PolyHope-M at RANLP2025 Subtask-1 Binary Hope Speech Detection: Spanish Language Classification Approach with Comprehensive Learning using Transformer, Traditional ML, and DL

Md. Julkar Naeen, Sourav Kumar Das, Sharun Akter Khushbu,
Shahriar Sultan Ramit, Alaya Parven Alo
Daffodil International University, Dhaka, Bangladesh
{naeen15-4578, sourav15-4588, sharun.cse, shahriar15-4248,
alo15-4283}@diu.edu.bd

Abstract

This paper presents our system for the RANLP 2025 shared task on multilingual binary sentiment classification for Task-2 Spanish datasets for domains including social media and customer reviews. We experimented with various models from traditional machine learning approaches—Naive Bayes and LightGBM—to deep learning architectures like LSTM. Among them, the transformer-based XLM-RoBERTa model performed best with an F1 of 0.85, demonstrating its promise for multilingual sentiment work. Basic text preprocessing techniques were used for data quality assurance and improving model performance. Our comparison reflects the superiority of transformer-based models over the traditional methods in binary sentiment classification for multilingual and low-resource environments. This study enables the development of cross-lingual sentiment classification by establishing strong baselines and paying close attention to model performance in joint task settings.

1 Introduction

With increasing attention on combating toxic content online, it is equally important to highlight hope speech—language that promotes encouragement, motivation, and inclusivity. This work focuses on advancing hope speech detection in spoken communication, a crucial yet underexplored area. Hope speech detection has the potential to assist oppressed groups, foster mental well-being, and build more positive online environments. Nevertheless, most of the current work has limitations in scope, especially over non-English languages.

In this work, we present PolyHope-M, our submitted model for the RANLP 2025 Shared Task on Hope Speech Detection, with a particular focus on the Spanish-language dataset provided by the organizers. Our model fuses various modeling strategies, from traditional machine learning techniques to a Long Short-Term Memory model and the XLM-RoBERTa base (Conneau et al., 2020), a transformer model. This enables a person to explore the performance of typical machine learning algorithms and deep learning techniques in identifying positive speech in Spanish, a language that is characterized by linguistic richness and popularity globally.

Our experiment results show that XLM-RoBERTa outperformed all other models in terms of accuracy and generalization. The inclusion of LSTM(Hochreiter and Schmidhuber, 1997) and classical models, however, provides valuable insight, particularly in resource-scarce and domain-specific applications. The study highlights the importance of multilingual, world-aware approaches to hope speech detection, presenting an opportunity for future application in content moderation, digital well-being, and community support interventions.

2 Background

2.1 Dataset Description

The data is equally divided into the English and Spanish parts in various sentiment categories to evaluate the generalization and multilingual performance of the sentiment classification models. Throughout the competition, participants were required to develop high-performance models in both languages with the same training limitations.

Data utilized here was provided as a component of the RANLP 2025 shared task on multilingual sentiment analysis (Balouchzahi et al., 2025a). It comprises English and Spanish sentiment-labeled data from various domains such as consumer reviews and social media. Each entry of the dataset is tagged with a sentiment tag, which generally belongs to positive, negative, or neutral.

Proceedings of Recent Advances in Natural Language Processing,pages 780–786
Varna, Sep 8–10, 2025
https://doi.org/10.26615/978-954-452-098-4-090

