

## Article

# RUDA-2025: Depression Severity Detection Using Pre-Trained Transformers on Social Media Data

Muhammad Ahmad <sup>1</sup>, Pierpaolo Basile <sup>2</sup>, Fida Ullah <sup>1</sup>, Ildar Batyrshin <sup>1</sup> and Grigori Sidorov <sup>1,\*</sup>

<sup>1</sup> Instituto Politécnico Nacional (IPN), Centro de Investigación en Computación (CIC), Mexico City 07700, Mexico; mahmad.2024@cic.ipn.mx (M.A.); fullah-2022@cic.ipn.mx (F.U.); batyr1@cic.ipn.mx (I.B.)

<sup>2</sup> Dipartimento di Informatica, Università degli Studi di Bari Aldo Moro, 70125 Bari, Italy

\* Correspondence: sidorov@cic.ipn.mx

## Abstract

Depression is a serious mental health disorder affecting cognition, emotions, and behavior. It impacts over 300 million people globally, with mental health care costs exceeding \$1 trillion annually. Traditional diagnostic methods are often expensive, time-consuming, stigmatizing, and difficult to access. This study leverages NLP techniques to identify depressive cues in social media posts, focusing on both standard Urdu and code-mixed Roman Urdu, which are often overlooked in existing research. To the best of our knowledge, a script-conversion and combination-based approach for Roman Urdu and Nastaliq Urdu has not been explored earlier. To address this gap, our study makes four key contributions. First, we created a manually annotated dataset named Ruda-2025, containing posts in code-mixed Roman Urdu and Nastaliq Urdu for both binary and multiclass classification. The binary classes are depression and not depression, with the depression class further divided into fine-grained categories: Mild, Moderate, and Severe depression alongside not depression. Second, we applied first-time two novel techniques to the RUDA-2025 dataset: (1) script-conversion approach that translates between code-mixed Roman Urdu and Standard Urdu and (2) combination-based approach that merges both scripts to make a single dataset to address linguistic challenges in depression assessment. Finally, we employed 60 different experiments using a combination of traditional machine learning and deep learning techniques to find the best-fit model for the detection of mental disorder. Based on our analysis, our proposed model (mBERT) using custom attention mechanism outperformed baseline (XGB) in combination-based, code-mixed Roman and Nastaliq Urdu script conversions.

**Keywords:** depression analysis; mental health; code-mixed Roman Urdu; Nastaliq Urdu; deep learning; transfer learning; social media; Facebook; YouTube; Twitter



Received: 2 July 2025

Revised: 31 July 2025

Accepted: 7 August 2025

Published: 18 August 2025

**Citation:** Ahmad, M.; Basile, P.; Ullah, F.; Batyrshin, I.; Sidorov, G. RUDA-2025: Depression Severity Detection Using Pre-Trained Transformers on Social Media Data. *AI* **2025**, *6*, 191. <https://doi.org/10.3390/ai6080191>

**Copyright:** © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

## 1. Introduction

Depression is a long-term and recurring mental health disorder that requires continuous monitoring and treatment [1–4]. According to the World Health Organization, over 300 million people worldwide suffer from depression [5], and it is becoming increasingly prevalent over time. It differs from regular mood swings and temporary emotional reactions that individuals encounter in daily life [6]. Psychiatrists typically use structured diagnostic interviews and specialized questionnaires to assess a patient's symptoms and make clinical decisions [7]. One common tool, the Patient Health Questionnaire-9 (PHQ-9),

asks individuals to rate how often they experience symptoms like hopelessness or a lack of interest in activities over the past two weeks, with the total score indicating the severity of depression [8]. However, these traditional methods are often time-consuming for healthcare professionals and place a heavy burden on mental health systems. Additionally, they rely on patients actively seeking help, which many may never do [9]. Given the growing role of health information technology in improving healthcare efficiency [10], developing automated methods for detecting depression is both necessary and beneficial.

The rapid rise of social media platforms such as Facebook, X (formerly Twitter), Reddit, YouTube, and Instagram has generated vast amounts of multimodal data, offering new opportunities for advancing mental health research as well as related domains like hate speech [11], hope speech, and named entity recognition (NER) [12]. Individuals often turn to these platforms to share their thoughts and emotions, especially when barriers such as stigma, limited awareness, or financial constraints prevent access to clinical care. This creates a valuable space for leveraging automated tools to identify mental health concerns [13], detect harmful content such as hate speech, promote positive communication through sentiment analysis [14], and extract relevant information using NER techniques.

Artificial intelligence (AI) and natural language processing (NLP) enable scalable and efficient monitoring of mental health by analyzing textual content and, in some cases, facial expressions. NLP techniques are used to detect depressive cues in language [15,16] and assess the emotional tone of online communication [17–22], while AI models identify linguistic markers associated with depression to support early diagnosis and timely intervention [23]. These technologies reduce the burden on healthcare systems and expand access to mental health support [24–26]. However, their development must consider critical ethical issues such as privacy and data security. As these tools continue to evolve, they hold the potential to facilitate more proactive and inclusive approaches to mental health care.

Depression detection using NLP has gained considerable attention in high-resource languages; however, code-mixed Roman Urdu and Nastaliq Urdu remain largely overlooked due to their unique linguistic characteristics, limited resources, and scarcity of high-quality annotated datasets. Urdu, the national language of Pakistan and spoken by over 230 million people globally [27], holds significant cultural and historical importance in literature, poetry, and social media communication. In particular, code-mixed Roman Urdu is widely used across South Asia, especially on social media and messaging platforms, as a popular medium for expressing personal thoughts and emotions [28]. However, it poses significant challenges for NLP models because of its non-standardized spelling, multiple word variations, and structural inconsistencies. On the other hand, Nastaliq Urdu—written in the Perso-Arabic script and deeply rooted in cultural and linguistic heritage—offers greater structural consistency but remains underrepresented in computational linguistics research. Consequently, existing depression detection models often fail to capture the linguistic diversity and informal nature of user-generated content in these forms of Urdu. Addressing these complexities is therefore essential for developing effective and inclusive automated depression assessment systems that can better serve diverse linguistic communities.

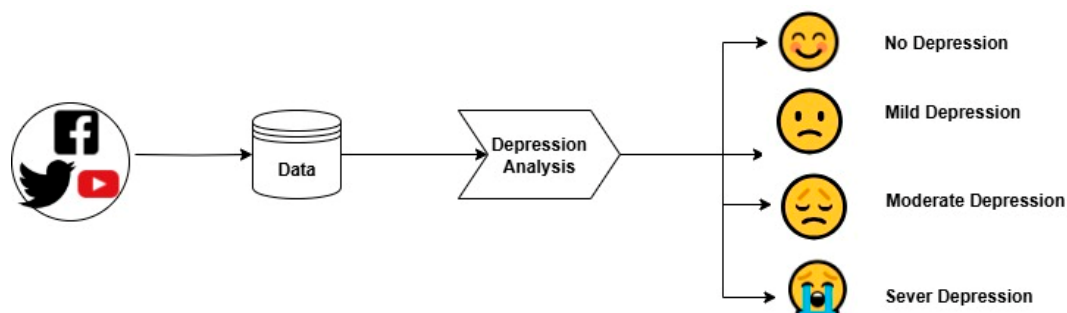
In pursuit of this objective, our study developed two distinct datasets named Roman Urdu and Urdu Depression Analysis-2025 (RUDA-2025): one in code-mixed Roman Urdu (Urdu written in the Roman script) and the other in Nastaliq Urdu (Urdu written in the Perso-Arabic script). To enrich these datasets, we employed two innovative strategies: a script-conversion approach and a combination-based approach. In the script-conversion method, we manually translated code-mixed Roman Urdu into Nastaliq Urdu and merged it with the original Nastaliq corpus; similarly, we converted Nastaliq Urdu into Roman Urdu and integrated it with the original Roman Urdu dataset. This bidirectional translation created unified bilingual corpora for each script. In parallel, the combination-based method

merged both the Roman Urdu and Nastaliq Urdu datasets into a single, more diverse corpus. To further strengthen our datasets, we applied data augmentation techniques, particularly focusing on the minority class, to support better performance in machine learning models.

Building on these enhanced datasets, we proposed a novel custom attention mechanism integrated with BERT to improve depression severity detection. We evaluated this approach through rigorous comparisons with baseline models and ablation studies. The results were highly promising: our method achieved 96% accuracy for binary classification on both the combination-based and Roman Urdu translation datasets and 95% accuracy on the Nastaliq Urdu translation dataset. For multiclass classification, DistilBERT achieved 80% accuracy on the combination-based dataset, mBERT reached 74% on Roman Urdu translation, and DistilBERT attained 81% on the Nastaliq Urdu translation dataset. These outcomes highlight the potential of our custom attention mechanism to bridge linguistic gaps, offering an effective solution for depression detection across diverse scripts and languages.

The contributions of this paper are as follows:

1. To the best of our knowledge, combination-based approach and script-conversion approaches were not explored earlier for depression detection in code-mixed Roman Urdu and Nastaliq Urdu datasets.
2. We developed a manually annotated code-mixed Roman Urdu and Nastaliq Urdu dataset with expert guidance for both binary and multiclass classification. In binary classification, posts are labeled as “depressed” or “non-depressed,” while in multiclass classification, “depressed” posts are further categorized into mild, moderate, and severe levels of depression (see Figure 1).
3. We conducted 60 different experiments to train and evaluate machine learning models using TF-IDF, deep learning models utilizing pre-trained word embeddings such as FastText and GloVe, and transformer-based models leveraging advanced contextual embeddings. In addition, we incorporated a custom attention mechanism within the transformer architecture to further enhance the model’s ability to capture complex linguistic patterns in the dataset.
4. Transformer-based models achieve the best performance over baselines in both binary and multiclass classification tasks. In binary classification, combination-based and Roman translation mBERT models reached 96% accuracy, while Nastaliq Urdu translation achieved 95%. For multiclass classification, combination-based DistilBERT attained 80% accuracy, mBERT achieved 74% for code-mixed Roman Urdu, and, for Nastaliq Urdu, DistilBERT achieved 81%.



**Figure 1.** Social media depression intensity analysis.

## 2. Literature Review

Previous studies on depression and emotion detection have explored a range of modalities, languages, and computational strategies, but a systematic convergence of bilingual

and script-diverse data remains largely absent. Sattar et al. [29] developed a code-mixed Roman Urdu dataset to distinguish between “Depressed” and “Not-Depressed” classes, showing that deep learning models, especially when combining character- and word-level features, outperformed traditional machine learning techniques. Similarly, working in Roman Urdu, Mohmand et al. [30] created a four-class depression dataset from 25,004 social media posts, applying a pre-trained BERT model to achieve remarkably high performance, further emphasizing the power of transfer learning in under-resourced language contexts. However, both these efforts are constrained to the Roman Urdu script and single-task classification, without exploring deeper multilingual or multiscript dynamics.

