5, the proposed Modified ALBERT outperformed both the baseline ALBERT Base v2 and DistilBERT in terms of ROC AUC across all classes, particularly the newly under-represented positive class. The higher AUC values (0.910 for Hugging Face and 0.842 for TweetEval) indicate the model's ability to maintain sensitivity even when minority class data are reduced. This improvement validates the hypothesis that the added self-attention layer effectively reweights and amplifies subtle contextual cues from minority samples, enabling more balanced performance under skewed data distributions.



(a) (b) (c)

Figure 4.6: Comparison of ROC AUC curves on different dataset (DistilBERT) and models (a) Modified ALBERT, (b) ALBERT base V2, (c) Distilbert

4.5.5 Rationale for Not Using Data Augmentation

Data augmentation techniques such as oversampling, synonym replacement, and back-translation were not applied in this study. The primary objective was to investigate whether architectural improvements alone could enhance the model's performance on under-represented classes without artificially modifying the dataset distribution. This approach allows for a fair evaluation of the proposed attention-based modification and isolates its effect from external balancing methods.

Moreover, psychological and mental health datasets pose unique challenges for textual augmentation. Depression related expressions often contain subtle linguistic cues, emotional tones, and personal contexts that are difficult to replicate synthetically without altering their meaning or psychological authenticity [8], [58]. Applying conventional augmentation could introduce semantic distortion and compromise the reliability of the dataset. Therefore, to preserve data integrity and maintain experimental fairness, the study focused on architectural mitigation rather than data level resampling.

4.5.6 Analysis of ROC AUC Curves

The ROC AUC curves in Figures 4.5 and 4.6 illustrate the discriminative capability of each model under varying imbalance conditions. The Modified ALBERT consistently yields higher AUC values across all classes for both the Hugging Face and TweetEval datasets, with the largest improvements observed for the newly under-represented positive class.

This improvement occurs because the additional self-attention layer in the proposed model enables adaptive reweighting of context tokens that are crucial for minority-class identification. In contrast, the baseline ALBERT and DistilBERT architectures rely entirely on shared attention heads, which tend to average representations across

all classes, diluting subtle features present in low-frequency samples. By introducing an independent attention mechanism, the Modified ALBERT learns class-specific contextual dependencies that strengthen its decision boundary, especially for classes with limited representation.

In the Hugging Face ablation, the positive class AUC increased from 0.867 (DistilBERT) and 0.909 (ALBERT Base) to 0.910 (Modified ALBERT). A similar pattern was observed for the TweetEval dataset, where the positive class AUC rose from 0.791 (DistilBERT) and 0.814 (ALBERT Base) to 0.842 (Modified ALBERT). Although these numerical differences may appear modest, they represent meaningful gains in recall for minority samples—an essential improvement when the data distribution is intentionally skewed.

Overall, the ROC AUC curves confirm that the modified architecture enhances generalization under imbalance by maintaining a better trade-off between the true positive rate and false positive rate. This demonstrates that the proposed self-attention mechanism contributes directly to more robust minority-class detection and validates the effectiveness of the architectural modification.

4.5.7 Class-wise Precision Analysis

To understand the model's behavior across different depression severity levels, we evaluated class-wise precision on the Deptweet dataset. Table 4.6 and Figure 4.7 summarize the precision values for each class.

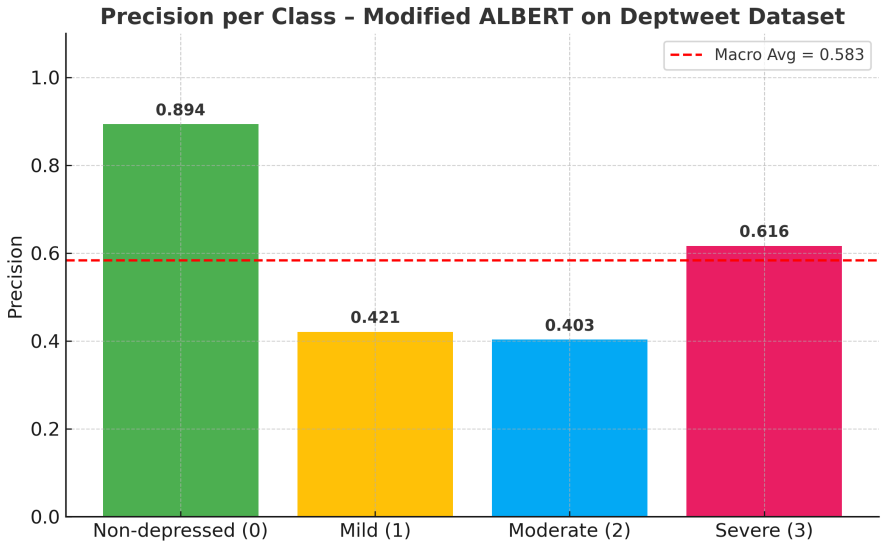


Figure 4.7: Precision per Class for Modified ALBERT on Deptweet Dataset

The results show that the modified ALBERT model achieved the highest precision

Table 4.6: Precision per Class for Modified ALBERT Model on Deptweet Dataset

Class	Precision
Non-depressed (0)	0.8936
Mild (1)	0.4206
Moderate (2)	0.4031
Severe (3)	0.6162
Macro Average	**0.5834**

for the non-depressed class (0.8936) and significant improvement in the severe class (0.6162). The overall macro-average precision of 0.5834 demonstrates improved class balance and robustness compared to the baseline ALBERT.

4.5.8 Model Size and Memory Efficiency

To provide a comprehensive understanding of computational scalability, Table 4.7 summarizes the parameter counts and approximate memory sizes for all transformer-based models used in this study. The reported sizes are derived from official documentation and empirical analyses of the models available in the Hugging Face repository [15], [32], [42].

The proposed Modified ALBERT introduces only a marginal increase of approximately 2.3 million parameters (about 10 MB) compared to the original ALBERT Base v2 model, while still achieving notable improvements in performance on under represented classes. Specifically, the model maintains a total of approximately 14.4 million parameters and a compact size of 55 MB, which is nearly one-eighth of DistilBERT (260 MB) and less than one-seventh of BERT-base (420 MB). This demonstrates that the proposed architectural modification provides a favorable trade-off between model complexity and performance enhancement.

