

Hope Speech Detection in code-mixed Roman Urdu tweets: A Positive Turn in Natural Language Processing

Muhammad Ahmad¹, Muhammad Waqas², Ameer Hamza², Ildar Batyrshin¹, Grigori Sidorov^{1,*}

¹ Centro de Investigación en Computación, Instituto Politécnico Nacional (CIC-PN), Mexico City 07738, Mexico.

² Department of Software Engineering, the Islamia University of Bahawalpur, 63100, Pakistan
Correspondence: sidorov@cic.ipn.mx

Abstract. Hope is a positive emotional state that involves the expectation of desirable outcomes in the future, whereas hope speech is any form of communication that promotes optimism, resilience, and support, especially in challenging or adverse contexts. Although the detection of hope speech has gained increasing attention within Natural Language Processing (NLP), existing research predominantly focuses on high-resource languages and standardized scripts, often overlooking informal and underrepresented forms such as Roman Urdu. To the best of our knowledge, this is the first study to tackle hope speech detection in code-mixed Roman Urdu by introducing a carefully annotated dataset, thereby filling a critical void in the literature and laying the groundwork for inclusive NLP research in low-resource, informal language varieties. Specifically, this study makes four key contributions: (1) it presents the first work on hope speech detection in Roman Urdu by introducing a multi-class annotated dataset that includes Generalized Hope, Realistic Hope, Unrealistic Hope, and Not Hope categories. (2) It explores the psychological underpinnings of hope and systematically analyzes its linguistic expressions in Code-mixed Roman Urdu, ensuring that the dataset construction is grounded in both theoretical insight and real-world language use. (3) It designs, implements, and evaluates a custom attention-based transformer architecture optimized to handle the syntactic and semantic variability of Roman Urdu, and employs 5-fold cross-validation for robust performance assessment. Based on the results, our proposed model, XLM-R, achieves the highest performance with a cross-validation score of 0.78, outperforming the baseline SVM (0.75) using TF-IDF and BiLSTM (0.76) leveraging pretrained GloVe embeddings. This corresponds to a 4% improvement over baseline and a 2.63% gain compared to BiLSTM. (4) We perform a t-test on our top-performing model to determine whether the observed performance improvement is statistically significant or due to chance. Based on statistical analysis, our proposed model proves to be effective.

Keywords: hope speech, social media, data mining, code-mixed Roman Urdu, twitter, X, machine learning, deep learning, transfer learning.

1 Introduction

Hope serves as a vital element in human psychology, empowering people to confront and endure life's difficulties [1]. The idea of hope has been construed in multiple ways across scholarly work, consistently emphasizing its essential role in life and its link to positive expectations for the future [2], [3]. It motivates people to act on their goals and can act as a protective factor against feelings of hopelessness and mental distress [4]. Furthermore, hope functions as a cognitive-emotional mechanism that supports individuals in coping with adverse or unforeseen circumstances [5].

The growing field of Natural Language Processing (NLP) has recently begun to explore the detection and analysis of hope speech within social media, marking it as an emerging and intriguing research direction. Although significant work has been done on different topics such as health care [6] [7], emotion recognition [8-10], hate speech detection [11-13], question answering [24], cyberbullying [26] and abusive language identification [8] [25], the focused investigation into hope speech remains largely unexplored.

Although hope speech has been examined in multiple languages including English [15], Standard Urdu [14] Arabic [20], Spanish [16], Bengali [17], and several Dravidian languages, [18], [19]—there has been little to no scholarly focus on its detection and analysis in the Code-mixed Roman Urdu language. This represents a notable and pressing research gap in the current literature. Roman Urdu, which is Urdu written using the Latin (Roman) script instead of its native Perso-Arabic script, is widely used informally in digital communication, particularly on social media, messaging apps, and forums. While Urdu itself is spoken by over 230 million people worldwide [23]—mainly in Pakistan and India—Roman Urdu is not a distinct spoken language but rather a writing style that has become popular among Urdu speakers in over 20 countries due to the accessibility of Latin-script keyboards. The language plays a big role in cultural, political, and social talks, especially in South Asia.

Social media platforms have significantly contributed to the proliferation of code-mixed Roman Urdu. The extensive use of code-mixed Roman Urdu across different platforms including YouTube, Instagram, X (formerly Twitter), Facebook, TikTok, and many online forums has become a major source of user-generated content that holds substantial potential for NLP research. However, this rising volume of textual content, Roman Urdu remains still unexplored within the NLP domain.

Roman Urdu presents unique challenges for Natural Language Processing (NLP) due to its lack of standardized orthography, frequent code-mixing with English, and informal syntactic structure. A single word has multiple spellings—for instance, “kaise,” “kesay,” and “kese” all represent “how”—which complicates core NLP tasks such as tokenization, language modeling, and classification. These challenges are further compounded by the scarcity of annotated datasets, language-specific tools, and dedicated NLP models for Roman Urdu. In the context of hope speech detection, the difficulties are even more pronounced. Unlike hate speech, which often contains explicit offensive language, hope speech is typically subtle, context-dependent, and metaphorical. It often appears in poetic or abstract expressions such as “Andheray mein bhi hum ek doosray ke liye roshni banay rahain” (Even in darkness, we remain light for

one another), “Umeed ka diya kabhi bujhne na dena” (Never let the lamp of hope extinguish), or “Kal behtar hoga, bas bharosa rakho” (Tomorrow will be better, just have faith). These expressions convey positivity through cultural and emotional nuance that is difficult for automated systems to interpret. Therefore, developing effective NLP solutions for Roman Urdu—particularly for tasks like hope speech identification—requires models capable of understanding not just lexical patterns but also deeper contextual and cultural meanings embedded in the language.

To achieve this objective, researchers have turned to a range of computational approaches. Traditional machine learning methods like Support Vector Machines (SVM), and Logistic Regression have been used with carefully crafted features such as n-grams, part-of-speech tags, and sentiment scores. These techniques try to capture the linguistic patterns that might signal hopeful content. More recently, deep learning approaches, especially transformer-based models like BERT (Bidirectional Encoder Representations from Transformers), RoBERTa, and XLM-R, have shown promising results in hope speech detection tasks. These models are capable of learning nuanced patterns in language by leveraging large pre-trained corpora and fine-tuning them on hope speech datasets. Fine-tuning on domain-specific corpora, such as social media posts during humanitarian crises, has enabled these models to better understand the emotional and cultural context of hopeful messages.