In contrast, Martinez-Navarro et al. [31] focused on speech-based emotion recognition using multiple datasets, including one in code-mixed Roman Urdu, with the system achieving strong performance in multilingual audio contexts. This audio-centric approach, though distinct in modality, reinforces the value of supporting emotional analysis in diverse language settings. Likewise, Qasim et al. [32] employed both content- and context-based NLP methods on Reddit data to classify varying depression severity levels, showing how transformer-based models can successfully detect subtle linguistic cues in mental health discourse. Their focus on English-only data and standard NLP pipelines highlights a limitation in extending these insights to low-resource, multilingual settings like Urdu.

Beyond textual analysis, Chhabra et al. [33] applied machine learning techniques to EEG data for mental stress classification, demonstrating high accuracy using entropy features and SVM classifiers. While methodologically distinct, their physiological approach also underlines the multidisciplinary interest in stress and depression detection. Meanwhile, Sharma et al. [34] surveyed hate speech detection in South Asian languages and highlighted the scarcity of resources and tools for low-resource scripts such as Nastaliq Urdu. Their work, though focused on hate speech, echoes the broader challenge of computational support for multilingual and multi-script environments.

In another dimension, Abbas et al. [35] localized and validated a stigma belief scale into Urdu, providing important cultural and linguistic grounding for mental health assessments in the Pakistani context. Yet, this effort remained survey-based, without tapping into the computational capabilities now available for large-scale data processing.

Amidst this growing body of literature, a clear gap emerges: while some studies have used Roman Urdu [29,30] and a few have mixed languages like English and Roman Urdu [36,37], none have ventured into systematically combining Roman Urdu with Standard Urdu written in Nastaliq script, particularly using script-conversion methods. Our proposed work directly addresses this limitation by constructing a novel bilingual, script-converted, and code-mixed dataset that integrates both Roman Urdu and Nastaliq Urdu. By converting Standard Urdu to both Roman and Nastaliq forms and then merging them, we provide a harmonized dataset that captures real-world language diversity far more comprehensively than any previous effort.

Table 1 presents a comparison of prior studies on depression severity detection from social media discourse, highlighting several key limitations. Most research focused either on English [38–40] or single-script Roman Urdu [29,30], addressing only binary or multiclass classification in isolation. While a few works combined languages such as English and Roman Urdu [36,37], they lacked script-conversion techniques and failed to incorporate low-resource scripts like Nastaliq Urdu. Similarly, studies involving Urdu or Punjabi [41] did not employ dataset integration or classification tasks, leaving depression detection in linguistically diverse, low-resource settings largely unexplored.

**Table 1.** Prior work in Roman Urdu and Nastaliq Urdu.

| Ref.     | Language                              | Combination | Script-Conversion | Binary Class | Multiclass |
|----------|---------------------------------------|-------------|-------------------|--------------|------------|
| [29]     | Roman Urdu                            | No          | No                | Yes          | No         |
| [30]     | Roman Urdu                            | No          | No                | No           | Yes        |
| [36]     | English + Roman Urdu                  | Yes         | No                | No           | Yes        |
| [37]     | Roman Urdu + English                  | Yes         | No                | Yes          | Yes        |
| [41]     | Punjabi + Standard Urdu               | No          | No                | No           | No         |
| [38]     | English                               | No          | No                | No           | Yes        |
| [39]     | English                               | No          | No                | Yes          | No         |
| [40]     | English                               | No          | No                | No           | Yes        |
| Proposed | Code-Mixed Roman Urdu + Nastaliq Urdu | Yes         | Yes               | Yes          | Yes        |

In contrast, our work uniquely integrates code-mixed Roman Urdu and Nastaliq Urdu, introduces a bidirectional script-conversion approach, and constructs a combined bilingual dataset to enhance linguistic diversity. Crucially, we address both binary and multiclass classification within a unified framework using a custom attention-enhanced mBERT model. To the best of our knowledge, this is the first study to jointly leverage multi-script, code-mixed data and advanced NLP techniques for depression severity detection, effectively filling a critical gap and advancing inclusive mental health assessment in low-resource languages.

### 3. Methodology and Design

#### 3.1. Construction of Dataset

For the construction of our dataset, we sourced data from three major social media platforms: Twitter API v2, YouTube API v3, and Facebook, spanning the period from May 2020 to May 2024. This specific time frame was selected because it encompasses the global outbreak and aftermath of the COVID-19 pandemic—a period marked by widespread lockdowns, social isolation, economic uncertainty, and significant disruptions to daily life, all of which had a profound impact on mental health. This period is particularly relevant for studies related to depression, as it reflects a unique socio-psychological environment in which the prevalence of anxiety, loneliness, and depressive symptoms increased notably across diverse populations.

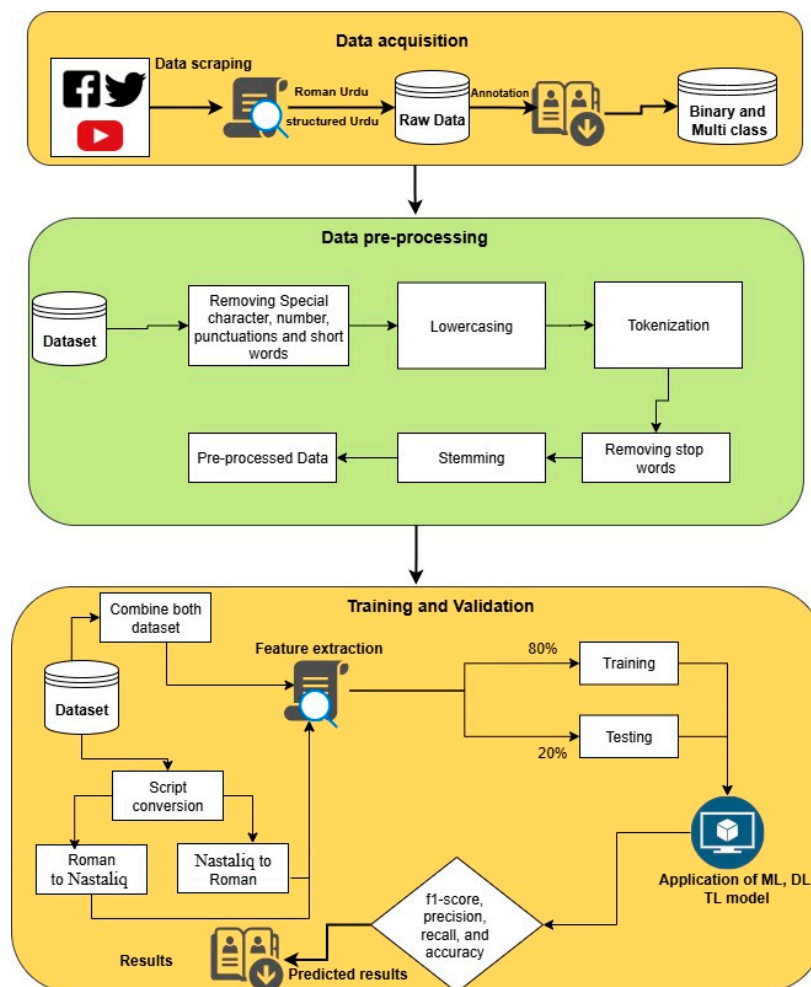
To collect data from Twitter, we utilized the Tweepy API, a Python library that enables developers to access tweets, user data, trends, and other features of the Twitter platform. To distinguish between depression and non-depression content, we employed a variety of targeted keywords, covering terms related to “non-depression,” “mild depression,” “moderate depression,” and “severe depression.” This allowed us to extract tweets containing relevant terms and gather real-time data on user engagement, sentiment, and discussions around these topics.

For YouTube, we employed the YouTube API to collect video-related user data, particularly user comments associated with our depression-related keywords. These comments provided valuable insights into public discourse surrounding mental health in a video-based social environment. We specifically targeted videos focused on mental health awareness, personal experiences with depression, motivational content, expert talks, and videos addressing the psychological impact of the COVID-19 pandemic. This selection ensured the inclusion of diverse perspectives and emotionally expressive content, which is essential for effective depression analysis.

In the case of Facebook, data was manually collected, focusing on user posts and comments that included the same depression-related keywords in both code-mixed Roman

Urdu and Nastaliq Urdu. This approach allowed us to capture a diverse set of content types, including personal experiences, supportive messages, and community responses.

By collecting a total of 40,000 social media posts and comments across these three platforms, we aimed to build a comprehensive and robust dataset. This dataset not only tracks how depression-related content is discussed but also explores variations in engagement across text-based, video-based, and socially interactive formats. Figure 2 provides a graphical representation of our data collection process and the construction of the dataset, which includes mental health-related content in both code-mixed Roman Urdu and Nastaliq Urdu.



**Figure 2.** Architecture of proposed methodology and design.

### 3.2. Preprocessing

After collecting user posts, we carefully analyzed and preprocessed the data, as this step was crucial. Social media data often contain noise, links, and hashtags, which can lead to incorrect results for machine learning and deep learning models. In this study, the process begins with text cleaning, where special characters, numbers, punctuation, and short words are removed. Additionally, duplicate tweets and tweets with fewer than 20 characters were also removed. Next, all text is converted to lowercase to ensure consistency in a language. Tokenization is performed to break the text into individual words or subwords. After tokenization, we apply stop word removal, followed by stemming or lemmatization. After these preprocessing steps, the cleaned and standardized text is ready for further processing and model training. As a result of these techniques, only 8000 code-mixed Roman Urdu and Nastaliq Urdu sentences remained for annotation, as shown in Figure 2. Following preprocessing, an 80-20 split was employed in all experiments, with 80% of the data used for

training and validation and the remaining 20% reserved for testing. For machine learning models, we applied feature extraction using term frequency-inverse document frequency (TF-IDF) for whole-word representations. For deep learning models, we used pretrained word embeddings, such as FastText and GloVe. Additionally, for transfer learning, we applied advanced language-based contextual embeddings to capture hidden patterns in both code-mixed Roman Urdu and Standard Urdu without modifying the vocabulary.

### 3.3. Annotation

Annotation involves assigning predefined labels to each sample to create a structured dataset to train supervised learning models. In the context of depression classification, the process begins with labeling each sample as either depression or non-depression in a binary classification. After that, the depression samples are further classified into three unique categories, such as mild depression, moderate depression, and severe depression, depending on the user's context. To ensure well-defined and reliable guidelines, we consulted three psychiatrists, who helped us in developing these classification guidelines based on their expert insights and clinical criteria.

### 3.4. Annotation Procedure

Our labelling process followed a strict and structured approach to ensure accuracy and consistency throughout the annotation procedure. We initially trained eight native Urdu-speaking master's students from the Department of Psychology at The Islamia University of Bahawalpur, Pakistan, trained under the supervision of a qualified psychologist. The training focused on understanding emotional and psychological cues, as well as familiarizing the annotators with the annotation guidelines for both code-mixed Roman Urdu and Nastaliq Urdu texts.

