Classification of Depressive Content in Decentralized Social Networks: A Case Study on Bluesky

Removed for double-blind review
Removed for double-blind review

ABSTRACT

Depression is a major global mental health concern, and early detection through social media has become increasingly relevant. This study introduces a machine learning framework to identify depressive tendencies in posts from Bluesky, an emerging and underexplored decentralized social network. We curated and manually annotated a dataset of depression-related posts with expert input and evaluated several transformer-based models, including BERT-base, RoBERTa-base, and MentalBERT. Our final pipeline integrates linguistic features with engagement metadata in a stacked ensemble, achieving an F1-score of 80.5% on Bluesky and 90.8% on Twitter, demonstrating strong cross-platform generalization. These results suggest that deep learning models, when enriched with contextual signals, can effectively capture depressive language patterns across diverse platforms. While promising, the system is not intended to replace clinical diagnosis, but rather to serve as an auxiliary tool to support mental health professionals. Any real-world deployment would require rigorous clinical validation, ethical oversight, and integration with human-centered support mechanisms.

KEYWORDS

depression detection, early detection, social media analysis, Bluesky, decentralized social networks, BERT, RoBERTa, MentalBERT, deBERTa, transformer models, ensemble learning, digital mental health

1 INTRODUCTION

Depressive disorder, or depression, affects approximately 5% of adults worldwide and remains one of the leading causes of disability [50]. Beyond mood fluctuations, it severely impacts daily functioning, relationships, and overall quality of life [3, 43].

Despite its prevalence, the diagnosis of Major Depressive Disorder (MDD) still depends heavily on subjective self-reports and clinical observation. Field trials of the DSM-5 indicate low inter-rater reliability, with Cohen's kappa values as low as 0.28 [30, 42], exposing inconsistencies in clinical assessments and underscoring the need for more objective, scalable diagnostic tools [42].

In recent years, social media has emerged as a promising source of behavioral signals for early depression detection at scale. Prior studies—especially on Twitter—have linked linguistic markers such as first-person singular pronouns, negative emotion words, and cognitive rigidity to depressive symptoms [41, 49]. While CLEF eRisk has advanced research in early risk prediction, its scope remains limited to centralized platforms. Our work addresses this gap by exploring early detection in decentralized social networks, which exhibit distinct user dynamics and information flows.

Bluesky, a decentralized social network launched in 2024, introduces a federated architecture, open protocol ecosystem, and community-based moderation—all of which influence how users communicate and engage online [36, 44]. Its user base is predominantly young, with the 18–24 age group representing the most significant demographic segment [17]. Despite its growth, no previous research has explored mental health expression or detection in this new ecosystem.

This study addresses this gap by investigating the influence of platform architecture and user demographics on depression expression in decentralized environments. We hypothesize:

> *The decentralized structure and younger demographic of Bluesky shape the linguistic expression of depression differently than centralized platforms like Twitter.*

To test this hypothesis, we curated and manually annotated a dataset of Bluesky posts with support from clinical psychologists. Using a lexicon grounded in clinical frameworks such as the DSM-5 and PHQ-9 [46], we captured both explicit and implicit depressive signals. We then proposed a hybrid detection pipeline combining: (i) a classical TF–IDF + SVM model, (ii) a metadata-based classifier using XGBoost, and (iii) fine-tuned transformer architectures (BERT-base, RoBERTa-base, and MentalBERT). Unlike prior work, we integrated engagement metadata and conducted ablation studies to assess its predictive value.

We further evaluated model generalization through external validation on a Twitter dataset. Our findings demonstrate that linguistic markers of depression remain transferable across platforms, while platform-specific features (e.g., user behavior, moderation style) influence how depression is expressed.

This research advances digital mental health and NLP by exploring depression expression in a novel sociotechnical context. Our key contributions are:

- The first manually annotated dataset of depression-related posts from Bluesky, a decentralized social network with a predominantly young user base.
- A hybrid detection pipeline that combines linguistic signals with engagement metadata, enabling context-aware modeling tailored to emerging online environments.
- Empirical evidence that linguistic markers of depression are stable across platforms, while expression patterns vary with platform architecture and audience demographics.
- Open-source resources, ethical safeguards, and linguistic insights that support reproducibility and responsible AI deployment in mental health NLP.

In: Proceedings of the Brazilian Symposium on Multimedia and the Web (WebMedia'2025). Rio de Janeiro, Brazil. Porto Alegre: Brazilian Computer Society, 2025.

ISSN 2966-2753

2 RELATED WORK

Over the past decade, the detection of mental health conditions from social media has advanced substantially, transitioning from traditional classifiers to deep learning architectures [7, 31].

2.1 Traditional Machine Learning Approaches

Early approaches to depression detection relied on handcrafted features combined with supervised algorithms. Ahmed et al. applied sentiment analysis with Naïve Bayes and a hybrid NBTree model on Twitter data, reporting high accuracy (97.3%)—though strongly dependent on TextBlob polarity scores [1]. D'Cruz et al. applied Naïve Bayes, Random Forest, and SVMs to Reddit and Twitter posts for depression prediction, achieving balanced precision and recall [13]. However, such models lack contextual understanding and are prone to errors in the presence of sarcasm, figurative language, and informal expressions.

2.2 Transformer-Based Architectures

Transformer models such as BERT and RoBERTa have transformed the landscape of depression detection [4]. MentalBERT, pre-trained on mental health forums, improved performance on domain-specific corpora [26], though it often underperforms on general social media text due to domain mismatch. Recent work highlights the promise of DeBERTa [24], which incorporates disentangled attention, yet remains underutilized in mental health contexts. Ensemble methods combining multiple transformers have demonstrated gains in robustness and predictive power [48].

2.3 Multimodal and Metadata-Enhanced Models

Recent studies have combined textual embeddings with auxiliary features—such as posting time, emoji usage, and engagement metrics (e.g., likes and replies)—to improve depression detection [28]. While metadata adds predictive value, it may also reflect confounding patterns tied to platform-specific interaction dynamics [7], requiring careful interpretation.

