

TF-IDF to Transformers: Benchmarking Classical and XLNet Approaches for Hope Speech Detection

Usha Raman Adapa^{1,*†}, Sheshi Sree Sama^{1,†}

¹Texas Tech University, Lubbock, TX, USA

Abstract

Detecting hope in social media text is a vital task for understanding emotional resilience, supporting mental health monitoring, and fostering constructive online discourse. This paper explores both classical machine learning and transformer-based approaches (XLNet) for the automatic detection of hope or lack of hope in social media texts, including both binary and multiclass classification tasks. All experiments were conducted on English language data. Subtask 1 addresses binary classification (hope vs. not hope), while Subtask 2 focuses on fine-grained multiclass categorization (generalized hope, realistic hope, unrealistic hope, not hope, and sarcasm). We initially experimented with traditional models such as Logistic Regression, SVM, and XGBoost using n-gram enriched TF-IDF features. These models demonstrate competitive performance, highlighting their suitability for lightweight and resource-constrained scenarios. To further improve classification accuracy, we fine-tuned a pre-trained XLNet transformer using supervised training with class imbalance handling. The XLNet-based model achieves higher F1 scores in most classes and demonstrates superior generalization in the development set, reinforcing the value of transformer-based architectures in nuanced emotion classification tasks. Our findings offer a robust benchmark for hope speech classification and highlight trade-offs between efficiency and expressiveness in real-world NLP tasks.

Keywords

Hope Speech Detection, Emotion Classification, TF-IDF, XLNet, Transformer Models, Machine Learning, Natural Language Processing, Sarcasm Detection, Multiclass Classification, Social Media Text Mining

1. Introduction

Hope is a vital psychological resource that fosters resilience, emotional well-being, and motivation. In recent years, the detection of hope speech in social media has emerged as a key task in the broader field of computational affective analysis, with applications ranging from digital mental health to social inclusion. Unlike traditional sentiment categories like joy or sadness, hope often appears in nuanced, implicit, or even sarcastic expressions, making its automatic identification especially challenging.

Initial efforts to define and detect hope speech began with the HopeEDI project, which proposed a multilingual dataset focused on Equality, Diversity, and Inclusion (EDI) [1]. This was followed by a series of shared tasks at IberLEF from 2021 to 2025, each expanding the conceptual and technical scope of the task. For instance, the IberLEF 2023 HOPE task explored

IberLEF 2025, September 2025, Zaragoza, Spain

*Corresponding author.

†These authors contributed equally.

✉ uadapa@ttu.edu (U. R. Adapa); shsama@ttu.edu (S. S. Sama)



© 2025 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

multilingual hope detection across several languages [2], while IberLEF 2024 emphasized both optimistic discourse and expected or sarcastic expressions [3].

Hope is a cognitive and emotional state characterized by the anticipation of positive outcomes, even amidst uncertainty. In the context of social media discourse, it may be explicitly expressed through motivational language or implicitly conveyed through expectations, desires, or even sarcastic comments. Following the definition adopted in the PolyHope shared task [4], hope is defined a spectrum of expressions including: Generalized Hope - broad expressions of positivity; Realistic Hope - goal oriented and feasible expectations; Unrealistic Hope - improbable or wishful sentiments; Not Hope - neutral or irrelevant expressions and Sarcasm - hope like expressions intended ironically. This granularity reflects a growing recognition of the complex ways hope is expressed and perceived in digital communication [4]. Prior work in this space has shown that such categories are not only linguistically diverse but also culturally situated and context-dependent [5].

To address these challenges, a variety of modeling approaches have been explored—from traditional machine learning with lexical features to transformer-based deep learning models [6]. This paper builds on these advancements, evaluating classical and transformer-based methods (specifically XLNet) for both binary and multiclass hope speech detection in English-language data.

2. Literature Review

The computational modeling of hope speech has evolved significantly over the past few years. One of the earliest large-scale initiatives in this space was HopeEDI, which introduced a multilingual dataset aimed at promoting equality, diversity, and inclusion through automated detection of hopeful language [1]. This work laid the foundation for further shared tasks that expanded the focus from binary hope detection to fine-grained, multiclass categorization.

The IberLEF 2023 shared task on multilingual hope detection explored variations across languages and cultural contexts [2], while the 2024 iteration examined hope in both its motivational and sarcastic dimensions [3]. This continued focus on evolving challenges within the field is also reflected in the organization of subsequent shared tasks, such as the upcoming IberLEF 2025 [7]. These studies emphasized the semantic complexity of hope-related discourse and the importance of contextual modeling.

Hope speech detection has progressed from binary classification to a nuanced, multiclass paradigm informed by theoretical and empirical studies. The baseline system proposed by Butt et al. [4] defines hope across five distinct categories—generalized, realistic, unrealistic, not hope, and sarcasm—based on psychological plausibility and linguistic framing. This definition is grounded in affective computing research and reflects real-world complexities in emotional communication. The original PolyHope dataset incorporated this classification framework, annotating thousands of social media posts across English and Spanish using expert-curated guidelines [8]. This provided a robust benchmark for evaluating a wide range of models, from traditional machine learning to deep neural networks. Such distinctions were operationalized through expert-labeled datasets and comprehensive annotation guidelines. This multi-class framing was inspired by prior work such as PolyHope’s two-level detection system, which

separated coarse hope detection from nuanced subcategory classification [9].

Transformer-based models have shown strong performance in this domain. For instance, Sidorov et al. [6] analyzed the efficacy of BERT and XLNet on datasets involving both regret and hope speech, demonstrating that transformers significantly outperform traditional classifiers in handling figurative and subtle sentiment. Moreover, these models benefit from pretraining on large corpora, allowing them to capture pragmatic and cultural cues often missed by feature-based systems.

Recent developments have also explored hope speech in specific social contexts. García-Baena et al. [5] focused on the LGTB community, illustrating how hope manifests uniquely within marginalized groups. This work, along with multilingual and inclusive modeling initiatives like those described in Chakravarthi et al. [10], highlights the importance of culturally sensitive hope speech detection systems.