Following the training, we provided a set of 400 social media posts to these eight annotators. Their labeled outputs were reviewed with the assistance of the psychologist to assess clinical relevance and consistency. Based on this assessment, we found that five annotators produced largely consistent and high-quality labels. To further refine the team, a second set of 400 new samples was given to these five annotators. After another round of evaluation—again in consultation with the psychologist—we observed that only three of them consistently demonstrated high-quality annotation aligned with psychological understanding.

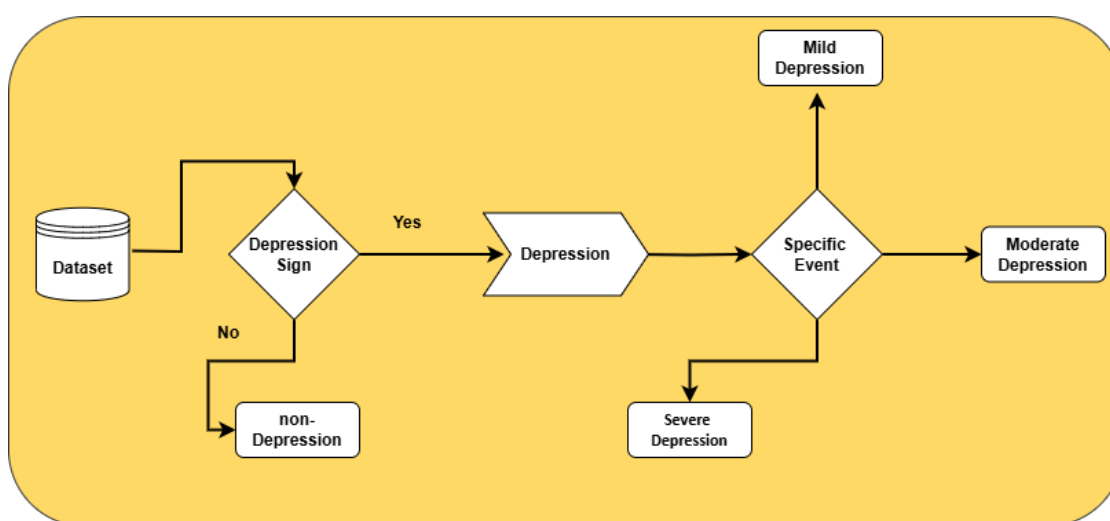
These three annotators were then selected to build the final RUDA-2025 dataset. To manage the process efficiently, separate Google Forms were created for each annotator, allowing us to track their progress and supervise their work effectively. All annotators were required to strictly follow the defined procedure and adhere to the provided guidelines. Although all annotators received the same set of guidelines, their individual interpretations ultimately determined whether the qualities described were present in a tweet—either for binary or multiclass classification, as shown in Table 2.

To address annotator disagreements and ensure consistency, we adopted a majority voting approach. Each sample was independently annotated by all three annotators. In cases where all annotators agreed, the label was directly assigned. For borderline cases, where two annotators agreed while one disagreed, the majority label was assigned to the sample. In rare instances of complete disagreement or persistent ambiguity, review meetings were held, and such cases were either discussed with guidance from a psychologist or excluded from the final dataset to maintain quality.

**Table 2.** Dataset overview.

| Level               | Urdu Tweet                                                                                         | Code-Mixed Roman Urdu Tweet                                                                                                                                                              | Binary | Multiclass |
|---------------------|----------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------|------------|
| Not Depression      | میں آج بہت خوش ہوں، موسم بھی بہت اچھا ہے!                                                          | Aaj ka din bohat acha guzra, friends ke saath outing ka maza aya!.                                                                                                                       | 0      | 0          |
|                     | I am very happy today, and the weather is also very nice                                           | Today went really well; I had a great time outing with friends.                                                                                                                          |        |            |
| Mild Depression     | دل تھوڑا اداس ہے، لیکن سب ٹھیک ہو جائے گا.                                                         | Mujhe kuch weeks se khushi mehsoos nahi hoti, lekin kaam waghera karta rehta hoon. . . maybe it's just a stress                                                                          | 1      | 1          |
|                     | The heart feels a little sad, but everything will be okay                                          | I haven't felt happy for a few weeks, but I keep doing my work and everything . . . maybe it's just stress . . .                                                                         |        |            |
| Moderate Depression | کچھ دنوں سے زندگی عجیب سی لگ رہی ہے، کچھ بھی اچھا نہیں لگ رہا                                      | Roz marra ke kaam mushkil lagte hain, concentration nahi hoti aur aksar sochta hoon ke sab bekaar hai                                                                                    | 1      | 2          |
|                     | Life has been feeling strange for the past few days; nothing seems to feel right                   | Everyday tasks feel difficult, I can't concentrate, and I often think that everything is pointless                                                                                       |        |            |
| Severe Depression   | اندھیرا ہی اندھیرا ہے، روشنی کی کوئی امید نہیں۔ شاید میرے جانے سے ہی سب مسائل ختم ہو جائیں گے۔     | Aaj kal lagta hai life end ho gayi hai. Har cheez bas ruk gayi hai. Na kisi se baat karne ka dil hai, na kuch work ka. Sab kuch darkness sa ho gaya hai. Lgtai hai mar jana chaye        | 1      | 3          |
|                     | It's all darkness, there's no hope for light. Perhaps all the problems will end with my departure. | These days it feels like life has ended. Everything just seems to have come to a stop. I don't feel like talking to anyone or doing any work. Everything feels dark. It feels like dying |        |            |

Through this rigorous and iterative selection and training process, we developed a reliable and high-quality dataset for mental health depression analysis. The final dataset supports both binary and multiclass classification tasks in code-mixed Roman Urdu and Nastaliq Urdu, comprising 8321 samples in code-mixed Roman Urdu and 7163 samples in Nastaliq Urdu, resulting in a total of 15,484 valid samples. Annotators were compensated at a rate of \$0.03 USD per post, ensuring fair remuneration for their efforts. Figure 3 illustrates the overall annotation workflow, while Table 3 provides an overview of the dataset distribution. Figure 3 illustrates the annotation procedure of the RUDA-2025 dataset for both binary and multiclass classification tasks.

**Figure 3.** Annotation procedure of our dataset.

### 3.5. Inter-Annotator Agreement

Inter-annotator agreement (IAA) is a crucial metric for evaluating the consistency and reliability of annotations while accounting for chance agreement. In this study, annotation quality was measured using two widely recognized statistics: Cohen’s kappa coefficient for the binary classification task and Fleiss’ kappa score for the multiclass classification task. The binary classification yielded a Cohen’s kappa of 0.85 (85%), indicating almost perfect agreement, while the multiclass task resulted in a Fleiss’ kappa of 0.79 (79%), reflecting substantial agreement among annotators. Despite these strong metrics, disagreements did occur, particularly between closely related classes such as mild and moderate depression, where symptoms can overlap. For instance, one annotator labeled the tweet “*mujhe kuch acha nahi lagta lekin zindagi barbad bhi nahi lagti*” (Nothing feels good to me, but life doesn’t seem completely ruined either) as mild, while another marked it as moderate due to the tone of emotional fatigue. To address such inconsistencies, we conducted a meeting with a psychologist, refined our annotation guidelines, and applied majority voting to finalize contentious labels. These steps helped reduce ambiguity and improve the reliability of our annotated dataset.

**Table 3.** Interpretation of the Kappa values.

| Kappa Value (K) | Agreement Level          |
|-----------------|--------------------------|
| $\leq 0.00$     | Poor Agreement           |
| 0.01–0.20       | Slight Agreement         |
| 0.21–0.40       | Fair Agreement           |
| 0.41–0.60       | Moderate Agreement       |
| 0.61–0.80       | Substantial Agreement    |
| 0.81–1.00       | Almost Perfect Agreement |

### 3.6. Dataset Statistics

Figure 4 presents a comparative analysis of Urdu language statistics across three datasets: translation of Nastaliq Urdu, translation of code-mixed Roman Urdu, and a combination of both. Each bar chart summarizes key linguistic metrics, including total posts, vocabulary size, total words, average words per sentence, and average sentences per post. All three datasets contain the same number of posts (15,484), ensuring a fair comparison. The Nastaliq Urdu translation shows the largest vocabulary (8302 unique words) and the highest total word count (431,388), along with an average of 27.86 words per sentence—indicating more expressive or verbose content in the standard script. In contrast, the Roman Urdu dataset, while matching in post count, has a smaller vocabulary (12,314) and fewer total words (339,342), with slightly shorter average sentences (21.44 words). The combined dataset, which merges both scripts, strikes a balance with a richer vocabulary (14,145) and 364,525 total words, averaging 23.27 words per sentence. Interestingly, the average sentences per post hover around 1 across all datasets, highlighting a consistency in structural length. This comparison not only underscores the linguistic richness and variance between scripts but also reflects how combining them can lead to more comprehensive and diverse data for natural language processing tasks. Figure 5 shows the label distribution, while Figure 6 presents a word cloud of the most frequent words extracted from the RUDA-2025 dataset.

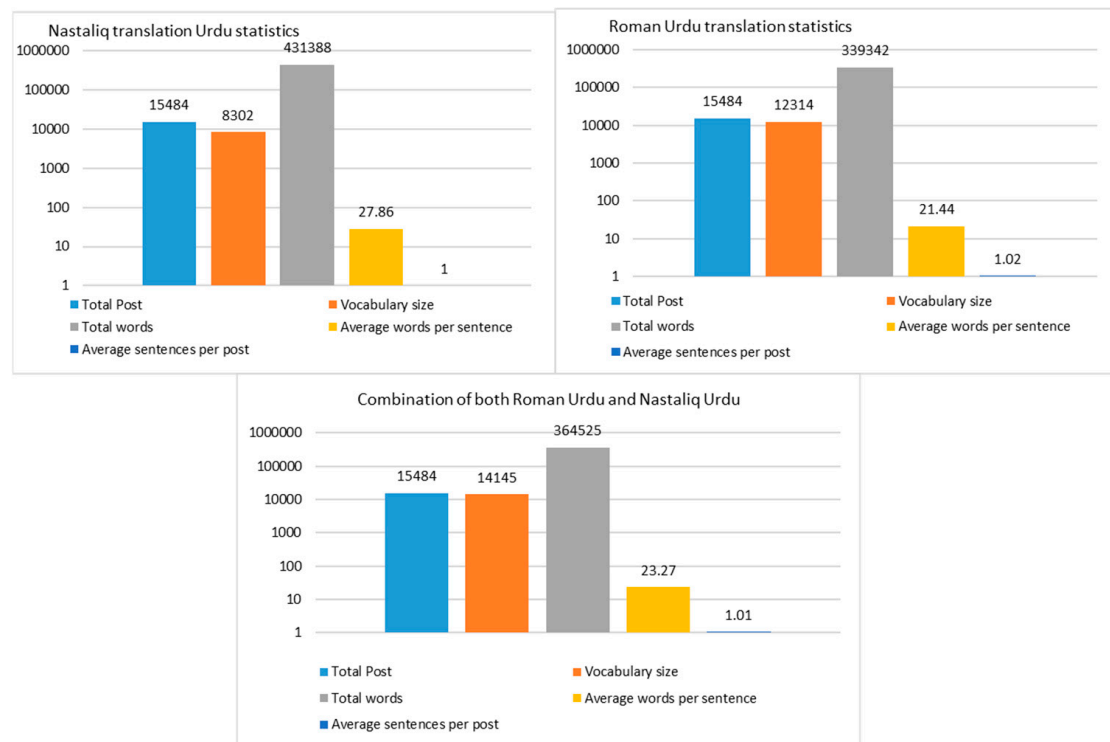


Figure 4. Code-mixed Roman, Nastaliq, and combination-based Urdu datasets statistics.