2.4 Datasets and Annotation Practices

The majority of studies use datasets from Twitter, Reddit, DAIC-WOZ, or the CLEF eRisk initiative. These are often labeled automatically—via keyword heuristics or self-disclosure—introducing annotation noise and demographic bias. Expert-led manual annotation remains rare but is essential for capturing subtler markers of depression such as metaphor, temporality, and indirect self-disclosure. Moreover, few datasets report inter-annotator agreement, limiting reproducibility [21].

2.5 Temporal Modeling and Early Detection

The CLEF eRisk shared tasks, initiated in 2017, emphasize the role of temporality by incrementally releasing user posts, encouraging early risk detection from partial user histories [33–35]. This framework has motivated the development of sequential models—such as LSTMs, GRUs, or transformers with temporal embeddings [5]—that can capture temporal dynamics in users' language over time.

While temporal modeling remains a valuable direction for future research, especially for longitudinal symptom tracking, our study focuses on how decentralized architecture and younger demographics shape the linguistic expression of depression. Given this goal, our models treat posts independently, as our primary interest lies in understanding platform- and audience-specific language patterns, rather than sequence-based progression.

2.6 Recent Advances and Algorithmic Bias

Recent directions include explainable AI and potential multiclass classification. Cha et al. [9] applied binary models to Korean social media, suggesting future use in broader mental health tasks. Hasan et al. compared transformers and LSTMs on Reddit, reporting F1-scores above 96% [23]. Yet, reviews still highlight challenges like sample bias, inconsistent preprocessing, and class imbalance [7].

2.7 Brazilian NLP Initiatives

Although our models were developed using English data, Brazilian NLP offers useful parallels. Resources such as BERTimbau provide high-quality contextual embeddings for Portuguese [45], while corpora like UStanceBR support stance detection in social contexts [39]. The IDPT shared task has also explored depression detection in Portuguese-language platforms, underscoring the feasibility of culturally adapted pipelines. Exploring how depressive expression manifests in Brazilian platforms remains an open direction for multilingual expansion.

2.8 Summary of Gaps Addressed

Although prior research has significantly advanced depression detection on centralized platforms like Twitter and Reddit, no previous study has examined how depressive content is expressed in decentralized environments. Platforms like Bluesky introduce novel dynamics—such as federated architecture, community-based moderation, and a demographically distinct user base—that may influence both how users communicate and how mental health risks are signaled.

Our study addresses this overlooked space by investigating the sociotechnical factors that shape depressive language on Bluesky and evaluating whether existing linguistic markers transfer across platforms with divergent architectures and audiences.

3 DATASET CHARACTERIZATION

3.1 Data Collection and Ethics

We collected public posts from Bluesky using Apify's scraping tool, focusing exclusively on non-private content to ensure compliance with ethical standards in digital mental health research [12]. All personally identifying information (e.g., usernames, URLs) was removed during preprocessing, and only anonymized data were retained.

Nonetheless, ensuring complete anonymization of linguistic data remains a significant challenge, given the potential for re-identification through unique writing patterns and stylistic cues [51]. To address this risk, we implemented strict privacy-preserving protocols and restricted data sharing to fully anonymized or aggregated representations, thereby minimizing any potential harm to individuals.

Initial post retrieval was based on a lexicon of depression-related expressions (Table 1), informed by clinical instruments such as the PHQ-9 [46], the DSM-5, and linguistic patterns described in prior studies like DEPTWEET [27]. This hybrid approach ensured both clinical relevance and coverage of informal expressions. It included explicit markers (e.g., "I want to die") and subtler cues (e.g., "feeling empty inside"), while minimizing false positives.

Category	Words/Expressions
General depressive thoughts	depress, alone, stress, I'm trash, desperate, gonna cry, solitude, emptiness, hate myself, feel useless, helpless, tired of everything, have no future, deep sadness, meaningless, dead inside, mentally exhausted, guilt, fear, emotional fatigue, feel empty inside, diagnosed depression, depressive thoughts
Suicidal and extreme thoughts	suicidal thoughts, kill myself, I just want everything to stop, no one would notice if I disappeared, life is pointless, hate my life, I'm done trying, want to die, I'm a burden, I wish I could disappear, I give up, wish I was dead, hate who I've become, no one needs me, no way out, end of the line

Table 1: Lexicon of depression-related terms used for initial data retrieval.

3.2 Manual Annotation

Two clinical psychologists independently annotated the dataset. To ensure consistency, a calibration phase involved jointly reviewing 200 posts to align labeling criteria, resolve ambiguities, and standardize edge cases—such as ironic or metaphorical expressions (e.g., "I wish I could disappear" used humorously).

Inter-annotator agreement was calculated on a random sample of 100 posts, yielding a Cohen's Kappa of 0.81, which indicates substantial agreement between annotators. This result supports the methodological soundness of our human-in-the-loop annotation process. For comparison, prior work such as Atapattu et al. [2] reports a Fleiss's Kappa of 0.82 in a similar mental health corpus.

3.3 Dataset Statistics

Following preprocessing, we excluded posts with highly ambiguous or metaphorical language, such as ironic expressions of distress. The final dataset includes **4,898** public English-language posts, with **712** (14.5%) labeled as depressive.

3.4 Engagement, Temporal, and Linguistic Patterns

Engagement Analysis. Figure 1 shows that over **60% of depressive posts received zero likes**, while non-depressive posts attracted higher engagement. Prior studies suggest that reduced interaction may stem from both social withdrawal [14] and stigma-related inhibition of emotional expression [41].

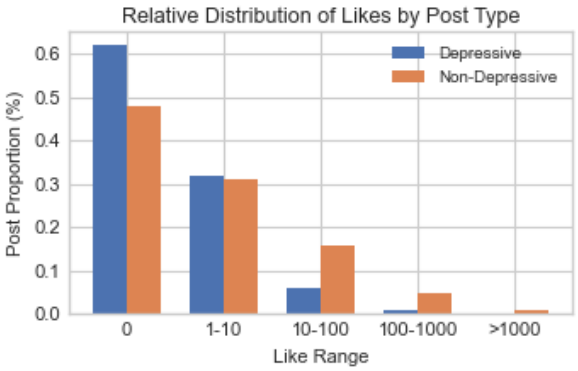

Figure 1: Distribution of like counts for depressive vs. non-depressive posts.

