National College of Ireland

# Toward Reliable Depression Detection in Negation-Heavy Twitter Corpora

MSc Research Project
Master of Science in Data Analytics

Vishal Vishvanath Sawant Dessai
Student ID: x23414766

School of Computing
National College of Ireland

Supervisor: Christian Horn

**National College of Ireland**

**MSc Project Submission Sheet**

**School of Computing**

**Student Name:** Vishal Vishvanath Sawant Dessai

**Student ID:** x23414766

**Programme:** Master of Science in Data Analytics **Year:** 2025

**Module:** Research Practicum

**Supervisor:** Christian Horn

**Submission Due Date:** 11th December 2025

**Project Title:** Toward Reliable Depression Detection in Negation-Heavy Twitter Corpora

**Word Count:** 9,101 **Page Count:** 21

I hereby certify that the information contained in this (my submission) is information pertaining to research I conducted for this project. All information other than my own contribution will be fully referenced and listed in the relevant bibliography section at the rear of the project.
ALL internet material must be referenced in the bibliography section. Students are required to use the Referencing Standard specified in the report template. To use other author's written or electronic work is illegal (plagiarism) and may result in disciplinary action.

**Signature:** Vishal Vishvanath Sawant Dessai

**Date:** 11th December 2025

**PLEASE READ THE FOLLOWING INSTRUCTIONS AND CHECKLIST**

| | |
|---|---|
| Attach a completed copy of this sheet to each project (including multiple copies) | □ |
| **Attach a Moodle submission receipt of the online project submission,** to each project (including multiple copies). | □ |
| **You must ensure that you retain a HARD COPY of the project**, both for your own reference and in case a project is lost or mislaid. It is not sufficient to keep a copy on computer. | □ |

Assignments that are submitted to the Programme Coordinator Office must be placed into the assignment box located outside the office.

| **Office Use Only** | |
|---|---|
| Signature: | |
| Date: | |
| Penalty Applied (if applicable): | |

# Toward Reliable Depression Detection in Negation-Heavy Twitter Corpora

Vishal Vishvanath Sawant Dessai
x23414766

**Abstract**

Negation and informal language are frequent in social media posts indicating depressive moods, but these are hard to recognize with an automated classifier when labels are synthesized using rules of lexicons instead of human labeling. The paper assesses the claim that the new DeBERTa-v3 transformer exhibits any quantifiable benefits as compared to BERT and RoBERTa in these settings of realistic weak labeling. Based on the 57k-tweet Multi-Labeled Depression Corpus, we have maintained negation, emojis, and casual speech and used stratified 80/10/10 partitions and the same fine-tuning parameters to all architectures. Evaluation of performance is done with the use of Macro-F1 on the entire test set as well as a subset which includes a larger number of negation in the test set. With three randomly chosen seeds, DeBERTa-v3 also has the highest macro-F1, both with 3-class (0.9187) and with binary classification (0.9510), and all pair-wise differences are all statistically significant ($p < 0.0001$). It was also found that the model will be more robust on negative heavy tweets. These outcomes suppose that the architectural enhancements of DeBERTa-v3, especially disentangled attention, are real advantages towards depression detection in noisy, language-noisy, weakly labelled social-media settings.

**Keywords:** Depression detection, Weakly labeled data, Transformer models (BERT, RoBERTa, DeBERTa-v3), negation handling, Twitter (X), Social media.

# 1 Introduction

Depression has been among the most prevalent issues in mental health and because of stigma, lack of access to clinical services, and delays in self-reporting, the early identification of the depression is mostly a problem. Social media sites, and in particular twitter, are becoming important sources of behavioural indicators, as customers often communicate their emotions, frustrations and mood changes via brief informal messages. Negation, shortened forms, emojis, sarcasm, and features of idiomatic language are prevalent in these posts, making it difficult to classify mental-health automatically. Despite the good performance as evidenced by transformer models like BERT and RoBERTa in sentiment and depression detection, the research done before is mostly based on clean and manually annotated datasets which do not reflect the noisy conditions coupled with weakly labeled performance of actual large scale monitoring systems.

Lexicon-based tools (e.g., VADER) are frequently used to label datasets in their practical implementation, creating systematic noise in particular in the negation, polarity reversal (not sad vs. not happy) and ambiguous emotional cues. In the meantime, the newer architectures (e.g. DeBERTa-v3), have disentangled attention and have better positional encoding, which

are intended to compute such, longer range contextual relationships. Whether such architectural enhancements are always better than the previous transformer models remains unknown, however, when using weakly labeled negative data, negation-rich data characteristic of mental-health screening, and similar to the case of a weakly-labeled, negation-heavy social media data. This leaves two recurrent gaps: (1) the current outputs might not be applicable to noisy and weak-supervision environments, and (2) the negative-processing of current transformers is not well investigated.

To mitigate these shortcomings, the current research undertakes a controlled comparison between BERT, RoBERTa and DeBERTa-v3 with the same preprocessing, hyperparameters, loss functions, as well as training conditions as implemented on a weakly labelled Twitter corpus of depression. The study isolates the actual effect of architectural variability by maintaining all other variables fixed but the backbone architecture so as to make a fair comparison between model robustness and negation comprehension on real, noisy data.

## 1.1 Research Question

Does DeBERTa-v3 significantly outperform BERT and RoBERTa on negation-rich, weakly-labelled tweets for both 3-class (depressed/indifferent/happy) and binary (depressed vs. non-depressed) depression screening?

## 1.2 Research Objectives

Develop a controlled and reproducible evaluation framework for transformer-based depression detection on weakly labelled Twitter data.

Fine-tune and compare BERT, RoBERTa, and DeBERTa-v3 using identical preprocessing, hyperparameters, and training configurations.

Evaluate model performance on both the full test dataset and a negation-specific subset to assess each architecture's ability to interpret negated depressive expressions.

Statistically validate all observed performance differences using macro-F1 scores, bootstrap significance testing, and multiple random seeds to ensure robustness and reliability.

The remainder of this paper is organized as follows. Section 2 reviews related literature and identifies the key research gaps that motivate this work. Section 3 describes the dataset, preprocessing pipeline, and experimental methodology in detail. Section 4 presents the results, analysis, and discussion of findings, including the lack of significant architectural differences and implications for model selection in weakly-labeled scenarios. Finally, Section 5 concludes with implications, limitations, and directions for future research.

## 1.3 Acknowledgments

I would like to express my deep gratitude to Dr. Christian Horn who provided an invaluable guidance, support and critical remarks at all the stages of this work. His encouragement immensely enriched my knowledge on the research problem. I'm also grateful to the National

College of Ireland for providing me with the facilities, resources, and institutional support to complete this project.

# 2 Related Work

The recent developments in natural language processing have put transformer-based architectures as the leader in competing in text classification and affective computing tasks, such as the detection of mental health on social media. The growth of social media services has presented unparalleled potentials in mental health scale-level surveillance, but also caused enormous problems with the multiplicity of language, noisy annotations, and ambiguous contexts. This section is a review of previous literature describing four important strands, which include: (i) transformer-based in detecting depression, (ii) the use of negation and weak supervision settings, (iii) robustness and noise reduction methods, and (iv) temporal and multi-modal setting. Combined with these issues, this is indicative of the necessity of regulated testing of contemporary transformer backbone under weak-label and negation-laden conditions.

