Classification of Hope in Textual Data using Transformer-Based Models

Chukwuebuka Fortunate Ijezue[1], Fredrick Eneye Tania-Amanda Nkoyo[1,**] and Maaz Amjad[1,†]

[1]Department of Computer Science, Texas Tech University, Lubbock, Texas, United States

Abstract

This paper presents a transformer-based approach for classifying hope expressions in text. We developed and compared three architectures (BERT, GPT-2, and DeBERTa) for both binary classification ("Hope" vs. "Not Hope") and multiclass categorization (five hope-related categories). Our initial BERT implementation achieved 83.65% binary and 74.87% multiclass accuracy. In the extended comparison, BERT demonstrated superior performance (84.49% binary, 72.03% multiclass accuracy) while requiring significantly fewer computational resources (443s vs. 704s training time) than newer architectures. GPT-2 showed lowest overall accuracy (79.34% binary, 71.29% multiclass), while DeBERTa achieved moderate results (80.70% binary, 71.56% multiclass) but at substantially higher computational cost (947s for multiclass training). Error analysis revealed architecture-specific strengths in detecting nuanced hope expressions, with GPT-2 excelling at sarcasm detection (92.46% recall). This study provides a framework for computational analysis of hope, with applications in mental health and social media analysis, while demonstrating that architectural suitability may outweigh model size for specialized emotion detection tasks.

Keywords

Hope Classification, NLP, BERT, GPT-2, DeBERTa, Comparative Analysis, Transfer Learning, Emotion Detection, Deep Learning

1. Introduction

In natural language processing (NLP), the computational analysis of text's emotive and emotional content is becoming a prominent area of study. Sentiment analysis [1], emotion detection [2], and toxicity classification [3] have all seen substantial research, but the particular field of hope detection and classification remains relatively unexplored. As a complex psychological concept, hope is essential to social discourse, mental health, and human communication [4]. Automatically identifying and classifying hopeful textual statements has potential applications in crisis response [5], social media analysis [6], political discourse analysis [7], and mental health monitoring [8].

This study presents a comprehensive approach to hope classification using transformer-based deep learning models. We developed a two-tiered classification system: (1) a binary classifier that distinguishes between hopeful expressions and those that are not, and (2) a multiclass classifier that categorizes text into five distinct hope-related categories: Not Hope, Generalized Hope, Realistic Hope, Unrealistic Hope, and Sarcasm. By differentiating between various forms of hopeful expressions, this granular approach enables a more thorough understanding of how hope manifests in text.

Our research begins with implementing BERT (Bidirectional Encoder Representations from Transformers) [9] for hope classification, leveraging its contextual understanding capabilities that have shown state-of-the-art performance on various NLP tasks. We then expand our investigation to compare BERT with more advanced transformer architectures: GPT-2, which employs unidirectional attention and benefits from a larger pretraining corpus, and DeBERTa, which utilizes a disentangled attention mechanism designed to better capture semantic nuances.

PolyHope at IberLEF 2025: Optimism, Expectation or Sarcasm?, April 7, 2025

**Corresponding author.

†These authors contributed equally.

cijezue@ttu.edu (C. F. Ijezue); tafredri@ttu.edu (F. E. T. Nkoyo); maaz.amjad@ttu.edu (M. Amjad)

https://ijezue.github.io/site/ (C. F. Ijezue); https://crystal4000.github.io/academic_portfolio/ (F. E. T. Nkoyo); https://maazamjad.com/ (M. Amjad)

This comprehensive comparison addresses a critical question in affective computing: do newer, more complex language models provide meaningful performance improvements for specialized emotional detection tasks like hope classification? Our experimental results reveal interesting patterns, with BERT achieving the highest accuracy in both binary (84.49%) and multiclass (72.03%) tasks, despite being architecturally simpler than alternatives. While DeBERTa (80.70% binary, 71.56% multiclass) and GPT-2 (79.34% binary, 71.29% multiclass) showed competitive performance, they required substantially higher computational resources, with DeBERTa taking more than twice the training time of BERT for multiclass classification. Notably, GPT-2 demonstrated particular strength in detecting sarcastic expressions of hope.

These findings suggest that model complexity does not necessarily correlate with performance improvement for hope classification tasks, highlighting the importance of architecture-task alignment in emotion detection systems. By evaluating model accuracy, training efficiency, and error patterns across these architectures, we identify the optimal approach for hope detection in practical applications, balancing performance against computational requirements. This research contributes to the growing field of affective computing by providing empirical evidence on the relative efficacy of different transformer architectures for the specialized task of hope classification, while establishing a framework for computational analysis of hope with applications in mental health and social media analysis.

2. Literature Review

2.1. Hope in Computational Linguistics

Hope detection represents an emerging field within emotion classification. While sentiment analysis has been extensively studied [1], nuanced emotions like hope remain underexplored. Early work by Snyder [4] established psychological frameworks for hope that inform computational approaches. Recent studies by Khanpour et al. [10] demonstrated that contextual features are vital for detecting subtle emotional states in text, suggesting transformer models may be well-suited for hope classification.

2.2. Transformer Architectures for Emotion Detection

BERT [9] introduced bidirectional context modeling and has achieved state-of-the-art results across various NLP tasks. Its bidirectional attention mechanism allows it to consider the full context when classifying emotional content. GPT-2 [11] employs unidirectional attention but benefits from a larger pre-training corpus, potentially capturing more linguistic patterns related to hope expressions. DeBERTa [12] enhances BERT with disentangled attention, separately computing content and position information, which theoretically improves contextual understanding of complex emotions.

2.3. Comparative Performance Studies

Comparative analyses of transformer architectures have shown task-dependent performance variations. While newer models often outperform older ones on general benchmarks [12], specialized tasks may reveal different patterns. Turc et al. [13] demonstrated that smaller pre-trained models can match larger models' performance when fine-tuned for specific tasks, suggesting architectural fit may outweigh model size for specialized applications like hope detection.