Temporal Analysis. Figure 2 shows that posting activity peaks notably between **11 PM and 1 AM**, suggesting a concentration of user activity during late-night hours. This pattern may reflect *circadian rhythm disruptions*, which are commonly linked to depressive symptoms and sleep irregularities [20]. Prior studies on platforms such as Reddit and Twitter have reported similar findings, where users exhibiting depressive behavior tend to post more frequently during late-night or early-morning hours—possibly due to insomnia, rumination, or emotional distress during quieter periods of the day [6].

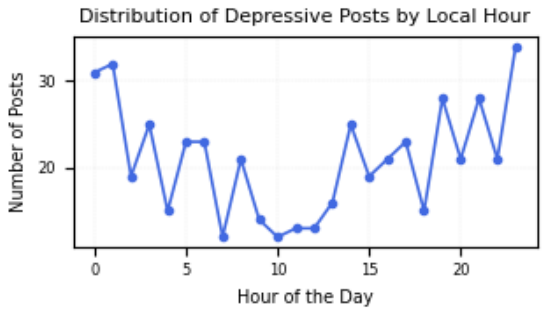

Figure 2: Post frequency by local time.

Linguistic Patterns. The word cloud in Figure 3 highlights tokens such as "tired", "hate", and "suicidal". However, as these terms were part of the initial retrieval lexicon, their frequency alone cannot reliably distinguish depressive intent. This limitation underscores the need for context-aware models to interpret sarcasm, exaggeration, and metaphor—frequent in social media discourse.

Figure 3: Most frequent words across the collected posts.

4 METHODOLOGY

This section presents the full pipeline for detecting depressive content in Bluesky posts, with emphasis on transparency, reproducibility, and alignment with best practices in NLP and digital mental health research. The methodology comprises six stages: (i) preprocessing and feature engineering, (ii) classical NLP baseline, (iii) metadata-based baseline, (iv) transformer-based models, (v) stacked ensemble learning, and (vi) reproducibility practices.

4.1 Preprocessing and Feature Engineering

The anonymized dataset was loaded from CSV format and underwent a series of preprocessing steps inspired by prior work on social media-based depression detection [15, 22]:

- Text entries were cast to string, and posts containing only whitespace or null values were removed.
- Emojis were removed during preprocessing, which may limit the model's sensitivity to affective signals often conveyed through symbolic expression.
- Platform-specific artifacts such as URLs, mentions (`@`), and hashtags were removed.
- Contractions were expanded (e.g., "I'll" → "I will"), glued terms separated (e.g., `mentalhealth` → `mental health`), and common abbreviations normalized (e.g., "kms" → "kill myself").
- Tokenization and lowercasing were consistently applied across all models.

These preprocessing decisions were validated via preliminary experiments and grounded in literature that stresses linguistic nuance in mental health NLP.

4.2 Baseline Model: TF-IDF + SVM

As a classical baseline, each post was vectorized using Term Frequency–Inverse Document Frequency (TF-IDF) with uni- and bi-grams (minimum document frequency = 5). A linear Support Vector Machine (SVM) classifier was then trained to perform the classification. This approach has been robustly employed in depression detection on social media, achieving up to 81.79% accuracy in recent studies [37]. This well-established baseline serves as a crucial reference point for assessing the added value of modern language models.

4.3 Metadata-Based Baseline: XGBoost

Given the predictive potential of user engagement features [11], we trained an XGBoost classifier using only metadata listed in Table 2. To mitigate class imbalance (14.5% depressive posts), random oversampling was applied to the training set only.

Repost Count	Like Count	Reply Count	Hour of Posting
Month	Has Image?	Has Link?	Has Hashtag?

Table 2: Metadata features used in the XGBoost baseline.

4.4 Transformer-Based Language Models

For text-based classification, we fine-tuned three transformer models using the HuggingFace `transformers` library, leveraging pre-trained weights and adapting them to our dataset. A classification head was added to each model and trained using cross-entropy loss on binary labels. All models were trained under identical experimental settings to enable fair comparison.

- **BERT-base uncased** [16], a widely used general-purpose model for contextual embeddings.
- **MentalBERT** [26], pre-trained on mental health forums, tailored for domain-specific nuance.
- **RoBERTa-base** [32], an optimized BERT variant with superior training strategies.

To handle class imbalance, oversampling was applied only to the training set. Table 3 summarizes hyperparameter choices.

Parameter	Value
Optimizer	AdamW
Learning Rate	2×10^{-5}
Batch Size	8
Epochs	5 (early stopping on validation loss)
Train/Test Split	80/20 (fixed seed)
Oversampling	Applied only to training set

Table 3: Hyperparameters used for transformer fine-tuning.

4.5 Stacked Ensemble

To combine strengths of different models, we implemented a stacked ensemble as follows:

(1) Independently train base models (TF-IDF SVM, XGBoost, BERT, MentalBERT, RoBERTa).
(2) Use each model's predicted probability as input feature.
(3) Train a logistic regression meta-learner on these features.
(4) At inference, base model outputs are aggregated via the meta-learner to produce the final prediction.

Stacking has shown to improve robustness in depression detection tasks by integrating heterogeneous signals [47].

4.6 Evaluation Metrics and Error Analysis

All models were evaluated on a held-out test set using standard classification metrics: accuracy, precision, recall, F1-score, and Area Under the Receiver Operating Characteristic Curve (AUROC). Given the inherent class imbalance in our dataset, special emphasis was placed on metrics that reflect performance on the minority (depressive) class. In line with recent best practices for mental health detection on social media, we prioritized F1-score and AUROC over accuracy, as the latter can obscure meaningful disparities across classes [8].

To complement the quantitative evaluation, we also conducted a qualitative error analysis, which revealed systematic patterns in both false positives and false negatives. These included the misclassification of posts describing physical symptoms as depressive, and the overlooking of retrospective or sarcastic expressions of distress.