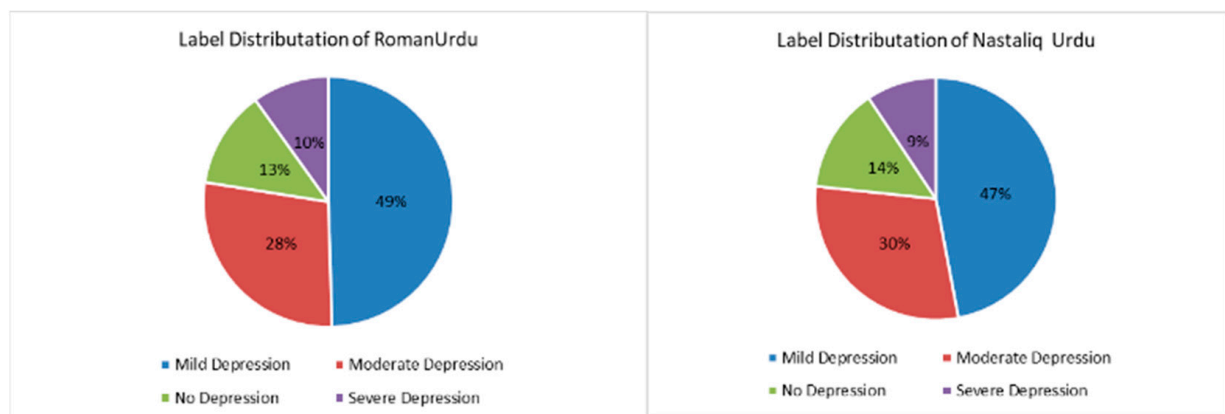


Figure 5. Label distribution of code-mixed Roman Urdu and Nastaliq Urdu dataset.



Figure 6. Word cloud of frequent words from the dataset.

### 3.7. Ethical Considerations

The dataset for this study was created using publicly available social media posts, for which informed consent is generally not required due to their public nature. However, we

recognize the sensitive nature of mental health data and have taken several measures to ensure ethical compliance. All data were carefully anonymized and aggregated to protect user privacy and prevent identification of individuals. We also acknowledge potential risks, such as misdiagnosis, stigmatization, or unintended psychological impact that may arise from automated depression detection systems. To mitigate these, we emphasize that our models are intended as supportive tools rather than diagnostic instruments and should be used with caution. The study adheres to ethical principles of transparency, fairness, and respect for user dignity, and we advocate for ongoing ethical oversight in deploying such technologies in real-world settings.

### 3.8. Translation Steps

The goal of the script conversion process is to unify posts written in multiple forms of Urdu—specifically Roman Urdu and Nastaliq Urdu—into a consistent representation. For this purpose, we constructed two distinct corpora corresponding to each script, as detailed below:

- For the first corpus, we converted code-mixed Roman Urdu posts into standard Nastaliq Urdu script.
- Post-conversion, the text was manually reviewed to correct script inconsistencies and preserve semantic meaning.
- For the second corpus, posts originally written in Nastaliq Urdu were transliterated into code-mixed Roman Urdu. After transliteration, the content was manually refined to correct phonetic variations and ensure readability. The resulting Roman Urdu text was then merged with the original Roman Urdu dataset for further analysis.

The pseudo-code of both approaches is presented in Table 4.

**Table 4.** Pseudocode overview for multi-script depression detection using RUDA-2025 corpus.

| Step No. | Operation                                                                   | Description                                                                      |
|----------|-----------------------------------------------------------------------------|----------------------------------------------------------------------------------|
| 1        | multi-script-depression-detection (RomanUrdu-Dataset, NastaliqUrdu-Dataset) | Main procedure to detect depression using both Roman and Nastaliq Urdu datasets. |
| 2        | Pre-Processing (RomanUrdu-Dataset) → D1                                     | Clean and preprocess Roman Urdu dataset.                                         |
| 3        | Pre-Processing (NastaliqUrdu-Dataset) → D2                                  | Clean and preprocess Nastaliq Urdu dataset.                                      |
| 4        | Script-Convert-RomanToUrdu (D1) → D1U                                       | Convert Roman Urdu to Nastaliq Urdu using script conversion.                     |
| 5        | Script-Convert-UrduToRoman (D2) → D2R                                       | Convert Nastaliq Urdu to Roman Urdu.                                             |
| 6        | Merge (D1, D1U) → Com-1                                                     | Combine original and converted Roman Urdu.                                       |
| 7        | Merge (D2, D2R) → Com-2                                                     | Combine original and converted Nastaliq Urdu.                                    |
| 8        | Merge (Com-1, Com-2) → Com-All                                              | Final combined dataset from both scripts.                                        |
| 9        | mBERT-Tokenizer (Com-All) → Tokenized-Data                                  | Tokenize combined dataset using multilingual BERT tokenizer.                     |
| 10       | fine-tune-mBERT-with-Attention (Tokenized-Data, 80-20) → Model              | Fine-tune mBERT with attention mechanism using an 80-20 train-test split.        |
| 11       | generate-results (Model) → confusion-matrix                                 | Generate confusion matrix for evaluation.                                        |
| 12       | compute-accuracy (confusion-matrix) → accuracy                              | Calculate accuracy.                                                              |
| 13       | compute-precision (confusion-matrix) → precision                            | Calculate precision.                                                             |
| 14       | compute-recall (confusion-matrix) → recall                                  | Calculate recall.                                                                |
| 15       | compute-f1 (confusion-matrix) → f1-score                                    | Calculate F1-score.                                                              |
| 16       | End Procedure                                                               | End of main procedure.                                                           |

### 3.9. Application of Models

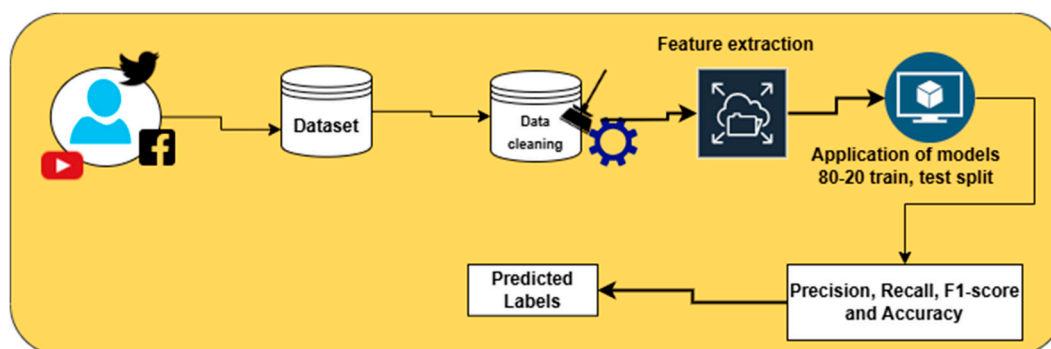
In this section, we discuss the entire process of model application during the training and testing phases for our text classification task. We started with a raw dataset that

underwent cleaning and preprocessing to remove noise and ensure usability. The data was labeled using a majority voting technique to maintain high annotation accuracy. For traditional machine learning, we employed four models—support vector machine (SVM), random forest (RF), XGBoost (XGB), and logistic regression (LR)—with TF-IDF features to capture important word patterns. TF-IDF statistically measures the importance of words within documents relative to the entire corpus.

For deeper semantic understanding, we also used advanced deep learning models such as bidirectional long short-term memory (BiLSTM) and convolutional neural networks (CNN) with pretrained word embeddings like FastText and GloVe. These models helped uncover complex hidden patterns beyond surface-level word statistics.

To further improve performance, we incorporated three pretrained transformer-based language models, including multilingual BERT (mBERT), which generates contextual embeddings that capture nuanced linguistic information. Notably, we enhanced the transformer models by integrating a custom attention mechanism on top of mBERT's token-level outputs. Instead of solely relying on the [CLS] token representation, this attention layer learns to weigh the importance of individual tokens dynamically, allowing the model to focus more effectively on key parts of the input text relevant to depression detection. Specifically, the custom attention operates on the sequence of hidden states produced by mBERT's last layer, computing attention scores for each token embedding and generating a weighted sum that forms a context vector. This context vector, summarizing the most informative tokens, is then passed to a feedforward classifier to predict depression labels. This addition proved particularly valuable for handling code-mixed and script-converted data, where important cues may be distributed unevenly across tokens.

After training, we evaluated all models using Precision, Recall, F1-score, and Accuracy, enabling us to identify the most effective approach for our classification problem. Figure 7 illustrates the entire workflow of model application, training, and testing phases utilized in this study.



**Figure 7.** Application of models training and testing phase.

Table 5 presents the models and their corresponding hyperparameters used for depression analysis in Nastaliq Urdu and code-mixed Roman Urdu. We explored three learning approaches: transformer-based models, machine learning algorithms, and deep learning architectures. Under the Transformer category, mBERT, DistilBERT, and Google Electra were fine-tuned using a learning rate of  $2 \times 10^{-5}$ , 3 epochs, a batch size of 64, with AdamW optimizer and CrossEntropyLoss. These models were chosen based on their superior performance. For traditional machine learning, we used SVM (random state 42, linear kernel,  $C = 1.0$ ), random forest (100 estimators, max depth of 6, random state 42), logistic regression (random state 42, max\_iter = 1000,  $C = 0.1$ , solver = 'liblinear'), and XGBoost (100 estimators, max depth of 6, learning rate 0.1, gamma 0.2). Deep learning models included BiLSTM and CNN, both trained with a learning rate of 0.1, 5 epochs,

embedding dimension of 300, and batch size of 32. The BiLSTM used 128 LSTM units, while the CNN had 128 filters with a kernel size of 5. Hyperparameters for all models were tuned using grid search to achieve optimal performance.