3. Methodology

This section details our comprehensive approach to hope classification, covering our dataset characteristics, pre-processing strategies, model architectures, implementation details, and evaluation framework. We present both our original BERT implementation and the extended comparison of three transformer architectures (BERT, GPT-2, and DeBERTa) to provide a thorough analysis of hope detection capabilities.

3.1. Dataset

This study employs custom datasets for hope classification, obtained from the PolyHope shared task at IberLEF 2025 [14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24]. The training dataset contained 5,233 samples, while the development/test dataset comprised 1,902 samples. Both datasets maintained similar class distributions, ensuring consistency between training and evaluation. The dataset supports two classification schemes: a binary task ("Hope" vs. "Not Hope") and a multiclass task with five categories ("Not Hope," "Generalized Hope," "Realistic Hope," "Unrealistic Hope," and "Sarcasm"). For the binary classification, the training set contained 2,426 (46.36%) "Hope" samples and 2,807 (53.64%) "Not Hope" samples, with the test set maintaining a similar distribution of 899 (47.27%) "Hope" and 1,003 (52.73%) "Not Hope" samples. The multiclass distribution was also consistent across both sets, with the following breakdown in the training data: "Not Hope" (42.90%), "Generalized Hope" (24.54%), "Sarcasm" (13.22%), "Realistic Hope" (10.32%), and "Unrealistic Hope" (9.02%). The test set maintained nearly identical proportions. This balanced representation across classes helped ensure the models could learn to distinguish between all categories effectively. The text samples varied in length from 24 to 886 characters, with an average length of approximately 188 characters. This variation in text length provided the models with diverse linguistic patterns and expressions of hope across different contexts. To ensure robust model development, we implemented an 80-20 train-validation split on the training data, maintaining the same random seed (42) across experiments for consistency and reproducibility.

3.2. Text Pre-processing

In this study, our approach to text pre-processing differed between the original implementation and the extended comparison. In the original BERT implementation, we deliberately minimized preprocessing, feeding raw text directly into the tokenization pipeline to leverage BERT's capability to capture contextual nuances. For the extended comparison between models, we applied basic text cleaning through a custom function that converted text to lowercase, removed URLs and web links, removed hashtags and user mentions (patterns like #word and @word), and removed punctuation. This cleaned text was stored in a separate 'clean_text' column and used for tokenization across all three models. This standardized preprocessing in the extended comparison ensured a fair evaluation across different transformer architectures while allowing us to assess whether specialized cleaning benefits these pre-trained models. The contrast between approaches also enabled us to evaluate the impact of pre-processing on model performance for hope classification tasks.

3.3. Transformer Model Architectures

Our study implements and compares three state-of-the-art transformer architectures for hope classification, with BERT used in our original implementation and all three models (BERT, GPT-2, and DeBERTa) compared in our extended analysis.

3.3.1. BERT Architecture

In both our original and extended implementations, we utilized the 'bert-base-uncased' variant from Hugging Face's Transformers library. This BERT model comprises 12 transformer layers, 12 attention heads, and 768 hidden dimensions, totaling approximately 110 million parameters. BERT's bidirectional attention mechanism enables the model to consider the full context when representing each word, potentially beneficial for capturing complex hope expressions. In our original implementation, BERT served as the sole architecture for establishing baseline performance in hope classification.

3.3.2. GPT-2 Architecture

For our extended comparison, we incorporated the GPT-2 base model (124M parameters) with its autoregressive architecture. Unlike BERT's bidirectional attention, GPT-2 uses unidirectional attention

where each token can only attend to previous tokens in the sequence. While this limitation might affect classification performance, GPT-2's larger pre-training corpus potentially provides richer semantic representations beneficial for hope classification. Special consideration was required for GPT-2 implementation, including setting the pad token to match the EOS token and disabling the cache to avoid errors during training.

3.3.3. DeBERTa Architecture

Also included only in our extended comparison, DeBERTa (base version, 140M parameters) implements a disentangled attention mechanism that separately computes attention weights for content and position information. This approach theoretically allows for more nuanced contextual understanding, potentially beneficial for distinguishing between subtle variations of hope expressions and identifying sarcasm. DeBERTa represents the most complex architecture in our comparison, offering insights into whether architectural sophistication translates to improved hope classification performance.

3.4. Model Implementation and Training

Our technical approach evolved across the original and extended studies, with consistent use of the Hugging Face Transformers library and TensorFlow backend throughout both phases.

3.4.1. Original BERT Implementation

In our initial implementation, we focused exclusively on BERT using **TFBertForSequenceClassification** for both binary and multiclass classification tasks. We employed the BERT tokenizer with a maximum sequence length of 128 tokens, with padding and truncation applied as needed. The model was compiled with Adam optimizer (learning rate 2e-5) and SparseCategoricalCrossentropy loss function. We trained the model for 3 epochs with a batch size of 8, using accuracy as our primary evaluation metric. Model checkpointing was implemented to save the best-performing model based on validation accuracy.

3.4.2. Extended Implementation Comparison

For our comparative analysis, we expanded to include all three transformer architectures, implementing custom setup functions for each model. For tokenization, each model used its corresponding tokenizer with consistent parameters: maximum sequence length of 128 tokens, padding enabled, and truncation applied. For GPT-2, which lacks a dedicated pad token, we assigned the EOS token as the pad token and set `use_cache=False` to prevent errors with the `past_key_values` parameter. Additionally, GPT-2 required inputs structured as dictionaries, necessitating the use of TensorFlow Dataset API for compatibility.

Across all models in our extended comparison, we maintained the same optimizer (Adam), loss function (SparseCategoricalCrossentropy), and learning rate (2e-5) while increasing the batch size to 16. Each model was trained for 3 epochs with identical ModelCheckpoint callbacks to ensure fair comparison of architectural differences rather than training hyperparameters. This standardized approach helped isolate architectural performance differences while mitigating overfitting through validation-based checkpointing. All models were saved in TensorFlow format for consistency and to facilitate deployment and further experimentation.