In summary, the current literature underscores two core needs: (1) robust classification models that can handle nuanced categories of hope, and (2) comprehensive, annotated datasets that reflect cultural, emotional, and linguistic diversity.

3. Methodology

Our methodology consists of five primary stages: data collection and preprocessing, feature extraction, model selection and training, evaluation, and prediction and output generation. We implemented two approaches—traditional machine learning classifiers and transformer-based deep learning models (XLNet)—to classify text into binary and multiclass categories of hope.

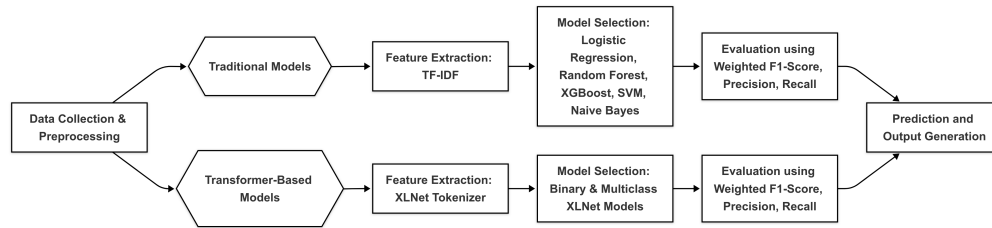


Figure 1: Flowchart of methodology

3.1. Data Collection and Preprocessing

The dataset used in this study comprises user-generated content from social media platforms. These texts often contain informal language, emojis, hashtags, mentions, and URLs, necessitating a robust preprocessing pipeline.

Text normalization included converting emojis into textual representations using `emoji.demojize()`, removing URLs, user mentions, and hashtags with regular expressions, eliminating non-alphabetic characters and punctuation, and lowercasing all text for consistency. Stopword removal was conducted using the NLTK library to reduce noise.

For the transformer-based model, we used the XLNet tokenizer, which segments inputs into context-aware subword tokens, preserving semantic integrity even for rare or misspelled words.

Unlike traditional pipelines, stopwords removal was selectively retained for transformers to limit clutter from non-informative tokens.

3.2. Feature Extraction

3.2.1. Traditional Models

Text was vectorized using Term Frequency–Inverse Document Frequency (TF-IDF), a widely used statistical measure that reflects word importance by down-weighting frequent terms across documents. TF-IDF is calculated as:

$$\text{TF-IDF}(t, d) = \text{TF}(t, d) \times \log \left(\frac{N}{\text{DF}(t)} \right), \quad (1)$$

where $\text{TF}(t, d)$ is the term frequency of term t in document d , $\text{DF}(t)$ is the number of documents containing term t , and N is the total number of documents.

3.2.2. XLNet Model

For transformer-based models, feature extraction is handled internally. Tokenized inputs from `XLNetTokenizerFast` were passed to the XLNet model, which generates dense contextualized embeddings through permutation-based autoregressive pretraining. These embeddings encode rich semantic information and were fine-tuned for the classification task.

Label encoding was applied for both binary and multiclass targets to convert categorical labels into numerical format. Separate encoders were used to ensure alignment across training and prediction phases.

3.3. Model Selection and Training

We implemented both classical machine learning models and transformer-based deep learning to evaluate performance across binary (Hope vs. Not Hope) and multiclass (Generalized Hope, Realistic Hope, Unrealistic Hope, Not Hope, and Sarcasm) settings. Implementations were done in Python using `scikit-learn`, `XGBoost`, and `Hugging Face Transformers`.

3.3.1. Traditional Machine Learning Models

- **Logistic Regression (LR):** A linear classifier that models the probability of a class using the logistic function:

$$P(y = 1|x) = \frac{1}{1 + e^{-\mathbf{w}^T \mathbf{x}}}, \quad (2)$$

where \mathbf{w} are the model weights and \mathbf{x} the input features. LR is efficient and interpretable.

- **Random Forest (RF):** An ensemble of decision trees using bagging and majority voting. Each tree is trained on a bootstrap sample, and splits are chosen to maximize information gain or Gini index. RF is robust to overfitting and noise.

- **XGBoost**: An optimized gradient boosting technique that sequentially builds trees by minimizing a regularized loss function:

$$\mathcal{L}(\phi) = \sum_i l(\hat{y}_i, y_i) + \sum_k \Omega(f_k), \quad (3)$$

where $\Omega(f_k) = \gamma T + \frac{1}{2} \lambda \|w\|^2$ is a regularization term. XGBoost is known for speed and high predictive accuracy.

- **Support Vector Machine (SVM)**: Finds the optimal hyperplane maximizing the margin between classes. Given training vectors x_i and labels y_i , the optimization is:

$$\min_{\mathbf{w}, b} \frac{1}{2} \|\mathbf{w}\|^2 \quad \text{s.t. } y_i(\mathbf{w}^T x_i + b) \geq 1. \quad (4)$$

Linear SVMs perform well in high-dimensional sparse feature spaces like TF-IDF.

- **Naive Bayes (NB)**: A probabilistic model using Bayes' theorem assuming feature independence:

$$P(c|x_1, \dots, x_n) \propto P(c) \prod_{i=1}^n P(x_i|c). \quad (5)$$

Multinomial NB is efficient and effective for text classification with sparse features.

Hyperparameter tuning was conducted via stratified k-fold cross-validation. Grid search was applied for SVM and LR, and performance was evaluated based on weighted F1-score.

3.3.2. Transformer-Based XLNet Models

We fine-tuned two models using the Hugging Face Transformers library:

- **Binary XLNet Model**: Implemented with `XLNetForSequenceClassification` using a sigmoid output. Loss was computed using `BCEWithLogitsLoss` and class weighting.
- **Multiclass XLNet Model**: Configured with a softmax activation and `CrossEntropyLoss` for five-class classification.

The XLNet architecture captures bidirectional dependencies by learning permutations of input tokens without masking, offering advantages over BERT for long or reordered sequences [?]. Both models were trained with the Adam optimizer, a linear learning rate scheduler, and dropout to prevent overfitting.