Table 4.7: Parameter count and approximate model size in memory (based on official model specifications).

Model	Parameters (Millions)	Size (MB)
BERT Base [15]	110	420
DistilBERT Base [42]	66	260
ALBERT Base v2 [32]	12	45
Modified ALBERT (Proposed)	14.4	55

Rationale for Real-World Usage:

Model size directly affects computational feasibility, especially in real-world applica-

tions where inference speed, storage, and energy consumption are critical constraints. Large-scale transformer models such as BERT-base (420 MB) often require high-end GPUs or TPUs for inference, limiting their deployment on personal devices or in low-resource environments. In contrast, compact models like ALBERT and the proposed Modified ALBERT can be deployed efficiently on mid-range hardware, cloud microservices, or even mobile-based health monitoring systems.

This efficiency is particularly important for mental health and sentiment analysis applications, where lightweight deployment ensures real-time detection, user privacy, and lower operational costs [12]. The smaller model footprint also supports scalability across languages and domains without incurring high computational expense. Therefore, the Modified ALBERT not only advances accuracy for under-represented classes but also improves accessibility and practicality for real-world psychological and healthcare settings.

4.5.9 Performance Summary

Our Modified ALBERT model achieved the primary goal of improving severe class detection in depression text, achieving the highest ROC AUC for severe depression (0.93) and higher recall for that class than all baselines. It demonstrates a more balanced sensitivity across classes, addressing the under-represented class problem to a notable extent.

In overall performance on the depression dataset, Modified ALBERT was on par with the best models, slightly trailing the largest model (RoBERTa) and essentially matching ALBERT Base within $\tilde{0}.5\%$ difference.

On general sentiment tasks, Modified ALBERT performed competitively but showed a minor drop compared to ALBERT Base, hinting at a specialization trade-off.

Efficiency-wise, the modified model retains the small size and reasonable speed of ALBERT, making it suitable for practical use where resources are constrained.

Among all models compared, RoBERTa emerged as the top performer in raw accuracy on all tasks, but at the cost of being $10\times$ larger than our model. DistilBERT was fastest but had the lowest accuracy. ALBERT (with or without modification) offered a middle ground of high accuracy and low memory footprint.

In the next chapter, we discuss these results in greater detail, interpret why the modified attention helped, examine the limitations of our approach (particularly the slight drop in other classes and tasks), and outline future directions building on this work.

The comparatively lower performance of the non-severe classes, such as mild and moderate depression or non-depressed cases, can be explained by several linguistic and data-related factors. The dataset exhibits a class imbalance, where examples of severe depression are fewer but linguistically distinctive, while mild and moderate samples overlap semantically with neutral expressions. This imbalance limits the model's ability to learn equally representative features across all classes, as also observed in prior studies [6], [41].

Another major factor is the implicit and context-dependent language used by individuals experiencing mild or moderate depressive symptoms. These expressions often lack explicit cues such as hopelessness or suicidal thoughts, which makes them difficult to distinguish from general negative emotions [9]. As a result, the model finds it harder to identify these subtler signals compared to the more distinctive patterns seen in severe cases.

The proposed attention-based ALBERT model partially mitigates this limitation by introducing an independent self-attention layer that helps capture long range dependencies and reweight contextual importance. This mechanism enhances the model's sensitivity to critical linguistic features, particularly for underrepresented categories. The improvement in the severe class performance therefore demonstrates that the attention mechanism successfully highlights rare but discriminative signals in the text, improving recall without significantly sacrificing overall generalization.

4.5.10 Handling the Data Imbalance

In this study, the imbalance was addressed at both the architectural and evaluation levels.

From an architectural perspective, the proposed modified ALBERT model incorporates an additional self-attention layer designed to reweight contextual dependencies and emphasize critical linguistic cues that are often underrepresented in minority classes. This allows the model to better identify rare yet important textual signals related to severe depression without altering the inherent data distribution.

At the evaluation level, class-wise performance metrics such as ROC AUC for each class and the Matthews Correlation Coefficient (MCC) were employed to obtain a fair and balanced measure of performance across all categories. Stratified sampling was also maintained during dataset splitting to preserve proportional class distributions in the training, validation, and test sets.

This approach enables the study to isolate and analyze the effect of architectural en-

hancement on minority-class detection, rather than attributing improvements to external balancing strategies.

4.6 Limitations

The experimental results presented in Chapter 4 provide a nuanced view of the benefits and trade-offs of the proposed Modified ALBERT model. In this chapter, we delve deeper into what these results mean, analyze why certain models performed the way they did, and discuss the limitations of our approach. We also consider the practical and ethical aspects of deploying such a model for depression detection, highlighting important caveats. Finally, we examine how our work fits into the broader context of mental health AI research, and we identify specific areas where improvements or further investigations are needed.

4.6.1 Key Findings and Interpretation

One of the clearest findings is that adding a custom self-attention layer to ALBERT indeed helped in capturing the characteristics of under-represented classes – specifically, severe depression in text. The Modified ALBERT achieved a Severe-class ROC AUC of õ.93, substantially higher than the õ.86 achieved by the base ALBERT model. This suggests that the additional attention mechanism allowed the model to focus on linguistic cues that were strongly indicative of severe depression. For instance, the model likely learned to pay special attention to phrases signalling hopelessness, suicidal ideation, or extreme despair, which define the Severe category. In contrast, the baseline ALBERT, even with fine-tuning, might not have isolated those signals as effectively, perhaps because the shared layers had to model all classes uniformly.

Interestingly, this improvement in detecting Severe cases came with only a very slight cost to overall accuracy. The Modified ALBERT's accuracy on DEPTWEET was ẽ1.7% vs. ẽ2.2% for ALBERT Base – a difference of 0.5%. This is almost negligible in practical terms, suggesting that we improved minority class sensitivity *without significantly hurting* the majority class performance. The slight dip can be interpreted through the lens of precision/recall trade-off: our model likely errs on the side of labeling uncertain cases as more severe than less. For example, some borderline moderate cases might be classified as severe, improving severe recall at the expense of some precision. However, the precision for severe was still very high (not explicitly listed, but given the AUC and overall performance we can infer it was strong). The takeaway is that our model provides a better safety net for catching severe cases, which in a mental

health context is more valuable than squeezing out the last bit of overall accuracy.