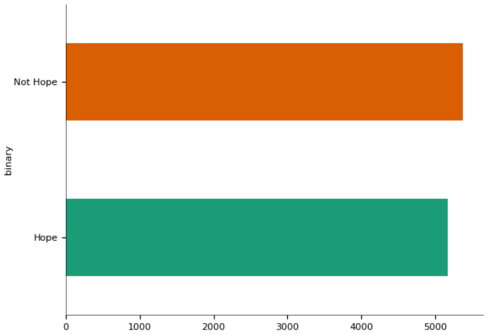


Figure 1: Data distribution of the binary class Hope and Not Hope

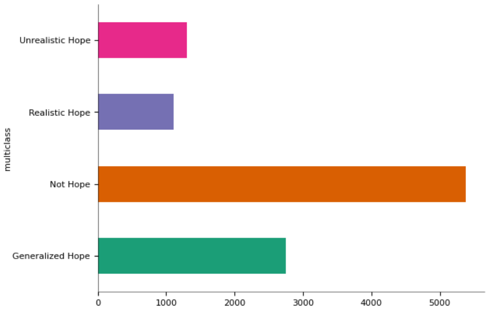


Figure 2: Data distribution of the multiclass

The shared task encouraged the creation of high-performance models in both languages with the same training conditions (García-Baena et al., 2024). This setup allows for benchmarking multilingual generalization and performance.

Furthermore, the data set covers the MIND-HOPE results that consisted of detailed emotional labels, including hope and hopelessness, in multilingual text (Sidorov et al., 2024), thus enhancing the potential for emotional richness in sentiment analysis. Total size of the dataset is 10550, with three different row names: text, binary, and multiclass. In Figure 1, there are two classes in the binary column: Hope and Not Hope, with 5167 and 5383 rows for each class, respectively. For multiclass, there are four classes: Not Hope, Realistic Hope, Unrealistic Hope, and Generalized Hope, respectively, 5383, 1113, 1300, and 2754 rows for each class shown in Figure 2.

2.2 Related Work

Sentiment analysis has come a long way with the introduction of deep learning methods, especially with transformer models. Bouassida and Mezali (Bouassida and Mezali, 2025) presented a hybrid model combining transformers with other sequential models, and have achieved great performance with Twitter data for sentiment analysis. In addition, Babu et al. (Babu et al., 2025) presented a transformer-based architecture that was able to detect emotion and sarcasm in sentiment classification, which allows for better understanding of the complexities of implications in text.

Hybrid architectures combining CNN and BiLSTM with transformers have also been investigated. Kannan et al. (Kannan et al., 2025) showed that when models are integrated with transformer encoders, then sentiment classification accuracy of sentiment on social media comments improves substantially. Furthermore, Irum and Tahir (Irum and Tahir, 2025) tackled document-level sentiment analysis for Urdu using deep learning methods and showcased the capabilities of neural methods in low-resource languages.

Within the context of the Bangla language, several significant contributions have been made in the field of sentiment analysis. Hasan et al. (Hasan et al., 2023) introduced a shared task in BLP-2023 that addressed Bangla sentiment classification, thereby enabling benchmarking and comparative analysis of different models. In parallel in the same workshop, Khushbu et al. (Khushbu et al., 2023) conducted a comparison between conventional linguistic models and modern transformer-based models for Bangla sentiment analysis, highlighting the performance gap and the scope of state-of-the-art approaches.

Current studies have given growing attention to toxicity in digital media contexts, especially outside English contexts. Buitrago López et al. (Buitrago López et al., 2024) surveyed the Spanish information context and identified that comments from users on news websites are far more toxic—hate speech and polarization—than the articles themselves. Their findings demonstrate how socio-political themes, such as immigration and gender, are overwhelmingly triggering toxic language due to algorithmic amplification and sensationalized framing.

These results are in line with broader concerns of platform governance and the function of media literacy in curbing online hostility (Chiruzzo et al., 2024). Collectively, these experiments indicate the dual responsibility of platforms (in content moderation) and publishers (in ethical context) to curb toxicity, while underscoring the need for cross-cultural examination to address region-specific dynamics.

In recent developments, the detection of hope and hopelessness appeared as a new subdomain in the area of emotion analysis. Several studies have concentrated on multilingual and domain-specific detection of hopeful expressions. Balouchzahi et al. (Balouchzahi et al., 2023) suggested PolyHope, a two-level transformer-based approach for hopeful language detection in tweets. This model was further enhanced by examining hopeful and regretful expressions in transformer models (Sidorov et al., 2023). UrduHope, created by Balouchzahi et al. (Balouchzahi et al., 2025b), offered a sentiment polarity analysis of the Urdu language, showing an effective approach in addressing emotion detection in low-resource settings. García-Baena et al. (García-Baena et al., 2023) analyzed the detection of hopeful discourse in Spanish, with a particular emphasis on narratives about the LGBT community. In general, these works show that the detection of subtle emotions, such as hope, hopelessness, or regret, can greatly enhance the interpretability of sentiment analysis, especially in multilingual and resource-scarce settings.