3.4. Evaluation Metrics

To evaluate model performance, we employed a comprehensive set of metrics that assess accuracy, class-wise discrimination, and balance:

- **Accuracy (Acc)**:

$$\text{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}, \quad (6)$$

where TP is true positives, TN is true negatives, FP is false positives, and FN is false negatives. This metric reflects overall correctness but can be misleading for imbalanced datasets.

- **Precision, Recall, and F1-Score:**

$$\text{Precision} = \frac{TP}{TP + FP}, \quad \text{Recall} = \frac{TP}{TP + FN}, \quad \text{F1-Score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \quad (7)$$

Precision measures exactness, recall measures completeness, and F1-score balances the two.

- **Macro-Averaged Metrics:** Calculated by averaging precision, recall, and F1-score across all classes equally, regardless of their frequency.
- **Weighted-Averaged Metrics:** Calculated by taking into account the proportion of each class in the dataset. More informative for imbalanced data.
- **Confusion Matrix:** Provides a tabular visualization of predictions vs. actual labels, helping diagnose class-specific errors.
- **ROC Curve and AUC (Area Under Curve):** Useful for binary classification tasks to assess trade-off between true positive rate and false positive rate.

The weighted F1-score was the primary metric for model selection, due to its robustness to class imbalance and balanced emphasis on both precision and recall.

3.5. Prediction and Output Generation

The best models from each approach were used to generate predictions on the test set. Outputs were stored as CSV files for both binary and multiclass classifications, enabling downstream tasks like visualization, interpretability analysis, and system integration.

This hybrid methodology enables a balance between explainability and contextual depth, leveraging classical models for efficiency and XLNet for semantic richness.

4. Results

This section presents an in-depth evaluation of classical TF-IDF-based models and transformer-based XLNet models for both binary (Hope vs. Not Hope) and multiclass (Generalized Hope, Realistic Hope, Unrealistic Hope, Not Hope, Sarcasm) hope speech classification. Results are presented across training, development (dev), and test sets, with a focus on generalization, per-class performance, and model robustness.

4.1. Binary Classification Results

4.1.1. Classical Models (TF-IDF)

All classical models demonstrated excellent training set performance (F1-scores > 0.92). This gap between train and dev performance is a classic symptom of overfitting in sparse feature spaces such as TF-IDF, especially when dealing with semantically nuanced text where surface features alone are insufficient. XGBoost and Random Forest emerged as the most stable, with dev F1-scores around 0.78. SVM offered a balanced performance with minimal variance, while Logistic Regression and Naive Bayes performed slightly lower.

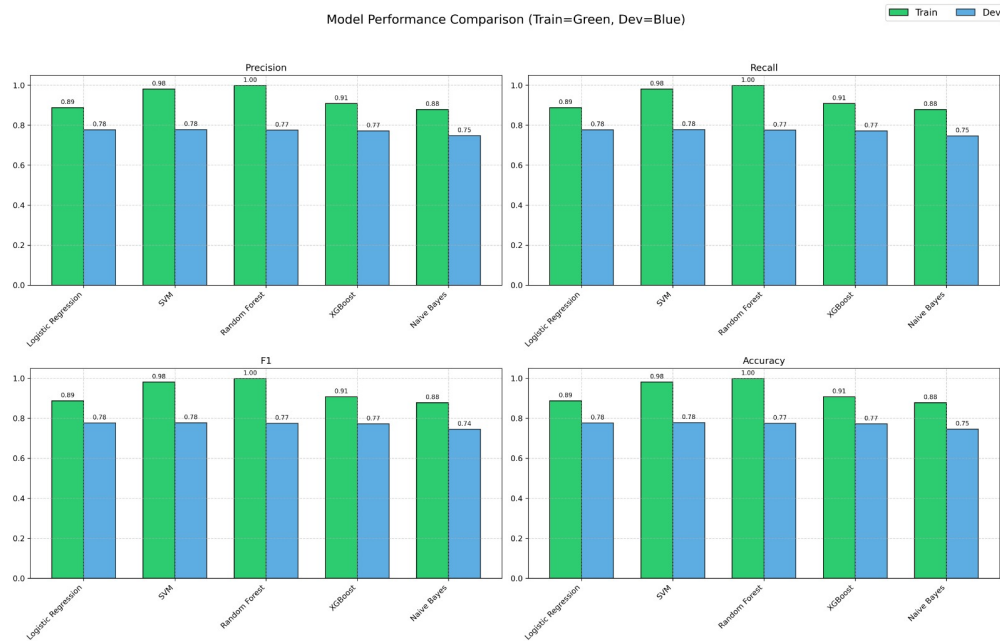


Figure 2: Binary classification: Model comparison using TF-IDF features

The confusion matrix for SVM (the top-performing classical model) indicated a strong ability to distinguish between Hope and Not Hope, with moderate false positives and negatives. Top predictive features included “hoping”, “hopeful”, “wish”, and “believe”, emphasizing lexical indicators of optimism. However, reliance on these terms occasionally led to false positives when hope-related words were used in sarcastic or negated contexts (e.g., “hoping this turns out to be another disaster”). ROC-AUC score for SVM was 0.86, showing solid discriminative capacity.

Table 1

Performance of SVM on Binary Classification

Performance Metric	Train	Dev	Test
Weighted Precision	0.9809	0.7783	0.8166
Weighted Recall	0.9809	0.7771	0.8164
Weighted F1	0.9809	0.7776	0.8159
Macro Precision	0.9806	0.7785	0.8168
Macro Recall	0.9810	0.7759	0.8138
Macro F1	0.9808	0.7766	0.8148
Accuracy	0.9809	0.7781	0.8164

4.1.2. Transformer-Based XLNet Model

XLNet outperformed all classical models, achieving near-perfect training scores (F1 0.994), and generalizing better on the dev set (F1 0.84). The loss curve illustrated stable learning.