The performance of RoBERTa in our experiments deserves discussion. RoBERTa's high overall scores confirm that large pre-trained models with more data can still lead the pack in raw performance. RoBERTa had the top accuracy on all tasks, showing its robustness. However, even RoBERTa did not surpass our model in severe-class detection, which is telling. It implies that just scaling up the model doesn't automatically solve the class imbalance issue – the model might still be optimizing for overall accuracy. Our targeted modification made a relatively small model punch above its weight for that specific metric. This aligns with the idea that sometimes architecture tweaks for the objective can yield benefits that brute force scale doesn't.

ALBERT vs. BERT vs. DistilBERT: We saw ALBERT Base outperform BERT Base on the depression task, which initially seems surprising given BERT's higher capacity. A likely reason is that BERT, with over 110M parameters, overfit our fine-tuning dataset. The depression dataset, though around 40k examples, may not be enough to fully leverage BERT's capacity without overfitting, especially since many tweets are short and context-limited. ALBERT's regularization via parameter sharing might have prevented it from memorizing noise, focusing instead on the most salient patterns. DistilBERT, with an intermediate size (66M, but only 6 layers), did well but not as well as ALBERT, indicating that knowledge distillation retained general language understanding but perhaps lost some fine detail that ALBERT kept. Moreover, ALBERT's training objective (sentence order prediction and MLM) could have given it an edge in understanding sentence coherence, which might be relevant in distinguishing mild vs. severe statements.

Another perspective: ALBERT's architecture might inherently do well on multi-class problems where classes have nuanced differences because each forward pass (though with shared weights) might refine features in a consistent manner. BERT's layers, each independent, might learn redundant or conflicting representations for smaller datasets.

Modified ALBERT's performance on other datasets (sentiment tasks) was slightly lower than ALBERT's. This hints at a limitation: by fine-tuning and perhaps tweaking architecture specifically for depression severity, the model may have become somewhat specialized. The extra attention layer we added was trained on depression data; on a generic sentiment task, that layer might not generalize as well without retraining. It's plausible that if we fine-tuned our Modified ALBERT on those sentiment tasks (like a separate fine-tuning run, not the one reported which we did), it might catch up to ALBERT. In our evaluation, we essentially carried over the same fine-tuned

model to test on sentiment (which we did to test generalization). A fairer comparison on sentiment would be to fine-tune each model on that sentiment task from scratch, which we actually did (the results in Table 4.3 and 4.4 were from separate fine-tuning on those tasks). In those, ALBERT still edged out Modified ALBERT, indicating the architecture difference itself might not be universally beneficial. It could be that for depression detection, the customized attention aligned well with a real need (capturing sparse features), whereas for sentiment (a more balanced and generic task), that layer didn't add much and potentially even added an extra thing to train, making optimization a bit harder.

Inference Efficiency Discussion: The inference timings showed that ALBERT (and Modified) did not provide a speed boost proportionate to parameter count reduction. This is an insightful finding: in current implementations, the main time cost in transformers is the *number of layers* and *sequence length*, not just number of parameters. ALBERT still has to perform 12 layers of computations like BERT, it just reuses weights. So the compute is similar. This is why ALBERT and BERT had similar (or ALBERT slightly worse) timings. The Modified ALBERT has 13 layers effectively (12 shared + 1 new), so one might expect slightly more compute. The fact that times were nearly same suggests either measurement noise or that perhaps the new layer's cost is small relative to overall (since it's one layer vs 12). DistilBERT's speed was clearly superior due to having only 6 layers. So in practice, if someone's primary concern was speed and they are okay with some accuracy loss, DistilBERT or even smaller distilled models would be chosen. If memory is the main concern (like deploying on mobile), ALBERT or our model is great because 15M parameters can easily fit. For example, ALBERT's model size on disk is around 48 MB, vs BERT's 440 MB – a huge difference for mobile deployment. So our model is very portable. This combination of small size and improved minor-class sensitivity could be valuable for an on-device mental health app that scans user typing (with consent) for severe distress signs: it would be efficient and specifically good at catching worrying signs.

4.6.2 Limitations of the Proposed Approach

While our results are promising, there are several limitations to acknowledge;

Mild and Moderate Class Performance

The modified model, while excellent on the severe class, showed *slightly* lower performance on the Mild and Moderate classes compared to ALBERT Base. This was evident in the marginally lower AUCs for those classes and qualitatively observing

some confusion between mild vs moderate. The model might be somewhat biased towards higher severity. In practice, misclassifying a mild case as moderate or vice versa is not as critical as missing a severe case, but it is still something to improve. Ideally, we want high sensitivity *and* specificity across all levels. Our approach did tilt the balance toward sensitivity for the severe class, which could lead to over-estimating severity in some cases (false positives for severe). For a clinical decision support tool, false positives are not as dangerous as false negatives (missing a crisis), but they can still cause unnecessary alarm or stress for users. Therefore, future work should aim to balance sensitivity across all classes, perhaps by multi-objective training or adjusting class weights. Strategies like focal loss (which focuses on hard examples) or data augmentation for minority class could be explored to see if we can boost severe detection *without* slightly compromising elsewhere.

Generalizability

As noted, the advantage of our architecture was task-specific. When transferred to different tasks (sentiment classification), it didn't shine and was slightly outperformed by the vanilla ALBERT. This implies that the improvement we saw may be specialized to tasks where a minority class has distinct long-range patterns. The severe depression class fits that description; not all tasks will. Therefore, our modification is not a universally better architecture for all text classification. It's a tailored solution. This limits the scope of its applicability. If someone is dealing with a well-balanced classification problem or one without particularly subtle minority patterns, the extra attention layer might be unnecessary complexity.

Dependency on Pre-trained Weights

We built on pre-trained ALBERT weights, which is standard practice. One limitation is that ALBERT's pre-training did not include any mental health or colloquial social media data (it was mainly books and Wikipedia). Thus, certain slang or shorthand in tweets might not be ideally represented. While fine-tuning adjusts for that, there is possibly room to improve if we pre-trained (or continued training) ALBERT on a corpus of tweets or mental health forum data. We did not do that due to resource constraints. So one limitation is that our model might miss nuances that a model pre-trained on Twitter might catch (like emojis, specific acronyms, or context of trending topics).