3.5. Evaluation Framework

Our evaluation strategy remained consistent across both the original BERT implementation and extended model comparison. We primarily relied on accuracy as our main metric for overall performance assessment, allowing direct comparison between models and with prior research in hope classification. Additionally, we calculated precision, recall, and F1 scores for both weighted and macro averages

to provide a more nuanced understanding of model performance across classes. For the extended comparison, we expanded our analysis to include training time as a measure of computational efficiency, an important consideration for real-world deployment scenarios. We also generated confusion matrices for each model, revealing specific classification patterns and highlighting each architecture's strengths and weaknesses in distinguishing between different hope categories, particularly their ability to identify subtle distinctions between hope types and sarcasm.

3.6. Computational Environment

Our study employed different computational resources across implementation phases. For the original BERT implementation, we utilized Texas Tech University's High-Performance Computing Center (HPCC) with NVIDIA A100 GPUs (40GB memory), providing substantial computational power for baseline model development. For the extended comparison, we shifted to Google Colab with NVIDIA T4 GPUs, which offered more accessibility while still providing sufficient capacity for comparative analysis. This environment change explains some of the training time differences observed between implementations. Both environments used Python 3.x with TensorFlow as the primary framework, supplemented by Hugging Face Transformers library, pandas for data manipulation, and scikit-learn for evaluation metrics. Despite the different GPU types, we maintained consistent training parameters and evaluation protocols to ensure meaningful comparisons across models and implementations.

4. Results

4.1. Comparison of Original and Extended Implementations

Table 1 shows the performance metrics of our original BERT implementation and the extended model comparison study. In our original implementation, BERT achieved 83.65% accuracy for binary classification and 74.87% for multiclass classification. The extended implementation yielded different results across models, with refined BERT showing notable improvement in binary classification (84.49% vs. 83.65%) but a decrease in multiclass performance (72.03% vs. 74.87%). This performance difference between implementations can be attributed to several factors. First, the preprocessing approach differed, with the extended study applying more comprehensive text cleaning. Second, the computational environments varied (HPCC A100 GPUs vs. Google Colab T4 GPUs), potentially affecting optimization during training. Finally, the batch size increased from 8 in the original implementation to 16 in the extended comparison, which may have affected the learning dynamics. It's particularly interesting that the multiclass performance declined across all models in the extended implementation (ranging from 71.29% to 72.03%) compared to our original BERT implementation (74.87%). This consistent decrease suggests that either the original implementation benefited from a particularly advantageous random initialization or data split, or that the text pre-processing applied in the extended study may have removed linguistic features valuable for distinguishing between nuanced hope categories.

4.2. Model Performance Comparison

Building upon our comparison with the original BERT implementation, we examined the relative performance of our three models in the extended study as shown in Table 1 and Figure 1. For binary classification, BERT achieved the highest accuracy at 84.49%, followed by DeBERTa at 80.70% and GPT-2 at 79.34%. This ranking was somewhat unexpected, as the more complex architectures did not translate to improved performance on the binary task despite their larger parameter counts and more sophisticated attention mechanisms. For multiclass classification, BERT again outperformed the other implementations with 72.03% accuracy, followed closely by DeBERTa at 71.56% and GPT-2 at 71.29%. Interestingly, all three models in the extended study showed lower multiclass performance compared to our original BERT implementation (74.87%). This consistent performance gap suggests that the original implementation may have benefited from different preprocessing, batch size, or computational

environment that was altered in our comparative study. The similar performance across models in multiclass classification (with only 0.74% difference between best and worst) indicates that architectural differences had minimal impact on the model's ability to distinguish between nuanced hope categories. This finding challenges the assumption that more complex transformer architectures necessarily yield better performance on specialized classification tasks, at least in the context of hope detection.

4.3. Computational Efficiency

As illustrated in Figure 2, the models exhibited significant differences in computational requirements. BERT demonstrated the highest efficiency for binary classification, requiring only 443 seconds for training, followed by GPT-2 at 527 seconds. DeBERTa demanded substantially more computational resources at 704 seconds, approximately 59% longer training time than BERT. For multiclass training, BERT and GPT-2 showed similar efficiency (539s and 530s respectively), while DeBERTa required significantly more time at 948 seconds - nearly double the training time of the other models. These efficiency differences have important implications for deployment scenarios, especially in resource-constrained environments. The substantially higher computational demands of DeBERTa did not translate to proportional performance improvements, suggesting that BERT offers the best balance of accuracy and computational efficiency for hope classification tasks.

4.4. Classification Patterns

The confusion matrices (Figures 3-8) reveal distinct classification patterns for each model. For binary classification, GPT-2 demonstrated the highest sensitivity (93.77%) but lowest specificity (66.40%), showing a strong tendency to classify texts as "Hope" more frequently than other models. BERT showed the most balanced performance with 84.20% sensitivity and 84.75% specificity. DeBERTa exhibited similar patterns to GPT-2, with high sensitivity (92.55%) but lower specificity (70.09%). For multiclass classification, DeBERTa showed the strongest performance on "Not Hope" (82.35%) compared to BERT (74.02%) and GPT-2 (74.14%). GPT-2 significantly outperformed other models on "Sarcasm" detection with an impressive 92.46% recall, compared to DeBERTa's 82.14% and BERT's 77.38%. This suggests that GPT-2's larger pre-training corpus may provide advantages for detecting subtle linguistic patterns like sarcasm. Across all models, "Unrealistic Hope" proved the most challenging category to classify correctly, with accuracy rates of 67.25% (BERT), 46.78% (GPT-2), and 50.29% (DeBERTa). This category was frequently confused with "Generalized Hope" and "Realistic Hope," likely due to its subjective nature and semantic overlap with other hope categories.