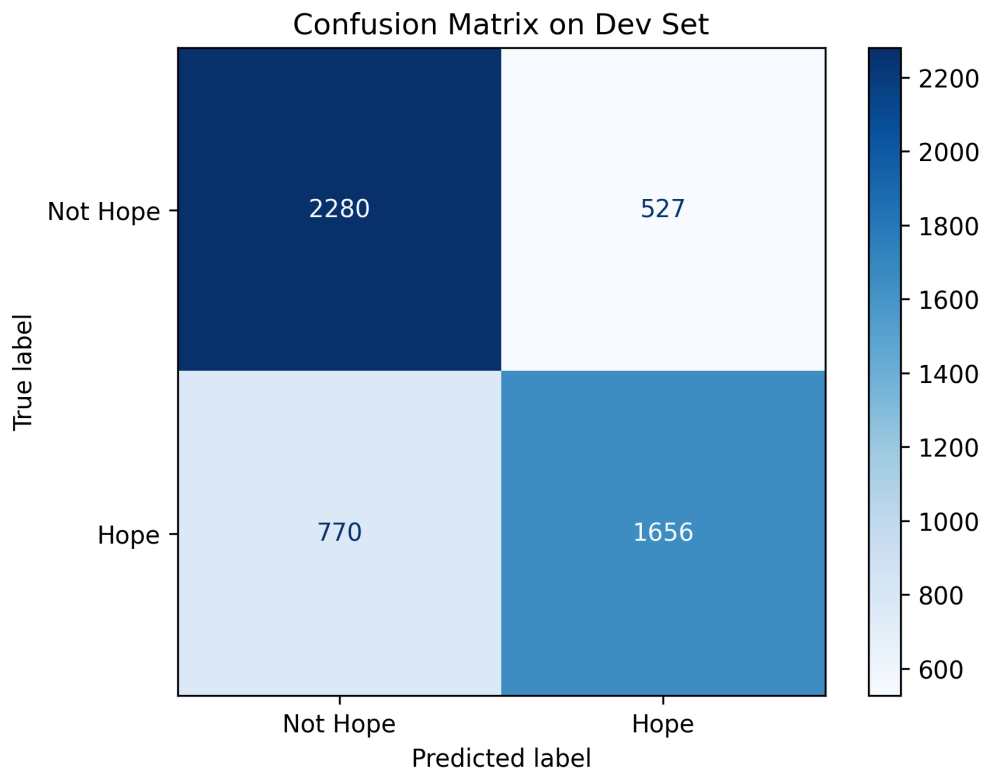


Figure 3: Confusion matrix for binary classification using SVM

Classical models showed strong training scores but lacked generalization. This performance gap reduction suggests that XLNet’s permutation-based attention mechanism captures the contextual nuances missed by linear models. In contrast to classical models’ reliance on isolated n-grams, XLNet effectively disambiguated complex constructs like sarcasm and double negatives through contextual embeddings. Interestingly, even XLNet struggled with samples where the expression of hope was highly metaphorical or embedded in indirect sentiment, which underscores the need for models trained on richer linguistic constructs or augmented with external knowledge. Confusion matrix analysis revealed strong balance between true positives and true negatives, with relatively fewer misclassifications.

4.2. Multiclass Classification Results

4.2.1. Classical Models (TF-IDF)

Among classical models, XGBoost led with a dev set accuracy of 0.69 and weighted F1-score of 0.67. Random Forest followed closely, while Logistic Regression and SVM were moderate (accuracy 0.64). Naive Bayes was the weakest, especially for sparse and rare classes.

F1-score breakdown indicated “Not Hope” was the easiest to classify (F1 0.9), followed by “Realistic Hope” (F1 0.77). “Generalized Hope” was moderately captured, while “Sarcasm” and