In this study, we created a robust hope speech dataset specifically for the code-mixed Roman Urdu language. To the best of our knowledge, there is currently no publicly available dataset dedicated to hope speech detection in code-mixed Roman Urdu, marking this as the first systematic effort to explore hope speech within code-mixed Roman-script languages which is overlooked in previous research. Our dataset fills this critical gap by capturing a wide range of expressions, encompassing both hopeful and non-hopeful sentiments from posts on X (formerly Twitter), and categorizing them into a multi-class setting that includes Generalized Hope, Realistic Hope, Unrealistic Hope, and Not Hope. This focus allows for a nuanced understanding of how hope manifests linguistically in social media content written in code-mixed Roman Urdu. To ensure the quality and reliability of the dataset, we adhered to well-established annotation guidelines inspired by the methodologies presented in [14] and [15], incorporating expert reviews and iterative refinements. The creation of this dataset not only provides a valuable resource for future research in hope speech detection but also sets a foundation for developing culturally and linguistically sensitive NLP models tailored to code-mixed Roman Urdu social media content.

This study makes the following contributions:

1. To the best of our knowledge, there is no prior work on hope speech detection in code-mixed Roman Urdu; therefore, we present the first-ever manually annotated, multi-class dataset specifically designed for hope speech detection in code-mixed Roman Urdu texts.
2. We explore the psychological basis of hope speech and conceptualize hope as a form of expectation, which informs the design of our dataset and the approach to classification tasks.

3. We designed, proposed, implemented, and evaluated an advanced language-based model using a custom attention mechanism optimized to handle the syntactic and semantic variability of Roman Urdu that helps users actively encourage and amplify positive and hopeful expressions on social media in Roman Urdu.
4. We conducted a comprehensive analysis and performance evaluation by employing various techniques along with visualization methods;
5. Based on the results, our proposed model, XLM-R, achieves the highest performance with a cross-validation score of 0.78, outperforming the baseline SVM (0.75) using n-gram TF-IDF and BiLSTM (0.76) leveraging pre-trained GloVe embeddings. The results show a 4% improvement over the baseline and a 2.63% gain compared to BiLSTM.
6. We performed a statistical t-test on our top-performing model to determine whether the observed performance improvement was statistically significant or merely due to chance.

The remaining sections of the study are as follows. The “Literature Survey” examines previous research on hope speech detection. The “Methodology and Design” section outlines the approach and system design. Findings are detailed in the “Results and Analysis” section. The “Limitations” section discusses the constraints of our proposed solution. Finally, the study concludes with “Conclusion and Future Work,” summarizing the key outcomes and suggesting directions for further research.

2 Literature Review

Ahmad et al. [20] created multilingual dataset named as Posi-Vox-2024, for hope speech detection in standard Urdu, English, and Arabic languages. They applied advance transformer models using a fine-tuned approach to classify hope speech into binary and multiclass categories. Their proposed model (Bert-based-uncased) outperformed traditional machine learning models such as Logistic Regression in both classification tasks.

Ahmad et al. [14] developed a multilingual hope speech dataset in English and standard Urdu and applied a unique multilingual technique to addressed multilingual challenges. They employed different machine learning models using TF-IDF, deep learning using pre-trained word embeddings such as Glove and FastText, and advance language-based transfer learning models to benchmark performance. Their proposed model such A BERT-based model outperformed traditional models, achieving 87% accuracy for English and 79% for Urdu.

Divakaran et al. [21] participated in the HOPE at IberLEF 2024 workshop task in English and Spanish hope speech. They employed multiple machine learning using TF-IDF and advance language-based models using pre-trained word embeddings. Their models achieved top-10 rankings in all tasks, with the highest macro F1 score of 0.82 in binary classification.

Balouchzahi et al. [15] created a hope speech dataset of English tweets and categorized into binary and multiclass classifications. They created detailed annotation guidelines to create dataset. They utilized different traditional machine learning using TF-IDF, deep learning using pretrained word embeddings and advance language-based models using advance contextual embeddings to bench mark the dataset. Their Results showed that contextual embeddings outperformed others, highlighting the dataset's quality and complexity.

Chakravarthi et al. [22] focused on promoting positivity rather than negativity by creating a multilingual dataset for hope speech detection using YouTube comments. They developed a custom deep learning-based model such as CNN using concatenated embeddings from T5-Sentence and compared its results with machine learning classifiers. Their CNN-based model shoed the bench mark performance, with macro F1-scores of 0.75 (English), 0.62 (Tamil), and 0.67 (Malayalam).

While previous studies have made significant strides in developing hope speech datasets in multiple languages—including Urdu written in the Perso-Arabic script—none have focused on Roman Urdu, a widely used but under-resourced script in digital communication. Unlike Ahmad et al. [14, 20] and Divakaran et al. [21], whose datasets center on formal scripts and established multilingual NLP frameworks, our study is the first to construct a dedicated hope speech dataset specifically for Roman Urdu, addressing a critical and previously overlooked gap. Moreover, while Balouchzahi et al. [15] and Chakravarthi et al. [22] applied a variety of traditional and deep learning techniques, they did not consider the unique linguistic and cultural nuances present in Roman-script languages. Our work also differs in its psychological framing of hope as a cognitive expectation in Roman Urdu discourse, providing deeper linguistic insight. Additionally, to the best of our knowledge, no existing research has employed a custom attention-based model specifically designed for code-mixed Roman Urdu. This tailored approach allows the model to better capture the informal structure, spelling variations, and code-mixed nature typical of Roman Urdu texts. As a result, it enhances contextual understanding, improves classification accuracy, and ensures more effective detection of subtle linguistic cues associated with hope speech—benefits that are often missed by generic multilingual models.

3 Methodology

3.1 Construction of Dataset

The data gathering stage was a crucial step in the construction of our code-mixed Roman Urdu hope speech dataset. We started by collecting a comprehensive list of keywords such as umeed (hope), acha waqt aayega (goo time will come), hausla (Courage), himmat na haaro (Don't lose courage), khuda pe bharosa (Trust in God), sabar karo (Be patient), zindagi haseen hai (Life is beautiful), Allah madad karega (God will help), positive raho (Stay positive), (umeed se raho) be hopeful, Kabhi haar mat mano (never give up), and mazboot raho (stay strong). These keywords are commonly used to express hope, encouragement, and optimism in Roman Urdu language. In parallel, to develop a balanced dataset, we also collected tweets that represent Not Hope speech—

tweets that reflect negativity, despair, frustration, or emotional hopelessness. For not hope, we used keywords such as umeed nahi hai (there is no hope), thak gaya hoon (I am tired), zindagi bekaar hai (life is useless), ab kuch acha nahi hoga (nothing good will happen now), haar gaya hoon (I have lost), kya faida (what's the use), Allah ne chhor diya hai (God has abandoned me), nafrat (hatred), dil toot gaya (heart is broken), and sab kuch khatam ho gaya (everything is over). In the second phase, we utilized the popular Twitter API named as Tweepy along with Python-code to extract tweets containing these keywords. We employed filters to collect only publicly available tweets, ensuring user privacy and compliance with Twitter's data usage policies. The tweets were collected over the period of January 2023 to February 2024 to ensure temporal diversity and cover different socio-political events, public reactions, and personal expressions. As a result, we targeted only original tweets—excluding replies, retweets, and promotional content—to ensure the authenticity and relevance of the dataset, which comprised 30,000 tweets. Figure 1 illustrates the methodology and design of the proposed study.