4.7 Reproducibility and Open Science Practices

To ensure reproducibility and transparency:

- **Code availability:** Scripts for data processing, training, and evaluation will be released on GitHub upon acceptance.
- **Dataset availability:** An anonymized subset and detailed reconstruction guide will be provided, respecting privacy constraints.
- **Hyperparameter disclosure:** Table 3 lists all values used.
- **Model checkpoints:** Fine-tuned models will be released post-approval.

These practices align with current reproducibility guidelines in machine learning [19, 40] and promote trustworthy research in sensitive domains like mental health.

5 RESULTS

5.1 Baseline Models Performance

Our initial baseline, a metadata-only **XGBoost** classifier, achieved strong performance for the non-depressive class, with **92% precision** and an **overall accuracy of 71%**. However, recall for depressive posts was limited to **51%**, indicating that a substantial portion of true positives remained undetected. These results suggest that engagement-related metadata can capture relevant behavioral patterns—such as posting frequency, interaction volume, or timing—but may lack the depth to fully represent psychological states. Rather than indicating clinical distress, such features often reflect usage habits specific to the platform's social dynamics, as noted in prior research on the limitations of social media data in mental health contexts [29]. This underscores the value of combining metadata with linguistic information, as pursued in our ensemble approach, to better detect nuanced signals of mental health risk.

As a textual baseline, we implemented a classical **TF-IDF + linear SVM** model, which achieved an **F1-score of 0.73**. While effective in capturing surface-level lexical associations, this model underperformed in handling contextual and pragmatic nuances—limitations typically addressed more effectively by transformer-based architectures.

5.2 Language Models Evaluation

To explore the advantages of contextualized language understanding, we fine-tuned three transformer-based models—**BERT**, **MentalBERT**, and **RoBERTa**—under identical experimental conditions. Their performance was evaluated against a **stacked ensemble** that combined both text-based and metadata-based signals.

To account for class imbalance, each model was tested with and without **random oversampling** during training. Table 4 presents the evaluation results across standard metrics, with the best-performing values per column highlighted in bold. Overall, transformer models showed substantial improvements in recall and F1-score, particularly when combined with oversampling and ensemble strategies.

With *oversampling*				
Models	Acc.	Prec.	Rec.	F1
BERT	**0.946**	0.830	0.775	0.802
MentalBERT	0.915	0.643	0.867	0.738
RoBERTa	0.935	0.757	0.790	0.773
Stacking	0.945	0.821	**0.791**	**0.805**
Without *oversampling*				
Models	Acc.	Prec.	Rec.	F1
BERT	0.937	**0.839**	0.681	0.752
MentalBERT	0.923	0.754	0.659	0.704
RoBERTa	0.924	0.668	0.920	0.774
Stacking	0.936	0.812	0.712	0.759

Table 4: Performance of transformer models evaluated with and without oversampling. Best scores per column are highlighted.

Key insights:

- **BERT with oversampling** achieved the best F1 among single models (0.802), balancing precision and recall.
- **RoBERTa** had the highest recall but more false positives, reducing precision.
- **MentalBERT** underperformed, likely due to domain mismatch with informal and sarcastic language on Bluesky.
- The **stacked ensemble (BERT + XGBoost)** achieved the highest F1 overall (0.805), validating the value of metadata—though it risks encoding platform-specific biases.

5.3 Ablation Study: Metadata Contribution

To evaluate the contribution of engagement metadata, we retrained the stacked ensemble using only textual model outputs. This resulted in a **6% drop in F1-score**, indicating that metadata provides complementary predictive signals. While useful, these gains should be interpreted with caution, as metadata may partially reflect platform-specific interaction patterns rather than underlying psychological states. This reinforces the importance of linguistic features as the primary source of mental health signals in our approach.

5.4 Qualitative Error Analysis

To better understand model behavior beyond quantitative metrics, we conducted a qualitative analysis of misclassified samples. This review revealed three recurrent patterns of error, which are further explored in the discussion.

False Positives: The model frequently classified posts describing physical symptoms as depressive, even when they lacked emotional content. Examples include:

- *"nauseous all day, can't eat"*
- *"my back hurts so bad I can't sleep"*
- *"migraine's killing me again"*
- *"still stuck in bed with a fever"*
- *"this cold is draining all my energy"*

These cases suggest that the model may conflate physical discomfort with psychological distress, over-relying on surface-level lexical patterns of suffering.

False Negatives: Posts referencing past depressive episodes or indirect indicators of risk were often ignored or downweighted. Representative examples include:

- *"I had depression through most of college"*
- *"used to cry myself to sleep every night"*
- *"back when I was in therapy, things were rough"*
- *"I got off meds a few months ago, finally"*
- *"was suicidal a few years ago, glad that's behind me"*

These examples highlight the model's sensitivity to present-tense disclosures, often overlooking retrospective self-reports that may still indicate vulnerability.

Sarcasm and Irony: Posts using irony or self-deprecating humor to mask distress were frequently misclassified as non-depressive. Illustrative cases include:

- *"love waking up wanting to disappear lol"*
- *"mental health? never heard of her"*
- *"doing great sweetie (barely holding it together)"*
- *"can't wait to fake a smile through the whole week :)"*
- *"nothing like existential dread with your morning coffee"*

Such expressions rely on pragmatic and cultural cues that are challenging for models to decode, especially in youth-oriented platforms like Bluesky, where humor is often used as a coping strategy.

While these cases represent a subset of the dataset, they provide concrete illustrations of failure modes that standard classification metrics may obscure. These findings are revisited in the discussion to examine their linguistic and contextual implications more deeply.

5.5 Cross-Platform Linguistic Insights

A comparative linguistic analysis highlighted differences between platforms:

- **Bluesky:** More explicit, self-reflective depressive disclosures (e.g., "I'm empty tonight"), likely influenced by a younger, niche user base and less algorithmic curation.
- **Twitter:** More sarcasm, meme culture, and indirect expressions shaped by broader demographics and content virality.

These findings support the hypothesis that platform architecture and audience shape the linguistic manifestation of depressive content.