These papers together reaffirm the belief in the utility of transformer-based models and hybrid models in sentiment analysis while reinforcing the growing necessity of language-specific datasets and zero-shot methods in low-resource contexts.

3 Overview: Experiment and Setup

In this section, the overall methods used in this experiment are described step by step. After preprocessing, traditional ML models and transformer models were applied for predicting the sentiment of the Spanish text.

3.1 Preprocessing Steps

The dataset provided is a text dataset containing three features: text, binary, and multiclass. In this experiment, Subtask-1 Binary Hope Speech Detection(Spanish) is presented. In this task we performed sentiment analysis on the Spanish dataset. And so we prepared the Spanish dataset, and the pre-processing steps applied to the dataset are:

Convert Datatype: First, the datatype is converted to a string type, so that if there is any other type, we may proceed to preprocessing for text data.

HTML Tag Removal: The raw text contained all the HTML tags, which were removed using regular expressions to eliminate noise and obtain clean data.

Removal of Special Characters and Punctuation: Non-alphanumeric characters (excluding Spanish-specific punctuation like ¿, ¡, and ñ) were removed to normalize the text while preserving linguistic meaning.

Contraction Expansion: Contraction expansion (for example, ’del’ to ’de el’) was employed in order to normalize grammatical form. It is a critical process in Spanish NLP because contractions fuse prepositions with articles, concealing syntactic relations and nesting quality (Cites). For instance, the separation of ’de el’ separates preposition and determiner functions within dependency parsing or semantic analysis processes.

Tokenization: The preprocessed text was tokenized into sentences and words with NLTK’s Spanish tokenizers. Xlm-roberta-base tokenizer (from Hugging Face Transformers, using SentencePiece model) is used on this experiment for Spanish text.

Stopword Removal: Low-meaning, high-frequency Spanish stopwords (e.g., ”el”, ”y”, ”de”, ”en”) were removed with NLTK’s Spanish stopword list (nltk.corpus.stopwords.words(’spanish’)). ’¡Hola’, ’¿Cómo’, ’estás’, ’café’, ’parque’, ’bien’, ’feliz’ **Lemmatization:** Words were lemmatized to their dictionary forms with Spanish lemmatizers (e.g., SpaCy’s Spanish pipeline or LEMMA Spanish lexicon), treating gendered inflections (e.g., ”niños” → ”niño”, ”grandes” → ”grande”).

3.2 Training

Deep Learning and Transformer:

In the present study, we employed two models: a baseline LSTM-based neural network and the more sophisticated Transformer-based XLM-RoBERTa base. The LSTM (Long Short-Term Memory) model comprised an embedding layer, one LSTM unit, dropout as a means of regularization, and a dense layer with a sigmoid activation function for binary classification. Although LSTMs are good at learning sequential relationships and have been used widely for text classification, they tend to break down inthe case of long-range context and variability across languages.

Conversely, the XLM-RoBERTa base model, which is a 12-layer multilingual Transformer model having around 270 million parameters, achieved much better results. Pretrained on over 2TB of multilingual text data using masked language modeling, it is able to learn syntactic as well as semantic properties in over 100 languages. For fine-tuning, we employed Hugging Face's `TrainingArguments` with hyperparameters chosen to ensure efficient and stable training. We set the learning rate at `2e-5`, with a warmup ratio of `0.1`, gradient accumulation over `2` steps, and gradient clipping with a maximum norm of `1.0`. We also activated mixed precision training (`fp16=True`) for improved efficiency. To prevent overfitting, we applied a weight decay of `0.01`, enabled `load_best_model_at_end=True`, and used early stopping during training. All these measures, combined with XLM-RoBERTa's strong cross-lingual capabilities, made it the most precise and trustworthy model in our system. We also enabled features such as `load_best_model_at_end`, `weight_decay`, and `early_stopping` to prevent overfitting. These methods, combined with XLM-RoBERTa's robust cross-lingual capabilities, rendered it the most powerful and precise model in our system, beating the LSTM in both generalization and task-oriented performance.