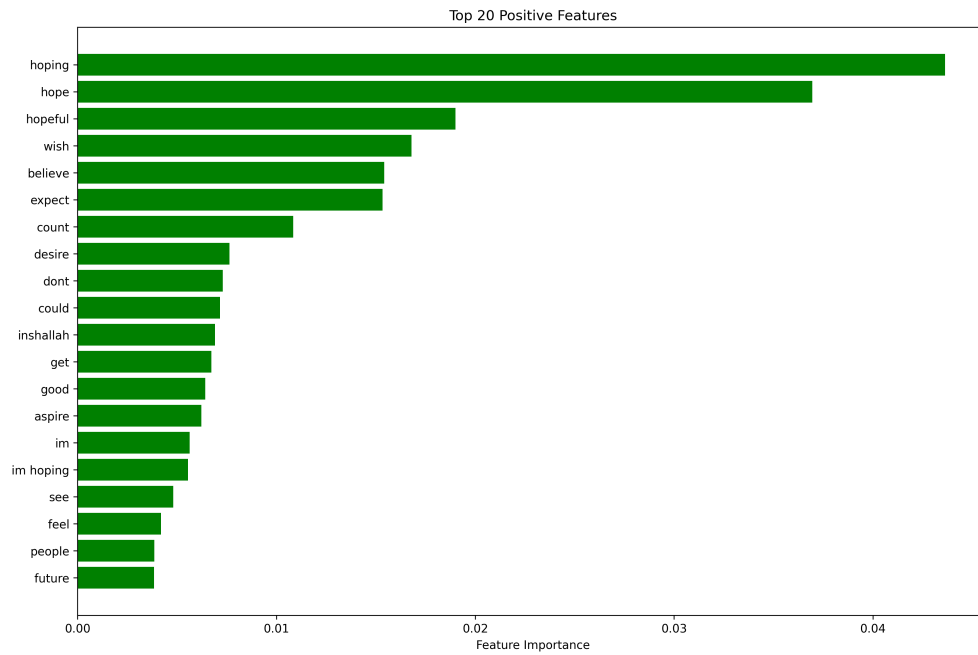


Figure 4: Top TF-IDF features in binary classification

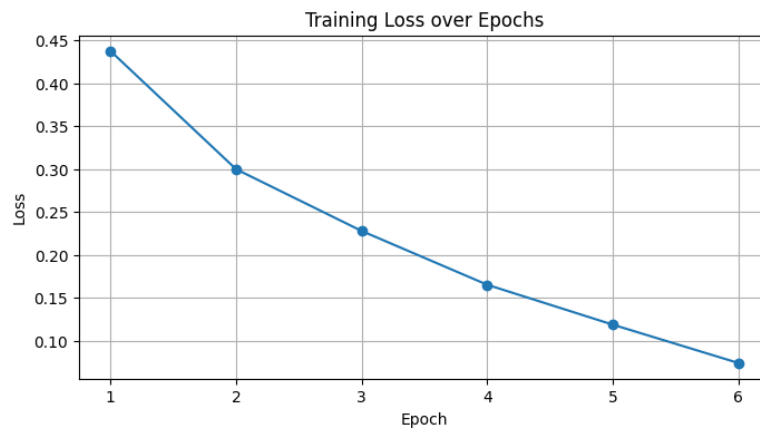


Figure 5: Training loss curve for binary XLNet model

“Unrealistic Hope” were the most challenging ($F1 < 0.5$), due to semantic overlap and class imbalance. This performance disparity is primarily attributed to two factors: (1) class imbalance, which led the model to under-prioritize underrepresented classes during training; and (2) semantic overlap between hope categories, especially when surface features (like “hope” or “wish”) appeared in both genuine and sarcastic texts.

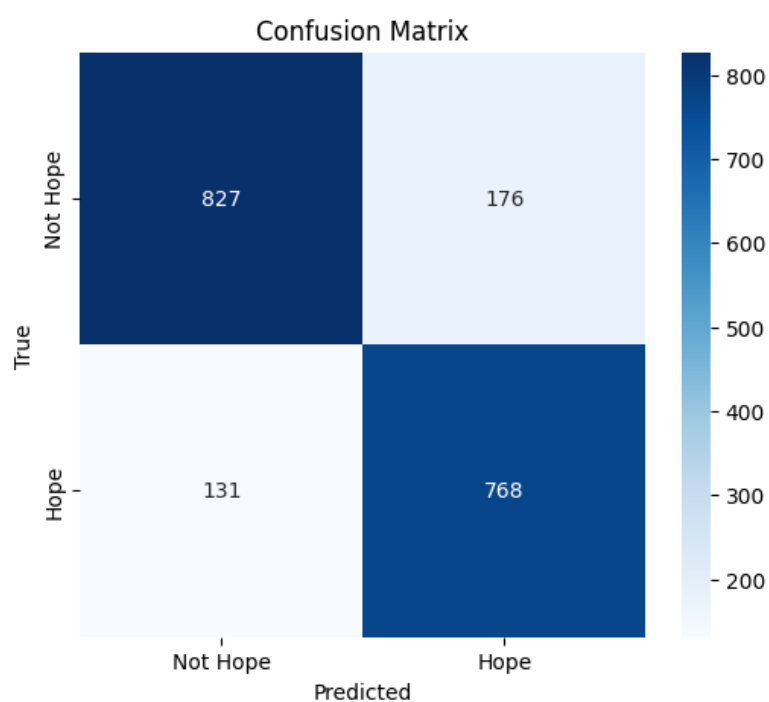


Figure 6: Confusion matrix (Dev set) for binary XLNet model

Table 2

Performance of XLNet on binary classification

Metric	Train	Dev	Test
Weighted Precision	0.9941	0.8398	0.85
Weighted Recall	0.9941	0.8388	0.84
Weighted F1	0.9941	0.8386	0.84
Macro Precision	0.9937	0.8384	0.85
Macro Recall	0.9944	0.8394	0.84
Macro F1	0.9940	0.8388	0.84
Accuracy	0.9941	0.8396	0.85

Table 3

Performance of XGBoost on multiclass classification

Performance Metric	Train	Dev	Test
Weighted Precision	0.8866	0.6715	0.6719
Weighted Recall	0.8852	0.6856	0.6874
Weighted F1	0.8837	0.6724	0.6756
Macro Precision	0.8937	0.6421	0.6391
Macro Recall	0.8597	0.5933	0.6202
Macro F1	0.8759	0.6094	0.6142
Accuracy	0.8852	0.6856	0.6874

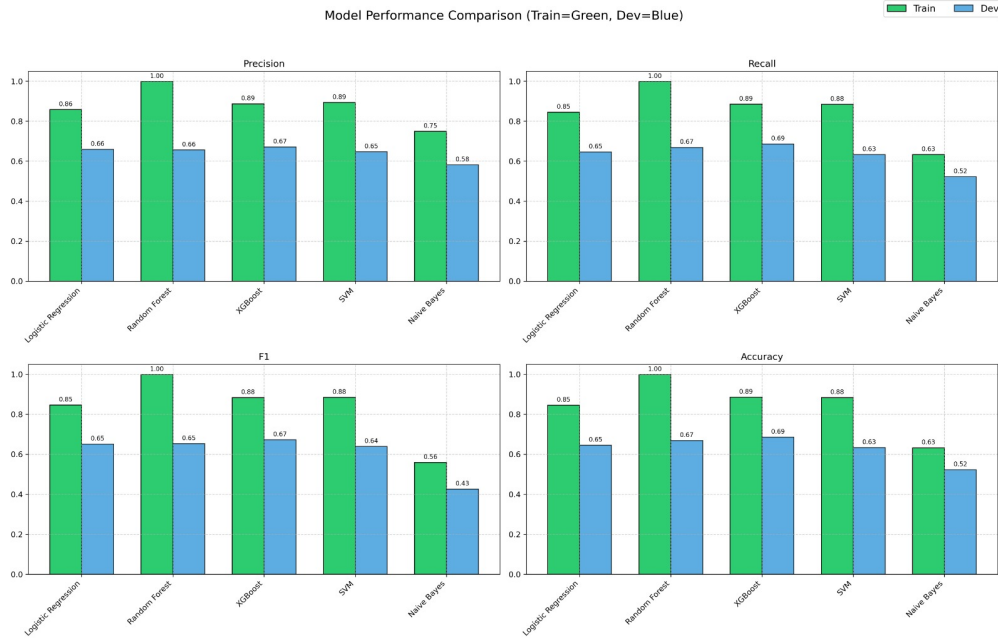


Figure 7: Multiclass classification: TF-IDF model comparison

4.2.2. Transformer-Based XLNet Model

The multiclass XLNet model exhibited a smoother training curve and achieved superior generalization. Despite close-class confusion (e.g., Generalized Hope vs. Realistic Hope), the model captured minority classes (e.g., Sarcasm) more accurately than classical models. XLNet significantly outperformed all classical models on multiclass classification. It maintained a better class-wise balance, especially for underrepresented labels. These results highlight that transformer models, when properly fine-tuned and class-weighted, can internalize subtle semantic and pragmatic cues that would otherwise be lost in feature-engineered models. Future work could explore label smoothing or focal loss to further reduce confusion between close categories.