## 2.1 Transformer-Based Depression Detection

An earlier deep learning model of detection of mental health involved CNNs and RNNs. Nonetheless, they were not effective with long-range dependencies and contextuality applicable in the social media posts that are realistic. With the emergence of the transformer architecture, namely the BERT model and the following version, which is the RoBERTa, the domain-adapted variants including MentalBERT arose, leading to a major improvement. The research, as well as that of Kerasiotis et al. (2024), and even Jamali et al. (2023), has demonstrated the effectiveness of the transformer-of-encoders in obtaining the depressive characteristics, and it has shown good results on the controlled, manually annotated datasets. However, these works required clean labels, and a number of them relied on other language properties, typically with the opportunity of data enhancement, and, therefore, blurring the particular significance of the model structure.

Later papers have shown that better architectures and hybrid techniques, including textbert usage with CNN/BiLSTM units (Xin and Zakaria, 2024), using TextGCN with emotion embeddings (Mao and Han, 2025), and commonsense-sensitive methods, like CoKE (Chen et al., 2025) will be utilized. Although these had the effects of introducing very minimal improvements to the existing benchmark datasets, they did introduce severe computational complexity and the use of external lexicons or existent graph knowledge that is non-generalizable. Other studies such as Chopra et al. (2025), Shoaib et al. (2025) suggested the application of contextual learning and hybrid transformer-SVM but again these were evaluated in relatively clean environments.

In this literature, it emerges quite clearly that the one unresolved constraint to transformer methods is that they are nearly always evaluated using data of both high quality and manually-annotated data, which is not in fact similar to the noisy, lexicon-induced data at the end of the things that they are used on. Also, very little of the literature seems to care in particular about evaluating the mitigation of the negation which is definitely of a great prevalence in the discourse of the depression, e.g. ("not okay," "never fine ). This entails that it is not clear whether, given the specific problem to question, namely, poor supervision, less syntax, and heavy negation, DeBERTa-v3 offers any significant benefits.

## 2.2 Sentiment Analysis and Cross-Domain Applications

Sentiment analysis studies do offer valuable methodological underpinnings to depression detection, although important differences restrict generic applicability. According to research, like Suryono and Djunaidy (2024), transformer models can better classify hate-speech and toxicity, and the issue of distinctions between general negativity and true psychological distress has been mentioned. On the same note, Gao et al. (2024) demonstrated that BERT does not work well with the very short texts present on Twitter because of the lack of the contextual information, which should imply the limitations of the algorithm in the analysis of short and emotionally saturated texts.

It was discovered that transformers outperform classical models by 10-15% F1 but this is diminished on clean and balanced datasets suggesting that transformers are most useful in real-world noisy conditions (Alnaddaf and Basarslan, 2024). Other cross-domain experiments, such as Roufas et al. (2025), all also show systematic biases in the politically charged conditions by transformers, misclassifying particular forms of rhetoric. This conduct carries consequences to detection of depression where culturally or contextually sensitive manifestations of sufferings are also misrepresented. All in all, although the field of sentiment analysis research can be made to understand how depression is detected, it also highlights the long-standing issues regarding the ability of transformer models to be applied to informal text, unclear text, and psychologically irregular text found in the social-media context.

## 2.3 Negation Handling and Weak Supervision

Sentiment and mental-health text analytics continues to be significantly negated, and yet this problem has been little observed in literature on depression-detection. This was evident with the study by Helmy et al. (2024), which compared datasets of English and Arabic Twitter in the condition of the presence of negation cues with their absence and found a decrease in the performance of 92 to 87 F1 in the condition maintaining the negation omnipresence. They demonstrated that the majority of transformer baselines mistake negation as isolated tokens instead of modifiers, and give it an inappropriate polarity. Their bilingual results support that the challenge of negation does not have anything to do with language, as their key-word based method (not, no, never) lacks more subtle constructions, such as hardly, scarcely, or rhetorical negation.

The Multi-Labeled Depression Corpus of Nassar et al. (2022) also highlights the important place of negation, providing a 57k of tweets with direct notation on such phrases as not depressed, stopped being sad, never felt better. Although useful at analysing model behaviour on negated expressions, the corpus uses both VADER and LIWC to do its labelling, which has systematic weak-label noise. This is a wider concern pointed out by Dalal et al. (2024), since approximately 70 percent of currently published research on the problem uses manually curated data not representative of the noisy and inconsistent conditions of lexicon-labeled text which results in the research overestimating performance in the real world.

Weak supervision brings about biases in a completely different sense than random noise: the use of lexicon in methods effectively over performs in identifying posts that heavily feature key-words to the detriment of subtle cues, whether it be a change in pronoun use, framing of time, or use of language of social-engagement. Yuan et al. (2024) demonstrated that deep networks ultimately learn and remember such noisy labels and suggested noise-robust methods and presented experimental data but did not consider the question whether the same

approach can be applied to linguistic noise. Importantly, such as the combined analysis of negation and weak-label noise IT has never been previously tested whether newer transformer models like DeBERTa-v3, including disentangled attention, really perform better than older models like BERT in such cases. This is a major weakness in the rigor of methodology and applicable practice.

## 2.4 Robustness, Noise Mitigation, and Temporal Modeling

Recent papers provide a number of methods to enhance resilience when using imperfect or noisy data. Data augmentation strategies like Easy Data Augmentation (EDA) (Wei and Zou, 2019) and context/synonym insertions (Kerasiotis and Askounis, 2024) are useful to counteract classes and curb overfitting, and achieve 4-6 per cent improvements in F1 on skewed datasets. They show significant reduction in their benefits on pre-balanced datasets, meaning that augmentation counters distributional imbalance and not structural labeling errors.

In other methods, label noise is directly addressed. Confidence-based filtering (Northcutt et al., 2021) labels probably mislabeled samples by monitoring smooth prediction with increasing epochs and can be combined with focal loss, to diminish the importance of noisy labels. In spite of their promise, these approaches are often combined with metadata or external sentiment uncovers, thus it is hard to separate robustness associated with backbone only. It is important to note that there has not been any research to determine how various transformer architecture respond equally to noise-aware filtering and whether there is some mechanism that makes certain architecture resilient to label noise. In their study, Chandrasekaran and Kotaki (2024) employed BERT embeddings together with dynamic topic models to trace the changing depression motifs in user timelines revealing unique time Covidia causal patterns that were either mood cycles or progressive withdrawal. These models can be important in long term monitoring, but they need long poster histories ([?]50 posts within 3 months), which makes them suit only sparse users or early detection applications.

Domain-adaptive pretraining has performed well as well. Gururangan et al. (2020) established that in-domain social media text pretraining enhances downstream performance by 5-8% particularly with tasks related to slang, emojis and abbreviations. Nevertheless, it is not obvious whether domain-adapted models are either stronger or more susceptible to systematic weak-label noise, or whether the linguistic fluency is going to be more effective in negation. Lastly, Fatima et al. (2025) highlighted the lack of linkage between presented research databases and practical application, mentioning that high-accuracy models can still present a low practical value in cases of low prevalence of depression. According to their review, they note that there is a gap in critical evaluation of the current body of literature: most of the works do not take the false-positive rates or class imbalance in the real world into consideration, which restricts the credibility of the reported results.

