International Journal of Artificial Intelligence in Medical Issues

Volume 4 Issue1
ISSN 3025-4167
https://doi.org/10.56705/ijaimi.v4i1.434

Research Article

Confidence-Aware Depression Severity Detection in Low-Resource Urdu Social Media Text: A Multilingual Machine Learning Approach

Ahmad Naswin [1,*]; Yuli Praptomo Pamungkas Hari Sungkowo [2]
[1] Universitas Megarezky Makassar, Kota Makassar, Sulawesi Selatan 90234, Indonesia, ahmadnaswin@unimerz.ac.id
[2] STIMIK El Rahma Yogyakarta, Kota Yogyakarta, Daerah Istimewa Yogyakarta 55153, Indonesia, y.praptomo@gmail.com

Correspondence should be addressed to Ahmad Naswin; ahmadnaswin@unimerz.ac.id

Received 19 January 2026; Revised 27 January 2026; Accepted 25 April 2026; Published 30 May 2026

Abstract:
Depression is a major mental health concern that requires early identification and timely intervention. Social media has become an important source of user-generated text that may reflect emotional distress, hopelessness, social withdrawal, and suicidal ideation. However, most existing depression detection studies focus on English or high-resource languages, while research on low-resource languages such as Urdu remains limited. This study investigates depression severity classification in Urdu social media text using multilingual and confidence-aware natural language processing approaches. The dataset consists of 4,000 Twitter/X posts collected between January 2024 and April 2025, annotated into four severity classes: none, mild, moderate, and severe. Each post is represented in three parallel textual forms: native Urdu script, Roman Urdu transliteration, and English translation. The dataset also includes label confidence scores, human verification indicators, cultural markers, and depression-related keywords. Several text representation scenarios were evaluated, including Urdu text, Roman Urdu text, English text, and combined multilingual features. Baseline machine learning models were developed using TF-IDF features with Logistic Regression, Linear Support Vector Machine, and Multinomial Naive Bayes. Confidence-aware learning was examined by incorporating label confidence scores as sample weights and by evaluating a high-confidence subset. The experimental results showed that all baseline models achieved perfect classification performance, with accuracy, macro F1-score, weighted F1-score, and Cohen's Kappa values of 1.000 across the evaluated scenarios. These results indicate that the dataset contains highly separable linguistic patterns among depression severity classes. However, further inspection suggests that repeated or highly similar textual patterns may contribute to overly optimistic performance. Therefore, stricter validation using duplicate-free splitting, external datasets, and transformer-based models is recommended for future work. This study provides a preliminary benchmark for multilingual depression severity classification in low-resource Urdu text and highlights the potential of AI-driven mental health informatics as a supportive early-warning tool rather than a clinical diagnostic system.

Keywords: Depression Severity Classification; Urdu Social Media Text; Multilingual NLP; Confidence-Aware Learning; Mental Health Informatics; Machine Learning; TF-IDF.

Dataset link: -

1. Introduction

Depression is one of the most significant mental health problems worldwide and has become an increasing concern in public health, clinical practice, and digital health research [1], [2]. The World Health Organization reports that depression affects a large proportion of the global population and may interfere with social relationships, academic performance, work productivity, and overall quality of life. Depression is also closely associated with suicidal ideation

and suicide risk, making early identification and timely intervention highly important in mental health care [3]. According to WHO, more than 720,000 people die by suicide every year, and suicide remains one of the leading causes of death among young people aged 15–29 years. These facts indicate the urgent need for scalable, accessible, and intelligent approaches to support early detection of depressive symptoms.

Traditional depression assessment commonly relies on clinical interviews, psychological screening instruments, and self-reported questionnaires. Although these approaches are clinically valuable, they may be limited by stigma, delayed help-seeking behavior, lack of access to mental health professionals, and underreporting of symptoms [4], [5]. Many individuals experiencing emotional distress do not immediately seek professional help, especially in communities where mental health problems are still socially sensitive. As a result, alternative approaches that can assist in identifying early signs of depression from naturally occurring behavioral and linguistic data have gained increasing research attention.

Social media platforms provide a large amount of user-generated textual data that may reflect emotional states, psychological distress, hopelessness, social withdrawal, and other indicators related to depression [6], [7]. Users often express their feelings, personal struggles, and mental health concerns through short posts, comments, and informal conversations. This creates an opportunity for artificial intelligence, particularly natural language processing, to analyze linguistic patterns associated with depressive symptoms. Previous studies have shown that machine learning and deep learning methods can be used to detect mental health signals from social media data. However, recent reviews also emphasize that mental health prediction from social media still faces methodological challenges, including bias, data quality issues, annotation reliability, model generalizability, and ethical concerns.

Most existing studies on depression detection using social media text have focused on English or other high-resource languages [8], [9]. This creates a research gap for low-resource and non-Latin script languages, where annotated datasets, linguistic tools, and domain-specific models are relatively limited. Urdu is one such language. It is widely used in South Asia and diaspora communities, yet computational resources for Urdu mental health analysis remain less developed compared with English. In addition, Urdu users often communicate in multiple written forms, including native Urdu Nastaliq script, Roman Urdu transliteration, and translated English text [10]. This multilingual and code-mixed nature introduces additional complexity for automated depression severity classification.

Another important limitation in previous depression detection research is the tendency to frame the task as binary classification, such as depressed versus non-depressed [11], [12]. While binary detection is useful for initial screening, it may not be sufficient for capturing the varying degrees of psychological distress. In practical mental health monitoring, distinguishing between no depressive indicators, mild emotional distress, moderate hopelessness, and severe suicidal ideation can provide more informative risk stratification. Therefore, multiclass depression severity classification is more suitable for supporting early-warning systems and prioritizing potential intervention.