Table 4

Performance of XLNet on multiclass classification

Metric	Train	Dev	Test
Weighted Precision	0.9662	0.7635	0.77
Weighted Recall	0.9662	0.7635	0.77
Weighted F1	0.9661	0.7635	0.77
Macro Precision	0.9607	0.7608	0.74
Macro Recall	0.9608	0.7605	0.73
Macro F1	0.9607	0.7605	0.73
Accuracy	0.9661	0.7634	0.77

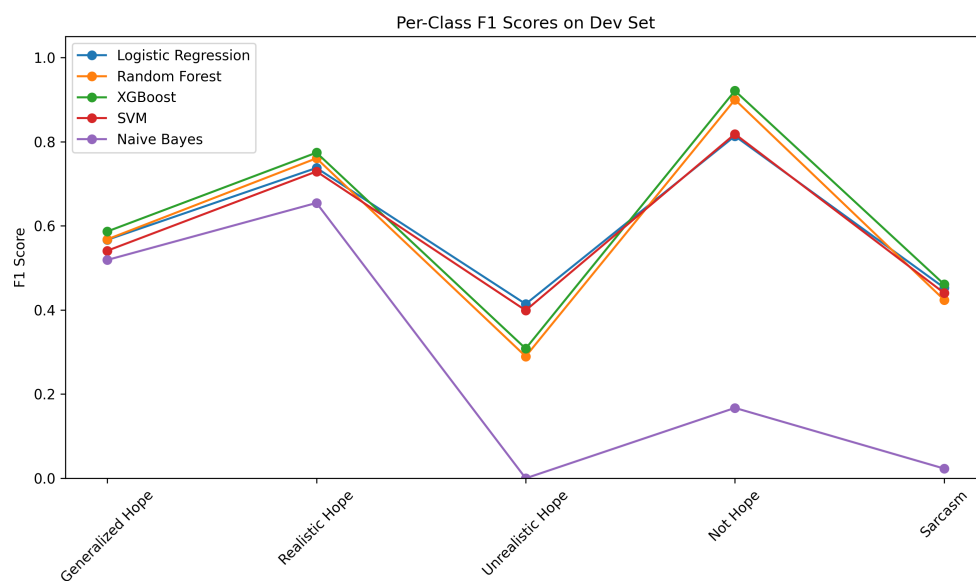


Figure 8: Per-class F1-scores for multiclass TF-IDF models

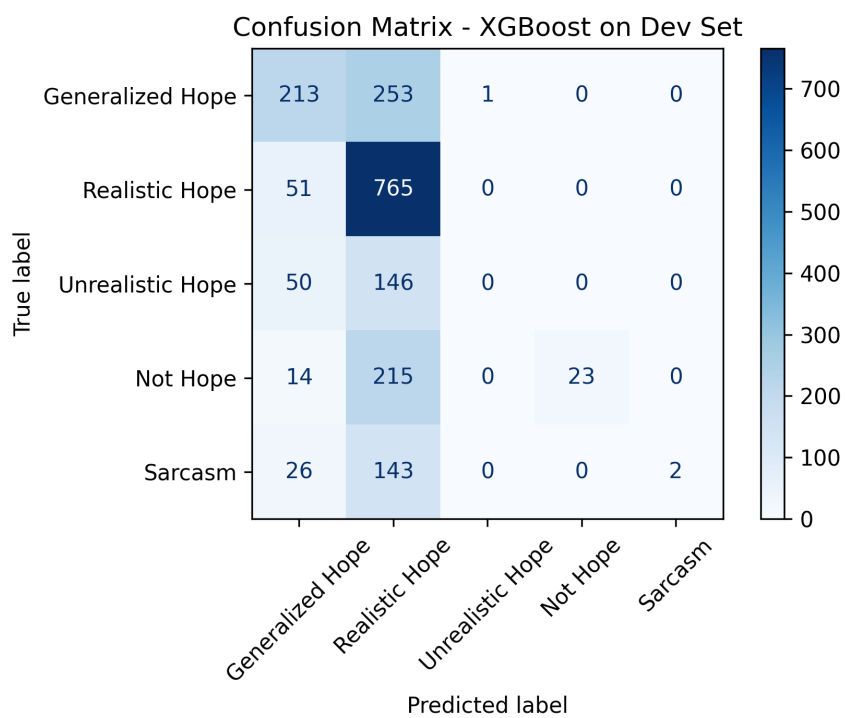


Figure 9: Confusion matrix: XGBoost multiclass predictions

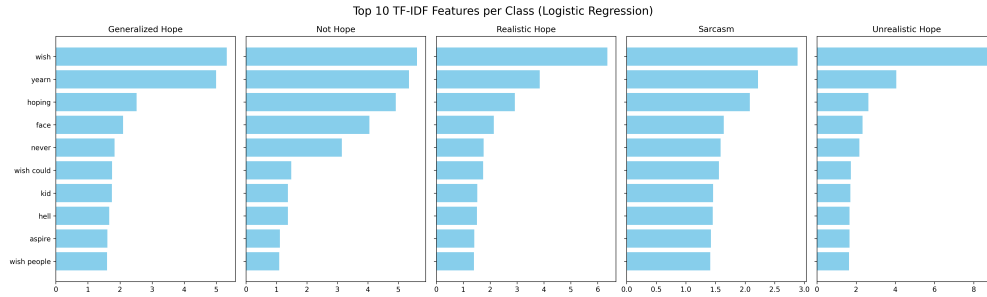


Figure 10: Top lexical features for multiclass classification

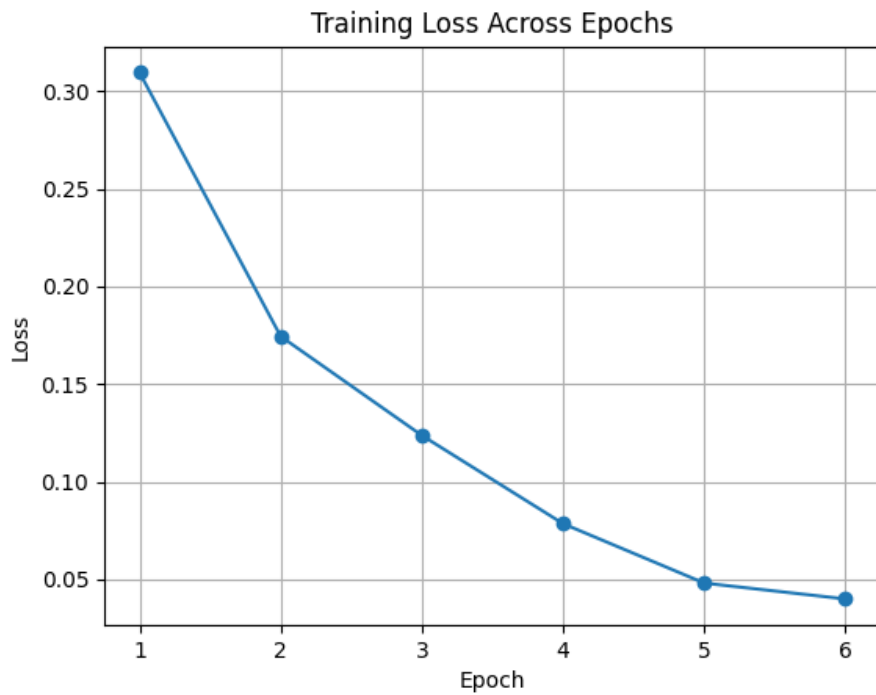


Figure 11: Training loss curve: XLNet multiclass model

5. Result Error Analysis

While both classical and transformer-based models yielded promising results in binary and multiclass hope speech classification, several sources of misclassification and model limitations were observed. This section analyzes the root causes of these errors and outlines targeted strategies for improvement.

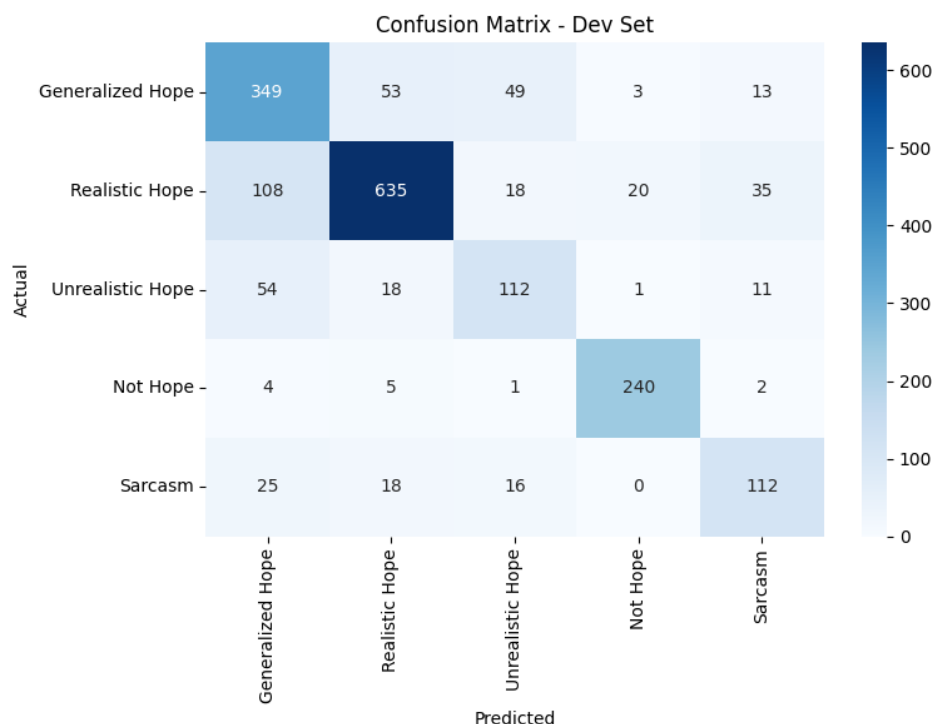


Figure 12: Confusion matrix: Multiclass XLNet on Dev set

5.1. Difficulties in Capturing Sarcasm and Unrealistic Hope