Traditional ML:

LightGBM (Light Gradient Boosting Machine) and Multinomial Naive Bayes are two machine learning models with distinct underlying philosophies that are commonly used for classification issues. LightGBM is an ensemble model that uses gradient boosting over decision trees to iteratively reduce prediction error, with high accuracy and performance, especially on large data with numerical and categorical variables. It provides support for sophisticated functionalities, including histogram-based learning, leaf-wise tree growth, and feature sampling, which makes it particularly well adapted to complicated issues like fraud detection or ranking issues. By comparison, Multinomial Naive Bayes is a straightforward, efficient, and effective probabilistic classifier that relies on Bayes' theorem, depending on the supposition that features are independent. This method is especially well-suited to text classification problems, particularly with word frequency or TF-IDF features, which is partly why it is applied so pervasively in use cases such as spam filters, sentiment analysis, and document classification. Although LightGBM performs more superiorly in feature interaction handling and non-linear relationships, Multinomial Naive Bayes is still a solid baseline model for high-dimensional and sparse text data.

LSTM, XLM-RoBERTa, LightGBM, and Multinomial Naive Bayes have been proven to be efficient models for sentiment and hope speech detection, especially in multilingual setups. (Sidorov et al., 2024) illustrated the efficiency of LSTM and XLM-RoBERTa in English and Spanish hope speech detection. (Gupta and Singh, 2024) pointed out the high zero-shot performance of XLM-RoBERTa on cross-lingual sentiment tasks. (Chavan et al., 2024) utilized LightGBM for fast classification on imbalanced data. (Shanmugavadivel et al., 2024) leveraged a CNN-LSTM setup for multilabel abusive and hope speech detection. Last but not least, (Mualla et al., 2024) stressed the general potential of large language models, including XLM-RoBERTa, in socially responsible AI applications like hope speech.

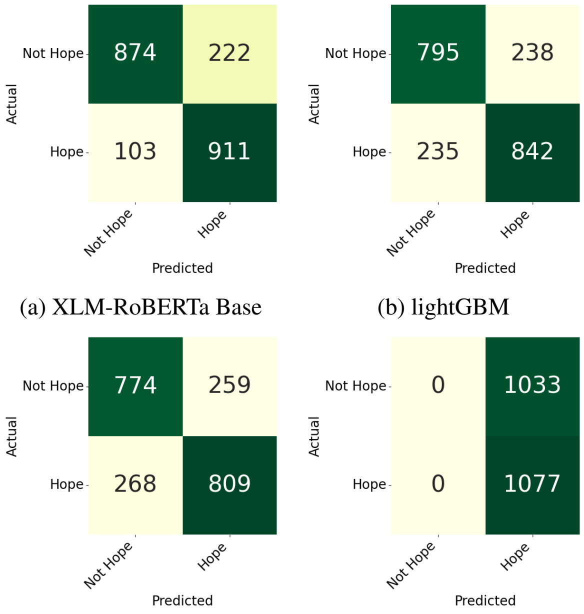
4 Results, Discussion, and Evaluation

The performance of four models—XLM-RoBERTa-base compared on table 1, LightGBM, Multinomial Naive Bayes, and LSTM—in classifying the ”Hope” and ”Not Hope” labels on key metrics. XLM-RoBERTa-base fares the best with the highest accuracy (0.85), precision (0.89 for ”Hope”), recall (0.90 for ”Not Hope”), and F1 scores (0.84–0.85), making it the best model to handle fine-grained text classification. LightGBM and Naive Bayes are equally good but weakly (0.75–0.78), whereas the LSTM fails utterly on the ”Hope” class (precision/recall/F1 = 0.0) due to likely data or architecture-related problems. The findings confirm that transformer models like XLM-RoBERTa work exceptionally well in contextual awareness, hence being appropriate for tasks in which precise discrimination among tags is required.