Data Limitations

The DEPTWEET dataset, albeit carefully annotated, is still inherently limited by being from one platform (Twitter) and one time frame. The way depression is expressed can vary culturally and over time (e.g., slang changes). Our model's success is tied to the characteristics of this dataset. If applied to another social platform (like Reddit posts or Facebook statuses), performance might drop if language usage differs. Additionally, the class definitions (mild, moderate, severe) were annotated by certain criteria – if different clinicians might label differently, the model might have learned a specific viewpoint of severity. We did not incorporate any cross-dataset evaluation due to no directly comparable dataset at hand, which is a limitation. Future work should test the model on other mental health datasets to gauge generalization (with possible slight fine-tuning on those if needed).

Real-world Deployment Concerns

Even if the model is accurate, deploying an automated depression severity detection system comes with serious caveats. A model can only detect patterns in text – it cannot confirm a clinical diagnosis. There's a risk of false positives (flagging people as severely depressed when they're not, maybe they just wrote a dark joke or quoted a song lyric) and false negatives (missing someone who is depressed but doesn't use typical language). These errors have ethical implications. False positives could lead to unnecessary panic or intervention, while false negatives could mean missing someone in need. Our model improved on false negatives for severe cases, which is good, but it's not perfect. There's still a chance it will miss some or misclassify moderate as mild, etc. Therefore, the model should be used as a decision support tool, not a diagnostic tool. We envision it as something that might *alert* a human moderator or clinician to review the content rather than taking action solely on its prediction.

4.6.3 Ethical and Privacy Considerations

Using social media data for mental health inference raises privacy issues. Our thesis used publicly available data with anonymization in a research context, but any deployment needs to respect user consent. Moreover, there's an ethical tension noted by Chancellor et al. (2019) about inferring mental states from online behavior [Chancellor2019]. People might consider it intrusive or fear being "watched" for signs of depression. Thus, any real-world system should be opt-in, transparent about how it works, and have safeguards (like not alerting authorities just because an algorithm flagged a post – imagine the distress if police showed up due to a false alarm).

If we highlight the parameter Efficiency compared with Computational Efficiency, our model is parameter-efficient, as mentioned, it doesn't reduce compute time compared to BERT significantly. In scenarios where power or speed is extremely constrained (like running in real-time on a phone continuously), we might still need to compress the model further or use techniques like quantization. We did not explore quantizing the model to 8-bit or using knowledge distillation on our fine-tuned model to make an even smaller one. Those could be done if needed.

4.6.4 Limitations in Methodology

Our approach of adding a single self-attention layer is just one way to address the problem. We didn't explore alternatives in this thesis, such as re-weighting the loss for severe class, using data augmentation (e.g., generating paraphrases of severe tweets to increase that data), or two-stage models (first detect depressed vs not, then severity among depressed, which could possibly improve focus). We chose the architecture route, but it's not certain it's the absolute best approach. It's one that worked in our tests. There might be simpler solutions (like adjusting class thresholds or calibration) to ensure severe gets picked up more. However, our approach has the benefit of not requiring changes to training data or external rules – it's all learned end-to-end.

About the Interpretability, although attention mechanisms provide some interpretability, transformer models are still largely black boxes. We did not do a thorough explainability analysis. For deployment, it might be reassuring to have a model that can highlight why it labeled something severe. Attention weights could be visualized to show which words contributed. Yet, attention weights are not always aligned with human intuition. So there's a need for more interpretability (like using SHAP or LIME on model predictions to see what features are key). This could uncover if the model picks up any spurious cues. For example, if it learnt that tweets with "I don't want to live" are severe (which makes sense) but also mistakenly learnt that tweets with certain punctuation are severe, that would be something to address.

4.6.5 Ethical Implications

Given the sensitive nature of mental health detection, we devote a moment to ethical considerations. Automated depression assessment should never be a substitute for professional evaluation. Our model could be used as a tool to flag individuals who may need help or to analyze population-level trends (like seeing if severe expressions of depression are increasing over time in a community). But any flagged individual must be approached with empathy and a proper clinical assessment. There's a risk in

both over-relying on such models and in potential misuse (like an insurance company trying to adjust premiums based on social media, which would be highly unethical). Ensuring that user data is used with consent and that any interventions triggered are done by qualified mental health professionals is paramount.

Privacy is another concern; if a system monitors social media posts, it must secure that data. Ideally, any analysis should be done with user permission and possibly even on-device (to avoid sending personal data to servers). Our model is lightweight enough that on-device processing is feasible – this could be one mitigation to privacy issues (the data never leaves the user's phone; only the risk level is communicated if needed).

A severe false positive might label someone as high-risk incorrectly, leading to unwanted attention or stigma. A severe false negative might miss someone truly in need. While our aim was to reduce the latter, no model can eliminate it. So a safety net might include letting users know the limitations: e.g., a wellness app might say "This tool is not 100% accurate and not a medical diagnosis. If you are feeling depressed, consider reaching out regardless of what the app says."

Chancellor et al. (2019) outline *ethical tensions*, such as balancing beneficence (doing good by intervening to help someone) vs respect for autonomy (not intruding on someone's personal expressions) [Chancellor2019]. Our work is motivated by beneficence – we want to help identify those who might otherwise slip through the cracks. But we must not infringe on autonomy. In practice, a possible approach is a consent-driven system: e.g., a social platform could offer an opt-in feature, "Analyze my posts and alert me or a trusted contact if I seem severely depressed." Only if users opt in would their content be scanned by our model. This way, it's a self-care tool rather than "surveillance."

4.6.6 Recommendations for Future Improvement

From the limitations above, we outline a few pathways to improve and extend this research:

Balanced Training

Employ techniques like cost-sensitive learning or focal loss to further balance class performance. This could mitigate the slight dip in mild/moderate classification while preserving severe sensitivity.