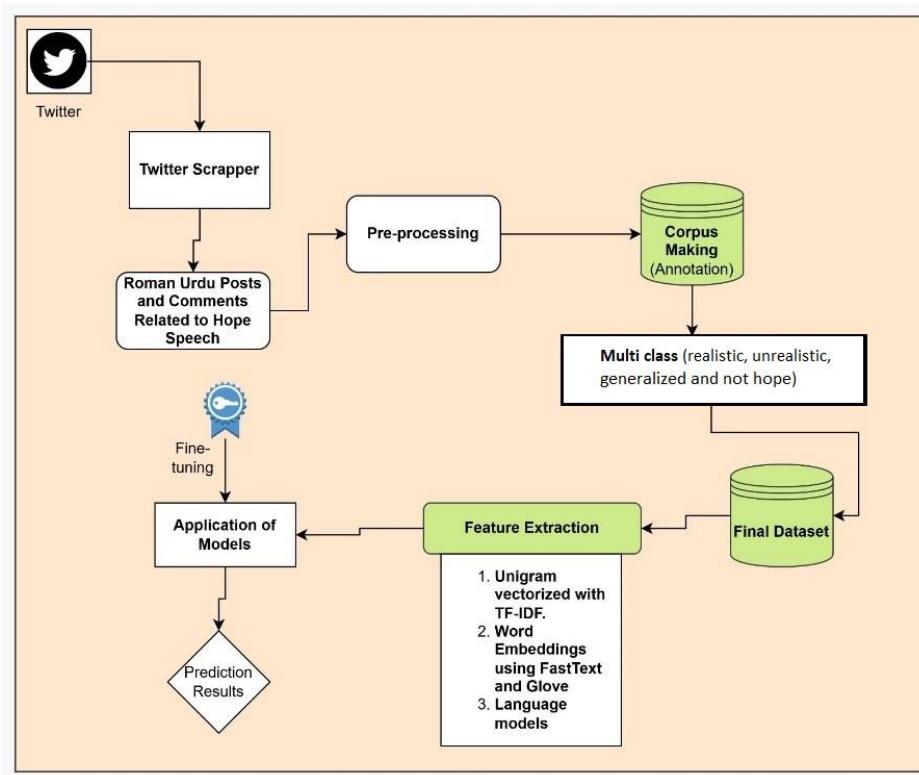


Fig. 1. Architecture of the proposed methodology and design

3.2 Pre-Processing

This section describes the preprocessing steps employed in this study to create a well-prepared code-mixed Roman Urdu dataset. The figure 2 illustrates the pre-processing pipeline applied to the dataset. The process begins with the raw text data. The first step involves removing hashtags, numbers, and punctuation marks to eliminate noise and irrelevant symbols. Next, short and repetitive character sequences such as bhttt (A lot)—often common in informal online text—are filtered out to improve text quality. Following this, hyperlinks embedded within posts are removed, as they don't contribute to meaningful linguistic features for NLP task. In the final stage, any post containing fewer than 15 characters was discarded, based on the assumption that such short texts are unlikely to convey substantial semantic value or capture the full intent of the message. After completing these steps, the remaining content forms the Pre-processed Dataset, which is cleaner, more consistent, and better suited for further analysis and modeling in hope speech detection tasks.

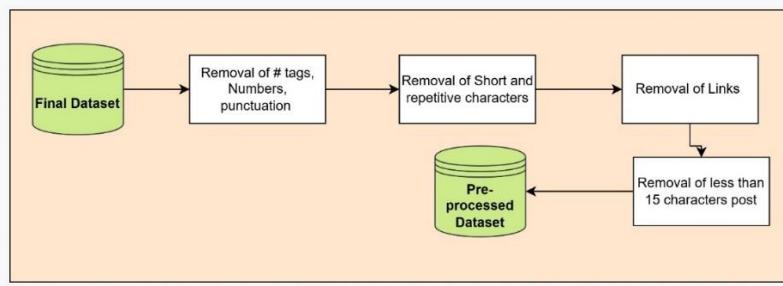


Fig. 2. Pre-processing steps utilized in this study.

3.3 Data Annotation

Annotation is the process of labeling a dataset into predefined categories so that machine learning models can understand and learn from it to make better decisions. In this study, we annotated each code-mixed Roman Urdu tweet into one of four categories such as Generalized Hope, Realistic Hope, Unrealistic Hope, or Not Hope. This step was crucial because machine learning models cannot interpret raw text on their own—they require labeled examples to learn the patterns associated with different types of language. For the annotation process, we selected postgraduate Computer Science students who had prior experience in dataset labeling, as referenced in [14] and [20], and who are native speakers of Urdu. This selection helped us to maintain linguistic accuracy and cultural relevance throughout the annotation process. The final label for each tweet was assigned using majority voting among the annotators.

3.4 Annotation Procedure

For the annotation process, we adopted the guidelines for multi class as realistic hope, unrealistic hope, generalized hope and not hope speech from the previously published papers [20] [15] to ensure consistency and clarity. We engaged three annotators, all postgraduate students and native speakers of Urdu language. Each annotator independently labeled the tweets according to the provided guidelines. To manage the annotation efficiently, we created separate Google Sheets for each annotator, as Google Sheets allowed real-time collaboration, easy progress tracking, and simple comparison of labels across annotators. To resolve disagreements, we scheduled weekly meetings where annotators discussed uncertain samples together. After these discussions, the final labels were assigned based on majority voting. This approach helped maintain high-quality and consistent labeling. Additionally, the annotators' prior experience with the dataset in [14] [20] was instrumental in guiding these decisions and ensuring linguistic and cultural accuracy throughout the annotation process. Table 1 presents representative examples from the code-mixed Roman Urdu Hope Speech Dataset, illustrating each of the four categories.