**Table 5.** Optimum VALUES FOR the hyperparameters of proposed models.

| Learning Approach              | Models                            | Hyperparameters                                                                                                                                                                                                | Grid Search Values                                                                                                             |
|--------------------------------|-----------------------------------|----------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------------------------------------------------------|
| Transformer + Custom Attention | mBERT, DistilBERT, Google Electra | Learning Rate, Epochs, Batch Size, Optimizer, Loss Function, Max Seq Length, Hidden Size, Intermediate Size, Attention Heads, Attention Dropout, Feedforward Dropout, Residual Connection, Positional Encoding | $2 \times 10^{-5}$ to $3 \times 10^{-5}$ , 3, 64, AdamW, CrossEntropyLoss, 128, 768, 2048–3072, 6–12, 0.1, 0.1, Yes, Sinusoida |
| Machine Learning               | SVM                               | Random state, Kernel, C value                                                                                                                                                                                  | 42, linear, 1.0                                                                                                                |
|                                | RF (Random Forest)                | n_estimators, max_depth, Random state                                                                                                                                                                          | 100, 6, 42                                                                                                                     |
|                                | LR (Logistic Regression)          | Random state, Max_iter, C value, Solver                                                                                                                                                                        | 42, 1000, 0.1, liblinear                                                                                                       |
|                                | XGB (XGBoost)                     | n_estimators, max_depth, Learning rate, Gamma                                                                                                                                                                  | 100, 6, 0.1, 0.2                                                                                                               |
| Deep Learning                  | BiLSTM                            | Learning rate, Epochs, Embedding_dim, Batch size, LSTM units                                                                                                                                                   | 0.1, 5, 300, 32, 128                                                                                                           |
|                                | CNN                               | Learning rate, Epochs, Embedding_dim, Batch size, Filters, Kernel size                                                                                                                                         | 0.1, 5, 300, 32, 128, 5                                                                                                        |

## 4. Results and Analysis

### 4.1. Experimental Setup

A series of experiments were conducted to fine-tune machine learning, deep learning, and transfer learning models in order to evaluate the proposed pipeline against baseline methods. The study utilized two depression analysis datasets: one in code-mixed Roman Urdu and the other in Nastaliq Urdu. An 80-20 split was employed across all experiments, where 80% of the data was used for training and validation and the remaining 20% was reserved for testing. The experiments were carried out on Google Colab using a Python 2.7 environment. Scikit-Learn [42] was used for machine learning models, while TensorFlow [43] and Keras [44] were employed for deep learning. Transformer-based models were fine-tuned using the Hugging Face Transformers library. Training was performed on an NVIDIA Tesla T4 GPU (2560 CUDA cores, 16 GB GDDR6 memory) supported by a seventh-generation Intel Core i5 processor (4 cores, 8 GT/s bus speed), along with 24 GB of RAM and 1 TB of storage. Further details of the datasets are provided in the following section.

### 4.2. Results for Machine Learning Models

In this section, we present the results of deep learning models on three datasets: the combined dataset, the Roman Urdu translation dataset, and the Nastaliq translation dataset. Given the linguistic complexity and orthographic diversity inherent in these text forms, we utilized TF-IDF to extract meaningful features from the corpus. For this purpose, we employed multiple classification algorithms and assessed their performance using standard evaluation metrics to determine their ability to effectively distinguish between classes within this multilingual and code-mixed context.

Table 6 presents the performance of various machine learning models using TF-IDF feature representations. In the case of the combination-based dataset, where inputs likely include a mix of Roman Urdu and Nastaliq Urdu, we see a noticeable variation in performance across the models. For the binary classification task, XGBoost (XGB) significantly outperforms the others, with a strong F1-score of 0.87 and an Accuracy of 0.88, showcasing its robustness in handling this mixed-language input. Logistic regression (LR) also performs reasonably well (F1: 0.76, Accuracy: 0.79), while SVM lags behind (F1: 0.63, Accuracy: 0.74), suggesting it struggles with the mixed-linguistic structure. For the multiclass scenario, all models show a general dip in performance, but again, XGB leads with an F1-score of 0.72,

indicating it is more adept at managing multiple categories, even in linguistically noisy settings. However, overall performance in multiclass tasks is lower, possibly due to the complexity of distinguishing between multiple nuanced categories in mixed-language data.

**Table 6.** Results for machine learning models.

| Class        | Language                  | Models | Precision | Recall | F1-Score | Accuracy |
|--------------|---------------------------|--------|-----------|--------|----------|----------|
| Binary class | Combination-based         | LR     | 0.78      | 0.79   | 0.76     | 0.79     |
|              |                           | SVM    | 0.55      | 0.74   | 0.63     | 0.74     |
|              |                           | XGB    | 0.88      | 0.88   | 0.87     | 0.88     |
|              | Roman Translation         | LR     | 0.81      | 0.82   | 0.79     | 0.82     |
|              |                           | SVM    | 0.57      | 0.75   | 0.65     | 0.75     |
|              |                           | XGB    | 0.86      | 0.86   | 0.85     | 0.86     |
|              | Nastaliq Urdu Translation | LR     | 0.79      | 0.81   | 0.79     | 0.81     |
|              |                           | SVM    | 0.81      | 0.75   | 0.64     | 0.75     |
|              |                           | XGB    | 0.92      | 0.92   | 0.92     | 0.92     |
| Multiclass   | Combination-based         | LR     | 0.57      | 0.58   | 0.57     | 0.58     |
|              |                           | SVM    | 0.57      | 0.57   | 0.56     | 0.57     |
|              |                           | XGB    | 0.72      | 0.73   | 0.72     | 0.73     |
|              | Roman Translation         | LR     | 0.59      | 0.59   | 0.59     | 0.59     |
|              |                           | SVM    | 0.63      | 0.58   | 0.57     | 0.58     |
|              |                           | XGB    | 0.69      | 0.7    | 0.69     | 0.7      |
|              | Nastaliq Urdu Translation | LR     | 0.58      | 0.58   | 0.58     | 0.58     |
|              |                           | SVM    | 0.62      | 0.58   | 0.56     | 0.58     |
|              |                           | XGB    | 0.78      | 0.78   | 0.78     | 0.78     |

The Roman Urdu translation dataset appears to provide cleaner and more structured input compared to the combination-based one. For binary classification, all models show improved performance over the combination setting. XGB once again leads, achieving a high F1-score of 0.85 and Accuracy of 0.86, with logistic regression close behind (F1: 0.79, Accuracy: 0.82). This suggests that Roman Urdu, when translated consistently, reduces ambiguity and aids in better classification. For multiclass classification, results are slightly more varied. Logistic regression and SVM hover around the 0.57–0.59 range in terms of F1-score, but XGB again stands out with an F1 of 0.69 and Accuracy of 0.70, indicating that even when dealing with multiple classes, Roman Urdu translation provides a relatively solid foundation for machine learning models to build upon.

Interestingly, the Nastaliq Urdu translation dataset consistently yields the best overall performance, especially in binary classification, where XGB achieves a perfect balance—F1-score and Accuracy both at 0.92. This suggests that when the text is fully and accurately translated into standard written Urdu, the linguistic richness and structure help the models—especially XGB—make more precise predictions. Logistic regression also performs strongly here, almost matching the Roman Urdu results (F1: 0.79, Accuracy: 0.81). Surprisingly, SVM underperforms in the binary task (F1: 0.64), possibly due to challenges in kernelizing the complex script and syntactic patterns of Nastaliq Urdu. For multiclass classification, the trend continues: XGB achieves the best performance again with an F1-score of 0.78, clearly outshining the other models. This consistent superiority implies that fully contextual and semantically rich translations in Nastaliq Urdu empower XGB to capture and classify intricate class distinctions more effectively.

#### 4.3. Results for Deep Learning Models

In this section, we present the results of deep learning models on three datasets: the combined dataset, the Roman Urdu translation dataset, and the Nastaliq translation dataset. To enhance the models' understanding of semantic and contextual nuances, we employed pretrained word embeddings from FastText and GloVe. These embeddings were integrated into the CNN and BiLSTM models to effectively represent the linguistic richness of Roman Urdu, Nastaliq Urdu, and their combinations.

Table 7 presents a comprehensive comparison of CNN and BiLSTM models using FastText and GloVe embeddings for both binary and multiclass text classification tasks across three language formats: combination-based, Nastaliq Urdu translation, and Roman Urdu translation. In binary classification, FastText consistently performed well, especially with Roman Urdu, where CNN achieved an impressive F1-score of 0.92, while BiLSTM slightly led on combination-based data with F1-scores reaching 0.90–0.91. GloVe also delivered strong binary performance, particularly on Roman Urdu (F1 up to 0.93 with CNN), though it struggled more with Nastaliq Urdu, where Precision dropped despite high Recall—indicating the model often misclassified negative instances.

**Table 7.** Results for deep learning models.

| Class      | Language                  | Models | Precision | Recall | F1-Score | Accuracy |
|------------|---------------------------|--------|-----------|--------|----------|----------|
| FastText   |                           |        |           |        |          |          |
| Binary     | Combination-based         | CNN    | 0.89      | 0.89   | 0.89     | 0.89     |
|            |                           | BiLSTM | 0.9       | 0.91   | 0.9      | 0.91     |
|            | Nastaliq Urdu Translation | CNN    | 0.88      | 0.88   | 0.88     | 0.88     |
|            |                           | BiLSTM | 0.9       | 0.89   | 0.89     | 0.89     |
|            | Roman Translation         | CNN    | 0.92      | 0.92   | 0.92     | 0.92     |
|            |                           | BiLSTM | 0.89      | 0.86   | 0.87     | 0.86     |
| Multiclass | Combination-based         | CNN    | 0.43      | 0.26   | 0.12     | 0.26     |
|            |                           | BiLSTM | 0.06      | 0.25   | 0.1      | 0.25     |
|            | Nastaliq Urdu Translation | CNN    | 0.06      | 0.25   | 0.1      | 0.25     |
|            |                           | BiLSTM | 0.06      | 0.25   | 0.1      | 0.25     |
|            | Roman Translation         | CNN    | 0.65      | 0.65   | 0.64     | 0.65     |
|            |                           | BiLSTM | 0.23      | 0.32   | 0.26     | 0.32     |
| GloVe      |                           |        |           |        |          |          |
| Binary     | Combination-based         | CNN    | 0.84      | 0.84   | 0.82     | 0.84     |
|            |                           | BiLSTM | 0.85      | 0.84   | 0.82     | 0.84     |
|            | Nastaliq Urdu Translation | CNN    | 0.56      | 0.75   | 0.65     | 0.75     |
|            |                           | BiLSTM | 0.56      | 0.75   | 0.65     | 0.75     |
|            | Roman Translation         | CNN    | 0.93      | 0.93   | 0.93     | 0.93     |
|            |                           | BiLSTM | 0.87      | 0.87   | 0.87     | 0.87     |
| Multiclass | Combination-based         | CNN    | 0.61      | 0.5    | 0.51     | 0.5      |
|            |                           | BiLSTM | 0.62      | 0.49   | 0.5      | 0.49     |
|            | Nastaliq Urdu Translation | CNN    | 0.48      | 0.26   | 0.13     | 0.26     |
|            |                           | BiLSTM | 0.46      | 0.26   | 0.13     | 0.26     |
|            | Roman Translation         | CNN    | 0.69      | 0.65   | 0.66     | 0.65     |
|            |                           | BiLSTM | 0.71      | 0.71   | 0.7      | 0.71     |

In the more complex multiclass classification setting, FastText's performance sharply declined, especially on combination-based and Nastaliq Urdu inputs, where F1-scores dropped to as low as 0.10–0.13, showing its limited ability to distinguish among several classes. GloVe performed relatively better in this scenario, particularly on Roman Urdu,

where BiLSTM achieved an F1-score of 0.70 and 0.71 accuracy, reflecting better adaptability to class diversity. The overall effect of the translation and script type was significant: Roman Urdu consistently led to higher model performance across both classification tasks, suggesting that its Latin-based script is easier for neural models to process and tokenize. In contrast, Nastaliq Urdu proved the most challenging due to its script complexity and possible inconsistencies in translation, leading to reduced Precision and poor multiclass prediction. Combination-based inputs had mixed results, indicating that while script diversity may offer broader language coverage, it can also introduce variability that affects model stability.

