International Journal of Computer Applications Technology and Research
Volume 14–Issue 10, 82 – 91, 2025, ISSN:-2319–8656
DOI:10.7753/IJCATR1410.1014

Implicit Negation based Polarity Shift Management using a Joint Optimization of LSTM based BERT and Contextual Back Translation Augmented with Seq2Seq Perturbations

Millicent Kathambi Murithi
Department of Computer Science
Murang'a university of Technology
Kenya

Aaron Mogeni Oirere
Department of Computer Science
Murang'a university of Technology
Kenya

Abstract: Sentiment analysis has become an indispensable tool for understanding consumer perceptions, yet its reliability is undermined by implicit negation, which triggers polarity shifts that reverse the intended sentiment. Despite progress in transformer-based and deep learning models, existing approaches remain limited in their ability to capture such subtle semantic inversions. We propose a hybrid deep learning architecture that combines an LSTM-enhanced BERT encoder with contextual back-translation and sequence-to-sequence perturbations for robust data augmentation. FastText and BERT embeddings are jointly leveraged with attention mechanisms to capture both long-range dependencies and nuanced contextual cues. The model was trained and validated on multiple mobile review datasets, and its performance was benchmarked against strong state-of-the-art baselines. Evaluation employed accuracy, Cohen's kappa, and Matthews's correlation coefficient, with statistical significance assessed through ANOVA and Kruskal–Wallis H statistic tests.The proposed framework consistently outperformed baseline models across all datasets, achieving superior scores on accuracy, kappa, and MCC. Statistical analyses confirmed that these improvements were significant. Visual inspection via attention heatmaps revealed enhanced sensitivity to negation cues, while box–whisker plots demonstrated greater robustness, with higher medians and reduced variance.These findings establish that integrating hybrid architectures with targeted augmentation strategies markedly improves the detection of polarity shifts induced by implicit negation. The approach enhances both accuracy and stability, offering a pathway toward more reliable sentiment analysis in linguistically complex contexts. Future research may extend this work by combining multiple augmentation strategies to further improve generalization across domains and languages.

Keywords: implicit negation, polarity shift, sentiment analysis, hybrid deep learning, transformers, data augmentation

1. INTRODUCTION

The explosive growth of the internet has generated vast volumes of user-generated content, offering immense opportunities for market research, opinion mining, and social media analytics (Kafi et al., 2019). This content, typically unstructured and dynamic, exhibits high volume, velocity, and variety, which poses significant challenges for extracting actionable insights. Natural Language Processing (NLP), particularly sentiment analysis, has emerged as a critical tool for identifying and classifying subjective opinions into positive, negative, or neutral polarities (Abirami & Gayathri, 2017; Khan et al., 2014). Sentiment analysis can be performed at the aspect level, distinguishing between explicit features, which are directly stated, and implicit features, which must be inferred from context. For instance, *"The Samsung phone has excellent features"* conveys explicit positivity, while *"Samsung phones require time to fully adapt to"* implicitly expresses negativity despite its surface-neutral tone (Bordoloi & Biswas, 2023). Empirical studies estimate that nearly 30% of consumer reviews lack explicit opinion words but still convey sentiment orientation (Li et al., 2021). A persistent challenge in this domain is polarity shift, the misclassification of sentiment due to negation, contrastive structures, or context-dependent meaning (Blázquez-López, 2022). Such errors undermine classifier reliability and interpretability, particularly in scenarios involving implicit negation, where sentiment is obscured by subtle linguistic cues. While deep learning methods and transformer-based models (e.g., BERT, RoBERTa) have advanced the field (Pipalia et al., 2020), and data augmentation techniques such as contextual back-translation and Seq2Seq paraphrasing have enriched training corpora (Satapathy et al., 2019), existing approaches still fall short in three respects. Firstly, they focus predominantly on explicit negation, leaving implicit negation underexplored. Secondly, they often remain domain-dependent, failing to generalize across datasets where polarity varies with context. Lastly, they rarely combine augmentation and hybrid deep architectures in a joint optimization framework for robust polarity-shift detection. This study, therefore, addresses these gaps by introducing a novel hybrid sentiment classification model tailored for implicit negation. The novelty of this work lies in the integration of transformer-based embeddings (BERT), recurrent sequence encoders (LSTM), and contextual augmentation (Seq2Seq perturbation + back-translation) for polarity shift management. It also lies on a joint optimization strategy that enhances both representational power and robustness against linguistic variation. Lastly, this new model's novelty lies on a comprehensive evaluation pipeline incorporating statistical significance testing and interpretability analysis, including error case studies. The study is guided by the following research questions:

1. How can a hybrid sentiment classification model that integrates transformer architectures, recurrent encoders, and augmentation techniques be developed to effectively detect polarity shifts arising from implicit negation?
2. How does the proposed hybrid model compare with existing state-of-the-art methods in detecting implicit negation across diverse benchmark datasets?
3. To what extent does the model maintain performance stability across domains and linguistic variations introduced through augmentation?