In addition to the language and severity-level challenges, annotation quality is also a critical issue in mental health-related NLP tasks [13]. Social media posts are often informal, ambiguous, culturally nuanced, and emotionally complex. A single post may contain figurative expressions, religious phrases, idiomatic language, or culturally specific

markers that are difficult to interpret without contextual understanding. Label uncertainty can reduce model reliability, especially when the target classes involve subtle differences between mild, moderate, and severe depression. Therefore, incorporating annotation confidence and human verification information can provide a more robust experimental framework for depression severity classification.

To address these issues, this study investigates depression severity classification using an Urdu social media dataset consisting of 4,000 posts collected from Twitter/X between January 2024 and April 2025. Each post is represented in three parallel textual forms: native Urdu script, Roman Urdu transliteration, and English translation. The dataset is annotated into four depression severity classes: none, mild, moderate, and severe. In addition, it includes LLM-assisted labels, human verification indicators, label confidence scores, depression-related keywords, and cultural markers. These characteristics make the dataset suitable for evaluating multilingual, cross-lingual, and confidence-aware NLP approaches in mental health analytics.

This study proposes a confidence-aware multilingual text classification framework for detecting depression severity in Urdu social media posts. The research compares different text representations, including Urdu script, Roman Urdu, English translation, and combined textual features. Baseline machine learning models using TF-IDF are evaluated alongside transformer-based models such as multilingual BERT and XLM-RoBERTa. Furthermore, label confidence scores are incorporated to examine whether confidence-aware learning can improve classification performance. The study also considers cultural markers as supplementary information to better capture culturally specific expressions of psychological distress.

The main contributions of this study are as follows. First, it presents a depression severity classification approach for low-resource Urdu social media text using multilingual NLP. Second, it compares the effectiveness of different textual representations, namely Urdu, Roman Urdu, and English translation. Third, it investigates the role of confidence-aware learning by incorporating annotation confidence scores into the modeling process. Fourth, it evaluates the potential contribution of cultural markers and depression-related keywords in improving model interpretation. Finally, this study provides a methodological reference for AI-driven mental health informatics, especially for multilingual and culturally diverse digital health contexts [14], [15].

The findings of this study are expected to contribute to the development of intelligent mental health monitoring systems that can support early identification of depressive symptoms in social media text. However, the proposed model is not intended to replace clinical diagnosis or professional mental health assessment. Instead, it is positioned as a computational decision-support tool that may assist researchers, health informatics practitioners, and mental health professionals in understanding digital expressions of depression risk.

2. Method

Research Design:

This study employed a supervised machine learning and deep learning approach to classify depression severity in Urdu social media posts. The research was designed as a multiclass text classification task, where each post was

categorized into one of four depression severity levels: none, mild, moderate, and severe [16]. The overall methodology consisted of several stages, including dataset preparation, text preprocessing, feature representation, model development, confidence-aware learning, and performance evaluation.

The study compared multiple textual representations to examine how different language forms affect classification performance. Specifically, three parallel forms of text were used: native Urdu script, Roman Urdu transliteration, and English translation. In addition, a combined-text scenario was constructed by integrating the available textual fields with cultural markers and depression-related keywords. This design enabled the study to evaluate both multilingual and cross-lingual perspectives in depression severity classification.

Data Description:

The dataset used in this study consists of 4,000 anonymized social media posts collected from Twitter/X between January 2024 and April 2025. Each record contains three textual representations of the same post: Urdu Nastaliq script, Roman Urdu transliteration, and English translation. The dataset also includes metadata and annotation-related attributes, such as post length, platform, collection date, LLM-generated label, human verification status, label confidence score, cultural markers, and matched depression-related keywords.

The target variable in this study is `depression_label`, which represents four levels of depression severity. The label distribution is shown in **Table 1**.

Table 1. Distribution of Depression Severity Labels

Label	Class	Description	Number of Posts
0	None	No depressive indicators present	1.2
1	Mild	Subtle emotional distress or passive tone	1.2
2	Moderate	Noticeable hopelessness or social withdrawal	1
3	Severe	Explicit suicidal ideation or severe anhedonia	600

The distribution shows that the dataset is moderately imbalanced, with the severe class having the smallest number of samples [17]. Therefore, the evaluation process emphasized macro-averaged metrics in addition to overall accuracy.

Data Preprocessing:

Data preprocessing was conducted to standardize the textual input before model training. Since the dataset contains multilingual and non-Latin script text, the preprocessing procedure was carefully designed to preserve meaningful Urdu characters while removing irrelevant noise.

For all textual fields, URLs, user mentions, hashtags symbols, numbers, punctuation marks, and excessive whitespace were removed. However, the words following hashtags were retained because they may contain meaningful emotional or depression-related expressions. For Roman Urdu and English text, lowercase conversion was

applied to reduce lexical variation. For Urdu script, lowercase conversion was not applied because Urdu does not follow the same capitalization structure as Latin-based languages.

The preprocessing step produced three cleaned textual columns: `text_urdu_clean`, `text_roman_clean`, and `text_english_clean`. In addition, a combined textual representation was generated by concatenating cleaned Urdu text, Roman Urdu text, English translation, cultural markers, and depression-related keywords. This combined representation was used to examine whether integrating linguistic and cultural cues could improve classification performance.