#### 4.4. Results for Transfer Learning

In this section, we present the results of deep learning models on three datasets: the combined dataset, the Roman Urdu translation dataset, and the Nastaliq translation dataset. Given the limitations of traditional word embeddings in capturing deep contextual relationships, we leveraged pretrained transformer-based models to enhance performance. These models are well-suited for handling the linguistic variability and structural complexities present in Roman Urdu, Nastaliq Urdu, and their combinations.

Table 8 presents a comparison of different transfer learning models—mBERT, DistilBERT, and Google Electra—applied to both binary and multiclass classification tasks using three datasets (combination-based, Roman Urdu translation, and Nastaliq Urdu translation), where model selection was based on their higher performance across evaluation metrics. In the combination-based input setting for binary classification (likely combining both Roman Urdu and Nastaliq Urdu translations), mBERT and DistilBERT perform excellently, achieving high scores across all metrics (Precision, Recall, F1-score, and Accuracy at 0.96 and 0.95, respectively). This indicates that both multilingual BERT models are highly effective for binary classification when diverse inputs are combined. In contrast, Google Electra performs significantly worse with only 0.62 Accuracy and 0.64 F1-score, likely due to its comparatively limited multilingual capabilities or poor generalization in this mixed setting.

**Table 8.** Results for transformer models.

| Class      | Language                  | Models         | Precision | Recall | F1-Score | Accuracy |
|------------|---------------------------|----------------|-----------|--------|----------|----------|
| Binary     | Combination-based         | Mbert          | 0.96      | 0.96   | 0.96     | 0.96     |
|            |                           | distil bert    | 0.95      | 0.95   | 0.95     | 0.95     |
|            |                           | Google Electra | 0.83      | 0.62   | 0.64     | 0.62     |
|            | Roman Translation         | Mbert          | 0.96      | 0.96   | 0.96     | 0.96     |
|            |                           | distil bert    | 0.94      | 0.94   | 0.94     | 0.94     |
|            |                           | Google Electra | 0.94      | 0.94   | 0.94     | 0.94     |
|            | Nastaliq Urdu Translation | Mbert          | 0.95      | 0.95   | 0.95     | 0.95     |
|            |                           | distil bert    | 0.94      | 0.94   | 0.94     | 0.94     |
|            |                           | Google Electra | 0.56      | 0.75   | 0.64     | 0.75     |
|            | Combination-based         | mbert          | 0.76      | 0.76   | 0.76     | 0.76     |
|            |                           | distil bert    | 0.8       | 0.8    | 0.8      | 0.8      |
|            |                           | Google Electra | 0.06      | 0.25   | 0.1      | 0.25     |
| Multiclass | Roman Translation         | mbert          | 0.75      | 0.74   | 0.74     | 0.74     |
|            |                           | distil bert    | 0.74      | 0.74   | 0.73     | 0.74     |
|            |                           | Google Electra | 0.31      | 0.31   | 0.31     | 0.31     |
|            | Nastaliq Urdu Translation | mbert          | 0.75      | 0.76   | 0.75     | 0.76     |
|            |                           | distil bert    | 0.81      | 0.81   | 0.81     | 0.81     |
|            |                           | Google Electra | 0.06      | 0.25   | 0.1      | 0.25     |

For multiclass classification using combined inputs, DistilBERT outperforms mBERT, achieving an F1-score and Accuracy of 0.80 compared to mBERT's 0.76. However, Google Electra performs very poorly (F1: 0.10, Accuracy: 0.25), suggesting its inefficiency in handling the complexity of multiclass classification with combined language inputs.

When using Roman Urdu translations, mBERT again achieves the highest scores (0.96 across all metrics), showing its robustness for binary classification in Roman Urdu. DistilBERT also performs well (0.94 on all metrics), and Google Electra surprisingly matches DistilBERT's performance in this case, which is an improvement compared to its results in the combination-based setup. This suggests Roman Urdu may align better with Google Electra's pretraining data.

In the multiclass setup with Roman Urdu, all models see a performance drop compared to binary classification. mBERT and DistilBERT have similar performance (F1-scores: ~0.74–0.75). However, Google Electra falls again, achieving only 0.31 in all metrics, showing it struggles to distinguish between multiple classes in Roman Urdu.

For binary classification using Nastaliq Urdu translations, mBERT scores slightly lower (0.95) than in the other settings but still maintains high performance. DistilBERT closely follows with 0.94 across the board. Notably, Google Electra shows an unusual pattern: while its Precision is low (0.56), its Recall is high (0.75), suggesting that it predicts the positive class too frequently, likely leading to more false positives and a middling F1-score of 0.64.

In this setting, DistilBERT outperforms all others, reaching the highest F1-score and Accuracy (0.81). This indicates its strong capacity to understand complex patterns in Nastaliq Urdu. mBERT follows with 0.75, while Google Electra once again shows very poor performance (F1: 0.10), reinforcing the model's unsuitability for multilingual multiclass tasks, especially in complex scripts like Nastaliq Urdu.

#### 4.5. Error Analysis

Table 9 presents a comprehensive comparison of different learning approaches—transfer learning, deep learning, and machine learning—across binary and multiclass depression detection tasks using three dataset configurations: combination-based, Roman Urdu translation, and Nastaliq Urdu translation.

**Transfer Learning:** Models like mBERT and DistilBERT performed exceptionally well across both binary and multiclass tasks. For binary classification, mBERT achieved outstanding results with an F1-score and Accuracy of 0.96 on both the combination-based and Roman Urdu translation datasets and slightly lower (0.95) on the Nastaliq Urdu translation. In the multiclass setting, DistilBERT and mBERT also showed promising performance. The Nastaliq Urdu translation performed best with DistilBERT (F1: 0.81), followed by the combination-based setup (0.80), while Roman Urdu translation with mBERT was slightly behind (0.74). This indicates that pretrained transformer models are highly effective for both types of classification tasks, especially when dealing with multilingual and code-mixed content.

**Deep Learning:** The performance of deep learning models, such as BiLSTM and CNN, varied across configurations. In binary classification, CNN with GloVe embeddings achieved the highest score (0.93) on the Roman Urdu translation dataset, while BiLSTM with FastText also performed well on the combination-based (0.90) and Nastaliq Urdu datasets (0.89). However, multiclass classification proved more challenging. BiLSTM with GloVe performed well on Roman Urdu translation (0.70), but CNN with GloVe yielded only moderate results on the combination-based dataset (0.51) and extremely low scores on Nastaliq Urdu translation (F1: 0.10), suggesting that deep learning models are sensitive

to script variations and struggle more with multiclass distinctions in low-resource or less-structured formats.

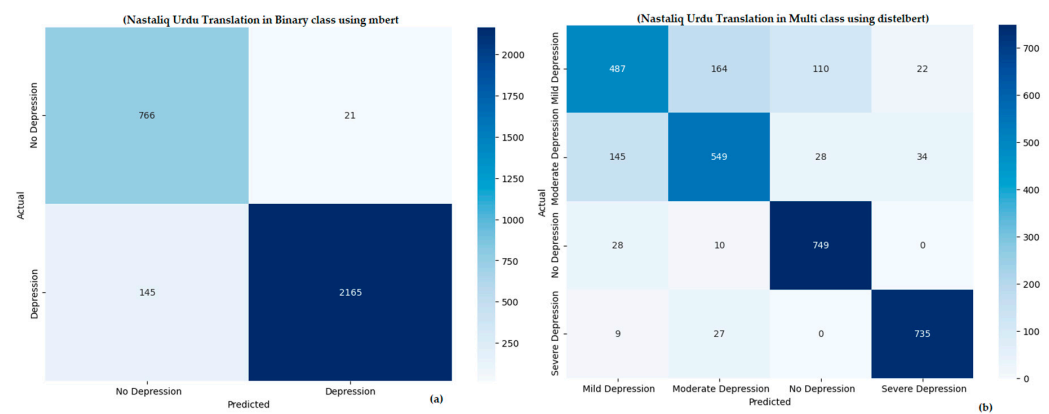
**Table 9.** Top-performing models in each learning approach.

| Task             | Language                  | Model             | F1-Score | Accuracy |
|------------------|---------------------------|-------------------|----------|----------|
| Transformer      |                           |                   |          |          |
| Binary           | Combination-based         | mBERT             | 0.96     | 0.96     |
|                  | Roman Translation         | mBERT             | 0.96     | 0.96     |
|                  | Nastaliq Urdu Translation | mBERT             | 0.95     | 0.95     |
| Multiclass       | Combination-based         | DistilBERT        | 0.80     | 0.80     |
|                  | Roman Translation         | mBERT             | 0.74     | 0.74     |
|                  | Nastaliq Urdu Translation | DistilBERT        | 0.81     | 0.81     |
| Deep learning    |                           |                   |          |          |
| Binary           | Combination-based         | BiLSTM (FastText) | 0.90     | 0.91     |
|                  | Roman Translation         | CNN (GloVe)       | 0.93     | 0.93     |
|                  | Nastaliq Urdu Translation | BiLSTM (FastText) | 0.89     | 0.89     |
| Multiclass       | Combination-based         | CNN (GloVe)       | 0.51     | 0.50     |
|                  | Roman Translation         | BiLSTM (GloVe)    | 0.70     | 0.71     |
|                  | Nastaliq Urdu Translation | CNN (FastText)    | 0.10     | 0.25     |
| Machine Learning |                           |                   |          |          |
| Binary           | Combination-based         | XGB               | 0.87     | 0.88     |
|                  | Roman Translation         | XGB               | 0.85     | 0.86     |
|                  | Nastaliq Urdu Translation | XGB               | 0.92     | 0.92     |
| Multi-class      | Combination-based         | XGB               | 0.72     | 0.73     |
|                  | Roman Translation         | XGB               | 0.69     | 0.70     |
|                  | Nastaliq Urdu Translation | XGB               | 0.78     | 0.78     |

Machine Learning: Using XGBoost (XGB), the traditional machine learning approach delivered stable and competitive results. In binary classification, XGB achieved the highest F1-score of 0.92 on the Nastaliq Urdu translation, followed by the combination-based (0.87) and Roman Urdu translation datasets (0.85). For multiclass tasks, performance was slightly lower but still consistent, with Nastaliq Urdu again leading (0.78), followed by the combination-based (0.72) and Roman Urdu translation (0.69). This shows that while not as powerful as transformers in general, XGBoost remains a reliable and interpretable model, especially effective when data is well-structured or when interpretability is a priority.