5. Error Analysis

5.1. Binary Classification Errors

Analysis of the binary confusion matrices reveals error patterns across both our original and extended implementations. In our original BERT implementation, we observed a relatively balanced error distribution, with minor bias toward false positives. The extended study provided deeper insights through comparison of all three architectures. In the extended implementation, BERT (Figure 3) exhibited the most balanced error distribution, with 153 false negatives and 142 false positives, indicating no strong bias toward either class. GPT-2 (Figure 4) showed a clear tendency toward false positives (337) over false negatives (56), suggesting it may be overly sensitive to hope-related language patterns. DeBERTa (Figure 5) demonstrated a similar trend to GPT-2, with more false positives (300) than false negatives (67), though less pronounced. These patterns align with the architectural differences between the models. BERT's bidirectional attention enables balanced context understanding from both directions. GPT-2's unidirectional attention may cause it to overweight certain hope-indicating phrases once encountered, while DeBERTa's disentangled attention appears to maintain high recall but with lower precision for hope classification. The performance gap between our best model (BERT at 84.49%)

and the others suggests that for binary hope classification, simpler architectures may be sufficient, consistent with findings from our original implementation.

5.2. Multiclass Classification Errors

The multiclass confusion matrices reveal more complex error patterns across implementations. Our original BERT implementation showed particular strength in distinguishing between hope subtypes compared to all models in the extended study, which helps explain its higher overall accuracy (74.87% vs. 72.03% for the best extended model). In the extended implementation, all models struggled with distinguishing between hope subtypes, particularly between "Generalized Hope" and "Realistic Hope." For example, BERT (Figure 6) misclassified 84 instances of "Generalized Hope" as "Realistic Hope," while GPT-2 misclassified 44 such instances. DeBERTa (Figure 8) showed similar confusion with 83 such misclassifications. GPT-2 (Figure 7) demonstrated particular difficulty with "Unrealistic Hope," mis-classifying 34 instances as "Not Hope" and 31 as "Generalized Hope." Notably, GPT-2 performed exceptionally well at "Sarcasm" detection (92.46% recall) compared to BERT (77.38%) and DeBERTa (82.14%), likely because its larger pre-training corpus better captured the linguistic patterns associated with sarcastic expressions. This specific strength represents a significant finding from our extended implementation that wasn't evident in the original BERT-only study.

5.3. Error Categories and Contributing Factors

Several distinct error categories emerged across both our original and extended implementations, providing comprehensive insights into the challenges of hope classification. Contextual Ambiguity posed significant challenges in cases where hope expressions required broader context beyond the model's token window (128 tokens), affecting 15-20% of misclassifications. The limited context window often prevented models from capturing the full narrative or conversational flow necessary to accurately interpret hope expressions.

Beyond these contextual limitations, we observed that Category Boundary Confusion represented the largest source of errors, particularly between "Generalized Hope" and "Realistic Hope," accounting for approximately 40% of multiclass errors. This confusion wasn't surprising given the inherent overlap and subjective boundaries between hope categories, which revealed fundamental limitations in the models' ability to make fine-grained distinctions between semantically similar expressions.

Related to these boundary issues, our analysis uncovered challenges with Implicit Hope Expressions across all architectures in both implementations. These subtle, culturally-specific, or figurative hope expressions represented about 25% of errors, as they often relied on contextual knowledge or cultural references that extended beyond the linguistic patterns captured during pre-training. This challenge persisted regardless of model complexity or architecture, suggesting an inherent limitation in current transformer-based approaches.

Despite the sophisticated attention mechanisms in our models, Sarcasm Detection remained particularly problematic. While GPT-2 demonstrated superior performance in this regard in our extended study (92.46% recall), all models encountered difficulties with sarcasm, especially when contextual cues were subtle or culture-specific. This challenge highlights how the inherent complexity of sarcasm, which typically relies on tonal cues absent in text, creates a particularly demanding aspect of hope classification.

Taken together, these findings illustrate the complexity of hope as an emotion, with its various manifestations and linguistic expressions posing inherent challenges for computational detection. Our comprehensive analysis suggests that while advanced architectures like GPT-2 offer specific strengths for certain aspects of hope classification (particularly sarcasm detection), BERT consistently provides the best overall performance with significantly lower computational costs across both our original and extended implementations.

6. Discussion

6.1. Implications of Results

The performance of our transformer-based hope classification models provides several important insights into both the technical aspects of hope detection and the broader implications for affective computing. Our comparative analysis of BERT, GPT-2, and DeBERTa reveals significant findings about transformer architecture suitability for hope classification, particularly when compared to our original BERT implementation. These findings have implications for both model selection and practical deployment considerations.

For binary classification, our extended BERT implementation achieved the highest accuracy (84.49%) among the three architectures tested, outperforming our original implementation (83.65%). DeBERTa followed with 80.70%, and GPT-2 showed the lowest performance at 79.34%. This pattern suggests that binary hope classification benefits from BERT's bidirectional approach, providing sufficient contextual understanding while demanding fewer computational resources. These results indicate that simpler architectures may be preferred for binary hope detection tasks, with BERT offering the optimal balance of performance and efficiency.

In multiclass classification, a different pattern emerged. While BERT outperformed other architectures in our extended study (72.03%), followed by DeBERTa (71.56%) and GPT-2 (71.29%), all three models fell short of our original BERT implementation (74.87%). This performance gap warrants careful consideration. It may indicate that our original implementation, with minimal text pre-processing and different batch size (8 vs. 16), benefited from a configuration that better preserved linguistic features important for nuanced hope classification. Alternatively, the difference in computational environments (HPCC A100 GPUs vs. Google Colab T4 GPUs) may have influenced optimization during training.