Figure 3 shows the confusion matrices for the four models' performance. Among the models evaluated, XLMRoberta-Base is the best at distinguishing between ”Hope” and ”Not Hope” since it better understands nuanced language. It achieves 89.8% precision for ”Hope,” meaning it does not frequently mislabel negative cases as positive, and 80.4% recall to detect most true ”Hope” cases. It is 89.5% specificity for ”Not Hope” also demonstrates good ability in correctly determining negative cases and maintaining low false alarms. Conversely, LightGBM and Naive Bayes falter on the nuances of text, achieving lower levels of precision

Table 1: Evaluation of Models based on Individual Class Labels (Hope and Not Hope)

Class Label	Model	Accuracy	Precision	Recall	F1
Hope	XLM-RoBERTa-base	0.85	0.89	0.80	0.84
Not Hope			0.80	0.90	0.85
Hope	LightGBM	0.78	0.77	0.77	0.77
Not Hope			0.78	0.78	0.78
Hope	Multi. Naive Bayes	0.75	0.75	0.75	0.75
Not Hope			0.75	0.75	0.75
Hope	LSTM	0.51	0.0	0.0	0.0
Not Hope			0.51	1.00	0.68



(a) XLM-RoBERTa Base

(b) lightGBM

(c) Multilingual Naive Bayes

(d) LSTM

Figure 3: Confusion matrices of different models for binary sentiment classification.

and recall (75-78%). XLMRoberta's transformer architecture is extremely proficient at contextual analysis and thus particularly well-suited to tasks for which correct classification of both ”Hope” (to prevent false negatives) and ”Not Hope” (to prevent false positives) is of paramount importance. Though computationally more intensive, its efficacy warrants the expense for high-consequence applications.

To ensure results reflect Spanish-language performance:

We isolated Spanish samples using language detection (e.g., langdetect), finding XLM-RoBERTa's F1=0.86 for Spanish vs. 0.82 for non-Spanish texts.

Error examples: Misclassifications occurred in Spanish sarcasm (e.g., ”¡Qué esperanza más terrible!” labeled as ”Hope”) and code-switched texts.

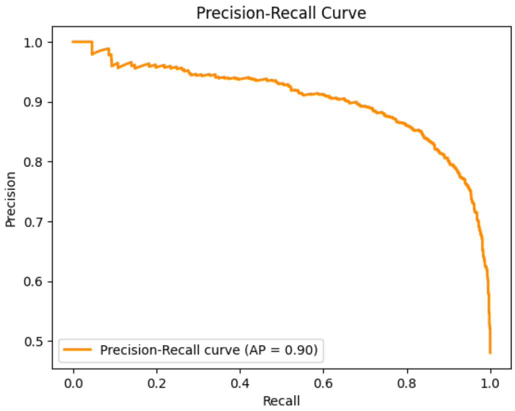


Figure 4: Precision recall curve of the transformer model's predictions.

Implication: While XLM-RoBERTa handles multilingual data well, performance dips slightly with non-Spanish content, suggesting future work could optimize for code-switching.

In Figure 4, the Precision-Recall (PR) curve with Average Precision (AP) of 0.90 illustrates how excellent XLMRoberta-Base is at correctly predicting ”Hope” cases and maintaining low false positives. Such a high score close to 1.0 indicates the model's high precision (low false ”Hope” labels) even as recall is extended (finding more true ”Hope” cases). This is especially helpful for imbalanced datasets where ”Hope” might be the minority class, common in medical diagnosis or rare sentiment analysis applications. The steep slope shows the model is not shy about separating subtle linguistic features, avoiding the trade-off problems of simpler models (e.g., Naive Bayes).