5.6 External Validation on Twitter

We tested the stacked ensemble on the Kaggle Twitter Depression dataset [25], obtaining:

- **Accuracy:** 90.0%
- **Precision:** 86.9%
- **Recall:** 95.1%
- **F1-score:** 90.8%

These results outperform the Kaggle leaderboard baseline (F1 = 0.87), demonstrating strong cross-platform generalization. Still, sarcasm-rich posts remained a challenge—highlighting a key avenue for future research involving explicit irony detection.

6 CONCLUSIONS AND FUTURE WORK

This study presented the first large-scale investigation of depression detection on **Bluesky**, a decentralized social media platform largely unexplored in mental health NLP research. We proposed a multi-stage detection pipeline that integrates classical NLP, metadata-based classifiers, transformer architectures, and a stacked ensemble—achieving strong performance while addressing the unique challenges posed by this novel environment.

Our experiments showed that transformer-based models consistently outperform traditional baselines. Metadata, when incorporated via stacking, provided complementary gains—albeit with potential for platform-specific bias. External validation on Twitter confirmed the **robust cross-platform generalization** of our approach, surpassing prior leaderboard benchmarks While core depression markers appear stable across platforms—enabling strong generalization—our qualitative results suggest that platform architecture and user context modulate the way depressive content is linguistically conveyed.

6.1 Key Contributions

This work advances the field of digital mental health and web-based NLP through four main contributions:

- Introduction of the **first expert-annotated dataset** of depressive posts from Bluesky, expanding mental health research to decentralized platforms.
- Empirical evidence that **linguistic markers of depression transfer** across platforms with distinct architectures and demographics.
- Demonstration that **metadata improves detection performance** but must be used cautiously due to its potential to encode non-linguistic patterns.
- Commitment to **open science and ethical AI**, through transparent practices and considerations for responsible deployment.

6.2 Limitations

Despite promising results, our study has some limitations that point to avenues for refinement:

- **Lack of emoji modeling.** Emojis—often used to convey affective nuance and contextual cues—were excluded during preprocessing. While this decision simplified the pipeline, it

may have limited the model's ability to capture implicit emotional signals. Future work could incorporate emoji-specific features or embeddings [38] to enhance affective sensitivity.
- **Binary classification scope.** Our approach focused on detecting the presence of depressive language, not estimating its clinical severity. Although appropriate for early risk identification, future extensions could explore ordinal or continuous labeling schemes grounded in clinical frameworks such as the PHQ-9 [27, 46].

Risks of Misuse. This research was conducted exclusively for academic purposes using anonymized data. Any real-world deployment of depression detection systems must be approached with caution. Such models should assist—but never replace—mental health professionals, and must incorporate rigorous privacy safeguards, mechanisms for informed consent, and ongoing ethical oversight. These concerns are particularly salient in decentralized environments, where user autonomy and trust are foundational.

6.3 Ethical Considerations

Ethics was a central pillar of this research. All data were anonymized and analyzed exclusively for academic use, in line with established guidelines for digital mental health studies. Future implementations should follow three principles:

- **Transparency:** Users must be clearly informed about data usage and model decisions.
- **Data minimization:** Only collect sensitive data when absolutely necessary.
- **Human-centered support:** Systems should guide empathetic interventions—not punitive or automated actions.

These guidelines ensure that technological advances empower, rather than endanger, vulnerable populations.

6.4 Future Directions

Our error analysis revealed recurring failure modes—such as somatic expressions, retrospective disclosures, and sarcastic remarks—that highlight important challenges for future research. Addressing these systematically may involve (i) distinguishing physical discomfort from psychological distress using external medical lexicons, (ii) employing temporal-aware architectures (e.g., LSTMs, transformers with time embeddings) to capture symptom trajectories, and (iii) incorporating sarcasm detection modules to improve pragmatic understanding in youth-oriented platforms.

This work lays the foundation for several promising research avenues:

- Integrate multimodal features (e.g., emojis, images, posting patterns) to capture deeper affective context.
- Extend classification beyond binary labels to estimate **severity of depressive symptoms**, using clinical frameworks like the PHQ-9 [27, 46].
- Investigate **domain-adaptive pretraining strategies** tailored to informal, irony-laden discourse common on social media platforms like Bluesky, to improve language model alignment and robustness.
- Explore **causal inference techniques** to disentangle genuine psychological signals from behavioral patterns induced by platform-specific interaction dynamics.
- Analyze depression expression across other decentralized networks to assess cross-platform variability.
- Explore advanced transformer architectures, such as **DeBERTa**, which have shown superior performance in general NLP tasks [24].

Based on the presented results, we reaffirm our central hypothesis: the decentralized structure and younger demographic of Bluesky influence how depression is linguistically expressed. While our findings show strong **cross-platform generalization**, they also reveal qualitative differences in how depressive content is articulated—suggesting that platform architecture and audience shape linguistic style more than the presence of depressive signals themselves.

Our findings strongly support this claim. Bluesky users tend to express depression more directly, while Twitter posts rely more on sarcasm and cultural references. These differences underscore the importance of context-aware models sensitive to platform architecture and user characteristics.

Ultimately, this research broadens the scope of mental health NLP by venturing into underexplored digital ecosystems, paving the way for future studies that balance technical innovation with ethical responsibility.

7 DISCUSSION

This study provides empirical and conceptual contributions to understanding how depression is expressed linguistically in decentralized social media platforms. It also highlights critical technical and ethical challenges for building robust classifiers in digital mental health.

7.1 Interpretation of Results

Our experiments confirm that transformer-based models (BERT, RoBERTa, MentalBERT) outperform both metadata-only and classical NLP baselines in capturing nuanced depressive cues. Fine-tuned BERT achieved the best single-model performance, while stacking yielded marginal gains by integrating engagement metadata.

Notably, MentalBERT underperformed—a result that extends beyond a simple domain mismatch. Although the model was pretrained on clinically focused mental health forums, its limited effectiveness on Bluesky empirically supports our central hypothesis: the linguistic expression of depression is modulated by platform-specific factors, including architecture and audience. The informal, irony-laden discourse prevalent on Bluesky—shaped by its decentralized design and younger user base—stands in contrast to the structured, support-seeking language typical of Reddit. Rather than contradicting cross-platform generalization, this finding highlights its boundaries: while core depressive markers may be stable, their stylistic and pragmatic manifestations vary significantly. As such, MentalBERT's performance underscores the need for models that are not only domain-informed, but also sensitive to the sociolinguistic context in which depression is expressed.