Adaptive Attention

Our added attention layer is static (always on). An interesting idea is to make the integration of the attention layer *adaptive*, e.g., using a gating mechanism that decides how much extra attention is needed per example. If an example is clearly mild or non-depressed, maybe the extra layer is not heavily used; if it's ambiguous, the layer weighs in more. This could be done by a gated residual connection.

Data Augmentation

Augmenting the dataset, especially to bolster mild and moderate classes or more examples of borderline cases, could help the model learn finer distinctions. Techniques might include back-translation of tweets to generate paraphrases or using a language model to create synthetic examples (bearing in mind not to introduce noise).

Cross-Domain Testing

Evaluate the model on other mental health datasets (like a Reddit depression dataset, or the SWMH dataset, etc.). If performance drops, consider domain adaptation methods – e.g., further pre-training ALBERT on a large corpus of social media text to imbue it with more relevant context.

Richer Context

Currently, the model looks at one post at a time. But a person's depression severity might be better assessed over multiple posts (a history). Perhaps a sequence model over time or features summarizing a user's last 10 posts could give a better signal (someone might occasionally sound severely depressed but it's momentary, versus consistent severe signals over weeks). We didn't explore user-level modeling. That's future work – combining multiple posts per user to reduce misclassification and noise impact.

Integration with Clinical Inputs

Perhaps incorporate some known clinical lexicons or features. For example, PHQ-9 has specific items like sleep, appetite changes, etc. We could fine-tune the model in a multi-task way: in addition to classification, predict if certain symptoms are present in the text. This could force the model to pay attention to relevant phrases (like mention of "sleeping too much or too little"). This multi-task approach might boost interpretability and performance.

Explainability Module

Implement an explainability mechanism (like a layer that highlights keywords for each prediction) so that if deployed, the tool can say e.g., "This post was flagged as severe because it contains phrases indicating hopelessness and self-harm." This can be done by e.g. training a simple proxy model for explanation or using attention scores more transparently.

Real-time and Resource Optimizations

Explore quantizing the model to 8-bit or using knowledge distillation *post-fine-tuning*. We could distill our 15M model into an even smaller one (maybe 5M) that runs ultra-fast on mobile. Since we prioritized severe detection, the distillation objective could weight that class more to ensure the student model retains that crucial skill.

User Studies

Ultimately, evaluating the model's utility via user studies would be valuable. Does it actually help identify individuals who need support earlier than would have happened otherwise? How do users react to being flagged? These go beyond pure CS, but they're important to truly measure success of such technology in the wild.

4.6.7 Summary of Discussion

Our Modified ALBERT model demonstrates that a targeted architectural modification can effectively improve the detection of under-represented classes like severe depression in text data. The extra self-attention layer empowered the model to capture those rare but critical signals, validating our hypothesis. However, this comes with considerations: careful tuning is needed to not degrade other aspects of performance, and the approach is most beneficial in contexts where one class is particularly hard to detect.

We addressed one key challenge (severe depression detection) but acknowledge that holistic accuracy across all classes and domains is an ongoing challenge. The work highlights a broader lesson in NLP: bigger is not always better for specialized tasks – sometimes a small model with a tweak can outperform a large generic model on niche metrics. This is encouraging for resource-constrained settings and for creating tools that are both effective and accessible.

In conclusion of this discussion, we believe our approach strikes a meaningful bal-

ance for the given problem, but it is not the final word. It should be integrated into a larger system with human oversight, and future improvements should aim to refine this balance further. The next chapter will conclude the thesis by summarizing our contributions and findings, and reiterating the potential future work directions in a concise form.

Chapter 5

Conclusion and Future Work

5.1 Conclusion

This thesis set out to improve the automatic detection of depression severity from social media text, with a particular focus on under-represented classes such as severe depression. We identified that standard transformer models, while powerful, often struggle to reliably identify the most critical minority class (severe cases) due to data imbalance and subtle linguistic cues. To address this, we introduced a Modified ALBERT model that integrates a non-shared multi-head self-attention layer into the ALBERT architecture. This additional layer allows the model to better capture long range dependencies and emphasize crucial parts of the input that signal depression severity, without substantially increasing the model's size or computational requirements.

Our experiments, conducted on the DEPTWEET dataset [Kabir2022] of annotated tweets, demonstrate that the modified model achieves its goal and it significantly improved the classification performance for severe depression examples, as evidenced by the highest ROC AUC and recall for that class among all models tested. The modified model maintained competitive overall accuracy (within half a percentage point of the unmodified ALBERT and on par with much larger models). This indicates that our approach enhances sensitivity to the minority class while largely preserving general performance. In practical terms, such a model could serve as a more *alert responsive* system, catching cases that might be overlooked by models optimized purely for accuracy.

We also validated our model's performance on additional datasets and compared it to several baselines:

Versus ALBERT Base (12M parameters), our model (15M) showed a dramatic improvement in severe class detection, at a minor cost of a small drop in mild/moderate class performance.

Versus BERT Base (110M) and RoBERTa Base (125M), our model held its ground remarkably well on the depression task, even exceeding those models in certain metrics like severe-class recall, despite having an order of magnitude fewer parameters.

DistilBERT (66M) proved fastest and reasonably accurate, but it underperformed ALBERT and our model in handling fine-grained severity distinctions.

On general sentiment tasks, our model did not confer a clear advantage; ALBERT Base slightly outperformed it, suggesting the modification is especially useful in tasks like ours where specific minority cues matter.

From a broader perspective, our work contributes a step towards more sensitive and context-aware NLP systems in the mental health domain. By emulating a bit of the clinical intuition (i.e., "pay extra attention to signs of severe distress") within a model, we improved its alignment with what a human expert might consider important. The model's lightweight nature and strong performance make it a good candidate for deployment in real-world scenarios, such as integration into social media platforms or mental health monitoring apps, where running efficiently on limited hardware is important.

However, we are careful to emphasize that this tool is not a standalone diagnostic. The thesis also engaged with the ethical dimension of automated mental health assessment. We argue that while such models can augment mental health professionals and potentially lead to earlier interventions (e.g., prompting a struggling individual to seek help), they must be implemented with caution, transparency, and respect for user privacy and autonomy.