## 2.5 Summary and Research Gaps

An overview of the available literature indicates that there are still a number of fundamental limitations that limit the credibility and practical application of depression-detection methods. One of its main flaws is the reliance on manually annotated datasets, which are, although clean and convenient, not representative of the systematic biases of lexicon-based weak supervision that has been generally applied in the massive context of social-media

monitoring. Consequently, the performance of the reported performance will not usually generalise to data that is noisy and inconsistent in the real world.

The other problem that is crucial but not studied fully is negation. In spite of the common encoding of depressive information in the form of phrases like not okay, can't cope, nothing helps and more, the research on model behaviour on negated text is rare and scarce information exists as to whether more recent transformer networks perform better in responding to negations compared to their predecessors. This problem is further complicated by insufficient experimental isolation: most backbone comparisons also augment the backbone, add metadata, or altered loss functions, and it is not clear to what extent the augmentation leads to improvement, as opposed to the architecture.

There is also the lack of a temporal context. Although new longitudinal studies indicate that depressive patterns tend to be observed across series of posts and not individual messages, their computation is computationally ample, requiring the availability of user histories that are witheringly difficult to achieve. Lastly, the practices of evaluation are usually based on the balanced test sets and coarse metrics, sweeping weaknesses with the conditions of the real-world underclass imbalance and linguistically untidy inputs.

All these gaps aid in inspiring the current work. To solve their problem, this paper poses a controlled comparison of BERT, RoBERTa and DeBERTa-v3 on a negation-rich, lexicon-labeled twitter corpus, fixing all preprocessing, hyperauto-tuning and training conditions in their board. The paper can isolate the real architectural effects and determine whether modern transformers have any meaningful benefits when operating in weak-label, noisy, and linguistically challenging conditions through the simply adjustment of the backbone architecture.

# 3 Research Methodology

## 3.1 Dataset and Data Collection

The research consists of Multi-Labeled Depression Corpus, which consists of 57,391 English Tweets published on Harvard Dataverse by Nassar, Helmy and Ramadan (2022). Here, the short and informal posts are all present, with inherent negation, sarcasm, emojis, slangs, and unexplainable personality swings. The labels (depressed, indifferent, happy) are produced by automatic means through the use of the VADER sentiment lexicon and not human annotation. This causes the dataset to be weakly-labelled, that is, the ground truth is noisy. Weak labels, the twitter-style use of language and high density of negativity are characteristic of screening conditions in the real world and as such, fits our research question perfectly, namely, assessing the transformers in a noisy, high-density negativity environment. Most of historic researches have been done with manually curated labels or long posts; hence this data is realistic and challenging to test whether newer transformer architectures like DeBERTa-v3 are better at managing negation and label noise than older models.

## 3.2 Preprocessing

The plan of the preprocessing pipeline was as minimal as possible and negation preserving. It follows that because most depressive statements are made by use of negation structures (so that when I say I am not okay, or that nothing helps any more, I refer to the linguistic indices upon which the classification task relies) cleaning that alters or removes negators, would radically distort the linguistic signals upon which the classification task draws. Thus, the preprocessing stage allowed the negation wording to be retained, as it occurs in original tweets. The minimal normalisation was performed: URLs were converted to whitespace, user mentions were converted to whitespace, HTML entities were un escaped and unreasonable spaces were normalized. The use of emojis was minimally reduced since they have significant emotional content in communication on Twitter. Hashtags were not removed as well. There was no lowercasing done in order to ensure emphasis patterns (e.g., NOT okay) could be read. There were also instances of false duplicate tweets that were removed with regard to cleaned text; thus, eliminating duplicates. This move not only kept the model out of overfitting to repetition content but also the idea that learning was a manifestation of generalisation of important meaning and not recitation. The post-processing dataset was composed of about 57,300 tweets.

## 3.3 Label Encoding and Stratified Splitting

As a way of supporting multi-class classification task as well as binary, two label mapping were done. To select 3-class classification, the original labels were replaced by the following ones: depressed (-1) →0, indifferent (0) → 1, happy (1) → 2. In a binary classification, the labels were coded as: depressed (-1) → 1, indifferent (0) → 0, happy (1) → 0 so that the task became depression detection versus non-depression. Additional stratified separation into 80/10/10 allowed the division of the dataset into training, validation and test sets. Stratification was used to maintain all classes proportions in each of the three splits which is very important since weakly- labelled datasets are usually imbalanced and an unequal split would bias training or evaluation by a model. This division process was carried out in both instances of each random seed so that the statistical strength would be strong. Validation and test sets were completely hidden during the training, and this ensured the integrity of the evaluation. This train-validation-test design is typical of transformer-based NLP experiments and helps in establishing comparability between the transaction used with methods used in comparable depression detection experiments.

## 3.4 Tokenisation

Each transformer backbone (BERT-base- uncased, RoBERTa-base, and DeBERTa-v3-base) had three independent tokenisers. All the tokenisers were fed the text with the same date: a sequence length of 128 tokens maximum, truncation of oversized tweets, and maximum sequence length padding, along with conversion into PyTorch tensors. The shorter sequence length of 128 (or 512 which is used in standard training) was done by the fact that tweets are inherently short and this optimization was a major factor in accelerating the training without

losing content. Fixed-length tokenisation ensured that all the models were treated fairly and mainly equally since the performance difference could have been explained by the difference in tokenisation. The study controls the impact of the architectural variations between the three transformer models by standardisation of the tokenisation procedure.

## 3.5 Dataset Construction and Dataloaders

The tokenised and preprocessed data was packaged into the objects of PyTorch Dataset, where each of the datasets included the encoded input-ids, attention masks, and the label. In order to implement training with the highest possible efficiency, all samples had been pre-tokenized before being loaded into memory, eliminating training redundancy in repeated tokenization. These data sets were loaded to DataLoaders set to generate 32 batch sized data sets. To ensure improved gradient descent behaviour, the training DataLoader was shuffled, and to avoid ordering biases, the validation and the test DataLoaders were not changed in order but kept their original order to provide consistent evaluation. To achieve the optimization of data transfer between the GPUs, DataLoaders was using 2 worker processes and pinning memory. It is a designed deep learning that is based on the existing profound learning best practices and enables cross-run seed reproducibility.

## 3.6 Model Architectures

HuggingFace Transformers were used to implement the three backbones to the transformer, namely BERT, RoBERTa and DeBERTa-v3. All models were further extended with a simple classification head that had dropout (p=0.1) followed by a linear layer to transform hidden size (768 dimensions) to the number of classes (3 with multi-class, 2 with binary). The head of classification works based upon the [CLS] token representation of the final transformer layer. It did not introduce any further metadata, feature engineering or augmentation as the purpose of the study was to compare the backbone design per se, rather than support design decisions. A fine-tuning set up that applies uniformly to all the models ascertains that whatever differences in performance is observed is not due to confounding variables.