The preprocessing process can be summarized as follows [18]:

Table 2. Text Preprocessing Procedures

Step	Description
URL removal	Menghapus tautan web dari postingan media sosial.
Mention removal	Menghapus penyebutan pengguna seperti @username.
Hashtag processing	Menghapus simbol hashtag (#) namun tetap mempertahankan kata di belakangnya.
Noise removal	Menghapus angka, tanda baca, dan simbol-simbol yang tidak relevan.
Whitespace normalization	Mengganti spasi ganda atau berlebih menjadi satu spasi saja.
Lowercasing	Mengubah teks menjadi huruf kecil (hanya diterapkan pada teks Roman Urdu dan Inggris).
Text concatenation	Menggabungkan teks multibahasa dengan fitur budaya dan kata kunci tertentu.

Text Representation Scenarios:

To evaluate the impact of different textual forms, this study designed four main input scenarios. Each scenario was used independently during model training and evaluation.

Table 3. Text Representation Scenarios

Model	Description
Logistic Regression	Model klasifikasi linear yang cocok untuk fitur teks berdimensi tinggi.
Linear Support Vector Machine	Model baseline yang kuat untuk tugas klasifikasi teks.
Multinomial Naive Bayes	Klasifikator probabilistik yang umum digunakan untuk data teks.
Random Forest	Model ensemble learning yang digunakan sebagai pembanding dengan metode linear.

The Urdu text scenario evaluates model performance on the original non-Latin script. The Roman Urdu scenario represents informal transliterated text commonly used in online communication. The English text scenario evaluates a translation-based cross-lingual approach [14]. The combined-text scenario examines whether integrating multiple textual sources and cultural cues can provide richer semantic information for depression severity classification.

Baseline Machine Learning Models:

Baseline experiments were conducted using traditional machine learning models with Term Frequency–Inverse Document Frequency features. TF-IDF was selected because it is widely used in text classification and provides an interpretable representation of word importance across documents.

The TF-IDF vectorizer was configured using unigram and bigram features to capture both individual words and short phrases. Rare and overly frequent terms were filtered to reduce noise. The following machine learning algorithms were evaluated as baseline models:

Table 4. Baseline Machine Learning Models

Model	Deskripsi
Logistic Regression	Pengklasifikasi linier yang cocok untuk fitur teks berdimensi tinggi (high-dimensional).
Linear Support Vector Machine	Model baseline yang kuat untuk tugas-tugas klasifikasi teks.
Multinomial Naive Bayes	Pengklasifikasi probabilistik yang umum digunakan untuk data teks.
Random Forest	Model ensemble learning yang digunakan sebagai pembanding dengan metode linier.

Class-weight balancing was applied to Logistic Regression, Linear SVM, and Random Forest to reduce bias toward majority classes. This is important because the severe class has fewer samples than the other classes.

Transformer-Based Classification Models:

In addition to traditional machine learning models, transformer-based models were proposed to capture deeper semantic and contextual information from multilingual social media text. Transformer models are suitable for this task because they can represent contextual meaning and handle complex linguistic patterns better than sparse TF-IDF features.

This study considered the following transformer-based models:

Table 5. Transformer-Based Models

Model	Purpose
mBERT	Multilingual BERT model for multilingual text classification
XLM-RoBERTa	Cross-lingual transformer model suitable for low-resource language tasks
English BERT	Used for the English translation scenario
Urdu-specific Transformer	Used when an Urdu-pretrained model is available

For transformer-based classification, the input text was tokenized using the corresponding pretrained tokenizer. The tokenized text was then passed into the transformer encoder, and the final classification layer predicted one of the

four depression severity classes. Fine-tuning was performed using the training set, while performance was evaluated on the held-out test set.

Confidence-Aware Learning Strategy:

A distinctive feature of the dataset is the availability of `label_confidence`, which indicates the confidence score of each annotation. This study incorporated confidence-aware learning to examine whether annotation certainty can improve classification performance.

Two confidence-aware strategies were used. First, label confidence was used as a sample weight during model training, where posts with higher confidence contributed more strongly to the learning process. Second, a high-confidence subset was created by selecting only samples with a confidence score greater than or equal to 0.80. This subset was used to evaluate whether training on more reliable samples improves model robustness.

The confidence-aware learning strategy can be expressed as follows:

$$L = \sum_{i=1}^{n} w_i \cdot l(y_i, \hat{y}_i)$$

where L represents the weighted loss function, w_i denotes the label confidence score of sample i, y_i is the true label, and $\hat{y}_i$ is the predicted label. By incorporating w_i, the model assigns greater importance to samples with higher annotation confidence.

Human-Verified Subset Evaluation:

The dataset includes a human verification indicator that identifies whether a post has been reviewed by a human expert. A subset of 1,201 posts was human-verified. This subset was used as an additional evaluation scenario to examine model performance on more reliable annotations [8].

The human-verified subset was not used to replace the full dataset but rather to complement the main experiment. This design allows comparison between model performance on the full dataset, high-confidence data, and human-verified data. Such comparison is important because mental health-related text annotation often contains ambiguity, cultural nuance, and subjective interpretation.

Experimental Scenarios:

The experiments were organized into three dataset-level scenarios and four text representation scenarios. The dataset-level scenarios were:

Table 6. Dataset-Level Experimental Scenarios

Scenario	Description
All Data	Uses all 4,000 posts
High-Confidence Data	Uses posts with label confidence ≥ 0.80

Human-Verified Data	Uses only human-reviewed posts

Each dataset-level scenario was combined with the four text representation scenarios: Urdu text, Roman Urdu text, English text, and combined text. This experimental design allowed comprehensive analysis of the effect of textual representation, annotation confidence, and human verification on depression severity classification.