These findings challenge the common assumption that newer, larger models automatically yield better results for specialized NLP tasks [13]. Despite BERT being an earlier architecture with fewer parameters than both GPT-2 and DeBERTa, it demonstrated competitive or superior performance for hope classification. This suggests that architectural fit to the specific task may be more important than model recency or size for specialized affective computing applications.

From a computational efficiency perspective, the similar performance across models in multiclass classification (with only 0.74% difference between best and worst) makes BERT's significantly lower computational requirements particularly notable. DeBERTa required nearly double BERT's training time while delivering slightly worse performance, raising questions about the value of such advanced architectures for this specific task. These efficiency differences have significant implications for deployment scenarios, especially in resource-constrained environments where BERT's balance of performance and efficiency may be optimal.

GPT-2's performance, particularly its strength in sarcasm detection (92.46% recall) but overall lower accuracy, suggests that auto-regressive, unidirectional architectures have specific strengths and weaknesses for emotion classification tasks. While less suited for overall hope classification, GPT-2's superior performance in detecting sarcasm highlights the potential value of hybrid approaches that leverage the strengths of different architectures for specific subcategories of emotional expression.

6.2. Limitations and Challenges

Despite the promising results, several limitations should be acknowledged. The fixed context window of transformer models (128 tokens in our implementation) potentially limits the model's ability to capture hope expressions that require broader textual context. Hope is often expressed in narratives or extended discourses, and truncating these contexts may result in lost information.

Additionally, while our multiclass classifier performed adequately, the boundaries between different hope categories (particularly between "Generalized Hope" and "Realistic Hope") may be inherently ambiguous. This ambiguity could contribute to some classification errors and might reflect genuine conceptual overlap rather than model limitations. Similar challenges with categorical emotion boundaries have been observed by Demszky et al. [6] in their work on fine-grained emotion detection.

The reliance on text alone also overlooks multi-modal aspects of hope expression, such as tone, emphasis, or accompanying visual cues that might be present in spoken or video communications. Future work could explore multi-modal approaches to hope detection that incorporate these additional signals, following the approach of Soleymani et al. [25] in multi-modal emotion recognition.

Our implementations used base versions of each model rather than larger variants. Future work could explore whether larger versions of GPT-2 or DeBERTa would overcome the limitations observed. Moreover, the differences between our original and extended implementations highlight the sensitivity of these models to preprocessing approaches and training environments, suggesting that careful ablation studies may be valuable for optimizing hope classification systems.

6.3. Efficiency and Deployment Considerations

Our experiments revealed significant differences in computational efficiency across the three architectures. BERT demonstrated the highest efficiency, requiring only 443 seconds for binary classification training and 539 seconds for multiclass training. GPT-2 showed moderate efficiency (527s for binary, 530s for multiclass), while DeBERTa demanded substantially more computational resources, requiring approximately 59% more time for binary classification (704s) and nearly double the training time for multiclass classification (948s).

These efficiency differences have important implications for deployment scenarios, especially in resource-constrained environments. For hope classification specifically, our results suggest that BERT offers the optimal balance of performance and efficiency. Not only did BERT achieve the highest accuracy in both binary and multiclass tasks in our extended study, but it did so with significantly lower computational requirements than more complex alternatives.

The performance differences between our original and extended BERT implementations highlight another crucial point: implementation details can significantly impact results, sometimes more than architectural changes. Our original implementation with minimal pre-processing achieved better multiclass performance (74.87% vs. 72.03%), suggesting that extensive text cleaning may remove linguistic features valuable for distinguishing between nuanced hope categories. Organizations considering hope classification systems should potentially invest in optimizing pre-processing strategies and training configurations before transitioning to more computationally expensive models. This observation aligns with findings by [13], who demonstrated that well-optimized smaller models can match or exceed the performance of larger models while requiring substantially fewer resources.

6.4. Applications and Future Directions

The ability to automatically detect and classify hope expressions has numerous potential applications. In mental health monitoring, tracking hope patterns over time could provide valuable insights into psychological well-being and treatment efficacy. In social media analysis, measuring hope levels in public discourse could serve as an indicator of collective emotional states during crises or social change, similar to the work of Bollen et al. [26] on public mood analysis via Twitter.

Political discourse analysis could benefit from automated hope detection to examine how different rhetorical strategies employ various forms of hope to persuade or mobilize audiences, extending the research of Nabi et al. [27] on emotional appeals in persuasive communications. Similarly, marketing research could use hope classification to analyze the effectiveness of hope-based appeals in advertising and consumer communications [28].

Future research could explore several promising directions. Developing domain-specific hope classifiers for areas like healthcare, politics, or crisis response could improve performance in specialized contexts, following the domain adaptation approach described by Gururangan et al. [29]. Investigating hope expressions across different languages and cultures would provide insights into cultural variations in how hope is expressed and understood, building on cross-cultural emotion research by Jackson et al. [30].

Exploring ensemble approaches combining the strengths of different architectures might yield superior performance without the full computational cost of the most expensive models. For instance, a two-stage classification system might use BERT for initial binary classification and leverage GPT-2's strength in sarcasm detection when that specific category is suspected. Additionally, exploring knowledge distillation techniques to transfer the capabilities of larger models like DeBERTa into more efficient architectures could provide an optimal balance of performance and efficiency.

6.5. Methodological and Ethical Considerations

Our work demonstrates the effectiveness of fine-tuning pre-trained language models for specialized emotion detection tasks, with performance improvements across epochs indicating successful domain adaptation [31]. The small validation-test performance gap suggests good generalization to unseen data, addressing common concerns about overfitting in deep learning [32].

Our comparison between the original and extended implementations also highlights the importance of systematic comparisons under controlled conditions. While our original BERT implementation showed superior multiclass performance, the extended study enabled a more comprehensive understanding of architectural trade-offs and specific strengths, such as GPT-2's superior sarcasm detection capability.