7.2 Platform-Specific Linguistic Patterns

Qualitative analysis revealed clear differences across platforms:

- **Bluesky:** More direct, self-revealing language (e.g., "I'm empty tonight"), reflecting a younger, niche demographic and less algorithmic curation.
- **Twitter:** More reliance on sarcasm, memes, and cultural references, shaped by its broader user base and viral content dynamics.

These contrasts suggest that while core depressive markers remain consistent, the way users express psychological distress may be influenced by age, moderation style, and platform architecture. This reinforces the value of demographically and context-aware models that can adapt to subtle linguistic and cultural variations across platforms.

7.3 Model Limitations and Error Sources

Building on the patterns observed in our qualitative error analysis, we interpret three core limitations of the model and their broader implications:

(1) **False positives from somatic expressions:** Posts describing physical symptoms (e.g., "nauseous," "can't sit up") were often misclassified as depressive. This suggests the model may overfit to surface-level expressions of suffering, without distinguishing between physical and emotional distress—an issue amplified in platforms like Bluesky, where self-disclosure is common and less moderated.
(2) **False negatives due to temporal context:** Posts in the past tense (e.g., "I had depression") were frequently downweighted, despite potential signs of ongoing vulnerability. The model struggled to interpret temporal cues, highlighting the need for temporal and discourse-aware modeling in mental health NLP.
(3) **Sarcasm and self-deprecating humor:** Ironic expressions such as "love waking up wanting to disappear lol" were often misclassified due to their humorous tone. These cases reveal the model's limitations in capturing pragmatic meaning, especially in informal or youth-driven contexts.

While challenges such as sarcasm, temporality, and multimodality have been widely recognized in broader NLP research [18], our results provide original empirical evidence on how these phenomena concretely affect false positives and false negatives—across both decentralized and centralized platforms. Specifically, sarcastic expressions and ambiguous temporal references were among the main sources of classification errors. These findings underscore the need for more adaptive and context-aware models capable of capturing the linguistic and pragmatic nuances of emerging online environments.

7.4 Implications for Digital Mental Health Research

This study offers two key implications for advancing mental health detection on social media:

- **Cross-platform transferability:** Our results show that core linguistic markers of depression—such as first-person singular pronouns, negative affect, and expressions of hopelessness—are remarkably stable across platforms with distinct architectures and user cultures. Notably, a model trained on Bluesky, a platform characterized by direct and emotionally explicit language, generalized effectively to Twitter, where depressive expression tends to be more indirect, sarcastic, or culturally coded. This suggests that while surface-level stylistic cues may vary, the underlying psychological and semantic signals of depression exhibit a form of universality that enables robust transfer learning across heterogeneous environments.
- **Contextualized modeling:** Despite the success of generalization, our error analysis and qualitative findings highlight the need for adaptive, context-aware models. Variations in demographic profiles, moderation norms, and communicative styles modulate how depression is expressed linguistically. One-size-fits-all models risk misinterpreting platform-specific discourse—such as irony, retrospective self-disclosure, or somatic metaphors—and may fail to detect at-risk individuals whose language deviates from expected norms. Future systems must balance generalizable core features with sensitivity to the sociotechnical context in which mental health cues are embedded.

7.5 Broader Research Significance

Our findings validate the feasibility and necessity of applying mental health NLP to decentralized networks. These environments offer unique windows into human expression but also demand heightened ethical and privacy sensitivity [10, 19].

By uniting technical innovation with ethical responsibility, this work sets a precedent for future research at the intersection of artificial intelligence, social platforms, and mental health.

Closing Perspective. Ultimately, this study demonstrates that decentralized networks like Bluesky provide a distinct and underutilized context for understanding mental health expression. By highlighting that linguistic styles of expressing depression differ—despite shared core markers—between centralized and decentralized platforms, this research opens promising avenues for developing ethically aligned, context-aware NLP tools that are sensitive to both universal markers and platform-specific discourse patterns.

REFERENCES

[1] Mohammed Faizan Ahmed, Mohammed Arsalan Ansari, Abdul Qadir, and Mohammed Jameel Hashmi. 2025. Depression Detection Using Machine Learning Techniques on X/Twitter Data. *International Journal of Innovative Technology and Exploring Engineering (IJITCE)* 13, 2s (2025), 225–232. https://doi.org/10.62647/IJITCE2025V13I2sPP225-232
[2] Tharindu Atapattu, Ushanthika Thayasivam, Katrina Falkner, Devanga Piyatissa, and Richi Nayak. 2022. EmoMent: Multilingual Mental Health Corpus with Fine-grained Emotion Annotations. In *Proceedings of the 29th International Conference on Computational Linguistics (COLING 2022)*. International Committee on Computational Linguistics, 6974–6987. https://aclanthology.org/2022.coling-1.609
[3] N. Bains and S. Abdijadid. 2023. Major Depressive Disorder. In *StatPearls [Internet]* (updated 2023 apr 10 ed.). StatPearls Publishing, Treasure Island (FL). Available from: https://www.ncbi.nlm.nih.gov/books/NBK559078/, Accessed: 2025-06-22.
[4] Anthony Bokolo and Chang Liu. 2024. Comparative performance of transformer-based models for detecting mental health conditions. *Electronics* 13, 20 (2024), 3980. https://www.mdpi.com/2079-9292/13/20/3980
[5] Ana-Maria Bucur, Adrian Cosma, Paolo Rosso, and Liviu P. Dinu. 2023. It's Just a Matter of Time: Detecting Depression with Time-Enriched Multimodal Transformers. (2023), 200–215. https://doi.org/10.1007/978-3-031-28244-7_13

[6] Fidel Cacheda, Daniel Fernandez, Francisco J. Novoa, and Victor Carneiro. 2019. Early detection of depression: social network analysis and random forest techniques. *Journal of Medical Internet Research* 21, 6 (2019), e12554. https://doi.org/10.2196/12554