Table 1. Examples selected from the code-mixed Roman Urdu Hope Speech Dataset to illustrate each category.

Category	Roman Urdu Tweet	English Translation
Not-Hope	Yeh issue kabhi theek nahi hogा, har cheez kharab ho rahi hai.	This issue will never get fixed; everything is getting worse.
Realistic Hope	Agar hum sab mil kar koshish karein to situation better ho sakti hai.	If we all try together, things can get better.
Unrealistic Hope	Jab Nawaz Sharif aaye ga, to har ghar mein dollars ki barsaat ho gi.	When Nawaz Sharif returns, every house will be showered with dollars
Generalized Hope	Achi thinking rakho, waqt zaroor behtar hoga.	Keep a positive mindset; time will surely get better.

3.5 Inter-annotator agreement

During the annotation process, annotators may occasionally disagree in their decisions. It is important to carefully examine and evaluate these differences to gain valuable insights from their collective input. To track the performance of annotators we calculated the inter-annotator agreement between the three annotators to ensure the reliability and consistency of the labeled dataset in Roman Urdu. We calculated Fleiss' Kappa score to assess the level of agreement beyond chance. Fleiss' Kappa is used to measure the

level of agreement between three or more raters when assigning categorical ratings to a set of items. The resulting average Kappa value was 0.81, which indicates a substantial level of agreement among the annotators. This high score demonstrates that the annotators were largely aligned in their understanding of the labeling criteria. We also calculate pair-wise inter-annotator agreement. Annotator 1 and Annotator 2 achieved the highest agreement with a score of 0.91, while Annotator 2 and Annotator 3 scored 0.75, and Annotator 1 and Annotator 3 scored 0.77 as shown in table 2. The overall average agreement was 0.81, reflecting a high level of consistency in the annotation process. Table 3 shows the general interpretation of the Kappa values.

Table 2. Pairwise inter-annotator agreement scores between three annotators.

Annotators	Agreement Score
Annotator 1 & Annotator 2	0.91
Annotator 2 & Annotator 3	0.75
Annotator 1 & Annotator 3	0.77

Table 3. Interpretation of the Kappa values.

Kappa value range	Interpretation
1.0	Perfect agreement
0.80 to 1.0	Substantial agreement
0.60 to 0.80	Moderate agreement
0.40 to 0.60	Fair agreement
< 40	Poor agreement

3.6 Ethical Concerns

Social media content related to racial, ethnic, religious, economic minorities, and individuals with disabilities contains sensitive information. To protect user privacy, all identifiable details—such as names and religious references—were removed, except in the case of public figures. No attempts were made to contact original content creators. The dataset will be made available only to researchers who agreed to comply with ethical guidelines for research.

3.7 Dataset Statistics

Figure 3 presents detailed statistics of the code-mixed Roman Urdu dataset used in this study for multi-class hate speech detection. It contains a total of 4,953 social media posts, comprising 108,249 words and 566,343 characters. The vocabulary size—indicative of the dataset's lexical richness—amounts to 13,581 unique words. On average,

each sentence in the dataset contains 19.15 words, while each post comprises 1.14 sentences, suggesting that most posts are short and concise. Additionally, the average number of characters per word is 5.23, and each post averages 114.3 characters. The logarithmic scale of the y-axis emphasizes the wide range in values, from linguistic averages to overall dataset sizes, offering a comprehensive overview of the dataset's structural properties. Figure 4 shows the word cloud representing the most frequently occurring words in the dataset, providing insights into the dominant vocabulary. Figure 5 illustrates the label distribution across each class, highlighting the balance or imbalance among the hope speech categories within the dataset.

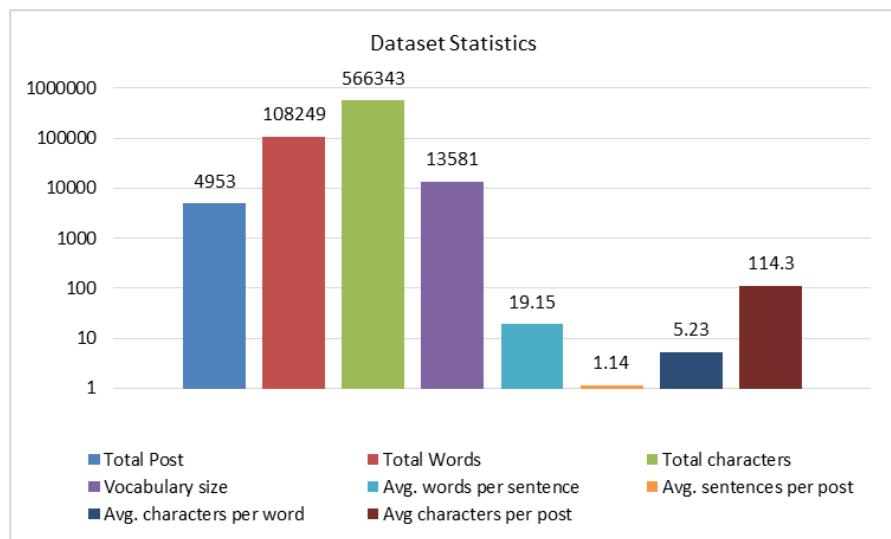


Fig. 3. Dataset Statistics



Fig. 4. Word cloud representing the most frequently occurring words in the dataset.

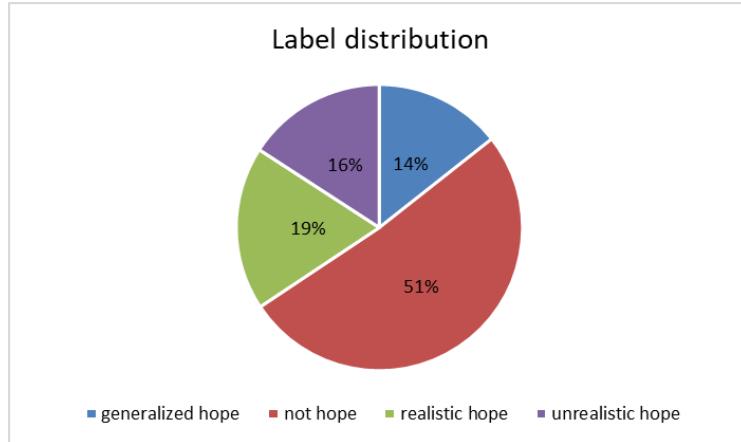


Fig. 5. Label distribution

3.8 Application of models Training and testing phase

Figure 6 presents the overall workflow for building and evaluating machine learning, deep learning, and advanced language-based models on the code-mixed Roman Urdu hope speech dataset. The process begins with the final dataset, which is evaluated using 5-fold cross-validation to ensure robustness and mitigate performance variance due to random data partitioning. In this setup, the dataset is split into five equal parts; in each iteration, four folds are used for training (the “learning set”), and one fold is used for evaluation (the “testing set”). This procedure is repeated five times, with each fold serving once as the testing set, and the results are averaged to obtain reliable performance metrics.