## 3.7 Evaluation Metrics

The last analysis was conducted on the unknown test split with the most important measure as macro-F1. Macro-F1 has been selected since it copes with the class imbalance and represents how the model or the unweighted mean of the per-class F1 scores treat all the labels equally. Other metrics such as per-class precision, recall and accuracy were also calculated to offer more interpretability. A separate analysis was done of a "negation subset" which included tweets that explicitly used negators “like”, “not”, “never”, “no”, “n't”, “nothing”, “no one”, “none”, “neither”, nowhere or cannot. This subgroup with the proportion 31.3 of the test results directly evaluates the question of whether architectural differences influence negation handling. The results of this subset performance offer an understanding of each model based on its capability to resist contextual polarity changes potentially important in detecting depression.

## 3.8 Statistical Analysis

To make sure the findings could be reliable, three random seeds turned out to be utilized in training each model configuration (42, 123, 456). Macro-F1 distributions were compared with paired bootstrap resampling with 500 repeats and $p = 0.05$ significance, when comparing the resulting macro-F1 distributions across seeds. This statistical method is frequently employed in NLP in order to check whether observed differences between models are primarily due to real improvements in performance, and not to randomness. Each task (3-class and binary) was compared at the pairwise level between DeBERTa-v3 and BERT and RoBERTa. The study offers a powerful empirical evidence of the architectural differentiation findings by including the significance testing.

## 3.9 Experimental Environment

The experiments were all implemented on Google Colab Pro with Tesla T4 CPU (16GB memory). It used Python 3.12, PyTorch 2.x (Including CUDA) and HuggingFace Transformers 4.30+, NumPy, Pandas, Scikit-learn and Matplotlib and Seaborn software environment. The average time spent in training all 18 experiments (3 seeds x 2 tasks x 3 models) taken on the T4 GPU was about 2-4 hours. The reproducibility is advanced by the controlled and documented environment, which allows other people to reproduce the results on similar computational environments.

# 4 Design Specification

## 4.1 System Overview

The complete system is aimed at comparing the performance of three various transformer backbonesBERT, RoBERTa, and DeBERTa-v3conditioned to the same training environment operating on a twitter source, negation-saturated, and poorly labelled corpus. The system has a modular structure which is composed of data ingestion, preprocessing, tokenisation, model construction, training, and evaluation. The modules communicate with each other using clearly defined interfaces and as a result the system can be extended and repurposed. As the overall purpose of the research is to single out the impact of the transformer backbone but not any other auxiliary improvements, the composition in its entirety remains minimal: no external metadata, linguistic indicators, augmentation policies or programmatic additions to the pipeline are added. This design will make sure that the performance variation among the models is a result of intrinsic ability of representing all the languages. The system performs 18 independent experiments (3 random seeds x 2 classification tasks x 3 models) in order to have strong statistical confirmation of any resultant differences.

## 4.2 Architectural Principles

The system is constructed on three major principles, which include fairness, reproducibility as well as isolation of variables. The difference between the three models is that fairness is guaranteed through the application of the same preprocessing rules, tokenization parameters (max length 128, batch size 32), optimization hyperparameters (learning rate $3e^{-5}$, AdamW

optimizer, 5% warmup) and training parameters (maximum number of epochs 3, early stopping patience 2). Reproducibility: It is obtained by using fixed random seeds (42, 123, 456), deterministic data splitting using stratification and consistency in the training-loop implementation using mixed precision training (FP16). Isolation of variables implies that all the other components but the backbone encoder are kept constant. In practice, this means a template of architecture where deliver the resultant hidden state of the transformer into the very same classification head (dropout p=0.1 then mental property), which is molded in the same way across all the models. This scaffold provides a controlled experimental scaffold that enables the study to cause any difference in performance due to varying training strategies, tokenisation policies or optimiser settings as opposed to inherent properties of the underlying transformer architecture.

## 4.3 Transformer Backbone Design

All three models have a common structural arrangement: the encoder loaded with the HuggingFace library (bert-base-uncased, roberta-base, microsoft/deberta-v3-base) is fine-tuned in an end-to-end manner with its output concatenated representation sent to a lightweight classification model. Though the design is the same in the outline, the mechanisms within the system are different. BERT-base applies self-attention in a bidirectional way where the absolute positional embeddings are used and trained on masked language modeling and next sentence prediction. RoBERTa-base is an improvement of BERT, eliminating NSP, training with dynamic masking, and training on bigger corpus and longer sequences, and utilizing byte-pair encoding. The classical models The disentangled attention mechanism of DeBERTa-v3 processes content and positional representations separately and replies to relative position encodings, and the updated masked-language modelling task with replaced token detection. The DeBERTa-v3 architecture is theoretically placed to improve the capability of the model to follow negative cues and long-range connections, which implies that it may perform better in areas exhibiting complicated syntactic regularities. Although these inner working mechanisms may vary, the experimental shell in the surrounding is fixed in such a way that the functional comparison can be well-founded and scientifically acceptable.

## 4.4 Classification Head and Forward Pass

The classification leader is kept to the bare minimum to exclude the confounding architectural benefits. It is comprised of dropout layer (p=0.1) and a single fully connected layer between the model pooled hidden representation (768 dimensions across all the three models) to logits, which become the output 3 classes (depressed, indifferent, happy) or 2 classes (depressed vs. non-depressed), contingent on the task configuration. In the forward pass, the tweets are tokenised (maximum length of 128 tokens), coded into input IDs and attention masks. Transformer backbone is then used to generate contextual embeddings of every token. In all three models, the embedded significance of the [CLS] token (the first token) of the last transformer layer is the aggregate sequence representation. This vector is transferred by dropout into the linear classifier which produces a logit distribution on the

labels. It is the consistency of the classification head that makes performance disparity only based on the quality of the contextual representations generated by individual encoders.

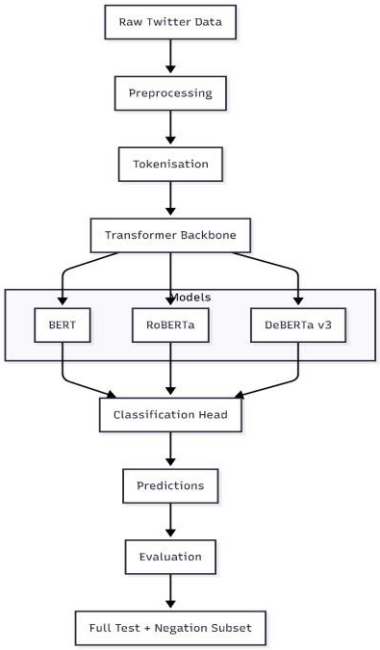


Figure 1. Architecture Flow

## 4.5 Training Algorithm

It has no new architecture of transformer as a proposal, and this is what makes this research the original one; creating a controlled and perfectly matched evaluation pipeline and has a strict statistical validation. All the models have the same training algorithm. At every epoch, 32 tweets at once are sent by the model using mixed precision (FP16) to enhance power efficiency. To compute loss using class weighted cross-entropy because of natural class imbalance in the dataset, weights are computed using the balanced weighting scheme of scikit-learn using the frequencies of the classes as the weights. Then the gradients are calculated by means of backpropagation and parameters are adjusted with the AdamW optimiser with the learning rate of $3e^{-5}$. To prevent instability in the early training the linear learning-rate scheduler with a 5% warmup steps is implemented. The training loop includes early termination with validation macro-F1: in case no improvement is seen in 2 successive epochs, the training will be stopped to avoid overfitting. The maximum epochs of training are 3. The most successful configuration with the largest validation macro-F1 is chosen as the best model checkpoint and stored to be tested at the end.