[7] Lin Cao, Rui Wang, and Yan Zhou. 2025. A systematic review of machine learning approaches for depression detection on social media. *Journal of Big Data Science* 10, 2 (2025), 1–18. https://jbds.isdsa.org/jbds/article/view/110/115

[8] Yuchen Cao, Jianglai Dai, Zhongyan Wang, Yeyubei Zhang, Xiaorui Shen, Yunchong Liu, and Yexin Tian. 2025. Machine Learning Approaches for Mental Illness Detection on Social Media: A Systematic Review of Biases and Methodological Challenges. *Journal of Behavioral Data Science* 5, 1 (2025). https://doi.org/10.35566/jbds/caoyc

[9] Junyeop Cha, Seoyun Park, and Jin-ah Sim. 2022. A lexicon-based approach to examine depression detection in social media: the case of Twitter and university community. *Palgrave Communications* 9 (2022), 1–10. Issue 1. https://doi.org/10.1057/s41599-022-01313-2

[10] Stevie Chancellor and Munmun De Choudhury. 2020. Methods in predictive techniques for mental health status on social media: a critical review. *npj Digital Medicine* 3, 1 (2020), 1–11. https://doi.org/10.1038/s41746-020-0233-7

[11] Stevie Chancellor, Zhiyuan Lin, Elyse Goodman, Stephanie Zerwas, and Munmun De Choudhury. 2016. Quantifying and predicting mental illness severity in online pro-eating disorder communities. *Proceedings of the 19th ACM Conference on Computer-Supported Cooperative Work Social Computing (CSCW)* (2016), 1171–1184. https://doi.org/10.1145/2818048.2819973

[12] Mike Conway and Daniel O'Connor. 2016. Social media, big data, and mental health: current advances and ethical implications. *Current Opinion in Psychology* 9 (2016), 77–82. https://doi.org/10.1016/j.copsyc.2016.01.004

[13] L. D'Cruz, V. Dubey, and P. Thakur. 2023. Depression prediction from combined Reddit and Twitter data using machine learning. In *2023 2nd International Conference for Innovation in Technology (INOCON)*. IEEE, 1–5. https://doi.org/10.1109/INOCON57975.2023.10101174

[14] Munmun De Choudhury, Scott Counts, and Eric Horvitz. 2013. Social media as a measurement tool of depression in populations. *Proceedings of the 5th Annual ACM Web Science Conference* (2013), 47–56. https://doi.org/10.1145/2464464.2464480

[15] Munmun De Choudhury, Michael Gamon, Scott Counts, and Eric Horvitz. 2013. Predicting depression via social media. In *Proceedings of the International AAAI Conference on Weblogs and Social Media (ICWSM)*. 128–137. https://ojs.aaai.org/index.php/ICWSM/article/view/14432

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding. https://arxiv.org/abs/1810.04805. Accessed: 2025-06-22.

[17] Exploding Topics. 2025. Bluesky Users: Age, Gender, and Demographic Insights (2025). https://explodingtopics.com/blog/bluesky-users. Accessed: 2025-08-06.

[18] Shafkat Farabi, Tharindu Ranasinghe, Diptesh Kanojia, Yu Kong, and Marcos Zampieri. 2024. A Survey of Multimodal Sarcasm Detection. In *Proceedings of the 33rd International Joint Conference on Artificial Intelligence (IJCAI-24)*. IJCAI. https://www.ijcai.org/proceedings/2024/0887.pdf

[19] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2021. Datasheets for datasets. *Commun. ACM* 64, 12 (2021), 86–92. https://doi.org/10.1145/3458723

[20] Anne Germain and David J. Kupfer. 2008. Circadian rhythm disturbances in depression. *Human Psychopharmacology: Clinical and Experimental* 23, 7 (2008), 571–585. https://doi.org/10.1002/hup.964

[21] Declan Grabb, Max Lamparth, and Nina Vasan. 2024. Risks from Language Models for Automated Mental Healthcare: Ethics and Structure for Implementation. *arXiv preprint arXiv:2406.11852* (2024). https://arxiv.org/abs/2406.11852

[22] Sharath C. Guntuku, David B. Yaden, Margaret L. Kern, Lyle H. Ungar, and Johannes C. Eichstaedt. 2017. Detecting depression and mental illness on social media: an integrative review. *Current Opinion in Behavioral Sciences* 18 (2017), 43–49. https://doi.org/10.1016/j.cobeha.2017.07.005

[23] Ahmed Hasan and Ritesh Kumar. 2025. Benchmarking transformer and LSTM models for depression detection on Reddit. *arXiv preprint arXiv:2507.19511* (2025). https://arxiv.org/abs/2507.19511

[24] Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTa: Decoding-enhanced BERT with disentangled attention. *arXiv preprint arXiv:2006.03654* (2021). https://arxiv.org/abs/2006.03654

[25] InfamousCoder. 2022. Depression: Twitter Dataset + Feature Extraction. https://www.kaggle.com/datasets/infamouscoder/mental-health-social-media. Accessed: 2025-06-22.

[26] Shaoxiong Ji et al. 2021. MentalBERT: Publicly Available Pretrained Language Models for Mental Healthcare. https://arxiv.org/abs/2110.15621. Accessed: 2025-06-22.

[27] Mohsinul Kabir, Tasnim Ahmed, Md. Bakhtiar Hasan, et al. 2022. DEPTWEET: A Typology for Social Media Texts to Detect Depression Severities. *Computers in Human Behavior* 139 (2022), 107503. https://doi.org/10.1016/j.chb.2022.107503

[28] Marios Kerasiotis, Loukas Ilias, and Dimitris Askounis. 2024. Depression detection in social media posts using transformer-based models and auxiliary features. *Social Network Analysis and Mining* 14, 196 (2024). https://doi.org/10.1007/s13278-024-01360-4

[29] Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, and Romain Billot. 2021. Machine Learning and Natural Language Processing in Mental Health: Systematic Review. *Journal of Medical Internet Research* 23, 5 (2021). https://doi.org/10.2196/15708

[30] Samuel M. Lieblich, David J. Castle, Christos Pantelis, Malcolm Hopwood, Allan H. Young, and Ian P. Everall. 2015. High heterogeneity and low reliability in the diagnosis of major depression will impair the development of new drugs. *BJPsych Open* 1, 2 (2015), e5–e7. https://doi.org/10.1192/bjpo.bp.115.000786