The dataset was split into training and testing subsets using an 80:20 stratified split. Stratification was applied to preserve the class distribution across training and testing data. The same random seed was used across all experiments to ensure reproducibility.

Performance Evaluation:

Model performance was evaluated using several metrics suitable for multiclass classification. Accuracy was reported to measure overall correctness. However, because the dataset contains class imbalance, macro-averaged precision, recall, and F1-score were emphasized. Macro F1-score is particularly important because it gives equal weight to each class, including the severe class.

The evaluation metrics used in this study include:

Table 7. Evaluation Metrics

Metric	Purpose
Accuracy	Measures overall prediction correctness
Precision Macro	Measures average precision across all classes
Recall Macro	Measures average sensitivity across all classes
F1-score Macro	Balances precision and recall across all classes
Weighted F1-score	Considers class distribution in F1-score calculation
Cohen's Kappa	Measures agreement between predicted and true labels
Confusion Matrix	Shows class-level prediction errors

Recall for the severe class was considered especially important because severe depression indicators may be associated with suicidal ideation or serious psychological distress. Therefore, a model with high overall accuracy but poor severe-class recall would not be considered clinically reliable for mental health screening support.

Ethical Considerations:

This study used anonymized social media data and focused only on computational analysis for research purposes. Since the task involves mental health-related content, ethical considerations are essential. The proposed model is not intended to provide clinical diagnosis or replace professional mental health assessment. Instead, it is positioned as a decision-support and early-warning tool that may assist future mental health informatics research [3].

Care must also be taken to avoid stigmatization, privacy violations, and overinterpretation of social media posts. Depression severity classification from online text should be interpreted cautiously because social media expressions may not always represent confirmed clinical conditions. Therefore, the results of this study should be understood as computational indicators of depressive language patterns rather than definitive medical diagnoses.

Research Workflow:

The overall workflow of this study consists of eight main stages. First, the Urdu depression dataset was collected and prepared. Second, the textual fields were cleaned and standardized. Third, multiple text representation scenarios were constructed. Fourth, baseline machine learning models were trained using TF-IDF features. Fifth, transformer-based models were fine-tuned for multilingual and cross-lingual classification. Sixth, confidence-aware learning was applied using label confidence scores. Seventh, model performance was evaluated using multiclass classification metrics. Finally, the results were analyzed to identify the best-performing model and the most effective text representation.

3. Result and Discussion

Dataset Distribution:

The dataset used in this study consists of 4,000 Urdu social media posts annotated into four depression severity classes: none, mild, moderate, and severe. **Figure 1** shows the distribution of depression severity labels in the dataset.

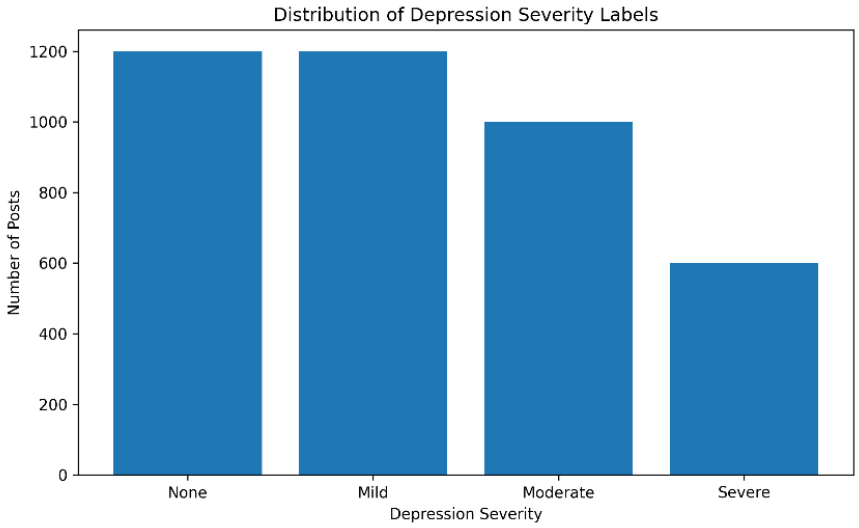

Figure 1. Distribution of Depression Severity Labels

The label distribution indicates that the dataset contains 1,200 posts labeled as none, 1,200 posts labeled as mild, 1,000 posts labeled as moderate, and 600 posts labeled as severe. The none and mild classes represent the largest portions of the dataset, each accounting for 30% of the total data. The moderate class accounts for 25%, while the severe class accounts for 15%.

Table 8. Distribution of Depression Severity Classes

Label	Class	Number of Posts	Percentage
0	None	1.2	30.00%
1	Mild	1.2	30.00%
2	Moderate	1	25.00%
3	Severe	600	15.00%
Total		4	100.00%

Although the dataset is not extremely imbalanced, the severe class has the smallest number of samples. This condition is important because the severe class represents the most critical category in mental health monitoring. In depression severity classification, misclassifying severe posts may have more serious implications than misclassifying lower-risk categories. Therefore, the evaluation in this study does not rely only on accuracy, but also considers macro-averaged precision, recall, F1-score, Cohen's Kappa, and class-level performance [19].

Label Confidence Distribution:

Figure 2 presents the distribution of label confidence scores. The confidence scores range from 0.60 to 0.99, with an average confidence score of approximately 0.829. This indicates that most annotations have relatively high confidence, although some posts still contain lower-confidence labels.