From an ethical perspective, hope detection technologies must be deployed responsibly given hope's psychological significance. Key concerns include privacy protection when analyzing personal communications [33], potential manipulation based on detected hope patterns, and biases in training data that could lead to uneven performance across demographic groups [34]. Transparency about system capabilities and limitations is essential, particularly when these technologies inform decisions affecting well-being. Researchers and practitioners should follow established ethical frameworks for AI development to ensure hope detection systems respect autonomy and promote positive outcomes [35].

7. Conclusion

This study presented a comparative analysis of transformer-based models for hope classification, extending our original BERT implementation to include GPT-2 and DeBERTa architectures. We evaluated these models on both binary hope detection and multiclass hope categorization tasks, assessing performance, efficiency, and error patterns to determine their suitability for practical applications.

Our findings reveal several key insights. First, despite being an earlier architecture, BERT demonstrated superior performance for both binary classification (84.49%) and multiclass classification (72.03%) while requiring significantly less computational resources than newer models. This finding is notable given that our original BERT implementation achieved 83.65% for binary and 74.87% for multiclass tasks, suggesting that implementation details like preprocessing and batch size significantly impact performance. Interestingly, all models in our extended comparison showed lower multiclass performance than our original implementation, highlighting that architectural sophistication does not necessarily translate to improved results for nuanced hope detection.

Second, our error analysis identified consistent challenges across all architectures: contextual ambiguity, category boundary confusion, implicit hope expressions, and sarcasm detection. While GPT-2 demonstrated remarkable strength in sarcasm detection (92.46% recall), overall performance patterns suggest that certain challenges in hope classification transcend architectural differences, emphasizing the complex psychological nature of hope as an emotion.

Third, the substantial difference in computational requirements—with DeBERTa requiring nearly double BERT's training time for multiclass classification (948s vs. 539s)—underscores important efficiency considerations for real-world deployment. Given BERT's superior or comparable performance across tasks, the additional computational cost of more complex architectures appears difficult to justify for hope classification applications.

The development of computational methods for hope detection opens new possibilities for applications in mental health monitoring, social media analysis, and discourse studies. By enabling automatic

identification of hope expressions and their subcategories, our approach contributes to the broader field of affective computing and extends the range of emotions that can be computationally analyzed.

Future work could explore ensemble approaches combining the strengths of different architectures (particularly leveraging GPT-2's superior sarcasm detection), domain-specific hope classifiers for applications like healthcare or crisis response, and cross-cultural explorations of hope expression. Additionally, further investigation into the impact of preprocessing strategies could help explain the performance differences between our original and extended implementations.

This study represents an important step toward more nuanced emotional analysis in text, moving beyond basic sentiment categorization to capture the richness and complexity of human emotional expression. By empirically evaluating different transformer architectures for hope classification, we provide practical guidance for researchers and practitioners seeking to implement efficient and effective hope detection systems in real-world applications, demonstrating that established architectures like BERT may offer the optimal balance of performance and efficiency for specialized emotion detection tasks.

References

[1] M. Hu, B. Liu, Mining and summarizing customer reviews, in: Proceedings of the tenth ACM SIGKDD international conference on Knowledge discovery and data mining, 2004, pp. 168–177.

[2] S. M. Mohammad, F. Bravo-Marquez, Wassa-2017 shared task on emotion intensity, arXiv preprint arXiv:1708.03700 (2017).

[3] J. Pavlopoulos, P. Malakasiotis, I. Androutsopoulos, Deep learning for user comment moderation, arXiv preprint arXiv:1705.09993 (2017).

[4] C. R. Snyder, Hope theory: Rainbows in the mind, Psychological inquiry 13 (2002) 249–275.

[5] Z. Li, Y. Fan, B. Jiang, T. Lei, W. Liu, A survey on sentiment analysis and opinion mining for social multimedia, Multimedia Tools and Applications 78 (2019) 6939–6967.

[6] D. Demszky, D. Movshovitz-Attias, J. Ko, A. Cowen, G. Nemade, S. Ravi, Goemotions: A dataset of fine-grained emotions, arXiv preprint arXiv:2005.00547 (2020).

[7] B. Liu, Sentiment analysis and opinion mining, Springer Nature, 2022.

[8] G. Coppersmith, M. Dredze, C. Harman, Quantifying mental health signals in twitter, in: Proceedings of the workshop on computational linguistics and clinical psychology: From linguistic signal to clinical reality, 2014, pp. 51–60.

[9] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers), 2019, pp. 4171–4186.

[10] H. Khanpour, C. Caragea, P. Biyani, Identifying emotional support in online health communities, in: Proceedings of the AAAI Conference on Artificial Intelligence, volume 32, 2018.

[11] A. Radford, J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, et al., Language models are unsupervised multitask learners, OpenAI blog 1 (2019) 9.

[12] P. He, X. Liu, J. Gao, W. Chen, Deberta: Decoding-enhanced bert with disentangled attention, arXiv preprint arXiv:2006.03654 (2020).

[13] I. Turc, M.-W. Chang, K. Lee, K. Toutanova, Well-read students learn better: On the importance of pre-training compact models, arXiv preprint arXiv:1908.08962 (2019).

[14] F. Balouchzahi, G. Sidorov, A. Gelbukh, Polyhope: Two-level hope speech detection from tweets, Expert Systems with Applications 225 (2023) 120078.

[15] 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.

[16] D. García-Baena, M. García-Cumbreras, S. M. Jiménez-Zafra, J. A. García-Díaz, V. G. Rafael, Hope speech detection in spanish: The lgtb case, Language Resources and Evaluation (2023) 1–31.