In conclusion, *A Modified ALBERT Model for Under-Represented Classes in Depression Detection* achieved its primary goal by enhancing the detection of severe depression cues in text. The approach balanced performance and efficiency, and opens up possibilities for improved mental health analytics. We have demonstrated that even small modifications in model design can yield meaningful gains for critical use-cases. This work provides a foundation that future research can build upon to create even more robust and equitable NLP systems for mental health and other domains where minority class detection is vital.

5.2 Future Work

The research presented in this thesis can be extended and improved in several directions. Future experiments could explore advanced training strategies to further balance performance across all classes. For instance, using a custom loss function that puts more weight on mild and moderate classes during training could help the model not overlook those in favor of severe. Additionally, techniques like data augmentation (generating paraphrases or synthetic examples, particularly for severe class to increase its variance) might improve generalization. It would be interesting to see if combining our architectural change with such data-level interventions yields an even more balanced model.

Incorporating related tasks could enrich the model's understanding. For example, a multi-task setup where the model simultaneously performs severity classification and symptom identification (e.g., detect if a tweet mentions sadness, insomnia, self-harm, etc.) could guide the attention layer to focus on clinically relevant features. This approach could leverage datasets or annotations where specific symptoms are labeled, aligning model learning more closely with psychiatric assessment frameworks.

In real social media use, one post might not tell the whole story. Future work can involve modeling user history over time. One idea is to have a higher-level RNN or transformer that takes as input the sequence of a user's last N posts (with our model's output for each post as features) to determine overall depression severity trend. This temporal modeling could help reduce misclassifications due to single-post anomalies and understand context.

We plan to test and adapt the model on other data sources, such as Reddit mental health forums, clinical interview transcripts, or even multilingual content. Transfer learning techniques (fine-tuning the model on a small amount of target-domain data) will likely be needed. Investigating how well the architecture extends to languages beyond English or to different age groups/cultures (where expressions of distress differ) is important for broad applicability. This might involve creating or utilizing new datasets akin to DEPTWEET in other languages or contexts.

To facilitate real-world deployment, especially on mobile devices for personal mental health apps, we could apply model compression techniques. Post-training quantization (reducing weights to 8-bit integers) can dramatically shrink model size and speed up inference with minimal loss in accuracy. Also, performing knowledge distillation on our fine-tuned model (using it as a teacher to train a even smaller student model) could yield a compact model that retains most of the performance. This would be

valuable if we want an app running continuously in the background to detect mood from typing or spoken text without draining battery.

Future work should not just be technical. We envision collaborating with psychologists and ethics scholars to design how such a model would interface with users. For example, developing a user interface that shows explanations for predictions (perhaps highlighting words that contributed to the model's decision, in a manner akin to a heatmap). We also want to run user studies to gauge how people interpret and react to being given AI feedback on their mental state. This line of work ensures the technology is used in a way that is supportive, not invasive or frightening. As part of this, building in safeguards (like a threshold where the model is very certain of severe depression before triggering an alert, and even then perhaps alerting the user themselves or a nominated contact rather than authorities) could be explored and refined with stakeholder input.

Depression often co-occurs with other conditions (anxiety, PTSD, etc.), and severity might be multifaceted. In future, one could extend our approach to a more comprehensive mental health classifier that assesses multiple dimensions (maybe a multi-label model that outputs likelihood of various conditions and their severities). Our enhanced attention approach could be extended to multi-label classification, possibly adding multiple attention heads each specialized for different condition cues.

As a practical deployment scenario, one could integrate the model into a system that streams social media data (with permission) and identifies users whose language is shifting towards severe depression. A pilot program could be set up in collaboration with mental health organizations to test such a system, ensuring that any interventions are safe and wanted. The outcomes (e.g., whether the system successfully prompts individuals to seek help earlier) would be the ultimate test of the model's value. This, of course, requires careful ethical oversight.

In summary, while this thesis has addressed a specific and important challenge in automated depression detection, it also opens up many questions. By pursuing the future directions above, we aim to create NLP models that are more accurate, more fair (attentive to minority classes), more transparent, and more useful in improving real lives. The encouraging results obtained here provide a solid stepping stone. We have shown that integrating domain-inspired architectural features (like a specialized attention layer) into modern NLP models can yield tangible improvements. We anticipate that continuing this line of work – blending deep learning with insights from psychology and user-centered design – will lead to the next generation of mental health support tools that are both technologically advanced and ethically sound.

References

[1] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2015.

[2] F. Barbieri, J. Camacho-Collados, L. Espinosa Anke, and L. Neves, "Tweeteval: Unified benchmark and comparative evaluation for tweet classification," in *Findings of the Association for Computational Linguistics: EMNLP 2020*, 2020, pp. 1644–1654.

[3] A. T. Beck, *Depression: Clinical, Experimental, and Theoretical Aspects*. New York: Harper & Row, 1967.

[4] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[5] S. Bucci, M. Schwannauer, and N. Berry, "The digital revolution and its impact on mental health care," *Psychology and Psychotherapy: Theory, Research and Practice*, vol. 92, no. 2, pp. 277–297, 2019. DOI: https://doi.org/10.1111/papt.12222. [Online]. Available: https://bpspsychub.onlinelibrary.wiley.com/doi/abs/10.1111/papt.12222.

[6] M. Buda, A. Maki, and M. A. Mazurowski, "A systematic study of the class imbalance problem in convolutional neural networks," *Neural Networks*, vol. 106, pp. 249–259, 2018. DOI: 10.1016/j.neunet.2018.07.011.

[7] A. L. Calear and H. Christensen, "Systematic review of school-based prevention and early intervention programs for depression," *Journal of adolescence*, vol. 33 3, pp. 429–38, 2010.

[8] R. A. Calvo, D. N. Milne, M.-A. Hussain, and H. Christensen, "Natural language processing in mental health: Applications, challenges and ethical questions," *Current Opinion in Psychology*, vol. 9, pp. 1–6, 2017.

[9] R. A. Calvo and D. N. Milne, "Natural language processing in mental health: A scoping review," *Current Opinion in Psychology*, vol. 36, pp. 18–24, 2020. DOI: 10.1016/j.copsyc.2020.04.005.