Overall, this evaluation highlights the strengths of transformer-based models for high-performance depression detection in multilingual and script-diverse settings, the nuanced capabilities of deep learning, and the robustness of traditional machine learning methods. Due to the popularity and strong performance of XGBoost in classical machine learning, we included it as a baseline since it remains one of the top-performing traditional models and provides a meaningful benchmark for comparison with deep learning approaches.

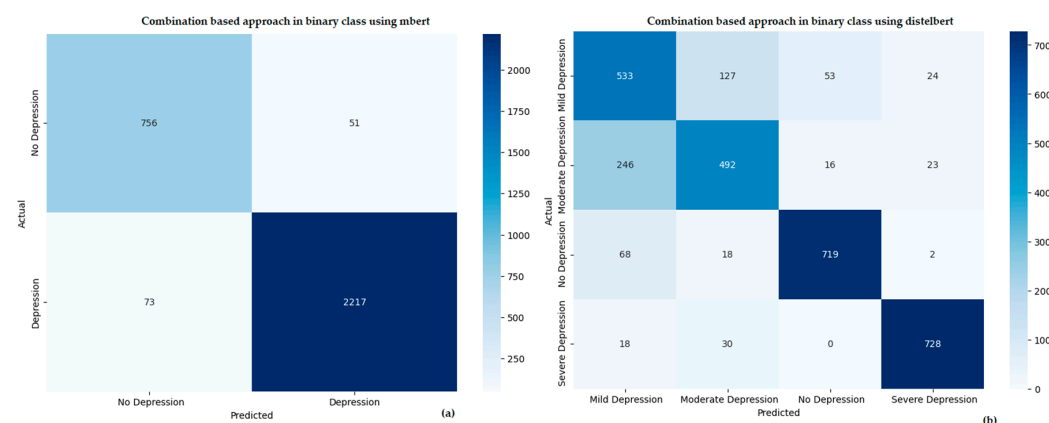
Figure 8 presents two confusion matrices that evaluate the performance of models on Nastaliq Urdu translation datasets for depression classification tasks. In Figure 8a, the confusion matrix displays results for the binary classification task using mBERT. The model correctly classified 766 instances as “No Depression” and 2165 instances as “Depression.” However, it misclassified 145 depressed tweets as non-depressed and 21 non-depressed tweets as depressed. This shows strong performance with relatively low misclassification, particularly in identifying depressed cases.



**Figure 8.** Shows the confusion matrices of mbert in binary class and distil bert in multiclass classification using the Nastaliq translation-based approach.

In Figure 8b, the confusion matrix corresponds to the multiclass classification task using DistilBERT. The model differentiates among four categories: Mild, Moderate, No Depression, and Severe Depression. While the model performs well for “No Depression” (749 correct) and “Severe Depression” (735 correct), it shows notable confusion between Mild and Moderate Depression—with 164 “Mild” cases predicted as “Moderate” and 145 “Moderate” cases predicted as “Mild.” This overlap suggests annotation ambiguity and linguistic similarity between the two classes, aligning with the reviewer’s concern about classification challenges in fine-grained emotional categories.

Figure 9 shows confusion matrices for the combination-based approach used in both binary and multiclass depression classification tasks. In Figure 9a, the confusion matrix represents the binary classification results using mBERT. The model achieved strong performance by correctly identifying 756 “No Depression” and 2217 “Depression” instances. Only 73 depressed tweets were misclassified as non-depressed, and 51 non-depressed tweets were labeled as depressed—indicating high Precision and Recall, particularly for the depression class.

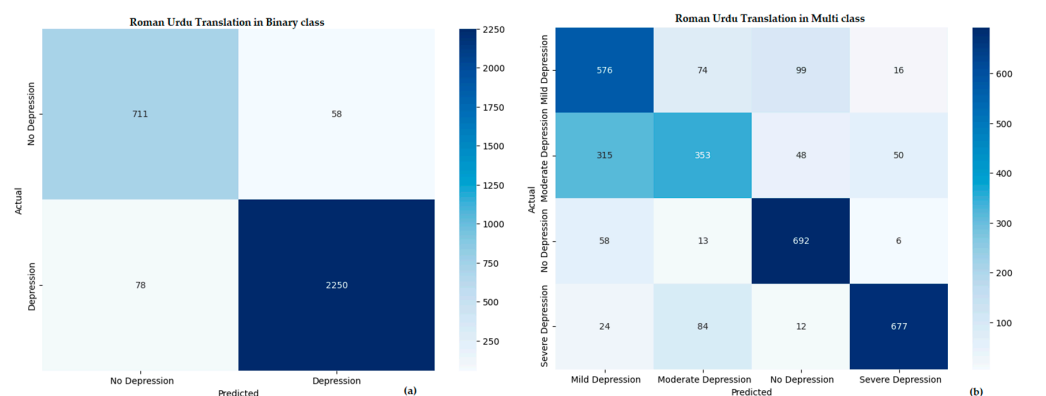


**Figure 9.** Shows the confusion matrices of mbert in binary class and distil bert in multiclass classification using the combination-based approach.

In Figure 9b, the confusion matrix shows multiclass classification results using DistilBERT. The model successfully identified “No Depression” (719) and “Severe Depression” (728) with high Accuracy. However, similar to the previous figure, there is significant confusion between “Mild” and “Moderate Depression.” For example, 246 instances of “Moderate Depression” were incorrectly labeled as “Mild,” and 127 “Mild” cases were predicted as “Moderate.” This suggests persistent challenges in distinguishing these two closely related

emotional states, likely due to subjective language and overlapping symptom descriptions in user-generated content.

Figure 10 presents two confusion matrices illustrating the performance of depression classification models applied to Roman Urdu text translations. Figure 10a shows the results for a binary classification task distinguishing between “Depression” and “No Depression.” The model correctly identified 2250 instances of depression and 711 instances of no depression, while misclassifying 78 depressive and 58 non-depressive cases, indicating strong performance. Figure 10b extends this to a multiclass classification scenario, differentiating between “Mild Depression,” “Moderate Depression,” “No Depression,” and “Severe Depression.” The confusion matrix reveals a more complex classification landscape with frequent misclassifications, particularly between mild and moderate depression (e.g., 315 actual “Mild Depression” samples were predicted as “Moderate Depression”). Despite this, the model still achieved strong predictions for “No Depression” (692 correct) and “Severe Depression” (677 correct). Overall, the binary model demonstrates higher accuracy, while the multiclass model captures finer-grained distinctions at the cost of increased confusion between similar categories.



**Figure 10.** Shows the confusion matrices of mbert in binary class and distil bert in multiclass classification using the code-mixed Roman Urdu approach.

## 5. Conclusions

This study presents a significant advancement in the field of automated depression detection by directly addressing several key limitations in prior research, particularly the lack of support for low-resource languages and script diversity. While earlier studies have primarily focused on single-script datasets—most often Roman Urdu or Standard Urdu independently—our work introduces a novel, script-converted, bilingual dataset that unifies Roman Urdu and Nastaliq Urdu. This script-level harmonization, absent in existing literature, allows for more inclusive modeling of real-world linguistic expression, especially as used on social media platforms in South Asia. The improved performance of script-conversion and combination-based methods can be theoretically attributed to the resolution of orthographic variance between scripts. By unifying the script representation, the models are able to generalize better across diverse user inputs, reducing noise and ambiguity in feature extraction. This linguistic normalization enhances semantic consistency, leading to more accurate and robust classification.

Moreover, previous research has tended to limit itself to either binary or multiclass classification without offering a framework that accommodates both. Our approach fills this methodological gap by designing a dataset and system architecture that supports both classification types, enabling more nuanced identification of depression severity levels. By conducting comprehensive experiments across machine learning, deep learning, and multilingual transformer-based models, we have not only validated our dataset but also

demonstrated the clear advantages of script-aware and language-integrated modeling—something missing from earlier efforts.

Importantly, this study moves beyond proof of concept toward practical application, showing that culturally and linguistically tailored NLP solutions can play a meaningful role in early detection and intervention for mental health issues in underserved populations. Our findings underscore the need to embed local language and script variation into mental health technologies to ensure broader reach and higher accuracy. Future research will focus on expanding the dataset to include a wider variety of social media platforms and longer time frames beyond the COVID-19 period to improve the model's generalizability and reduce sampling bias. Additionally, we plan to integrate clinical and psychological expertise into the annotation protocols to enhance label accuracy and address class ambiguities. We also intend to perform an in-depth qualitative analysis of misclassified examples to uncover underlying linguistic and contextual challenges specific to Roman Urdu and Nastaliq Urdu. Furthermore, advanced deep learning techniques and contextual embeddings will be explored to further improve classification performance.

**Author Contributions:** Conceptualization, M.A. and P.B.; methodology, M.A., P.B., F.U. and I.B.; software, M.A.; validation, G.S. and F.U.; formal analysis, M.A.; investigation, I.B.; resources, G.S. and I.B.; data curation, M.A. and P.B.; writing—original draft preparation, M.A. and P.B.; writing—review and editing, M.A. and F.U.; visualization, M.A. and F.U.; supervision, I.B.; project administration, I.B. All authors have read and agreed to the published version of the manuscript.

**Funding:** This research did not receive any external funding.

**Institutional Review Board Statement:** Not applicable.

**Informed Consent Statement:** Not applicable.

**Data Availability Statement:** The dataset used in this study is not publicly available due to ongoing research but can be provided upon reasonable request. Interested researchers should contact the first author at mahmad2024@cic.ipn.mx, Centro de Investigación en Computación, Instituto Politécnico Nacional (CIC-PN), Mexico City 07738, Mexico. Requests must include a detailed description of the intended use and the requester's institutional affiliation.