The most significant challenge was accurately distinguishing Sarcasm and Unrealistic Hope, particularly in the multiclass setting. Sarcasm, by nature, depends on tone, context, and sometimes contradiction between literal words and intended meaning—elements often absent from textual input alone. Even advanced models like XLNet struggled with sarcastic expressions that contained hope-related words used ironically, such as “I just love hoping for miracles that never happen.”

The following examples are not from the actual test data set but are created and tested to illustrate representative cases of misclassification.

Example 1: “Just waiting for a flying unicorn to fix everything. Fingers crossed.”

Predicted: Generalized Hope Actual: Sarcasm

Reason: The presence of hopeful phrases misled the model, but sarcasm was conveyed through absurdity and emoji tone, which the model misinterpreted as genuine optimism.

Example 2: “I know my lottery ticket will solve my student loans. Manifesting hard.”

Predicted: Realistic Hope Actual: Unrealistic Hope

Reason: The model classified this as realistic due to the serious context (loans), but failed to infer the low plausibility and sarcasm embedded in “manifesting” and emoji use.

Similarly, Unrealistic Hope was frequently confused with Generalized Hope or Not Hope. This is because recognizing the implausibility of a hopeful statement (e.g., “I hope unicorns

fix climate change”) often requires world knowledge or commonsense reasoning—areas where traditional classifiers and even transformers with limited pretraining context fall short.