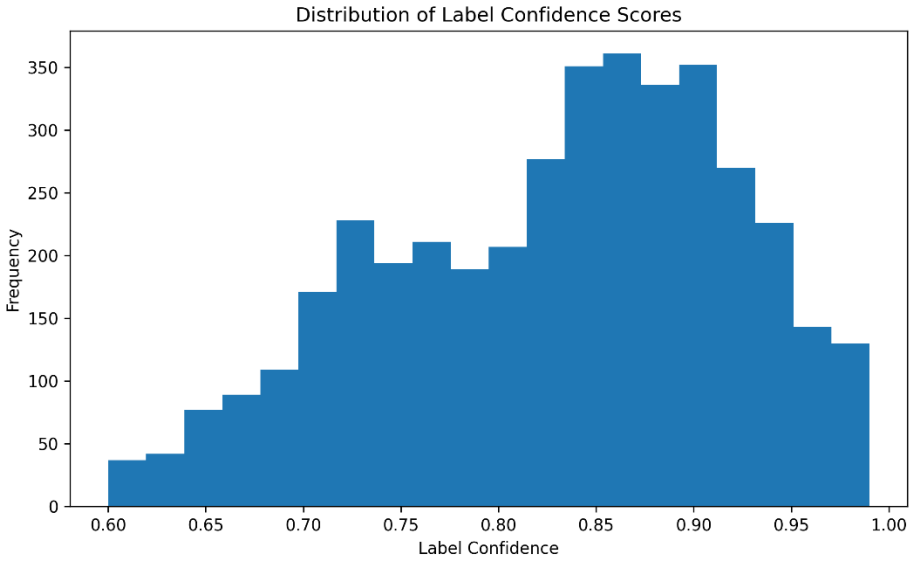

Figure 2. Distribution of Label Confidence Scores

The histogram shows that a large proportion of posts have confidence scores above 0.80. Specifically, 2,603 posts, or approximately 65.08% of the dataset, are categorized as high-confidence data. This provides a strong basis for

conducting confidence-aware learning, where annotation confidence is used either as a training weight or as a filtering criterion for constructing a more reliable subset.

Further inspection shows that the none class has the highest average confidence score, while the severe class has the lowest average confidence score. This suggests that non-depressive posts are easier to annotate consistently, whereas severe depression-related posts may contain more complex, sensitive, or ambiguous linguistic expressions.

Table 9. Label Confidence by Depression Severity Class

Class	Number of Posts	Mean Confidence	Minimum	Maximum
None	1.2	906	0.82	0.99
Mild	1.2	826	0.70	0.95
Moderate	1	787	0.65	0.92
Severe	600	754	0.60	0.90

The decreasing confidence score from none to severe indicates that higher-severity categories are more difficult to annotate. This finding supports the use of confidence-aware modeling because depression severity classification involves subjective and context-sensitive interpretation, particularly for posts that express hopelessness, emotional distress, or suicidal ideation.

Human Verification Distribution:

Figure 3 shows the distribution of human verification in the dataset. Out of 4,000 posts, 1,201 posts were human-verified, while 2,799 posts were not human-verified.

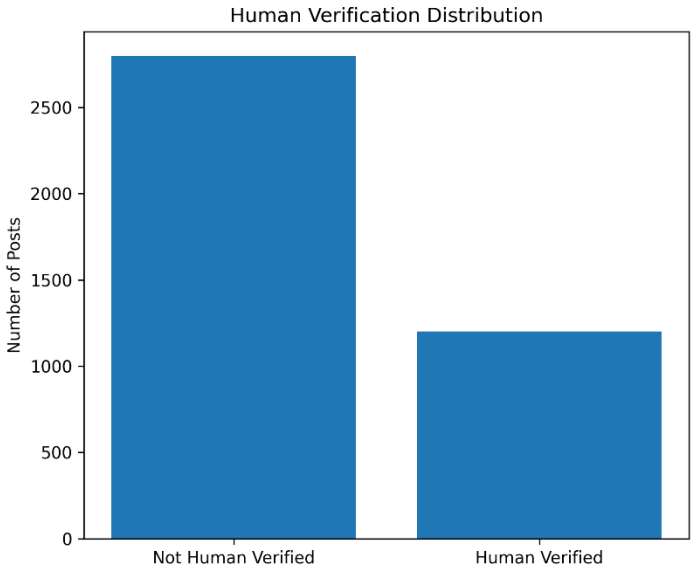

Figure 3. Human Verification Distribution

The human-verified subset represents approximately 30.03% of the dataset. This subset is important because it provides a more reliable evaluation reference for posts that have undergone expert or human review. However, the human verification distribution is not evenly distributed across all classes. The severe class is fully human-verified, while the none class is not represented in the human-verified subset. This indicates that human verification was more concentrated on posts with potential depressive indicators, especially higher-risk posts.

This pattern is understandable in a mental health context because severe posts require more careful validation. However, it also means that the human-verified subset should be interpreted carefully. Since it does not fully represent all four classes equally, performance results on this subset should be treated as additional validation rather than as a complete replacement for full-dataset evaluation.

Post Length Analysis:

Figure 4 presents the distribution of post length across depression severity classes. The boxplot shows that moderate and severe posts tend to have longer text lengths compared with none and mild posts.