2. RELATED WORKS

2.1 Sentiment analysis and polarity shifts

Sentiment analysis has evolved into a central task in NLP, enabling the automatic classification of subjective opinions into polarity classes such as positive, negative, or neutral (Khan et al., 2014). While early methods relied on lexicon-based approaches, advances in machine learning and deep learning have driven substantial performance gains (Abirami & Gayathri, 2017). Beyond document-level polarity, aspect-based sentiment analysis (ABSA) has emerged, focusing on sentiments expressed toward specific features of an entity (Bordoloi & Biswas, 2023). This distinction is critical for handling implicit sentiment**,** where opinion cues are not overtly expressed but inferred through contextual meaning (Li et al., 2021). A persistent challenge in sentiment analysis is the polarity shift problem, in which a classifier misinterprets sentiment due to negation, irony, or contextual reversal. For example, a superficially positive statement may mask implicit negativity, leading to systematic misclassification (Blázquez-López, 2022). Studies have shown that polarity shifts undermine both accuracy and interpretability, particularly when negation is implicit rather than explicit (Abuhammad & Ahmed, 2024).

2.2 Advances in deep learning for sentiment analysis

Recent years have witnessed the adoption of transformer-based architectures such as BERT, RoBERTa, and DistilBERT, which leverage contextual embeddings to capture nuanced semantic relationships (Pipalia et al., 2020). These models outperform traditional recurrent or convolutional architectures on benchmark sentiment tasks by effectively modeling long-range dependencies. Cross-lingual extensions like multilingual BERT (mBERT) have further expanded sentiment analysis to multilingual domains, demonstrating scalability across diverse datasets. Hybrid models that combine transformers with recurrent networks (e.g., LSTM encoders) have also been explored to strengthen temporal sequence modeling, particularly for tasks requiring fine-grained contextual disambiguation. However, while these architectures improve baseline performance, they remain sensitive to polarity-shifting cues, particularly implicit negation.

2.3 Data augmentation and robustness in NLP

To address data sparsity and improve generalization, data augmentation has become an essential strategy in sentiment analysis. Back-translation introduces paraphrases across languages to expand training corpora, while contextual word replacement leverages pre-trained embeddings for semantic-preserving substitutions (Sudheer et al., 2025). Seq2Seq perturbation, in particular, generates syntactic and lexical variations while maintaining sentiment polarity, thereby enriching the diversity of training samples (Satapathy et al., 2019). Although these augmentation methods improve robustness, most studies treat them as standalone enhancements rather than integrating them with hybrid deep learning models. As a result, improvements remain incremental, and their potential to mitigate implicit polarity shifts is underexplored.

2.4 Interpretability and transparency in sentiment classification

Beyond raw performance, explainability has become increasingly important for ensuring trust in sentiment analysis systems. Tools such as SHAP, LIME, and Integrated Gradients have been applied to visualize feature contributions and improve transparency (Sudheer et al., 2025). While these methods shed light on classifier decision-making, they have rarely been used to specifically analyze error cases of polarity shifts, leaving a critical interpretability gap.

2.5 Gap analysis

Taken together, the related works highlights significant progress in sentiment analysis through deep learning, transformer architectures, data augmentation, and interpretability methods. However, three key gaps remain. Firstly, there is negation Scope. Most studies primarily address explicit negation, leaving implicit negation insufficiently modeled. Secondly, there are integration limitations where the existing models rarely integrate augmentation strategies (e.g., Seq2Seq perturbations and back-translation) with hybrid architectures (transformers + recurrent encoders) in a unified framework. Lastly, there are gaps in robustness and interpretability where few works systematically evaluate performance stability across domains or provide transparent error analyses of polarity-shift misclassifications. These gaps underscore the need for a jointly optimized hybrid model that leverages augmentation, deep contextual embeddings, and recurrent modeling to address polarity shifts caused by implicit negation, while also incorporating interpretability and cross-domain robustness.