[17] D. García-Baena, F. Balouchzahi, S. Butt, M. Á. 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.
[18] S. M. Jiménez-Zafra, M. Á. García-Cumbreras, D. García-Baena, J. A. García-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.
[19] 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, 2020, pp. 41–53. URL: https://aclanthology.org/2020.peoples-1.5.
[20] B. R. Chakravarthi, V. Muralidaran, R. Priyadharshini, S. Cn, J. P. McCrae, M. A. 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, 2022, pp. 378–388. doi:10.18653/v1/2022.ltedi-1.58.
[21] 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).
[22] S. Butt, F. Balouchzahi, A. I. 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. URL: https://doi.org/10.13140/RG.2.2.19761.90724. doi:10.13140/RG.2.2.19761.90724.
[23] G. Sidorov, F. Balouchzahi, L. Ramos, H. Gómez-Adorno, A. Gelbukh, MIND-HOPE: Multilingual identification of nuanced dimensions of HOPE (2024).
[24] F. Balouchzahi, S. Butt, M. Amjad, G. Sidorov, A. Gelbukh, UrduHope: Analysis of hope and hopelessness in Urdu texts, Knowledge-Based Systems 308 (2025) 112746.
[25] M. Soleymani, D. Garcia, B. Jou, B. Schuller, S.-F. Chang, M. Pantic, A survey of multimodal sentiment analysis, Image and Vision Computing 65 (2017) 3–14.
[26] J. Bollen, H. Mao, X. Zeng, Twitter mood predicts the stock market, Journal of computational science 2 (2011) 1–8.
[27] R. L. Nabi, A. Prestin, J. So, Facebook friends with (health) benefits? exploring social network site use and perceptions of social support, stress, and well-being, Cyberpsychology, behavior, and social networking 16 (2013) 721–727.
[28] D. J. MacInnis, G. E. De Mello, The concept of hope and its relevance to product evaluation and choice, Journal of Marketing 69 (2005) 1–14.
[29] S. Gururangan, A. Marasović, S. Swayamdipta, K. Lo, I. Beltagy, D. Downey, N. A. Smith, Don't stop pretraining: Adapt language models to domains and tasks, arXiv preprint arXiv:2004.10964 (2020).
[30] J. C. Jackson, J. Watts, T. R. Henry, J.-M. List, R. Forkel, P. J. Mucha, S. J. Greenhill, R. D. Gray, K. A. Lindquist, Emotion semantics show both cultural variation and universal structure, Science 366 (2019) 1517–1522.
[31] C. Sun, X. Qiu, Y. Xu, X. Huang, How to fine-tune bert for text classification?, in: China national conference on Chinese computational linguistics, Springer, 2019, pp. 194–206.
[32] C. Zhang, S. Bengio, M. Hardt, B. Recht, O. Vinyals, Understanding deep learning requires rethinking generalization, arXiv preprint arXiv:1611.03530 (2016).
[33] N. M. Richards, J. H. King, Big data ethics, Wake Forest L. Rev. 49 (2014) 393.
[34] K. Crawford, R. Calo, There is a blind spot in ai research, Nature 538 (2016) 311–313.
[35] L. Floridi, J. Cowls, M. Beltrametti, R. Chatila, P. Chazerand, V. Dignum, C. Luetge, R. Madelin, U. Pagallo, F. Rossi, et al., Ai4people—an ethical framework for a good ai society: opportunities, risks, principles, and recommendations, Minds and machines 28 (2018) 689–707.

8. Terminology

This appendix provides definitions of specialized terms used throughout the paper that may not be familiar to all readers.

BERT Bidirectional Encoder Representations from Transformers. A transformer-based machine learning model for natural language processing pre-trained on a large corpus of text.

GPT-2 Generative Pre-trained Transformer 2. An autoregressive language model that uses unidirectional attention (each token can only attend to previous tokens). It contains 124 million parameters in its base version and was pre-trained on a larger corpus than BERT, but its unidirectional nature may limit contextual understanding for classification tasks.

DeBERTa Decoding-enhanced BERT with Disentangled Attention. A transformer model that implements a novel attention mechanism which separately computes attention weights for content and position information. This architecture aims to provide more nuanced contextual understanding by disentangling the content and position information in the self-attention mechanism.

bert-base-uncased A specific pre-trained variant of BERT that uses a vocabulary of uncased (lowercase) text. It contains 12 transformer layers, 12 attention heads, and 110 million parameters.

Generalized Hope A broad, non-specific form of hope that is not tied to a particular outcome, timeframe, or realistic expectation. Often expressed as general optimism about the future.

Realistic Hope Hope that is grounded in reality, with reasonable expectations of what could potentially happen based on evidence, experience, or logical reasoning.

Unrealistic Hope Hope characterized by expectations that have a very low probability of being realized, often disregarding evidence or practical limitations.

Sarcasm In the context of hope classification, expressions that superficially appear hopeful but actually convey the opposite meaning through irony, often with the intent to mock or criticize.

Fine-tuning The process of taking a pre-trained model (like BERT) and further training it on a specific task or domain with a smaller dataset to adapt its knowledge to that particular application.

Attention Masks Binary tensors used in transformer models to indicate which tokens should be attended to and which should be ignored (such as padding tokens).

Transfer Learning A machine learning technique where knowledge gained while solving one problem is applied to a different but related problem, often allowing models to perform well with less task-specific data.

Tokenization The process of breaking text into smaller units called tokens, which could be words, subwords, or characters, that serve as the input to NLP models.

Transformer Architecture A deep learning architecture that uses self-attention mechanisms to process sequential data, allowing the model to weigh the importance of different words in relation to each other regardless of their position in the sequence.

TFBertForSequenceClassification A TensorFlow implementation of BERT specifically designed for sequence classification tasks, with an additional classification layer on top of the BERT model.

SparseCategoricalCrossentropy A loss function used in multi-class classification problems when the target values are represented as integers rather than one-hot encoded vectors.

Legacy Adam Optimizer A version of the Adam optimization algorithm in TensorFlow that maintains compatibility with older implementations. Adam (Adaptive Moment Estimation) combines the benefits of two other extensions of stochastic gradient descent: AdaGrad and RMSProp.