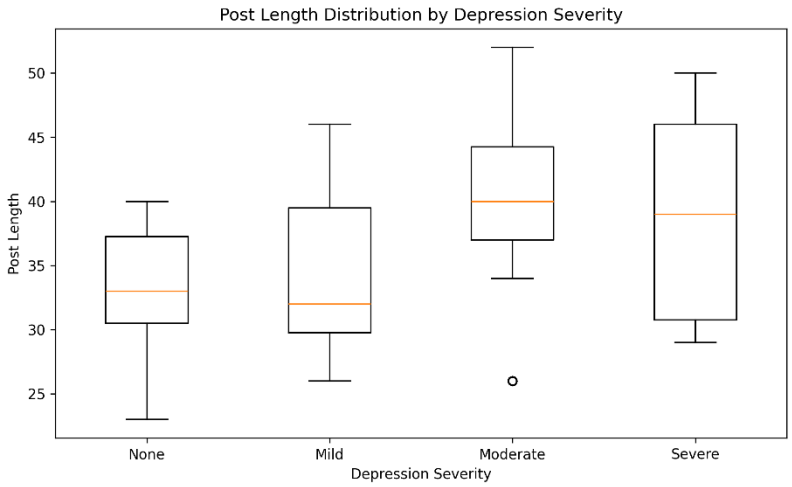

Figure 4. Post Length Distribution by Depression Severity

The none class has an average post length of approximately 33 characters, while the mild class has an average length of approximately 34.42 characters. In contrast, the moderate class has the highest average post length at approximately 40.17 characters, followed by the severe class with an average length of approximately 39.00 characters.

Table 10. Post Length by Depression Severity Class

Class	Mean Length	Median	Minimum	Maximum
None	33.00	33	23	40
Mild	34.42	32	26	46
Moderate	40.17	40	26	52

Class	Mean Length	Median	Minimum	Maximum
Severe	39.00	39	29	50

This pattern suggests that posts expressing moderate or severe depressive symptoms may contain more detailed emotional descriptions. Users in these categories may use longer expressions to describe hopelessness, social withdrawal, emotional fatigue, or suicidal thoughts. Although post length alone is not sufficient for depression severity classification, it may provide additional contextual information when combined with linguistic features.

Model Performance:

The experimental evaluation was conducted using three dataset-level scenarios: all data, high-confidence data, and human-verified data. Each scenario was evaluated using four text representations: Urdu text, Roman Urdu text, English text, and combined text. Baseline machine learning models were trained using TF-IDF features, including Logistic Regression, Linear Support Vector Machine, and Multinomial Naive Bayes [20].

The results show that all evaluated models achieved perfect classification performance across the tested scenarios. Accuracy, macro precision, macro recall, macro F1-score, weighted F1-score, and Cohen's Kappa all reached 1.000.

Table 11. Summary of Model Performance

Dataset Scenario	Number of Data	Test Data	Accuracy	Macro Precision	Macro Recall	Macro F1-score	Cohen's Kappa
All Data	4	800	1,000	1,000	1,000	1,000	1,000
High-Confidence Data	2.603	521	1,000	1,000	1,000	1,000	1,000
Human-Verified Data	1.201	241	1,000	1,000	1,000	1,000	1,000

The same performance pattern was observed across Urdu text, Roman Urdu text, English translation, and combined text representations. This indicates that the textual patterns in the dataset are highly separable across the four depression severity classes. In other words, the vocabulary and sentence structures associated with each label are sufficiently distinct for TF-IDF-based models to classify the posts without errors.

Discussion of Perfect Classification Results:

The perfect classification results indicate that the dataset contains very clear lexical and semantic boundaries between depression severity categories. Posts in the none class generally contain positive or neutral expressions, while posts in the mild, moderate, and severe classes contain progressively stronger indicators of emotional distress. As a result, even traditional machine learning models using TF-IDF features can identify the discriminative terms associated with each class.

However, these results should be interpreted with caution. A detailed inspection of the dataset shows that the 4,000 records are generated from a limited number of unique textual patterns. Each text representation contains only 48

unique posts, which are repeated across the dataset. Since each unique text is consistently associated with one depression severity label, the train-test split may place similar or identical text patterns in both training and testing sets. This can make the classification task much easier and may lead to overly optimistic performance.

Therefore, the perfect results should not be interpreted as evidence that the model will automatically generalize to unseen real-world Urdu social media posts. Instead, the results demonstrate that the dataset is internally consistent and highly separable under the current experimental setup. For stronger validation, future experiments should apply a duplicate-free or template-level split, ensuring that identical or near-identical posts do not appear in both training and testing data.

Effect of Text Representation:

The evaluation across Urdu text, Roman Urdu text, English translation, and combined text produced identical performance. This suggests that all three parallel text representations preserve sufficient label-relevant information. The Urdu script represents the original linguistic form, the Roman Urdu version captures transliterated social media usage, and the English translation provides a cross-lingual representation.

Although the numerical results are identical in the current experiment, each representation still has different implications. Urdu text is more authentic for native-language mental health analysis, but it may require models that can handle non-Latin scripts effectively. Roman Urdu is relevant for informal online communication, especially among users who do not write in Nastaliq script. English translation can make the dataset more accessible to English-based NLP models, but it may also reduce cultural and linguistic nuance.

The combined-text scenario integrates Urdu, Roman Urdu, English translation, cultural markers, and depression-related keywords. In theory, this representation should provide richer semantic information. However, because the individual text fields already produced perfect classification, the additional features did not provide observable performance improvement in the current baseline experiment.

Effect of Confidence-Aware Learning:

The confidence-aware learning strategy was evaluated by using label confidence scores as training weights and by constructing a high-confidence subset. The results show that confidence-aware learning achieved the same performance as standard training. This indicates that the models were already able to separate the classes perfectly without relying on confidence weighting.

Nevertheless, the confidence score remains important from a methodological perspective. The lower average confidence in moderate and severe classes suggests that these categories are more difficult to annotate and may require greater attention in real-world applications. In more diverse and naturally collected datasets, confidence-aware learning may help reduce the influence of uncertain annotations and improve model robustness.