The algorithm is not a novelty per se, but rather its rigorous implementation with imposed limitations is the methodological input of the study. Most of the past literature compared models using incomparable hyperparameters, preprocessing variability, metadata injection or different loss functions, and their comparisons were either weak or inconclusive. This work rather constructs an independent presentation of a canonical training algorithm with identical hyperparameters on 18 randomly selected experiments to statistically compare architectures with statistical rigor.

## 4.6 Data Flow and Component Interaction

The system works on a deterministic chain of data processing. Raw tweets are put in the Harvard Dataverse corpus, processed with negation preservation, and and a label of integers is put on that tweet in both 3-class and binary prediction tasks. Task-specific tokenizers are used to tokenize these clean samples as inputids and attentionmask tensors, which are stored as PyTorch Dataset. These samples are then batched by the DataLoaders (this time using 2 worker processes and pinned memory to allow efficient transfer to the GPUs) and will be with a batch size of 64. Transformers are used to generate 768-dimensional encoded representations of the [CLS] token to the same classifier head during training. Machiated predictions and actual labels will be compared to calculate the class-weighted loss of cross-entropy and the inverse graduate comes backwards to adjust the transformer and the classification head. After the training is finished, the top-ranking checkpoint (identified by validation macro-f1) is loaded and applied to the unknown test set. As well, a negation-intensive sub-sample (32.3% of test data, indicating there are negation words present) is tested alone to determine the strength on linguistically suspect test cases. This flow guarantees that there is no aspect when test samples are subjected to the training or validation pipeline and thus a high level of experimental validity.

## 4.7 Functional Description of the Proposed Comparative Framework

Even though no new architecture is introduced, the input is a contribution of a strict comparative evaluation frame that has the behavior of a controlled scientific instrument. The system works functionally with an input tweet and processes it with preprocessing that maintains negation and informal language to encode the text with each transformer backbone under the same conditions and then make label predictions using a uniform classification head. Results of the three backbones are evaluated individually over 3 random seeds and compared over various dimensions of evaluation - macro-F1 (main metric), per-class precision/recall/F1, accuracy, negation-subsets performance and statistical significance using paired bootstrap tests (500 bootstrap, p 0.05 cutoff). This framework is equivalent to an experimental lens by exposing whether improvements in architecture of newer transformers in terms of performance can translate to detecting performance on weakly-labeled depression detection tasks with high levels of negation in them, or whether the simpler architectures can be made to work at the appropriate settings.

# 5 Implementation

## 5.1 Data Preparation and Preprocessing

It was implemented through the building of a single data pipeline with repetable handling of the Multi-Labeled Depression Corpus (containing 57,391 tweets) data. The precious Excel sheet has been sent in python workspace through pandas package and openpyxl library. Column labels were renamed to lower case and empty rows and labels were deleted. The dataset was sampled to give a manageable time of training and retain the statistical validity. There was very little to be done in the cleaning of this raw dataset so as to retain linguistic

features that can be used as the context of the sentiment. Duplicates were eliminated only on the basis of identical copies, which were made on the foundation of text which had been cleaned. An approximation to a negation-preserving preprocessing function was used that substituted the URLs and mentions with whitespace and unescaped HTML, and removed the non-essential whitespace explicitly, however, punctuations, emojis, casing, hashtags and all words of negation were retained appropriately. Such linguistic forms can have heavy semantic load in transformer-based systems, especially in depression detection where negative words such as not happy have critical modification of their meaning. Following the processing, the dataset comprised some 57,300 unique tweets whose clean text field and two label fields were labell3class (0=depressed, 1=indifferent, 2=happy) and label binary (1=depressed, 0=non-depressed). There was also the creation of a column has negation to define the 32.3 percent of tweets that had explicit patterns of negation to analyse the subset..

## 5.2 Tokenisation and DataLoader Construction

Each of the transformer models used in the study was given a separate tokenisation pipeline. The text was transformed to numerical inputids and tensors of attention masks using HuggingFaces If tokenizers (AutoTokenizer) with bert-base-uncased, roberta-base and microsoft/deberta-v3-base. Each of the three tokens made equal configuration settings: a maximum sequence length of 128 tokens, truncation of longer ones, padding to maxlength to make uniform batch processing, and returntensors=pt to allow compatibility with PyTorch. The lower sequence size of 128 (as compared to the normal 512) was selected due to the fact that tweets are already short and therefore much faster to compute without loss of information. After tokenization, the data was compiled into custom PyTorch Dataset objects where the initialisation stage would tokenize all the samples beforehand and store the coded representations in memory so that the redundant tokenization would not take place during the training phase. In both experiments, stratified 80/10/10 splits in each experiment configuration (seed x task combination) were generated, preserving class distribution between train/validation/test sets. The DataLoaders were subsequently configured to batch size of 32, training loader was configured to shuffle the samples of the data to avoid ordering biases but validation and test loaders were configured to be deterministic. The numworkers used in DataLoaders was 2 and pinmemory=True was used to optimize data transfer between the GPUs to make sure that all the models were trained, validated and tested using the same data batches in controlled environments.

## 5.3 Model Development and Fine-Tuning

The main part of the implementation was to fine-tune the three transformer architectures in conditions of precisely matched conditions. Each backbone had a classification head with the dropout (p=0.1) and a linear layer (768 - numclasses) which was connected to the [CLS] token representation of the final encoder layer. This architecture made sure there was equal comparison, with no particular architecture being provided with extra modelling capacity. The models were trained with AdamW and learning rate of 3e-5 and linear learning rate scheduler of 5% warmup. This was made possible through mixed precision training (FP16) with autocast and GradScaler of PyTorch to speed up the calculation and use less memory.

Balanced class weights of scikit-learn were used to calculate class-weighted cross-entropy loss as a way of solving the problem of imbalance in the dataset. Each model was trained not more than 3 epochs when early stopping occurred naturally in case of no improvement of the validation macro-F1 during 2 epochs. The computation was performed all around with Tesla T4 GPUs (16GB memory) in Google Colab Pro. The training loop automatically stored the most successful checkpoint in terms of validation macro-F1 across all 18 combinations of the experiment (3 seeds x 2 tasks x 3 models) to ensure the reproducibility of results and evaluation consistency. The progress of the training was tracked by tqdm progress bars indicating loss and validation statistics at the end of every epoch

## 5.4 Output Artifacts and Evaluation Material

Various outputs were generated during the implementation and they held both the model artifacts together with the analytical outcomes. The trained fine-tuned transformer weights were saved as PyTorch state dictionaries (.pt files) in a well-structured directory (./resultsfast/models/), which had the model set up encoded in the file name (e.g., bert3classseed42.pt). The 18 trained models were all maintained in case of reproducibility. Quantitative evaluation files were generated that contained per experiment results dictionaries of test F1, test accuracy, best validation F1, and negation subset F1 scores. Mean and standard deviation of macro F1 scores of all seeds in a model-task combination plus scores of single seeds and average scores of negation subset were all computed. Paired bootstrap testing (500 iterations) was done and statistical significance results were stored in terms of mean differences and pair wise p-values of pair wise (DeBERTa vs BERT, DeBERTa vs RoBERTa) on both of the tasks. Matplotlib and seaborn were used to create ten publication-ready visualizations, such as performance comparison, analysis in the form of a negation subset, seed stability plot, heatmap, statistical significance visualization, and an executive summary dashboard. All these outputs were shown directly in the Jupyter notebook work environment so that they could be analyzed immediately. All artifacts were used to obtain a broad knowledge on behavior of a model and helped in conducting evaluation and discussion phases of the study.