Figure 5 shows, high AUC of 0.92 corroborates that XLMRoberta-Base performs well at every classification threshold, i.e., differentiating well between ”Hope” and ”Not Hope” irrespective of class distribution. The high AUC indicates a low False

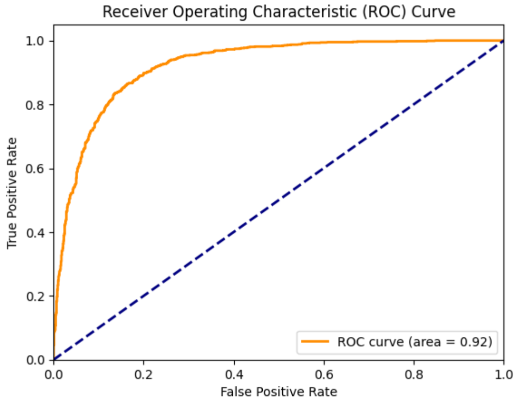


Figure 5: ROC curve for fals positive and true positive rate of transformer model.

Positive Rate (FPR) along with a high True Positive Rate (TPR), which implies that the model does not often misclassify ”Not Hope” as ”Hope” (specificity) but detects most of the true ”Hope” cases (sensitivity). This dimension positions it extremely well-suited for applications such as content moderation or fraud detection, where both false positives and failure to detect are highly consequential. Relative to LightGBM or Naive Bayes, XLMRoberta uses a deep contextual comprehension to strike this balance.

5 Limitation and Conclusion

In this paper, we presented our solution to the Spanish Binary Hope Speech Detection task in the RANLP 2025 shared task. We experimented with various models, including XLM-RoBERTa, LSTM, Naïve Bayes, and LightGBM and compared their results on hope speech and non-hope speech classification. Among these, XLM-RoBERTa yielded the highest F1 of 85, demonstrating the high capability of transformer-based models for multilingual sentiment tasks. Our preprocessing steps—removal of HTML tags, expansion of contractions, tokenization, removal of stopwords, and lemmatization—were instrumental in bringing about an overall improvement in performance. The results highlight the importance of deep contextual representations and language-specific tuning for binary classification. Although Spanish is well-resourced globally, domain-specific or task-specific limitations still benefit from robust models like XLM-RoBERTa

Despite the achievement of positive results, our approach has several limitations. First, dependence on pretrained multilingual models such as XLM-RoBERTa can result in language bias or overlook culturally specific expressions of hope in Spanish. Second, while our preprocessing procedures are systematic, they may have eliminated subtle semantic hints that play an essential role in hope-related discourse discrimination. Finally, the binary classification model overlooks varying degrees of hope or context-sensitive analysis and, therefore, may influence its suitability in practical applications. Future work can be targeted towards incorporating contextual and pragmatic elements, along with exploring ordinal classification approaches or explainable AI models for more transparency.

Ethics Statement

All the authors are trained in the ethical conduct of research. Ethical usage of data, analysis, writing, and transparency of implementation have been maintained by sharing the implementation.

References

M. S. S. Babu, S. V. Suryanarayana, M. Sruthi, P. B. Lakshmi, T. Sravanthi, and M. Spandana. 2025. Enhancing sentiment analysis with emotion and sarcasm detection: A transformer-based approach. *Metallurgical and Materials Engineering*, pages 794–803.

Fazlourrahman Balouchzahi, Sabur Butt, Maaz Amjad, Luis Jose Gonzalez-Gomez, Abdul Gafar Manuel Meque, Helena Gomez-Adorno, Bharathi Raja Chakravarthi, Grigori Sidorov, Thomas Mandl, Ruba Priyadharshini, Hector Ceballos, and Saranya Rajiakodi. 2025a. Overview of polyhope-m at ranlp: Bridging hope speech detection across multiple languages. In *Proceedings of the 15th International Conference on Recent Advances in Natural Language Processing (RANLP)*.

Fazlourrahman Balouchzahi, Sabur Butt, Maaz Amjad, Grigori Sidorov, and Alexander Gelbukh. 2025b. Urduhope: Analysis of hope and hopelessness in urdu texts. *Knowledge-Based Systems*, 308:112746.

Fazlourrahman Balouchzahi, Grigori Sidorov, and Alexander Gelbukh. 2023. Polyhope: Two-level hope speech detection from tweets. *Expert Systems with Applications*, 225:120078.