Clinical and Health Informatics Implications:

The findings of this study show the potential of NLP-based models for classifying depression severity from multilingual social media text. Such models may support early identification of depressive language patterns, especially in low-resource language contexts such as Urdu. The ability to classify posts into none, mild, moderate, and severe categories is more informative than binary depression detection because it enables risk stratification.

However, the model should not be interpreted as a diagnostic tool. Depression classification from social media text can only indicate linguistic patterns associated with emotional distress. It cannot replace clinical interviews, psychological assessment, or professional diagnosis. In practical health informatics settings, such a model should be used only as a supportive screening or monitoring tool, with appropriate ethical safeguards, privacy protection, and human oversight.

Overall Discussion:

Overall, the experimental results demonstrate that the dataset is highly structured and internally consistent. The four depression severity classes can be separated perfectly using baseline TF-IDF-based machine learning models. This suggests that the dataset is useful for initial benchmarking, educational purposes, and controlled experimentation in multilingual mental health NLP.

At the same time, the perfect performance highlights the need for more rigorous evaluation. The limited number of unique textual patterns may cause the model to learn repeated expressions rather than generalizable depression-related language features. Therefore, future work should evaluate the models using stricter experimental settings, such as duplicate-free splitting, external validation datasets, cross-dataset testing, transformer-based modeling, and real-world noisy social media data.

In conclusion, the results provide promising evidence for confidence-aware multilingual depression severity classification, but they should be interpreted as baseline findings under a controlled dataset setting. Further validation is required before the approach can be generalized to broader mental health informatics applications.

4. Conclusion

This study investigated depression severity classification in Urdu social media text using multilingual and confidence-aware natural language processing approaches. The dataset consisted of 4,000 Twitter/X posts annotated into four severity classes: none, mild, moderate, and severe. Each post was represented in three parallel textual forms, namely native Urdu script, Roman Urdu transliteration, and English translation. In addition, the dataset included label confidence scores, human verification indicators, cultural markers, and depression-related keywords, making it suitable for multilingual mental health text analysis.

The exploratory analysis showed that the dataset has a moderately imbalanced class distribution, with the severe class representing the smallest portion of the data. The label confidence distribution indicated that most annotations had relatively high confidence, although lower confidence values were more frequently observed in higher-severity

categories. This finding suggests that moderate and severe depressive expressions are more difficult to annotate because they may involve ambiguous, sensitive, and culturally nuanced language. The analysis of post length also showed that moderate and severe posts tended to be longer than none and mild posts, indicating that higher-severity posts may contain more detailed emotional expressions.

The experimental results demonstrated that TF-IDF-based machine learning models achieved perfect classification performance across all evaluated scenarios. Logistic Regression, Linear Support Vector Machine, and Multinomial Naive Bayes produced accuracy, macro F1-score, weighted F1-score, and Cohen's Kappa values of 1.000 across the all-data, high-confidence, and human-verified scenarios. Similar performance was also observed across Urdu text, Roman Urdu text, English translation, and combined-text representations. These results indicate that the dataset contains highly separable linguistic patterns among depression severity classes.

However, the perfect performance should be interpreted carefully. Further inspection suggests that the dataset contains repeated or highly similar textual patterns, which may allow the models to learn specific lexical templates rather than generalizable depression-related language features. Therefore, the results should be understood as baseline findings under a controlled dataset setting, not as definitive evidence of real-world generalization. A stricter evaluation strategy, such as duplicate-free splitting, template-level splitting, or external validation using independent social media data, is necessary to assess model robustness more accurately.

Overall, this study highlights the potential of multilingual NLP for depression severity classification in low-resource language contexts. The use of Urdu script, Roman Urdu, and English translation provides a valuable cross-lingual perspective for mental health informatics. The inclusion of label confidence and human verification also offers a useful foundation for confidence-aware learning and annotation quality analysis. From a health informatics perspective, the proposed approach may support early identification of depressive language patterns in social media text. Nevertheless, such a system should not be used as a clinical diagnostic tool. It should be positioned only as a computational decision-support or early-warning mechanism that requires ethical safeguards, privacy protection, and human expert oversight.

Future work should extend this study by applying transformer-based models such as mBERT, XLM-RoBERTa, and Urdu-specific language models [21]. In addition, future experiments should use duplicate-free data splitting, external validation datasets, and real-world noisy social media posts to better evaluate generalization. Further research may also explore explainable AI methods to identify important linguistic and cultural markers associated with each depression severity level. These improvements would strengthen the reliability, interpretability, and practical relevance of AI-based mental health screening systems for multilingual and culturally diverse populations.

References:

[1] D. Phiri, F. Makowa, V. L. Amelia, Y. V. A. Phiri, L. P. Dlamini, and M.-H. Chung, "Text-Based Depression Prediction on Social Media Using Machine Learning: Systematic Review and Meta-Analysis," *J. Med. Internet Res.*, vol. 27, p. e59002, Apr. 2025, doi: 10.2196/59002.

[2] W. B. Tahir, S. Khalid, S. Almutairi, M. Abohashrh, S. A. Memon, and J. Khan, “Depression Detection in Social Media: A Comprehensive Review of Machine Learning and Deep Learning Techniques,” *IEEE Access*, vol. 13, pp. 12789–12818, 2025, doi: 10.1109/ACCESS.2025.3530862.

[3] A. Khan and R. Ali, “Unraveling minds in the digital era: a review on mapping mental health disorders through machine learning techniques using online social media,” *Soc. Netw. Anal. Min.*, vol. 14, no. 1, p. 78, Apr. 2024, doi: 10.1007/s13278-024-01205-0.