[31] Danxia Liu, Xing Lin Feng, Farooq Ahmed, Muhammad Shahid, and Jing Guo. 2022. Detecting and Measuring Depression on Social Media Using a Machine Learning Approach: Systematic Review. *JMIR Mental Health* 9, 3 (2022), e27244. https://doi.org/10.2196/27244

[32] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL] https://arxiv.org/abs/1907.11692

[33] David E. Losada, Fabio Crestani, and Javier Parapar. 2017. Overview of eRisk: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2017) (Lecture Notes in Computer Science, Vol. 10456)*. Springer, 346–360. https://doi.org/10.1007/978-3-319-65813-1_30

[34] David E. Losada, Fabio Crestani, and Javier Parapar. 2018. Overview of eRisk: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF2018) (Lecture Notes in Computer Science, Vol. 14662)*. Springer, 343–361. https://doi.org/10.1007/978-3-319-98932-7_30

[35] David E. Losada, Fabio Crestani, and Javier Parapar. 2019. Overview of eRisk 2019: Early Risk Prediction on the Internet. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction (CLEF 2019) (Lecture Notes in Computer Science, Vol. 11696)*. Springer, 340–357. https://doi.org/10.1007/978-3-030-28577-7_27

[36] Horea Matei. 2025. Bluesky vs Twitter: Which Platform Is the Future of Social Media? https://planable.io/blog/bluesky-vs-twitter/.

[37] Danish Muzafar, Furqan Yaqub Khan, and Mubashir Qayoom. 2023. Machine Learning Algorithms for Depression Detection and Their Comparison. *arXiv preprint arXiv:2301.03222* (2023). https://arxiv.org/abs/2301.03222

[38] Petra Kralj Novak, Jasmina Smailović, Borut Sluban, and Igor Mozetič. 2015. Sentiment of emojis. In *PLOS ONE*, Vol. 10. e0144296. https://doi.org/10.1371/journal.pone.0144296

[39] Camila Pereira, Matheus Pavan, Sungwon Yoon, Ricelli Ramos, Pablo Costa, Laís Cavalheiro, and Ivandrê Paraboni. 2024. UstanceBR: a social media language resource for stance prediction. *arXiv preprint arXiv:2312.06374*. https://arxiv.org/abs/2312.06374

[40] Joelle Pineau, Philippe Vincent-Lamarre, Koustuv Sinha, Vincent Larivière, Alina Beygelzimer, Florence d'Alché Buc, Emily Fox, and Hugo Larochelle. 2021. Improving reproducibility in machine learning research (a report from the NeurIPS 2019 reproducibility program). *Journal of Machine Learning Research* 22, 164 (2021), 1–20. http://jmlr.org/papers/v22/20-303.html

[41] S. Rai, E.C. Stade, S. Giorgi, A. Francisco, L.H. Ungar, B. Curtis, and S.C. Guntuku. 2024. Key language markers of depression on social media depend on race. *Proceedings of the National Academy of Sciences of the United States of America* 121, 14 (2024), e2319837121. https://doi.org/10.1073/pnas.2319837121

[42] Darrel A. Regier, William E. Narrow, Diane E. Clarke, Helena C. Kraemer, Susan J. Kuramoto, Emily A. Kuhl, and David J. Kupfer. 2013. DSM-5 Field Trials in the United States and Canada, Part II: Test-Retest Reliability of Selected Categorical Diagnoses. *American Journal of Psychiatry* 170 (2013), 59–70. https://doi.org/10.1176/appi.ajp.2012.12070999

[43] N. Rüsch, M.C. Angermeyer, and P.W. Corrigan. 2005. Mental illness stigma: concepts, consequences, and initiatives to reduce stigma. *European Psychiatry* 20, 8 (Dec. 2005), 529–539. https://doi.org/10.1016/j.eurpsy.2005.04.004 Epub 2005 Sep 19.

[44] Amanda Silberling. 2024. Bluesky Is Now Open for Anyone to Join — TechCrunch. https://techcrunch.com/2024/02/06/bluesky-is-now-open-for-anyone-to-join/. Accessed: 2025-03-22.

[45] Fábio C. Souza, Rodrigo F. Nogueira, and Roberto A. Lotufo. 2020. BERTimbau: Pretrained BERT Models for Brazilian Portuguese. In *Proceedings of the 9th Brazilian Conference on Intelligent Systems (BRACIS 2020)*. IEEE / Springer, 403–417. https://doi.org/10.1007/978-3-030-61377-8_28

[46] Robert L. Spitzer, Kurt Kroenke, and Janet B.W. Williams. 2001. Validity of a Brief Depression Severity Measure (PHQ-9). *Journal of General Internal Medicine* 16, 9 (2001), 606–613. https://doi.org/10.1046/j.1525-1497.2001.016009606.x

[47] Michael M. Tadesse, Hongfei Lin, Bo Xu, and Liang Yang. 2019. Detection of Depression-Related Posts in Reddit Social Media Forum. *IEEE Access* 7, 44883–44893. https://doi.org/10.1109/ACCESS.2019.2909180

[48] Ilija Tavchioski, Marko Robnik-Šikonja, and Senja Pollak. 2023. Detection of depression on social networks using transformers and ensembles. https://arxiv.org/abs/2305.05325. arXiv:2305.05325.

[49] Raluca Nicoleta Trifu et al. 2024. Linguistic Markers for Major Depressive Disorder: A Cross-Sectional Study Using an Automated Procedure. *Frontiers in Psychology* 15 (March 2024). https://doi.org/10.3389/fpsyg.2024.1355734
[50] World Health Organization. 2023. Depressive Disorder (Depression). https://www.who.int/news-room/fact-sheets/detail/depression. Accessed: 2025-06-22.
[51] Michael Zimmer. 2010. "But the Data Is Already Public": On the Ethics of Research in Facebook. *Ethics and Information Technology* 12, 4 (2010), 313–325. https://doi.org/10.1007/s10676-010-9227-5