**Acknowledgments:** The work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONAHCYT, Mexico, grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONAHCYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico, and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

**Conflicts of Interest:** The authors declare no competing interests.

## References

1. Kumar, R.; Mukherjee, S.; Choi, T.M.; Dhamotharan, L. Mining voices from self-expressed messages on social-media: Diagnostics of mental distress during COVID-19. *Decis. Support Syst.* **2022**, *162*, 113792. [[CrossRef](#)] [[PubMed](#)]
2. Yang, X.; Joukova, A.; Ayanso, A.; Zihayat, M. Social influence-based contrast language analysis framework for clinical decision support systems. *Decis. Support Syst.* **2022**, *159*, 113813. [[CrossRef](#)]
3. Guohou, S.; Lina, Z.; Dongsong, Z. What reveals about depression level? The role of multimodal features at the level of interview questions. *Inf. Manag.* **2020**, *57*, 103349. [[CrossRef](#)]
4. Costantini, L.; Pasquarella, C.; Odone, A.; Colucci, M.E.; Costanza, A.; Serafini, G.; Aguglia, A.; Murri, M.B.; Brakoulias, V.; Amore, M.; et al. Screening for depression in primary care with Patient Health Questionnaire-9 (PHQ-9): A systematic review. *J. Affect. Disord.* **2021**, *279*, 473–483. [[CrossRef](#)]

5. Pile, V.; Williamson, G.; Saunders, A.; Holmes, E.A.; Lau, J.Y. Active Ingredients for Addressing Youth Anxiety and Depression 3. Harnessing emotional mental imagery to reduce anxiety and depression in young people: An integrative review of progress and promise. *Lancet Psychiatry* **2021**, *8*, 836–852. [\[CrossRef\]](#)
6. Nelson, J.; Klumpparendt, A.; Doebler, P.; Ehring, T. Everyday emotional dynamics in major depression. *Emotion* **2020**, *20*, 179. [\[CrossRef\]](#)
7. Balahur, A.; Hermida, J.M.; Montoyo, A. Detecting implicit expressions of emotion in text: A comparative analysis. *Decis. Support Syst.* **2012**, *53*, 742–753. [\[CrossRef\]](#)
8. Cui, S.; Yin, Y.; Wang, D.; Li, Z.; Wang, Y. A stacking-based ensemble learning method for earthquake casualty prediction. *Appl. Soft Comput.* **2021**, *101*, 107038. [\[CrossRef\]](#)
9. Uban, A.S.; Chulvi, B.; Rosso, P. An emotion and cognitive based analysis of mental health disorders from social media data. *Future Gener. Comput. Syst.* **2021**, *124*, 480–494. [\[CrossRef\]](#)
10. Mahoney, J.; Le Moignan, E.; Long, K.; Wilson, M.; Barnett, J.; Vines, J.; Lawson, S. Feeling alone among 317 million others: Disclosures of loneliness on Twitter. *Comput. Human. Behav.* **2019**, *98*, 20–30. [\[CrossRef\]](#)
11. Silveira, B.; Silva, H.S.; Murai, F.; da Silva, A.P.C. Predicting user emotional tone in mental disorder online communities. *Future Gener. Comput. Syst.* **2021**, *125*, 641–651. [\[CrossRef\]](#)
12. Ahmad, M.; Farid, H.; Ameer, I.; Ullah, F.; Muzamil, M.; Jalal, M.; Hamza, A.; Batyrshin, I.; Sidorov, G. UE-NER-2025: A GPT-based Approach to Multilingual Named Entity Recognition on Urdu and English. *IEEE Access* **2025**, *13*, 111175–111186. [\[CrossRef\]](#)
13. Joshi, M.L.; Kanoongo, N. Depression detection using emotional artificial intelligence and machine learning: A closer review. *Mater. Today Proc.* **2022**, *58*, 217–226. [\[CrossRef\]](#)
14. Mehmood, K. On Multi-Domain Sentence Level Sentiment Analysis for Code-Mixed Roman Urdu. Ph.D. Dissertation, UNSW, Sydney, NSW, Australia, 2021.
15. Feuerriegel, S.; Maarouf, A.; Bär, D.; Geissler, D.; Schweisthal, J.; Pröllochs, N.; Robertson, C.E.; Rathje, S.; Hartmann, J.; Mohammad, S.M.; et al. Using natural language processing to analyse text data in behavioural science. *Nat. Rev. Psychol.* **2025**, *4*, 96–111. [\[CrossRef\]](#)
16. Lázaro, F.L.; Madeira, T.; Melicio, R.; Valério, D.; Santos, L.F. Identifying Human Factors in Aviation Accidents with Natural Language Processing and Machine Learning Models. *Aerospace* **2025**, *12*, 106. [\[CrossRef\]](#)
17. Nandwani, P.; Verma, R. A review on sentiment analysis and emotion detection from text. *Soc. Netw. Anal. Min.* **2021**, *11*, 81. [\[CrossRef\]](#)
18. Gaing, B.; Syal, V.; Padgalwar, S. Emotion detection and analysis on social media. *arXiv* **2019**, arXiv:1901.08458.
19. Kusal, S.; Patil, S.; Kotecha, K. Multimodal text-emoji fusion using deep neural networks for text-based emotion detection in online communication. *J. Big Data* **2025**, *12*, 1–25. [\[CrossRef\]](#)
20. Mobbs, R.; Makris, D.; Argyriou, V. Emotion Recognition and Generation: A Comprehensive Review of Face, Speech, and Text Modalities. *arXiv* **2025**, arXiv:2502.06803.
21. Kanaparthi, S.D.; Patle, A.; Naik, K.J. Prediction and detection of emotional tone in online social media mental disorder groups using regression and recurrent neural networks. *Multimed. Tools Appl.* **2023**, *82*, 43819–43839. [\[CrossRef\]](#) [\[PubMed\]](#)
22. Usman, M.; Ahmad, M.; Sidorov, G.; Gelbukh, I.; Tellez, R.Q. A Large Language Model-Based Approach for Multilingual Hate Speech Detection on Social Media. *Computers* **2025**, *14*, 279. [\[CrossRef\]](#)
23. Barrios González, E.; Tovar Vidal, M.; Contreras González, M. Detection of Depression Using Depression Indicators. *Comput. Sist.* **2025**, *29*, 135–143. [\[CrossRef\]](#)
24. Lauderdale, S.; Griffin, S.A.; Lahman, K.R.; Mbaba, E.; Tomlinson, S. Unveiling public stigma for borderline personality disorder: A comparative study of artificial intelligence and mental health care providers. *Personal. Ment. Health* **2025**, *19*, e70018. [\[CrossRef\]](#)
25. Lambert, S.I.; Madi, M.; Sopka, S.; Lenes, A.; Stange, H.; Buszello, C.P.; Stephan, A. An integrative review on the acceptance of artificial intelligence among healthcare professionals in hospitals. *npj Digit. Med.* **2023**, *6*, 111. [\[CrossRef\]](#) [\[PubMed\]](#)
26. Saragih, I.D.; Hsiao, C.T.; Fann, W.C.; Hsu, C.M.; Saragih, I.S.; Lee, B.O. Impacts of interprofessional education on collaborative practice of healthcare professionals: A systematic review and meta-analysis. *Nurse Educ. Today* **2024**, *136*, 106136. [\[CrossRef\]](#) [\[PubMed\]](#)
27. New World Encyclopedia. (n.d.). Urdu. New World Encyclopedia. Retrieved 23 November 2024. Available online: <https://www.newworldencyclopedia.org/entry/Urdu> (accessed on 1 January 2020).
28. Garg, M. Towards mental health analysis in social media for low-resourced languages. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2024**, *23*, 1–22. [\[CrossRef\]](#)
29. Sattar, H.; Munir, M.; Malik, M.K.; Nasar, Z. Detecting Empathy in Roman Urdu Using Transfer Learning. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2025**, *24*, 1–36. [\[CrossRef\]](#)
30. Mohmand, R.; Habib, U.; Usman, M.; Baili, J.; Nam, Y. A Deep Learning Approach for Automated Depression Assessment Using Roman Urdu. *IEEE Access* **2024**, *12*, 193387–193401. [\[CrossRef\]](#)

31. Martinez-Navarro, J.A.; Rubio-Espino, E.; Sossa-Azuela, J.H.; Ponce-Ponce, V.H.; Molina-Lozano, H.; Garcia-Sebastian, L.M. Comparison of Neural Networks for Emotion Detection. *Comput. Sist.* **2023**, *27*, 653–665. [CrossRef]
32. Qasim, A.; Mehak, G.; Hussain, N.; Gelbukh, A.; Sidorov, G. Detection of Depression Severity in Social Media Text Using Transformer-Based Models. *Information* **2025**, *16*, 114. [CrossRef]
33. Chhabra, H.; Chauhan, U.; Jain, P.; Sharma, L.D.; Dev, A. Machine Learning Assisted EEG Signal Classification for Automated Diagnosis of Mental Stress. In *Artificial Intelligence in Biomedical and Modern Healthcare Informatics*; Academic Press: London, UK, 2025; pp. 447–454.
34. Sharma, D.; Nath, T.; Gupta, V.; Singh, V.K. Hate Speech Detection Research in South Asian Languages: A Survey of Tasks, Datasets and Methods. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.* **2025**, *24*, 1–44. [CrossRef]
35. Abbas, R.; Bracke, P.; Delaruelle, K. Urdu Translation and Psychometric Validation of Stigma Beliefs Scale. *Res. Soc. Work. Pract.* **2025**, 10497315241307615. [CrossRef]
36. Rehmani, F.; Shaheen, Q.; Anwar, M.; Faheem, M.; Bhatti, S.S. Depression detection with machine learning of structural and non-structural dual languages. *Healthc. Technol. Lett.* **2024**, *11*, 218–226. [CrossRef] [PubMed]
37. Noor, S. Depression Detection and Explanation with Machine Learning. Master's Thesis, Stevens Institute of Technology, Hoboken, NJ, USA, 2021. Available online: <https://www.proquest.com/openview/b8e453dd2d2eab60a612de78377c0d59/1?pq-origsite=gscholar&cbl=18750&diss=y> (accessed on 1 January 2020).
38. Tavchioski, I.; Robnik-Šikonja, M.; Pollak, S. Detection of depression on social networks using transformers and ensembles. *arXiv* **2023**, arXiv:2305.05325. [CrossRef]
39. Khaparde, A.A.; Das, R.; Bhargava, R. Transformer based approach for depression detection. In Proceedings of the 2023 15th International Conference on Developments in eSystems Engineering (DeSE), Baghdad, Iraq, 9–12 January 2023; pp. 369–374.
40. Govindasamy, K.A.; Palanichamy, N. Depression detection using machine learning techniques on twitter data. In Proceedings of the 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS), Madurai, India, 6–8 May 2021; pp. 960–966.
41. Khokhar, I.A.; Garcia, M.I.M. Decoding Depression: A Psycholinguistic Lexical Analysis of the Urdu Speech. *Linguist. Forum J. Linguist.* **2024**, *6*, 48–70.
42. Pedregosa, F. Scikit-learn: Machine learning in python Fabian. *J. Mach. Learn. Res.* **2011**, *12*, 2825.
43. Abadi, M.; Barham, P.; Chen, J.; Chen, Z.; Davis, A.; Dean, J.; Zheng, X. {TensorFlow}: A system for {Large-Scale} machine learning. In Proceedings of the 12th USENIX Symposium on Operating Systems Design and Implementation (OSDI 16), Savannah, GA, USA, 2–4 November 2016; pp. 265–283.
44. Gulli, A.; Pal, S. *Deep Learning with Keras*; Packt Publishing Ltd.: Birmingham, UK, 2017.

**Disclaimer/Publisher's Note:** The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.