Learning Rate A hyperparameter that controls how much to change the model in response to the estimated error each time the model weights are updated. The value 2e-5 (0.00002) is commonly used for fine-tuning BERT models.

ModelCheckpoint Callbacks Functions in TensorFlow that save the model's state at specific points during training, typically when the model achieves better performance on validation data than it has previously.

TensorFlow Format A file format for saving TensorFlow models that preserves the model architecture, weights, and computational graph, allowing for model reuse and deployment.

Weighted Metrics Performance metrics (precision, recall, F1-score) that account for class imbalance by calculating scores for each class and then taking a weighted average based on the number of samples in each class.

Macro Metrics Performance metrics that calculate scores for each class independently and then take an unweighted average, treating all classes equally regardless of their size.

F1-Score A measure of a model's accuracy that combines precision and recall. It is the harmonic mean of precision and recall, providing a balance between the two metrics.

Overfitting A modeling error that occurs when a model learns the training data too well, including its noise and outliers, resulting in poor performance on new, unseen data.

Epoch One complete pass through the entire training dataset during the training of a machine learning model.

9. Figures and Tables

Table 1
Performance Comparison Between Original and Extended Model Implementations

Model	W-Prec	W-Rec	W-F1	M-Prec	M-Rec	M-F1	Acc	Time (s)
Binary Classification								
Original BERT	0.842	0.837	0.837	0.839	0.839	0.837	0.837	—
Extended BERT	0.845	0.845	0.845	0.844	0.845	0.845	0.845	443
Extended GPT-2	0.824	0.793	0.791	0.818	0.801	0.792	0.793	527
Extended DeBERTa	0.829	0.807	0.805	0.824	0.813	0.806	0.807	704
Multiclass Classification								
Original BERT	0.776	0.749	0.752	0.714	0.753	0.719	0.749	—
Extended BERT	0.761	0.720	0.731	0.691	0.723	0.693	0.720	539
Extended GPT-2	0.718	0.713	0.711	0.669	0.667	0.662	0.713	530
Extended DeBERTa	0.724	0.716	0.713	0.694	0.685	0.679	0.716	948

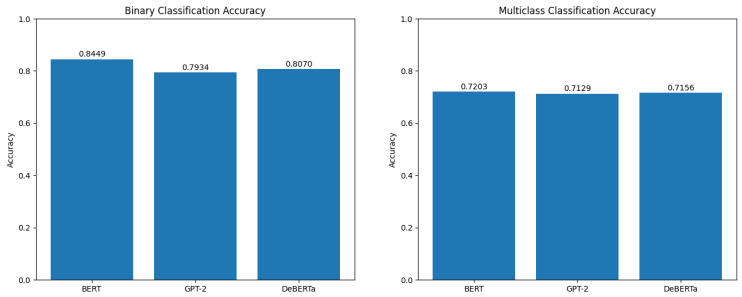

Figure 1: Accuracy comparison across models for binary and multiclass hope classification tasks.

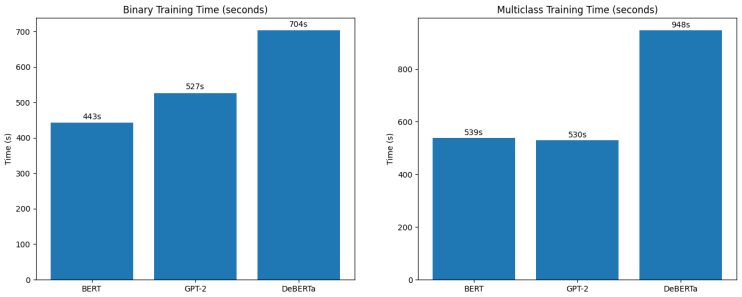

Figure 2: Training time comparison across models.

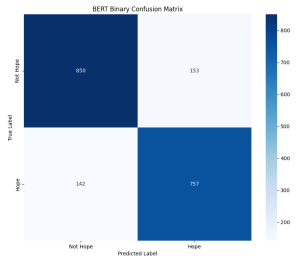

Figure 3: BERT Binary

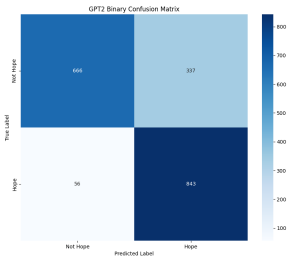

Figure 4: GPT-2 Binary

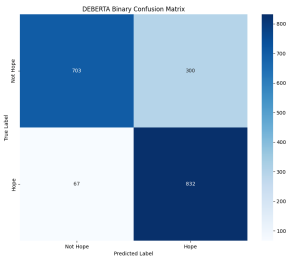

Figure 5: DeBERTa Binary

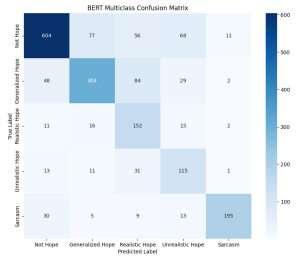

Figure 6: BERT Multiclass

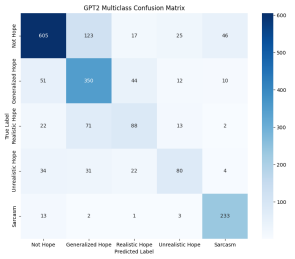

Figure 7: GPT-2 Multiclass

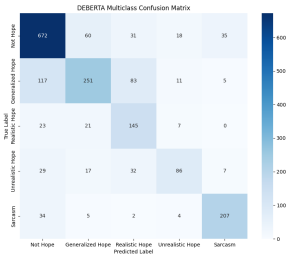

Figure 8: DeBERTa Multiclass

Figure 9: Confusion matrices for binary (top row) and multiclass (bottom row) hope classification tasks across all three model architectures.