3. MATERIALS AND METHODS

This study employed a multi-stage methodology encompassing dataset selection, preprocessing, model development, training, and evaluation. The primary objective was to construct a hybrid sentiment classification model capable of detecting polarity shifts arising from implicit negation with higher accuracy and robustness than existing approaches. Multiple benchmark datasets of mobile phone reviews were utilized, as they are well suited for implicit sentiment detection given their mixture of explicit and implicit opinion expressions. Each dataset was subjected to rigorous preprocessing, including text normalization, tokenization, stop-word removal, and lemmatization, to ensure consistency and reduce noise. To mitigate data sparsity and enhance the model’s ability to generalize, two complementary augmentation strategies were applied. Contextual back-translation was used to generate semantically equivalent paraphrases by translating reviews through intermediate languages, while Seq2Seq perturbations introduced syntactic and lexical variations that preserved sentiment polarity. Together, these techniques expanded the training data and exposed the model to diverse linguistic patterns. The proposed architecture combines the representational strength of transformers with the sequential modeling capacity of recurrent networks. Specifically, pre-trained BERT embeddings were used to capture deep contextual representations of text, which were then fed into an LSTM encoder to model temporal dependencies and sequential nuances. An attention mechanism was integrated to highlight polarity-shifting cues such as implicit negation. This hybrid design was chosen to balance semantic depth with sensitivity to subtle contextual variations. Training was conducted using mini-batch gradient descent with Adam optimization, incorporating dropout regularization to prevent overfitting. Hyperparameters such as learning rate, batch size, and sequence length were tuned empirically to maximize validation performance. Evaluation of the proposed model was performed against strong baseline classifiers, including standard transformer and recurrent architectures. Performance was measured using accuracy, Cohen's kappa, and the Matthews Correlation Coefficient (MCC), which provide complementary perspectives on classification reliability. To assess the robustness of the observed improvements, statistical significance testing was carried out using one-way ANOVA and the Kruskal–Wallis H test. Additional interpretability analyses, including attention heatmaps and case-specific error inspection, were conducted to better understand how the model responded to implicit negation. Through this methodological framework, the study ensured not only rigorous benchmarking but also transparent evaluation of the proposed hybrid architecture in comparison with state-of-the-art sentiment analysis models.

3.1 Experimental Design

This study adopted an experimental research design that combined both quantitative and qualitative approaches to construct and validate the proposed hybrid deep learning model. The quantitative dimension focused on measuring predictive performance across multiple benchmark datasets, while the qualitative dimension emphasized model interpretability and error analysis. Experimental results were systematically tabulated and subjected to rigorous statistical evaluation. Analysis of variance (ANOVA) and the Kruskal–Wallis H test were employed to assess whether the observed performance differences between the proposed architecture and baseline models were statistically significant. These complementary tests ensured robustness in the interpretation of results across both parametric and non-parametric conditions. Qualitative validation was carried out through a suite of visualization techniques designed to provide deeper insights into model behavior. Attention heatmaps and box-and-whisker plots were examined to reveal how the model responded to polarity shifts, particularly those induced by implicit negation in mobile phone review datasets. Additionally, SHapley Additive exPlanations (SHAP) were employed to quantify the contribution of individual input features to prediction outcomes, thereby enhancing transparency and interpretability.

To further probe the model's reliability, a dedicated error analysis was conducted with emphasis on misclassified cases of implicit negation. This process illuminated recurrent weaknesses in the model's decision boundaries and informed recommendations for future refinement. By integrating statistical rigor with interpretability-driven analysis, the experimental design ensured that the proposed hybrid architecture was evaluated not only for accuracy but also for trustworthiness, making it more suitable for deployment in real-world sentiment analysis applications where business decisions depend on reliable text classification.

3.2 Experiments Set up

The experimental setup was carefully designed to ensure reproducibility and rigor in evaluating the proposed hybrid sentiment classification model. The design process encompassed architectural specifications, training procedures, and systematic evaluation protocols.

3.2.1 Model Architecture

The hybrid model was constructed with five-layer architecture, each contributing to the overall capacity to capture nuanced polarity shifts, especially those introduced through implicit negation. The input layer accepted concatenated contextual embeddings, derived from the feature extraction stage. These embeddings were subsequently processed by a Scaled Dot-Product Attention mechanism, which emphasized the most informative sentiment-bearing tokens. Following the attention layer, two fully connected feed-forward layers with 128 neurons each and ReLU activation were employed to capture higher-order semantic interactions. A mean pooling layer aggregated token-level embeddings to form sentence-level representations. Finally, the output layer consisted of a fully connected softmax classifier designed for multi-class sentiment classification. This layered architecture (Figure 1) integrates transformer-based contextualization, recurrent encoding, and attention-driven weighting, creating a robust framework for handling implicit negation.