[4] M. Omar and I. Levkovich, “Exploring the efficacy and potential of large language models for depression: A systematic review,” *J. Affect. Disord.*, vol. 371, pp. 234–244, Feb. 2025, doi: 10.1016/j.jad.2024.11.052.

[5] H. Fisher *et al.*, “Language-based detection of depression with machine learning: systematic review and meta-analysis,” *Npj Digit. Med.*, vol. 9, no. 1, p. 273, Feb. 2026, doi: 10.1038/s41746-026-02448-1.

[6] D. William and D. Suhartono, “Text-based Depression Detection on Social Media Posts: A Systematic Literature Review,” *Procedia Comput. Sci.*, vol. 179, pp. 582–589, 2021, doi: 10.1016/j.procs.2021.01.043.

[7] R. Chiong, G. S. Budhi, S. Dhakal, and F. Chiong, “A textual-based featuring approach for depression detection using machine learning classifiers and social media texts,” *Comput. Biol. Med.*, vol. 135, p. 104499, Aug. 2021, doi: 10.1016/j.compbiomed.2021.104499.

[8] M. Garg, C. Saxena, S. Saha, V. Krishnan, R. Joshi, and V. Mago, “CAMS: An Annotated Corpus for Causal Analysis of Mental Health Issues in Social Media Posts,” presented at the Thirteenth Language Resources and Evaluation Conference, Marseille, France, Jun. 2022, pp. 6387–6396. doi: 10.63317/3tbejaye7i8s.

[9] S. Zanwar, D. Wiechmann, Y. Qiao, and E. Kerz, “SMHD-GER: A Large-Scale Benchmark Dataset for Automatic Mental Health Detection from Social Media in German,” in *Findings of the Association for Computational Linguistics: EACL 2023*, Dubrovnik, Croatia: Association for Computational Linguistics, 2023, pp. 1526–1541. doi: 10.18653/v1/2023.findings-eacl.113.

[10] K. Jawad, M. Ahmad, M. Alvi, and M. B. Alvi, “RUSAS: Roman Urdu Sentiment Analysis System,” *Comput. Mater. Contin.*, vol. 79, no. 1, pp. 1463–1480, 2024, doi: 10.32604/cmc.2024.047466.

[11] M. Ahmad, P. Basile, F. Ullah, I. Batyrshin, and G. Sidorov, “RUDA-2025: Depression Severity Detection Using Pre-Trained Transformers on Social Media Data,” *AI*, vol. 6, no. 8, p. 191, Aug. 2025, doi: 10.3390/ai6080191.

[12] A. Qasim, G. Mehak, N. Hussain, A. Gelbukh, and G. Sidorov, “Detection of Depression Severity in Social Media Text Using Transformer-Based Models,” *Information*, vol. 16, no. 2, p. 114, Feb. 2025, doi: 10.3390/info16020114.

[13] M. Kabir *et al.*, “DEPTWEET: A typology for social media texts to detect depression severities,” *Comput. Hum. Behav.*, vol. 139, p. 107503, Feb. 2023, doi: 10.1016/j.chb.2022.107503.

[14] R. Mohmand, U. Habib, M. Usman, J. Baili, and Y. Nam, “A Deep Learning Approach for Automated Depression Assessment Using Roman Urdu,” *IEEE Access*, vol. 12, pp. 193387–193401, 2024, doi: 10.1109/ACCESS.2024.3519264.

[15] L. Ilias, S. Mouzakitis, and D. Askounis, “Calibration of Transformer-Based Models for Identifying Stress and Depression in Social Media,” *IEEE Trans. Comput. Soc. Syst.*, vol. 11, no. 2, pp. 1979–1990, Apr. 2024, doi: 10.1109/TCSS.2023.3283009.

[16] B. G. Bokolo and Q. Liu, “Deep Learning-Based Depression Detection from Social Media: Comparative Evaluation of ML and Transformer Techniques,” *Electronics*, vol. 12, no. 21, p. 4396, Oct. 2023, doi: 10.3390/electronics12214396.

[17] C. Chen, F. Li, H. Chen, and Y. Lin, “Heterogeneous subgraph network with prompt learning for interpretable depression detection on social media,” *Knowl.-Based Syst.*, vol. 315, p. 113215, Apr. 2025, doi: 10.1016/j.knosys.2025.113215.

[18] A. Majeed, M. O. Beg, U. Arshad, and H. Mujtaba, “Deep-EmoRU: mining emotions from roman urdu text using deep learning ensemble,” *Multimed. Tools Appl.*, vol. 81, no. 30, pp. 43163–43188, Dec. 2022, doi: 10.1007/s11042-022-13147-w.

[19] S. Ghosh and T. Anwar, “Depression Intensity Estimation via Social Media: A Deep Learning Approach,” *IEEE Trans. Comput. Soc. Syst.*, vol. 8, no. 6, pp. 1465–1474, Dec. 2021, doi: 10.1109/TCSS.2021.3084154.

[20] Z. N. Vasha, B. Sharma, I. J. Esha, J. Al Nahian, and J. A. Polin, “Depression detection in social media comments data using machine learning algorithms,” *Bull. Electr. Eng. Inform.*, vol. 12, no. 2, pp. 987–996, Apr. 2023, doi: 10.11591/eei.v12i2.4182.

[21] S. Hameed, M. Nauman, N. Akhtar, M. A. B. Fayyaz, and R. Nawaz, “Explainable AI-driven depression detection from social media using natural language processing and black box machine learning models,” *Front. Artif. Intell.*, vol. 8, p. 1627078, Sep. 2025, doi: 10.3389/frai.2025.1627078.