The learning set is used to train various models, including traditional Machine Learning algorithms such as Support Vector Machine (SVM) with both linear and RBF kernels, XGBoost (XGB), Logistic Regression (LR), and Decision Tree (DT) using the unigram TF-IDF feature vector; Deep Learning models such as Bidirectional Long Short-Term Memory (BiLSTM) and Convolutional Neural Network (CNN) using pre-trained word embeddings including GloVe and FastText; and Transfer Learning models such as Bidirectional Encoder Representations from Transformers (mBERT), Robustly Optimized BERT Pretraining Approach (RoBERTa), Cross-lingual Language Model – RoBERTa (XLM-R), and Distilled BERT using advanced contextual embeddings to capture nuanced patterns in Roman Urdu hope speech.

These models are optimized to learn linguistic features indicative of hope speech in Roman Urdu, despite challenges like spelling variation and code-mixing. Once trained and evaluated across the cross-validation folds, the models generate predictions that are analyzed in the “Prediction Results” stage, comparing predicted labels with ground

truth to assess accuracy and reliability. The output consists of predicted values classifying the input text as expressing hope speech or not. This structured pipeline enables systematic experimentation and evaluation, supporting the identification of the most effective approaches for detecting positive discourse in Roman Urdu digital communication.

Table 4 shows the hyper-parameters employed to fine-tune our model on hope speech dataset. For transformer models—specifically mBERT, XLM-RoBERTA, and DistilBERT with a custom attention mechanism—we fine-tuned each model using a learning rate of 3×10^{-5} , for 3 epochs, with a batch size of 16, employing the AdamW optimizer and CrossEntropyLoss. Due to the high computational demands of these models, we evaluated them using a separate validation split rather than cross-validation. In the machine learning category, we used Support Vector Machines (SVM), Logistic Regression (LR), Decision Trees (DT), and XGBoost (XGB), tuning their hyper-parameters through grid search—for instance, testing different kernels and regularization strengths for SVM and LR, or various depths and learning rates for DT and XGB. For deep learning, we implemented BiLSTM and CNN models using pretrained embeddings, and adjusted parameters like embedding dimensions, LSTM units, filter sizes, and kernel size. To ensure a fair comparison and reliable performance estimates, we applied 5-fold cross-validation across all machine learning and deep learning models, which allowed each model to be trained and tested on different parts of the dataset and reduced the risk of over fitting. This comprehensive evaluation helped us identify the most suitable approach for accurately classifying hope speech in Roman Urdu.

Table 4. Hyper-parameters identified during the experimentation.

Learning Approach	Models	Hyper-parameters	Grid Search Values
Transformer	mBERT, XLM-RoBERTA, Distil Bert (with custom attention)	Learning rate, Epochs, Batch size, Optimizer, Loss Function	3×10^{-5} , 3, 16, AdamW, CrossEntropyLoss
Machine Learning	SVM	Kernel, C, gamma	'linear', 'rbf'; C: 1, 10; gamma: 'scale', 'auto'
	Logistic Regression (LR)	C, penalty, solver	C: 0.1, 1, 10; penalty: 'l2'; solver: 'lbfgs', 'saga'
	Decision Tree (DT)	max_depth, min_samples_split, min_samples_leaf, max_features	5, 10, 20, None; 2, 5, 10; 1, 2, 4; 'sqrt', 'log2', None
	XGBoost (XGB)	n_estimators, max_depth, learning_rate	100, 6, 0.3
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

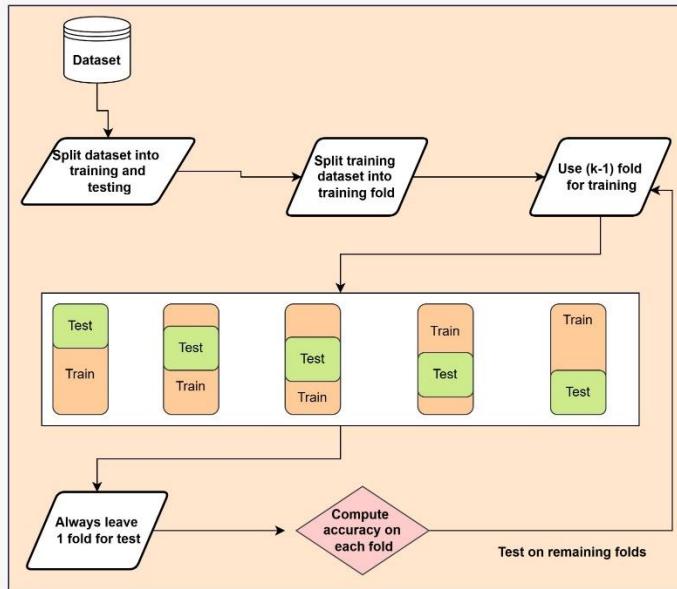


Fig. 6. Application of models, training and testing

4 Result and Analysis

In this section, we present the machine learning, deep learning and advance pre-trained language based transformer models used for the Roman Urdu multi-class hope speech detection task. For traditional machine learning, different models were trained using n-gram TF-IDF feature vectors to capture lexical and contextual patterns in the Roman Urdu text. We use TF-IDF because it helps identify and weigh the most important words in documents by reducing the impact of very common words and emphasizing the words that are more unique and informative to each document. For the deep learning models we employed popular pre-trained word embeddings including FastText and GloVe to better capture the semantic meaning of the text. We use GloVe because it helps capture the overall meaning of words based on how they appear together in large texts. FastText is useful because it looks at parts of words, so it can understand and represent even rare or new words better. Together, they make text understanding more accurate and flexible. Furthermore, we designed and implement and proposed a custom attention mechanism in pre-trained Transformer models to explore and enhance context understanding and improve classification performance across the four categories: generalized hope, realistic hope, unrealistic hope, and not hope. We used random oversampling to balance the class distribution by increasing the number of samples in minority classes. All models were evaluated using standard metrics—precision, recall, F1-score—and 5-fold cross-validation to compare their effectiveness and identify the most suitable approach for this multi-class classification task.