5.2. Effect of Class Imbalance

Another key source of error was the dataset’s inherent class imbalance. Classes such as Sarcasm and Unrealistic Hope had fewer training samples compared to dominant classes like Not Hope and Realistic Hope. This led to skewed model learning, favoring frequent labels and reducing recall for minority categories. Although class-weighted loss functions (e.g., $\text{CrossEntropyLoss}(\text{weight}=\dots)$) were used, they only partially mitigated this imbalance.

Example 3: “I hope scientists find a cure for being broke.”

Predicted: Generalized Hope Actual: Sarcasm

Reason: The model likely favored Generalized Hope due to class frequency and the presence of the phrase “I hope”, despite the humorous exaggeration signaling sarcasm.

5.3. Overfitting in Classical Models

Classical models like XGBoost and Random Forest exhibited signs of overfitting. High F1-scores on training data contrasted with lower development and test set performance, indicating that these models may have memorized patterns rather than generalized them. This is typical in high-dimensional feature spaces produced by TF-IDF vectorization, especially when using shallow models on limited data.

5.4. Generalization Strength of XLNet

In contrast, XLNet demonstrated better generalization across both tasks, maintaining a more consistent gap between training and development scores. Its strength lies in its permutation-based training objective and bidirectional context modeling, which enables it to capture subtle dependencies missed by classical models. However, it still faced difficulty with low-resource classes and irony due to the limited signal in the data itself.

In summary, while XLNet significantly outperformed traditional models in nuanced hope classification, it remains sensitive to dataset limitations, particularly class imbalance and subtle semantic phenomena like sarcasm. Addressing these limitations through data augmentation, specialized models, and smarter loss functions will be critical to further advancing hope speech detection.

6. Conclusion

This study presented a comparative analysis of classical machine learning models and XLNet for binary and multiclass hope speech detection. While traditional models like XGBoost offered strong baseline performance, they struggled with class imbalance and subtle categories such as Sarcasm and Unrealistic Hope.

XLNet consistently outperformed classical models across both tasks, demonstrating better generalization and sensitivity to nuanced expressions. However, challenges like sarcasm detection and underrepresented classes persist.

Future work will explore advanced transformers, context integration, and augmentation strategies to address these limitations. Overall, combining robust language models with balanced datasets and targeted optimization proves key to accurate and reliable hope speech classification.

Acknowledgements

This research was conducted as part of the HopeEDI shared task at IberLEF 2025. The authors would like to thank the organizers: Sabur Butt (Institute for the Future of Education, Tecnológico de Monterrey, Mexico), Fazlourrahman Balouchzahi (Independent Researcher, Mexico), Maaz Amjad (Texas Tech University, USA), Salud María Jiménez-Zafra (SINAI, Universidad de Jaén, Spain), Hector G. Ceballos (IFE, Tecnológico de Monterrey, Mexico), and Grigori Sidorov (CIC, Instituto Politécnico Nacional, Mexico), for providing valuable data and guidance throughout the competition.

Declaration on Generative AI

During the preparation of this work, the authors used OpenAI’s GPT-4 to assist with grammar and spelling checks, as well as for refining the structure and clarity of technical content. We reviewed and edited all outputs carefully and take full responsibility for the final content of this publication.

References

- [1] B. R. Chakravarthi, Hopeedi: A multilingual hope speech detection dataset for equality, diversity, and inclusion, in: *Proceedings of the Third Workshop on Computational Modeling of People’s Opinions, Personality, and Emotion’s in Social Media*, Association for Computational Linguistics, Barcelona, Spain (Online), 2020, pp. 41–53. <https://aclanthology.org/2020.peoples-1.5>.
- [2] S. M. Jiménez-Zafra, M. Garcia-Cumbreras, D. García-Baena, J. A. Garcia-Díaz, B. R. Chakravarthi, R. Valencia-García, L. A. Ureña-López, Overview of hope at iberlef 2023: Multilingual hope speech detection, *Procesamiento del Lenguaje Natural* 71 (2023) 371–381.
- [3] D. García-Baena, F. Balouchzahi, S. Butt, M. García-Cumbreras, A. L. Tonja, J. A. García-Díaz, S. M. Jiménez-Zafra, 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 (2024) 407–419.
- [4] S. Butt, F. Balouchzahi, A. Amjad, M. Amjad, H. G. Ceballos, S. M. Jiménez-Zafra, Optimism, expectation, or sarcasm? multi-class hope speech detection in spanish and english, ResearchGate, 2025. <https://doi.org/10.13140/RG.2.2.19761.90724>.

- [5] D. García-Baena, M. García-Cumbreras, S. M. Jiménez-Zafra, J. A. García-Díaz, R. Valencia-García, Hope speech detection in spanish: The lgtb case, *Language Resources and Evaluation* (2023) 1–31.
- [6] G. Sidorov, F. Balouchzahi, S. Butt, A. Gelbukh, Regret and hope on transformers: An analysis of transformers on regret and hope speech detection datasets, *Applied Sciences* 13 (2023) 3983.
- [7] J. Á. González-Barba, L. Chiruzzo, S. M. Jiménez-Zafra, Overview of IberLEF 2025: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2025), co-located with the 41st Conference of the Spanish Society for Natural Language Processing (SEPLN 2025)*, CEUR-WS. org, 2025.
- [8] S. Butt, F. Balouchzahi, M. Amjad, S. M. Jiménez-Zafra, H. G. Ceballos, G. Sidorov, Overview of polyhope at iberlef 2025: Optimism, expectation or sarcasm?, *Procesamiento del Lenguaje Natural* (2025).
- [9] F. Balouchzahi, G. Sidorov, A. Gelbukh, Polyhope: Two-level hope speech detection from tweets, *Expert Systems with Applications* 225 (2023) 120078.
- [10] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, S. Cn, J. P. McCrae, M. García, S. M. Jiménez-Zafra, R. Valencia-García, P. Kumaresan, R. Ponnusamy, D. García-Baena, J. García-Díaz, Overview of the shared task on hope speech detection for equality, diversity, and inclusion, in: *Proceedings of the Second Workshop on Language Technology for Equality, Diversity and Inclusion organized as part of ACL 2022, 2022*, pp. 378–388. <https://doi.org/10.18653/v1/2022.ltedi-1.58>.