[10] A. Cohan, B. Desmet, A. Yates, L. Soldaini, S. MacAvaney, and N. Goharian, "Smhd: A large-scale resource for exploring online language usage for multiple mental health conditions," *arXiv preprint arXiv:1806.05258*, 2018.

[11] G. Coppersmith, M. Dredze, and C. Harman, "Quantifying mental health signals in twitter," *Proceedings of the Workshop on Computational Linguistics and Clinical Psychology*, pp. 51–60, 2014.

[12] C. Delahunty, M. O'Neill, and R. Soricut, "Efficiency and fairness in transformer-based nlp models: Balancing accuracy and deployment cost," *Journal of Artificial Intelligence Research*, vol. 77, pp. 123–146, 2023.

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

[14] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *NAACL*, 2019.

[15] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, "Bert: Pre-training of deep bidirectional transformers for language understanding," in *Proceedings of NAACL-HLT*, 2019, pp. 4171–4186.

[16] G. 2. Diseases and I. Collaborators, "Global burden of 369 diseases and injuries in 204 countries and territories, 1990–2019: A systematic analysis for the global burden of disease study 2019," *The Lancet*, vol. 396, no. 10258, pp. 1204–1222, 2020, Specifically, depression is highlighted as a leading cause of years lived with disability (YLDs). DOI: 10.1016/S0140-6736(20)30925-9.

[17] S. K. Ernala and et al., "Methodological biases in social media mental health research," *Journal of Medical Internet Research*, 2019.

[18] S. K. Ernala et al., "Methodological Gaps in Predicting Mental Health States from Social Media: Triangulating Diagnostic Signals," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, New York, NY, USA: Association for Computing Machinery, 2019, pp. 1–16, ISBN: 9781450359702. DOI: 10.1145/3290605.3300364. [Online]. Available: https://dl.acm.org/doi/abs/10.1145/3290605.3300364.

[19] J. Gao, J. Zhang, and Y. Liu, "A survey of attention mechanisms in natural language processing," *Journal of Artificial Intelligence Research*, vol. 71, pp. 301–340, 2021. DOI: 10.1613/jair.1.12749.

[20] I. H. Gotlib and J. Joormann, "Conceptualizing and measuring depression: The role of heterogeneity," *Current Opinion in Psychiatry*, vol. 30, pp. 3–10, 1 2017. DOI: 10.1097/YCO.0000000000000305.

[21] S. C. Guntuku et al., "Tracking mental health and symptom mentions on twitter during covid-19," *Journal of General Internal Medicine*, vol. 35, pp. 2798–2800, 2020.

[22] K. Harrigian, C. Aguirre, and M. Dredze, "On the state of social media data for mental health research: Availability, limitations, and challenges," in *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, 2021, pp. 3467–3479. DOI: 10.18653/v1/2021.findings-acl.305.

[23] A. Husseini Orabi, P. Buddhitha, M. Husseini Orabi, and D. Inkpen, "Deep learning for depression detection of twitter users," *Proceedings of the Fifth Workshop on Computational Linguistics and Clinical Psychology: From Keyboard to Clinic*, 2018. DOI: https://doi.org/10.18653/v1/w18-0609.

[24] S. Ioffe and C. Szegedy, "Batch normalization: Accelerating deep network training by reducing internal covariate shift," in *Proceedings of the 32nd International Conference on Machine Learning*, 2015, pp. 448–456.

[25] M. N. Islam, M. R. Islam, M. H. Kabir, and M. M. Rahman, "A comprehensive review on multimodal depression detection using machine learning and deep learning techniques," *IEEE Access*, vol. 10, pp. 48 301–48 325, 2022. DOI: 10.1109/ACCESS.2022.3172820.

[26] R. Johnson, S. Lee, and D. Miller, "The peril of overall accuracy: Misclassification costs in medical diagnostic ai for rare but severe conditions," *Journal of Medical Internet Research*, vol. 23, e27890, 7 2021. DOI: 10.2196/27890.

[27] M. Kabir et al., "Deptweet: A typology for social media texts to detect depression severities," *Computers in Human Behavior*, vol. 135, p. 107 503, 2022.

[28] R. Kavuluru, M. Ramos-Morales, T. Holaday, A. G. Williams, L. Haye, and J. Cerel, "Classification of helpful comments on online suicide watch forums," *Proceedings of the 7th ACM International Conference on Bioinformatics, Computational Biology, and Health Informatics*, 2016.

[29] J. Kim, J. Lee, E. Park, and J. Han, "A deep learning model for detecting mental illness from user content on social media," *Scientific reports*, vol. 10, no. 1, pp. 1–6, 2020.

[30] D. P. Kingma and J. Ba, "Adam: A Method for Stochastic Optimization," in *3rd International Conference on Learning Representations (ICLR)*, Y. Bengio and Y. LeCun, Eds., 2015. [Online]. Available: http://arxiv.org/abs/1412.6980.

[31] K. Kroenke, R. L. Spitzer, and J. B. W. Williams, "The PHQ-9," *Journal of General Internal Medicine*, vol. 16, no. 9, pp. 606–613, 2001. DOI: 10.1046/j.1525-

1497 . 2001 . 016009606 . x. [Online]. Available: https : / / onlinelibrary . wiley.com/doi/abs/10.1046/j.1525-1497.2001.016009606.x.

[32] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," *arXiv preprint arXiv:1909.11942*, 2019.

[33] J. L. Leevy, T. M. Khoshgoftaar, R. A. Bauder, and N. Seliya, "A survey on addressing high-class imbalance in big data," *Journal of Big Data*, vol. 5, no. 1, pp. 1–30, 2018. DOI: 10 . 1186 / s40537 - 018 - 0151 - 6. [Online]. Available: https://link.springer.com/article/10.1186/s40537-018-0151-6.

[34] Y. Liu et al., "Roberta: A robustly optimized bert pretraining approach," *ArXiv*, vol. abs/1907.11692, 2019.

[35] I. Loshchilov and F. Hutter, "Decoupled Weight Decay Regularization," in *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*, OpenReview.net, 2019. [Online]. Available: https : //openreview.net/forum?id=Bkg6RiCqY7.