4.1 Machine learning Results

The table 5 presents a performance comparison of different machine learning models applied to a multi-class hope speech detection task using a Roman Urdu dataset. Among the five models evaluated—SVM with RBF kernel, XGBoost (XGB), Logistic Regression (LR), Decision Tree (DT), and SVM with a linear kernel—the Logistic Regression (LR) and SVM (linear) models perform the best overall, each achieving the highest cross-validation (C.V) score of 0.75. LR stands out with a strong balance between precision (0.70), recall (0.75), and the highest F1-score (0.73), suggesting it is highly effective in correctly identifying various classes of hope speech while maintaining consistency. SVM (linear) also performs competitively with the highest precision (0.76), showing it is particularly confident in its positive predictions, while matching LR's recall (0.75). In contrast, SVM (rbf), although showing decent precision (0.73) and recall (0.70), suffers from a lower F1-score (0.61), indicating an imbalance between precision and recall. XGB and DT show relatively weaker performance, especially DT which has the lowest precision (0.58), indicating a higher rate of false positives. Interestingly, while XGB performs slightly better than DT in terms of F1-score (0.63 vs. 0.61), both models trail behind LR and SVM (linear) across most metrics. These results suggest that linear models, especially LR and linear SVM, are more suitable for this Roman Urdu hope speech classification task, likely due to the textual nature and possibly high dimensionality of the input data, which linear models tend to handle well.

Table 5. Results for machine learning models.

Model	Precision	Recall	F1-score	C.V score
SVM (rbf)	0.73	0.7	0.61	0.7
XGB	0.62	0.67	0.63	0.67
LR	0.7	0.75	0.73	0.75
DT	0.58	0.67	0.61	0.67
SVM (linear)	0.76	0.75	0.71	0.75

4.2 Results for deep learning

Table 6 compares the performance of deep learning models—BiLSTM and CNN—using two different word embeddings, FastText and GloVe, for a multi-class hope speech detection task in Roman Urdu. Starting with FastText, both BiLSTM and CNN models achieve the same precision (0.76), but CNN outperforms BiLSTM slightly in recall (0.75 vs. 0.73), F1-score (0.75 vs. 0.70), and cross-validation score (0.75 vs. 0.73), indicating CNN's better balance between precision and recall with FastText embeddings. When using GloVe embeddings, performance improves overall. BiLSTM reaches equal precision and recall of 0.76 with a slightly improved F1-score of 0.72 and the highest C.V score of 0.76, showing better generalization. CNN with GloVe slightly edges out others by achieving the highest precision (0.77), F1-score (0.76), and a strong

recall (0.75), suggesting it is the most reliable and consistent model among all tested. Overall, models with GloVe embeddings perform marginally better than those with FastText, and CNN generally shows stronger and more stable results than BiLSTM across both embeddings. This highlights the effectiveness of CNN with GloVe for capturing meaningful features in Roman Urdu text for hope speech classification.

Table 6. Results for Deep learning models.

Model	Precision	Recall	F1-score	C.V score
FastText				
BiLSTM	0.76	0.73	0.7	0.73
CNN	0.76	0.75	0.75	0.75
GloVe				
BiLSTM	0.76	0.76	0.72	0.76
CNN	0.77	0.75	0.76	0.75

4.3 Results for Transformer

Table 7 shows the performance of three transformer-based models—mBERT, XLM-R, and DistilBERT—on a Roman Urdu multi-class hope speech detection task. Among them, XLM-R clearly stands out as the best-performing model, achieving the highest precision (0.78), recall (0.78), F1-score (0.77), and cross-validation (C.V) score (0.78). This consistent performance indicates that XLM-R is highly effective and reliable at identifying and generalizing hope speech across various classes in Roman Urdu, likely due to its stronger multilingual capabilities and deeper architecture. In comparison, mBERT shows moderate performance, with a decent recall of 0.71 but lower precision (0.65) and F1-score (0.67), suggesting it detects hope speech reasonably well but at the cost of more false positives. DistilBERT, while being a lighter and faster model, delivers the weakest results, especially with the lowest precision (0.58) and F1-score (0.61), although it maintains a recall of 0.70. This implies that while DistilBERT may capture many relevant instances, it lacks accuracy in predictions. Overall, XLM-R proves to be the most robust and balanced model for this complex, multilingual task, making it the top choice for hope speech classification in Roman Urdu.

Table 7. Results for transformer model

Model	Precision	Recall	F1-score	C.V score
mbert	0.65	0.71	0.67	0.71
XLM-R	0.78	0.78	0.77	0.78
DistilBert	0.58	0.7	0.61	0.7

4.4 Error Analysis

The table 8 summarizes the top-performing models from each learning approach—transformer-based, traditional machine learning, and deep learning—for the multi-class hope speech detection task in Roman Urdu. Among them, XLM-R leads overall with the highest precision (0.78), recall (0.78), F1-score (0.77), and C.V score (0.78), showcasing its exceptional ability to generalize across multilingual text due to its robust transformer architecture. From the traditional models, SVM with a linear kernel performs best, achieving a solid precision of 0.76 and recall of 0.75, resulting in an F1-score of 0.71 and C.V score of 0.75, making it a reliable choice especially when computational resources are limited. In the deep learning category, BiLSTM with GloVe embeddings stands out, delivering a balanced precision and recall (both 0.76), a respectable F1-score of 0.72, and a strong C.V score of 0.76, highlighting its capability to capture sequential dependencies in Roman Urdu text. While all three models show strong potential, XLM-R emerges as the most effective due to its consistently high performance across all metrics.

Table 8. Top performing models in each learning approach.

Model	Precision	Recall	F1-score	C.V score
XLM-R	0.78	0.78	0.77	0.78
SVM (linear)	0.76	0.75	0.71	0.75
BiLSTM (GloVe)	0.76	0.76	0.72	0.76

The table 9 presents the class wise performance of the XLM-R model in detecting types of hope speech in Roman Urdu across four categories: generalized hope, not hope, realistic hope, and unrealistic hope. The model performs exceptionally well on the "not hope" class, achieving perfect scores (Precision, Recall, and F1-Score all at 1.0), likely due to its large support size (508) and clearer distinction from hope-related categories. In contrast, it struggles with "generalized hope," showing low recall (0.28) and F1-Score (0.37), indicating difficulty in identifying this vague or overlapping category. The performance on "realistic hope" (F1: 0.59) and "unrealistic hope" (F1: 0.64) is moderate, suggesting the model can somewhat differentiate these more concrete types of hope. The overall accuracy of the model across all 990 samples is 78%, showing reasonable effectiveness in multi-class hope speech detection, though further improvement is needed for the less distinct classes. Figure 7 shows the confusion matrix for the XLM-R model, illustrating how often each class was correctly or incorrectly predicted. The confusion matrix is important because it provides a detailed breakdown of classification performance, revealing which classes are most frequently confused and helping identify areas where the model needs improvement.