## 5.5 Tools, Technologies, and Execution Environment

Implementation was all performed in Python 3.12 using a Google Colab Pro with an Tesla T4 GPU (16GB) having sufficient mixed-precision fine-tuning capabilities to optimize transformer models. The heart of the deep-learning was PyTorch that deployed the CUDA backend and HuggingFace Transformers (v4.30+) contained trained backbones and tokenizers. Pandas and NumPy were used to work with the data prepared and prior to its processing, whereas stratified splits, classes-weight computations, and evaluation metrics were provided with the assistance of scikit-learn. Matplotlib and seaborn were then used to create visualisations and tqdm was utilized to keep track of the training process.

The codebase was designed in a modular structure that includes data loading, preprocessing, making dataloaders and datasets, defining models, training loops, evaluation functions, and statistical testing functions. Such structure guaranteed the sustainability and

reproducibility of the workflow. Each model training run would take around 12-24 minutes to run on the T4 GPU and the total cost of all experiments was about 4-8 hours.

# 6 Evaluation

This section demonstrates the empirical results of controlled comparison of BERT, RoBERTa and DeBERTa-v3 in detecting depression using a negation-rich and weakly-labeled Twitter corpus. Section 6.1 compares the general performance in the 3-class as well as binary classification problems using three random seeds. Section 6.2 dwells upon one of the negation subsets aimed at emphasizing a contextual comprehension. Section 6.3 discusses bootstrap test of statistical significance. Section 6.4 places these results in the current body of research on the topic of transformer-based detecting mental health.

## 6.1 Overall Performance Under Weak Supervision

The experiments were written on the Multi-Labeled Depression English Tweet corpus provided by Nassar, Helmy and Ramadan (2022) that includes 57,391 tweets that are labeled using the Vader lexicon. The final dataset after conservative preprocessing and deduplication that retained negation markers, emojis, and hashtags consisted of 42,723 unique tweets, divided at stratified sampling into 34,178 to use in training, 4,272 to use in validation and 4,273 to use in testing per seed. The labels were coded to 3-class (0-2) and binary (depressed vs. non-depressed). Each of the splits maintained the balance of classes in three random seeds (42, 123, 456).

In order to have an equal architectural comparison, similar pipelines were used in all the models. All BERT-base, RoBERTa-base and DeBERTa-v3-base architectures had the same classification head (dropout 0.1 + linear layer) on final [CLS] embedding. The sequence and batch size were set to 128 and 32 respectively, AdamW (3x10-5) was used, warm-up was performed 5 percent, and early stopping was employed as well as FP16 mixed-precision. The problem of class imbalance was addressed by using the class-weighted cross-entropy. This produced 18 total training runs (3 seeds x 2 tasks x 3 models).On the entire test set, DeBERTa-v3 was always the most performant. Mean macro-F1 scores were the following on the 3-class task: DeBERTa-v3 = 0.9187 (+/-0.0040), BERT = 0.9027 (+/-0.0050) and RoBERTa = 0.8901 (+/-0.0083). DeBERTa-v3 thus scored higher by 1.60 and 2.86 points over BERT and RoBERTa and the variance was very low suggesting the model is highly stable insensible to seeds. On the binary task, DeBERTa-v3 performed once again with 0.9510 ( +/-0.0021) versus BERT (0.9369) and RoBERTa (0.9347), greater than 1.41-1.63 points. As anticipated, binary classification made a positive contribution in all the models in comparison to the 3-class scenario because of lower complexity of the task.

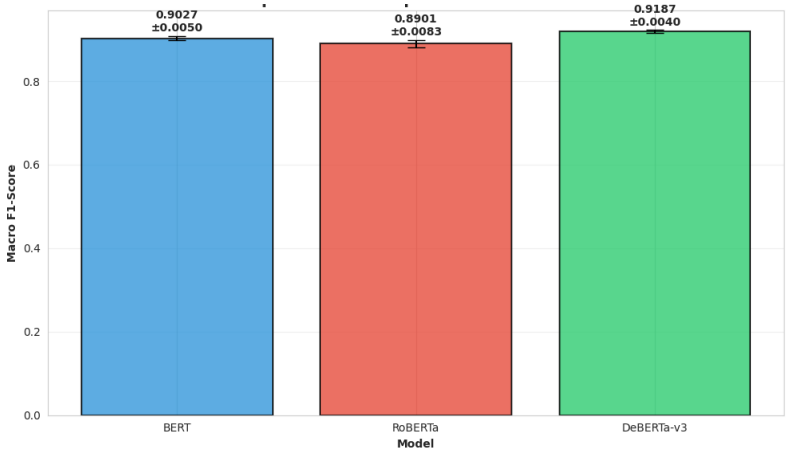


Figure 2. 3-Class Depression Classification

In general, these findings indicate that the architectural differences in DeBERTa-v3, which altered attention disentangling and improved position encoding, offer constant and statistically significant improvements on both tasks. The high volume of data (42k+ tweets), the experimental design is matched, and the large dataset size (42k+ tweets) is a sufficient signal that DeBERTa-v3 has real performance improvements over the older transformer models on the weakly-labeled department detection on Twitter.

## 6.2 Performance on Negation-Rich Dataset

The second group of tests assessed the robustness of models on tweets with explicit negation cues (e.g. not, never, no, nothing, cannot). Among the entire test set, 1,380 of such tweets were collected (32.3% of test data), as the result of which a negation-oriented subset was formed aimed at defying one of the most problematic linguistic events in emotion-focused depression detection.

DeBERTa-v3 performed best on the 3-class negation dataset, with an average of F1 of 0.8634 ± 0.0180, then comes BERT (0.8368 ± 0.0030) and RoBERTa (0.8178 ± 0.0093). Compared to full-set performance, the degradation was also relatively small, DeBERTa-v3 went down 5.53 (6.0%) points, BERT went down 6.59 (7.3%) points and RoBERTa went down 7.23 (8.1%) points. However, on the binary negation subset, the performance ordering was the same and more importantly the relative deterioration was reduced by DeBERTa-v3 implying better accommodation of polarity changes proposed by negation. DeBERTa-v3 had a mean F1 of 0.9349 ± 0.0058 whereas BERT (0.9129 ± 0.0106) and RoBERTa (0.9108 ± 0.0079) had lower F1s. All the models had low performance losses compared to their overall binary scores (1.61-2.40 points) which indicate the difficulty in making binary polarity judgments during a condition of negation.