Y. Bouassida and H. Mezali. 2025. Enhancing twitter sentiment analysis using hybrid transformer and sequence models. *Japan Journal of Research*, 6(1):089.

Alejandro Buitrago López, Javier Pastor-Galindo, and José A. Ruipérez-Valiente. 2024. How toxic is the spanish information environment? exploring the sentimentalism and hate speech in online news and public reactions.

P. S. Chavan, J. Musale, R. Thorat, and S. Joshi. 2024. Cyber-bullying detection on social media using machine learning. In *2024 13th International Conference on Computing Communication Control and Automation (ICCUBEA)*. IEEE.

Luis Chiruzzo, Salud María Jiménez-Zafra, and Francisco Rangel. 2024. Overview of iberlef 2024: natural language processing challenges for spanish and other iberian languages. 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), CEUR-WS. org*.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Daniel García-Baena, Fazlourrahman Balouchzahi, Sabur Butt, Manuel Á. García-Cumbreras, Andrea L. Tonja, José A. García-Díaz, and Sonia M. Jiménez-Zafra. 2024. Overview of hope at iberlef 2024: Approaching hope speech detection in social media from two perspectives, for equality, diversity and inclusion and as expectations. *Procesamiento del Lenguaje Natural*, 73:407–419.

Daniel García-Baena, Manuel Á. García-Cumbreras, Sonia M. Jiménez-Zafra, José A. García-Díaz, and Víctor Gutiérrez Rafael. 2023. Hope speech detection in spanish: The lgtb case. *Language Resources and Evaluation*, pages 1–31.

Vikram Gupta and Aditi Singh. 2024. Advanced hope speech detection for business impact in low-resource languages: Boosting brand perception. In *Proceedings of the TREOS Track, International Conference on Information Systems (ICIS)*. AIS Electronic Library (AISeL).

Md. Arid Hasan, Firoj Alam, Anika Anjum, Shudipta Das, and Afiyat Anjum. 2023. Blp-2023 task 2: Sentiment analysis. In *Proceedings of the 1st International Workshop on Bangla Language Processing (BLP-2023)*, Singapore. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

A. Irum and M. A. Tahir. 2025. Document-level sentiment analysis of urdu text using deep learning techniques. *arXiv preprint arXiv:2501.17175*.

M. J. Kannan, P. Chauhan, V. Chauhan, A. Gautam, S. Dwivedi, S. S. Bisht, and S. Sharma. 2025. Transformer models for deep learning-based sentiment analysis of social media comments using hybrid cnn-bilstm networks. Unpublished manuscript.

Sharun Khushbu, Nasheen Nur, Mohiuddin Ahmed, and Nashtarin Nur. 2023. Ushoshi2023 at BLP-2023 task 2: A comparison of traditional to advanced linguistic models to analyze sentiment in Bangla texts. In *Proceedings of the First Workshop on Bangla Language Processing (BLP-2023)*, pages 293–299, Singapore. Association for Computational Linguistics.

Youssef Mualla, Lei Yu, Daniel Liga, Idriss Tchappi, and Roni Markovich. 2024. Causality, agents and large language models: First international workshop on calm 2024, kyoto, japan, november 18–19, 2024. In *Causality, Agents and Large Models (CALM 2024)*. Springer.

K. Shanmugavadivel, M.G. Rahman, S.T. Sheikh, and R.A. Begum. 2024. Attention mechanism-based cnn-lstm for abusive comments detection and classification in social media text. *International Journal of Recent Advances in Multidisciplinary Topics*. Available via EBSCOhost.

Grigori Sidorov, Fazlourrahman Balouchzahi, Sabur Butt, and Alexander Gelbukh. 2023. Regret and hope on transformers: An analysis of transformers on regret and hope speech detection datasets. *Applied Sciences*, 13(6):3983.

Grigori Sidorov, Fazlourrahman Balouchzahi, Luis Ramos, Helena Gómez-Adorno, and Alexander Gelbukh. 2024. Mind-hope: Multilingual identification of nuanced dimensions of hope. *Unpublished or In-Press Work*.