Table 9. Class wise score for XLM-R model

Class	Precision	Recall	F1-Score	Support
generalized hope	0.53	0.28	0.37	142

not hope	1	1	1	508
realistic hope	0.52	0.68	0.59	183
unrealistic hope	0.63	0.65	0.64	157

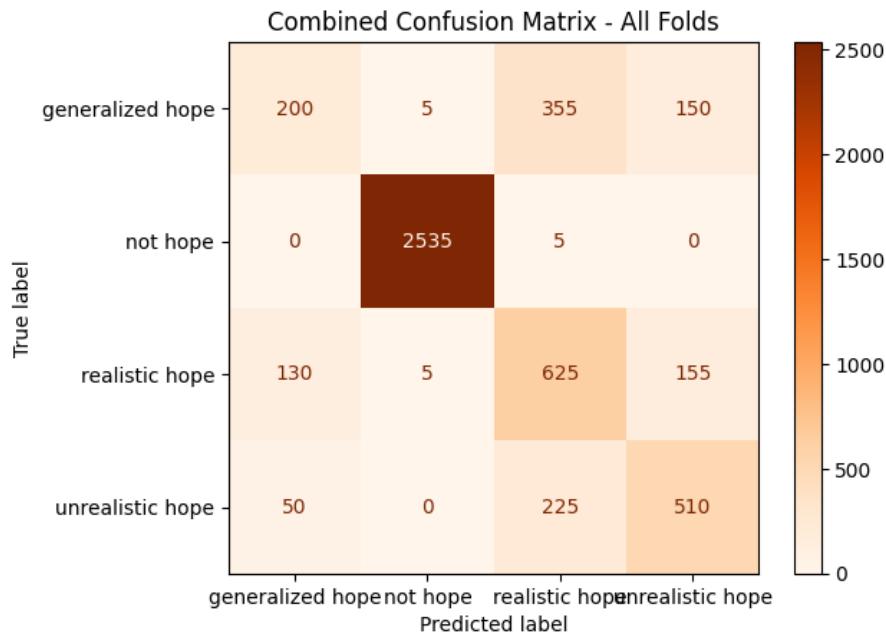


Fig. 7. Shows the confusion matrix of our proposed methodology XLM-R

A detailed statistical comparison using the paired t-test was conducted to evaluate whether the performance differences between XLM-R and the other top models—SVM (linear) and BiLSTM with GloVe—are statistically significant based on their F1-scores across fivefold experimental runs as shown in Table 10. The results show that XLM-R consistently achieved higher F1-scores (mean values around 0.78) compared to SVM (around 0.74) and BiLSTM (around 0.75). When comparing XLM-R with SVM, the t-test produced a t-statistic of 21.17 and a p-value of 2.94e-05, which is extremely small and far below the standard significance threshold of 0.05. This means there is a very low probability that the observed improvement in performance is due to random chance, strongly favoring XLM-R. Similarly, the comparison between XLM-R and BiLSTM yielded a t-statistic of 8.27 and a p-value of 0.00116, again indicating a statistically significant difference in favor of XLM-R. These results clearly demonstrate that XLM-R performs significantly better than both SVM and BiLSTM, and the consistency of its F1-scores across multiple runs further reinforces its robustness. Therefore, we can confidently conclude that XLM-R is the most effective and reliable model for multi-class hope speech detection in Roman Urdu, offering both high accuracy and statistical superiority over other learning approaches.

Table 10. Fold wise cross validation score of top performing models

Model	f1	f2	f3	f4	f5
SVM	0.7407	0.7477	0.7397	0.7455	0.7446
XLM-R	0.782	0.7805	0.776	0.783	0.7875
BiLSTM	0.7619	0.7548	0.7447	0.7636	0.7515

5 Limitations of proposed solution

Despite its contributions, our Roman Urdu hope speech dataset has a few notable limitations. One of the primary challenges is the lack of standardized spelling in Roman Urdu, which results in significant variation in how users express similar sentiments. For instance, hopeful messages like "Sab theek hojaye InshaAllah (Everything will be fine, God willing)", "Sab thk hoga inshAllah (All will be well, God willing)", or "Sab theek ho jaye ga inshaa Allah (Everything will turn out fine, God willing)" all convey the same idea but differ greatly in form, complicating both annotation and automated processing. Although we removed tweets with fewer than 15 characters to ensure meaningful content, some remaining tweets still lacked clear emotional or contextual depth, making it difficult to confidently assign a label. For example, a tweet like "Umeed hai ke mere sary enemies mar jain gy or ma akela mazy kroon ga (There is hope that all my enemies will die and I will enjoy being alone)" expresses optimism, but the subtlety of language can still lead to disagreement. Furthermore, while we followed rigorous annotation guidelines, hope is an inherently subjective concept, and annotators occasionally differed on nuanced cases such as "Zindagi mushkil hai, magar rukna nahi (Life is hard, but don't stop)." Although expert reviews helped refine the labels, some ambiguity remains. Lastly, our dataset is drawn exclusively from Twitter, which may not capture the full range of Roman Urdu expressions used on other platforms, potentially limiting the broader applicability of the system.

6 Conclusions and Future Work

In this study, we introduced the first annotated dataset for hope speech detection in Roman Urdu, a language form that is widely used in informal digital communication but remains underrepresented in Natural Language Processing (NLP) research. By capturing a rich variety of hopeful and non-hopeful expressions from X (formerly known as Twitter), and incorporating psychological insights into the conceptualization of hope, our work lays a strong foundation for future sentiment and emotion analysis in Roman-script languages. We designed and proposed an advanced deep learning framework leveraging XLM-R model integrated with a custom attention mechanism. Our experimental results demonstrate that the proposed approach achieves benchmark-level performance, significantly outperforming traditional machine learning models. This confirms the effectiveness of contextual language models in capturing the nuanced, metaphorical, and culturally embedded nature of hope speech in Roman Urdu. Future work

will expand in several promising directions. First, we aim to extend our existing dataset by incorporating more diverse samples from multiple platforms such as Facebook, Instagram, TikTok, and YouTube comments to better capture the linguistic and contextual variability of Roman Urdu. We also plan to include additional low-resource languages and underrepresented language varieties to broaden the scope of inclusive hope speech detection. Second, we will explore cross-lingual transfer learning, investigating how Roman Urdu hope speech models can be adapted or extended to related languages and scripts, such as Nastaliq Urdu, Hindi, and Hinglish. Lastly, we intend to fine-tune and evaluate large language models (LLMs) on our expanded dataset to assess their effectiveness in detecting nuanced hope speech across informal, multilingual, and code-mixed text settings.