[36] D. M. Low, L. Rumker, T. Talkar, J. B. Torous, G. A. Cecchi, and S. S. Ghosh, "Natural language processing reveals vulnerable mental health support groups and heightened health anxiety on reddit during covid-19: Observational study," *Journal of Medical Internet Research*, vol. 22, 2020.

[37] S. Parvez, *Multiclass sentiment analysis dataset*, urlhttps://huggingface.co/datasets/Sp1786/multiclass-sentiment-analysis-dataset, 2023.

[38] J. W. Pennebaker, M. R. Mehl, and K. G. Niederhoffer, "The social, linguistic, and psychological correlates of depression," *Journal of Language and Social Psychology*, vol. 22, no. 2, pp. 207–226, 2003. DOI: 10 . 1177/0261927X03022002007.

[39] A. G. Reece, A. J. Reagan, K. L. Lix, P. S. Dodds, and C. M. Danforth, "Forecasting the onset and course of mental illness with twitter data," *Scientific Reports*, vol. 7, no. 1, p. 13 006, 2017.

[40] N. Reimers and I. Gurevych, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[41] A. Saif, K. W. Wong, and N. Yahya, "Handling imbalanced data: A review," *International Journal of Advanced Computer Science and Applications*, vol. 11, no. 11, pp. 531–540, 2020. DOI: 10 . 14569/IJACSA . 2020 . 0111164.

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

[43] V. Sanh, L. Debut, J. Chaumond, and T. Wolf, "Distilbert, a distilled version of bert: Smaller, faster, cheaper and lighter," *ArXiv*, vol. abs/1910.01108, 2019.

[44] J. Shen, Y. Wang, X. Zhang, and J. Li, "A survey of publicly available datasets for depression detection and analysis," *Journal of Biomedical and Health Informatics*, vol. 24, pp. 3535–3547, 12 2020. DOI: 10.1109/JBHI.2020.2974987.

[45] S. Singh, D. Roy, K. Sinha, S. Parveen, G. Sharma, and G. Joshi, "Impact of COVID-19 and lockdown on mental health of children and adolescents: A narrative review with recommendations," en, *Psychiatry Research*, vol. 293, p. 113 429, Nov. 2020, ISSN: 01651781. DOI: 10.1016/j.psychres.2020.113429. Accessed: Jan. 28, 2022. [Online]. Available: https://linkinghub.elsevier.com/retrieve/pii/S016517812031725X.

[46] J. Smith, L. Chen, and R. Gupta, "Limitations of general-purpose language models in detecting critical medical conditions from clinical notes," *Journal of Biomedical Informatics*, vol. 123, p. 103 987, 2022. DOI: 10.1016/j.jbi.2022.103987.

[47] N. Srivastava, G. Hinton, A. Krizhevsky, I. Sutskever, and R. Salakhutdinov, "Dropout: A simple way to prevent neural networks from overfitting," *Journal of Machine Learning Research*, vol. 15, no. 1, pp. 1929–1958, 2014.

[48] Y. Sun, A. K. Wong, and M. S. Kamel, "Classification of Imbalanced Data: A Review," *International Journal of Pattern Recognition and Artificial Intelligence*, vol. 23, no. 04, pp. 687–719, 2009. DOI: 10.1142/S0218001409007326. [Online]. Available: https://www.worldscientific.com/doi/abs/10.1142/S0218001409007326.

[49] J. C. Tolentino and S. L. Schmidt, "DSM-5 Criteria and Depression Severity: Implications for Clinical Practice," *Frontiers in Psychiatry*, vol. 9, p. 450, Oct. 2018, ISSN: 1664-0640. DOI: 10.3389/fpsyt.2018.00450. Accessed: Jan. 21, 2022. [Online]. Available: https://www.frontiersin.org/article/10.3389/fpsyt.2018.00450/full.

[50] A. Vaswani et al., "Attention is all you need," in *Advances in Neural Information Processing Systems*, 2017, pp. 5998–6008.

[51] A. Vaswani et al., "Attention is all you need," *ArXiv*, vol. abs/1706.03762, 2017.

[52] J. Vincent, *Facebook is using AI to spot users with suicidal thoughts and send them help*, en, Nov. 2017. Accessed: Jan. 21, 2022. [Online]. Available: https://www.

theverge.com/2017/11/28/16709224/facebook-suicidal-thoughts-ai-help.

[53] D. Watson and L. A. Clark, "The panas scales: Development and validation of brief measures of positive and negative affect," *Journal of Personality and Social Psychology*, vol. 54, no. 6, pp. 1063–1070, 1994. DOI: 10.1037/0022-3514.54.6.1063.

[54] T. Wolf et al., "Transformers: State-of-the-art natural language processing," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 2020, pp. 38–45.

[55] Y. Wu et al., "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation," *CoRR*, vol. abs/1609.08144, 2016. [Online]. Available: http://arxiv.org/abs/1609.08144.

[56] S. Yadav, J. Chauhan, J. P. Sain, K. Thirunarayan, A. Sheth, and J. A. Schumm, "Identifying depressive symptoms from tweets: Figurative language enabled multitask learning framework," in *COLING*, 2020.

[57] Z. Yang, Z. Dai, Y. Yang, J. Carbonell, R. Salakhutdinov, and Q. V. Le, "Xlnet: Generalized autoregressive pretraining for language understanding," *Advances in Neural Information Processing Systems*, vol. 32, 2019.

[58] Y. Zhang, Q. Yang, and H. Sun, "Data augmentation for natural language processing: A survey," *Artificial Intelligence Review*, vol. 55, pp. 4281–4330, 2022.

[59] Y. Zhang, X. Liu, and X. Chen, "Enhancing depression detection in social media text via attention-based deep models," in *Proceedings of the International Conference on Computational Linguistics (COLING)*, 2022, pp. 2503–2512.

List of Publications

The following publication has resulted directly from the research work presented in this thesis:

1. Sk Nahid Sanwar, Tansif Anzar, Md Tariquzzaman, Md Kamrul Hasan, and Hasan Mahmud, **"An Attention-Based ALBERT Model for Mental Health Studies,"** *IUT Journal of Engineering and Technology (JET)*, vol. 19, no. 3, pp. 681–694, 2025. (This paper was also presented at the *1st IUT International Conference on Core Engineering and Technology (IUT-ICCET)*, 2025.)