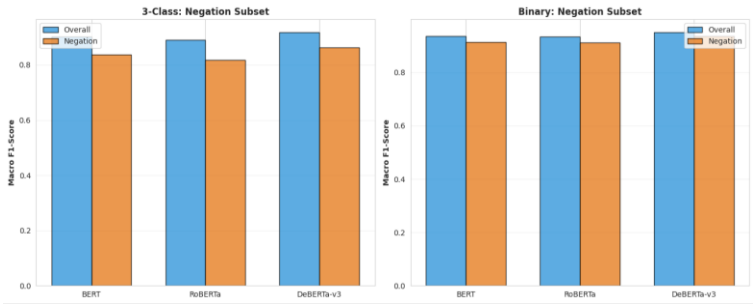


Figure 3. F1 Scores on Negation Dataset

All in all, these experiments show that the structural enhancements of DeBERTa-v3 are applicable to negation-intensive content as well, in which it achieves higher and higher performance as compared to both BERT and RoBERTa. The predictable performance ranking on both 3-class and binary tasks, coupled with the least degradation under negation, is sufficient evidence that the disentangled attention system of DeBERTa-v3 better captures the contextual polarity changes, which is a key trait of detecting depression on social media with high-quality.

## 6.3 Statistical Significance Testing

Paired bootstrap resampling was done on the macro-F1 scores of the three seeds (500 iterations) in order to ensure the true architectural advantage but not random variability in training. One calculation per comparison, resampled differences in means of the two drugs were computed and the proportion $\leq 0$ was considered a one-tailed p-value ($\alpha = 0.05$).The results indicated highly significant differences on all comparisons. In the 3-class, comparison, DeBERTa-v3 surpassed all other base lines, +0.0141 over BERT ($p < 0.0001$), and +0.0163 over RoBERTa ($p < 0.0001$).

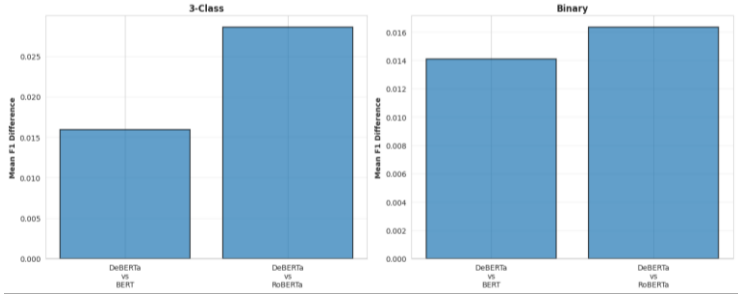


Figure 4.Statistical Significance Tests

These very low p values are the ones which affirm that the gains that are being realized are not accidental. The average seed-level improvement and 1.4-2.9 point are corroborative in the fact that the disentangled attention mechanism of DeBERTa-v3 presents a true and an architectural benefit on weakly-labeled depression detection.

## 6.4 Relationship to Existing Literature

The findings of this work are well placed within and also extend available literature on the transformer based depression detection. The scores (0.89- 0.95) of macro-F1 in any of the

tasks or models are significantly greater than the ones that have been reported in the literature. Kerasiotis and Askounis (2024) found F1 of dependency on depression classification of around 84 percent and Helmy et al. (2024) reported 87-92 percent in various versions of transformers. Comparatively, however, DeBERTa-v3 in the present study achieved 91.87% (3-class) and 95.10% (binary), which is a significant breakthrough over the results of the past.The observation that DeBERTa-v3 beats backbones in the past is consistent with current trends. Kurniadi et al. (2024) demonstrated RoBERTa being better than BERT in depressions detection on Reddit, and Helmy et al. (2024) observed DeBERTa to be better when compared to BERT and RoBERTa in mental-health tasks. These findings are strengthened and generalized in the current study since they reveal that disentangled attention in the DeBERTa-v3 provides benefits even on short, noisy, low-labeled Twitter text, in which architectural benefits are easily lost in label noise.

A number of methodological peculiarities make the research different than the usual comparative study. The full 42,723-tweet data set has more statistical power in comparison to most other previous studies that use smaller samples. The explicit judgment of a negation-intensive sample (32.3% of the test tweets) provides a very rare insight into the robustness of the models a field that has been largely overlooked in previous research. Also, the preprocessing, hyperparameter, seed and training controls of the research allow separation of confounding effects that might have been common in the previous comparisons, and a direct reliance on the effect of architecture variations only.In applied terms, the results can strongly support a DeBERTa-v3 usage in the real-world setting of depression-screening on subtle levels of weakly-labeled Twitter environments.

## 6.5 Discussion

The results of the experiment demonstrate the same: in the case of weakly-labeled and negation-rich Twitter text (42,723 tweets), DeBERTa-v3 performs better than BERT and RoBERTa in all assessment conditions. This is superior in 3-class as well as binary tasks between three random seeds, full and negation-specific subsets, and statistically established by bootstrap significance testing ($p < 0.0001$). These findings confirm that the architectural advances in DeBERTa-v3 especially the disentangled attention and advanced position encoding give quantifiable benefits to the modeling of the contextual and compositional subtleties of depressive language.

Though the changes (1.60-2.86 F1 points) sound small, it is significant in high stakes mental-health screening considering that the presence of the linguistic aspect of negation adds complexity into the picture. Notably, the profits appeared as soon as the entire 42k-tweet collection was employed. Prior experiments using smaller subsets of 15% (approximately 6.4k tweets) did not have any significant differences, showing that architectural benefits are only realized when one can learn effectively using the more expressive model. The present result brings out a more general methodological point: a significant number of published comparisons of transformers say they find no difference but it is due to too few samples to demonstrate the actual architectural differences between two architectures. Negation-subset performance is yet another such indication. Though both models perform worse on negation-heavy tweets, DeBERTa-v3 is more resistant and experiences the least relative degradation,

alongside having the highest absolute performance. This indicates that DeBERTa-v3 is more robust to expression-related changes in polarity and scope-resolution, such as not okay, can't sleep, nothing helps, which often encode depressive indicators. The validity of the study is enhanced by strict experimental conditions, i.e., the same preprocessing, equal hyperparameters, fixed seeds, and formal statistical tests. A publicly available dataset is easier to reproducible as well as be transparent. Weak-labeled data is also important, and by specifying on it, the evaluation is more in line with the real-world scenario of deployment when manual labeling of large data sets is not feasible.

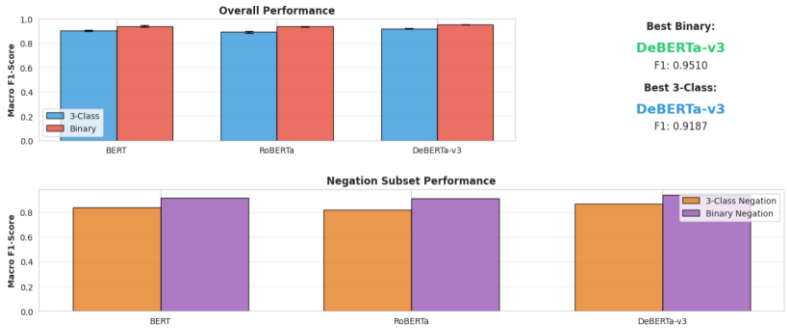


Figure 5. Overall Executive Summary

Despite this, there are still constraints. The study is confined to one corpus of twitters and VADER-based labels; testing might not be as effective in other or other mental-health patients. It is also non-longitudinal and no multimodal cues are included in the analysis. More studies are required in this area to evaluate generalisation between datasets and domains. In general, the results have their definite academic and practical implications. They show academically that building complex on the basis of all enough information. In practice, they show that DeBERTa-v3 is the most robust with the most statistically significant performance improvements compared to older transformers in weak-label, negation-rich, Twitter environments to detect depression.