7 Funding

This research received no external funding.

8 Data Availability Statement

Data will be made available upon request.

9 Acknowledgments

This work was performed with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, and grants 20241816, 20241819, and 20240951 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank CONACYT 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 support of Microsoft through the Microsoft Latin America PhD.

10 Conflicts of Interest

The authors declare no conflicts of interest.

References

1. Snyder, C. R. (Ed.). (2000). *Handbook of hope: Theory, measures, and applications*. Academic press.
2. D. Webb, Modes of hoping, *Hist. Hum. Sci.* 20 (3) (2007) 65–83.
3. V. Lohne, E. Severinsson, Hope during the first months after acute spinal cord injury, *J. Adv. Nurs.* 47 (3) (2004) 279–286.

4. Wetterauer, C., Schürmann, J., Winkler, L., Westermair, A. L., Biller-Andorno, N., Tschudin, S., ... & Trachsel, M. (2025). How to deal with the criterion of severe mental distress for late termination of pregnancy? A scoping literature review and a content analysis of clinical ethics consultations. *BMC Medical Ethics*, 26(1), 45.
5. Liu, B., Li, S., Xu, Y., Ding, S., & Ning, P. (2025). The mediating effects of parental hope and psychological resilience on social support and decision conflict in children with hypopspadias. *Journal of pediatric urology*, 21(1), 154-159.
6. Ahmad, M., Waqas, M., & Sidorov, G. (2025). Leveraging Large Language Models for Multi-Class and Multi-Label Detection of Drug Use and Overdose Symptoms on Social Media. arXiv preprint arXiv:2504.12355.
7. Sidorov, G., Ahmad, M., Ameer, I., Usman, M., & Batyrshin, I. (2025). Opioid Named Entity Recognition (ONER-2025) from Reddit. arXiv preprint arXiv:2504.00027.
8. Andalibi, N., & Buss, J. (2020, April). The human in emotion recognition on social media: Attitudes, outcomes, risks. In Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems (pp. 1-16).
9. Sujatha, R., Chatterjee, J. M., Pathy, B., & Hu, Y. C. (2025). Automatic emotion recognition using deep neural network. *Multimedia Tools and Applications*, 1-30.
10. Liu, Z., Zhang, T., Yang, K., Thompson, P., Yu, Z., & Ananiadou, S. (2024). Emotion detection for misinformation: A review. *Information Fusion*, 107, 102300.
11. Gandhi, A., Ahir, P., Adhvaryu, K., Shah, P., Lohiya, R., Cambria, E., ... & Hussain, A. (2024). Hate speech detection: A comprehensive review of recent works. *Expert Systems*, 41(8), e13562.
12. Rawat, A., Kumar, S., & Samant, S. S. (2024). Hate speech detection in social media: Techniques, recent trends, and future challenges. *Wiley Interdisciplinary Reviews: Computational Statistics*, 16(2), e1648.
13. Arya, G., Hasan, M. K., Bagwari, A., Safie, N., Islam, S., Ahmed, F. R. A., ... & Ghazal, T. M. (2024). Multimodal hate speech detection in memes using contrastive language-image pre-training. *IEEE Access*, 12, 22359-22375.
14. Ahmad, M., Ameer, I., Sharif, W., Usman, S., Muzamil, M., Hamza, A., ... & Sidorov, G. (2025). Multilingual hope speech detection from tweets using transfer learning models. *Scientific reports*, 15(1), 9005.
15. Balouchzahi, F., Sidorov, G., & Gelbukh, A. (2023). Polyhope: Two-level hope speech detection from tweets. *Expert Systems with Applications*, 225, 120078.
16. García-Baena, D., García-Cumbreras, M. Á., Jiménez-Zafra, S. M., García-Díaz, J. A., & Valencia-García, R. (2023). Hope speech detection in spanish: The lgbt case. *Language Resources and Evaluation*, 57(4), 1487-1514.
17. Nath, T., Singh, V. K., & Gupta, V. (2025). BongHope: An annotated corpus for Bengali hope speech detection. *International Journal of Information Technology*, 1-9.
18. M.D.S.S. Eswar, N. Balaji, V.S. Sarma, Y.C. Krishna, S. Thara, Hope speech detection in Tamil and english language, in: 2022 International Conference on Inventive Computation Technologies, ICICT, IEEE, 2022, pp. 51–56.
19. A. Hande, R. Priyadarshini, A. Sampath, K.P. Thamburaj, P. Chandran, B. Raja Chakravarthi, Hope speech detection in under-resourced Kannada language, 2021, arXiv e-prints, arXiv-2108
20. Ahmad, M., Usman, S., Farid, H., Ameer, I., Muzammil, M., Hamza, A., ... & Batyrshin, I. (2024). Hope Speech Detection Using Social Media Discourse (Posi-Vox-2024): A Transfer Learning Approach. *Journal of Language and Education*, 10(4 (40)), 31-43.
21. Divakaran, S., Girish, K., & Shashirekha, H. L. (2024). Hope on the horizon: Experiments with learning models for hope speech detection in spanish and english. In Proceedings of

- the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEURWS.org.
22. Chakravarthi, B. R. (2022). Hope speech detection in YouTube comments. *Social Network Analysis and Mining*, 12(1), 75.
 23. New World Encyclopedia. (n.d.). Urdu. New World Encyclopedia. Retrieved November 23. from (2024). <https://www.newworldencyclopedia.org/entry/Urdu>.
 24. Butt, S., Ashraf, N., Siddiqui, M. H. F., Sidorov, G., & Gelbukh, A. (2021). Transformer-based extractive social media question answering on TweetQA. *Computación y Sistemas*, 25(1), 23-32.
 25. Shushkevich, E., Cardiff, J., Rosso, P., & Akhtyamova, L. (2020). Offensive language recognition in social media. *Computación y Sistemas*, 24(2), 523-532.
 26. Rahman-Laskar, S., Gupta, G., Badhani, R., & Pinto-Avendaño, D. E. (2024). Cyberbullying Detection in a Multi-classification Codemixed Dataset. *Computación y Sistemas*, 28(3), 1091-1113.