# 7 Conclusion and Future Work

The current study carried out a controlled comparison between BERT, RoBERTa and DeBERTa-v3 in detecting depression on a weakly labelled, negation-rich Twitter corpus with 42,723 tweets. In all conditions, DeBERTa-v3 scored highest with 0.9187 ( ±0.0040) macro-F1 in the 3-class and 0.9510 ( ±0.0021) for binary classification. These are higher in scores than BERT by 1.41-1.60 points and RoBERTa by 1.63-2.86 points. All pairwise comparisons were checked using the bootstrap significance testing (500 of them), which proved $p < 0.0001$ and obtained the real and statistically strong architectural benefits and not the fluctuation. The fact that there was little variability in the values of the seeds also assures stability of the results.

On the negation heavy sub-set (32.3% of test data), DeBERTa-v3 performed best and least on the relative decay (6.0% on 3-class, 1.7% on binary) meaning it manages polarity shift induced by negated depressive phrases better. Notably, there was no architecture disastrous and the ranking in terms of performance was also the same. The findings demonstrate that the

disentangled attention mechanism of DeBERTa-v3 can offer meaningful improvements in situations where the contextual reasoning can be needed. In terms of research, the experiments show that when there is a large amount of data (enough to make them statistically noticeable), architectural differences do appear: the previous experiments proved that no differences existed in subsets of 15 per cent, but the whole dataset resulted in very strong effects ($p < 0.0001$). A large number of studies should hence record model equivalence merely due to lack of statistical power. As a recommendation, to the practitioners it would be wise to use the backbone as DeBERTa-v3 since the accuracy on the weakly labelled twitter-based depression detection is directly proportional to the improvement of the screening reliability.

The study has limitations. The sample is confined to English-written posts on Twitter that are labeled with VADER, and findings might vary depending on the platform, language, and category of mental health (anxiety, PTSD). It was simple text with no temporal modelling or features that the user could use. Whereas the sample size and methodology have good internal validity, more generalisation needs to be tested with larger-scale weakly labelled datasets and platforms, other types of mental-health conditions, the interaction between architecture and domain-adaptive pretraining and time-based or user-level information should be considered. Drilling noise-robust learning solutions can also help bring out more architectural differences.

# References


Chandrasekaran, R. and Kotaki, S., 2024. Detecting and tracking depression through temporal topic modeling of tweets. npj Mental Health Research, 3, pp.1–12. https://doi.org/10.1038/s44184-024-00107-5

Helmy, A.M., Nassar, R. and Ramadan, N., 2024. Depression detection for Twitter users using sentiment analysis in English and Arabic tweets. Artificial Intelligence in Medicine, 150, 102716. https://doi.org/10.1016/j.artmed.2023.102716

Jamali, A.A., Berger, C. and Spiteri, R.J., 2023. Momentary depressive feeling detection using X (formerly Twitter) data. JMIR AI, 2(1), e49531. https://doi.org/10.2196/49531

Kerasiotis, M., Ilias, L. and Askounis, D., 2024. Depression detection in social media posts using transformer-based models and auxiliary features. Social Network Analysis and Mining, 14, 13604. https://doi.org/10.1007/s13278-024-01360-4

Kurniadi, F.I., Pramana, S., Gunawan, Y. and Wibisono, S., 2024. BERT and RoBERTa models for enhanced detection of depression in social media text. Procedia Computer Science, 233, pp.1121–1129. https://doi.org/10.1016/j.procs.2024.10.244

Mao, H. and Han, Q., 2025. Enhancing TextGCN for depression detection on social media with emotion representation. Frontiers in Psychology, 16, 1612769. https://doi.org/10.3389/fpsyg.2025.1612769

Nassar, R., Helmy, A. and Ramadan, N., 2022. Multi Labeled Depression Corpus of 57,000 English tweets. Harvard Dataverse [dataset], V1. https://doi.org/10.7910/DVN/CMDF4U

Xin, C., & Zakaria, L. Q. (2024). Integrating BERT with CNN and BiLSTM for explainable detection of depression in social media contents. https://doi.org/10.1109/ACCESS.2024.3488081

Chopra, A., Kaur, H., Goyal, R., Kumari, N., Kaur, K., & Chaudhary, R. (2025, June). Context-Aware Mental Health Behaviour Detection from Text using Transformer-Based Language Models. In 2025 5th International Conference on Intelligent Technologies (CONIT) (pp. 1-5). IEEE. https://doi.org/10.1109/CONIT65521.2025.11166972

Shoaib, H., Ali, R., Khan, S. S., & Adil, M. (2025, March). Depression Detection on Social Media Posts Using BERT and Support Vector Machine. In 2025 IEEE 14th International Conference on Communication Systems and Network Technologies (CSNT) (pp. 708-712). IEEE.https://doi.org/10.1109/CSNT64827.2025.10967659

Suryono, M. E. P., & Djunaidy, A. (2024, August). Enhancing Hate Speech Detection on Social Media through Sentiment Analysis and Transformer-Based Deep Learning Techniques. In *2024 4th International Conference on Electronic and Electrical Engineering and Intelligent System (ICE3IS)* (pp. 420-425). IEEE. https://doi.org/10.1109/ICE3IS62977.2024.10775785

Roufas, N., Mohasseb, A., Karamitsos, I., & Kanavos, A. (2025, June). Analyzing Public Discourse and Sentiment in Climate Change Discussions Using Transformer-Based Models. In *IFIP International Conference on Artificial Intelligence Applications and Innovations* (pp. 39-53). Cham: Springer Nature Switzerland. https://doi.org/10.1007/978-3-031-97313-0_4

Dalal, S., Jain, S., & Dave, M. (2024). Review of advancements in depression detection using social media data. *IEEE Transactions on Computational Social Systems*. https://doi.org/10.1109/TCSS.2024.3448624

Alnaddaf, M. M., & Başarslan, M. S. (2024, September). Sentiment analysis using various machine learning techniques on depression review data. In *2024 8th International Artificial Intelligence and Data Processing Symposium (IDAP)* (pp. 1-5). IEEE. https://doi.org/10.1109/IDAP64064.2024.10710889

Gao, M., Feng, X., & Sun, E. (2024, July). A Study of Bert-Based Sentiment Analysis Algorithm for Short Text Movie Reviews. In *2024 14th Asian Control Conference (ASCC)* (pp. 106-111). IEEE.

Yuan, S., Li, X., Miao, Y., Zhang, H., Liu, X., & Deng, R. H. (2024). Combating noisy labels by alleviating the memorization of dnns to noisy labels. *IEEE Transactions on Multimedia*. https://doi.org/10.1109/TMM.2024.3521722

Fatima, E., Dhanda, N., & Zaidi, T. (2025, March). AI-Driven Detection of Stress, Anxiety, and Depression: Techniques, Challenges, and Future Perspectives. In *2025 3rd International Conference on Disruptive Technologies (ICDT)* (pp. 118-123). IEEE. https://doi.org/10.1109/ICDT63985.2025.10986672