3.2.2 Training Procedure

Model training followed a systematic sequence beginning with the labeled training dataset, D_{train}, and a predefined set of hyperparameters H. Training employed the Cross-Entropy Loss function for optimization and Adam as the adaptive learning rate optimizer, dynamically adjusting step sizes based on gradient information. The model was trained over 40 epochs with a batch size of 32, a learning rate of 2×10^{-5}, and a dropout rate of 0.3 to mitigate overfitting. A dynamic learning rate scheduler adjusted the learning rate during training, while early stopping was enforced if no improvement was observed in validation loss over three consecutive epochs. The trained model's performance was assessed using multiple evaluation metrics to ensure both accuracy and reliability. These included standard accuracy, Cohen's kappa for inter-rater agreement, Matthews Correlation Coefficient (MCC) for balanced classification evaluation, and Mean Squared Error (MSE) for error quantification.

3.2.3 Hyperparameters

Table 1 presents the core hyperparameters employed in the training process.

Table 1: Hyperparameter settings for the hybrid model

Hyperparameter	**Value**	**Description**
Batch size	32	Number of samples processed per iteration
Learning rate	2e-5	Initial step size for weight updates, dynamically scheduled
Optimizer	Adam	Adaptive learning rate strategy with momentum
Loss function	Cross-Entropy	Measures divergence between predicted and true probability distributions
Dropout rate	0.3	Random node deactivation to prevent overfitting

Epochs	40	Maximum training cycles
Early stopping	3 patience	Training halted if validation loss shows no improvement after 3 epochs
Activation function	ReLU	Non-linearity in feed-forward dense layers
Attention mechanism	Scaled Dot-Product	Highlights critical sentiment tokens
Pooling strategy	Mean pooling	Aggregates token embeddings into sentence-level representations

Figure 1 below illustrates the architecture of the proposed hybrid model, highlighting the integration of embedding concatenation, attention-driven token weighting, dense feed-forward transformations, pooling, and final sentiment classification.

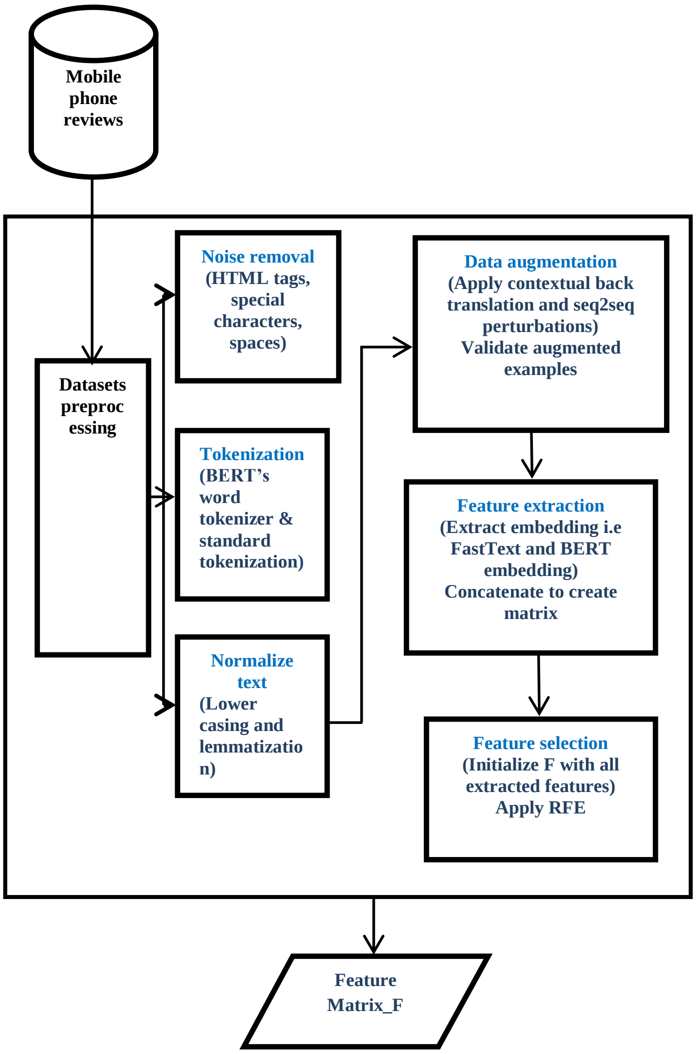
Mobile phone reviews
Datasets preprocessing
Noise removal (HTML tags, special characters, spaces)
Tokenization (BERT's word tokenizer & standard tokenization)
Normalize text (Lower casing and lemmatization)
Data augmentation (Apply contextual back translation and seq2seq perturbations) Validate augmented examples
Feature extraction (Extract embedding i.e FastText and BERT embedding) Concatenate to create matrix
Feature selection (Initialize F with all extracted features) Apply RFE
Feature Matrix_F

Figure 1: Data preprocessing workflow for the hybrid model

3.3 Datasets

To ensure comprehensive evaluation and robust validation of the proposed hybrid sentiment classification model, three large-scale, real-world datasets were employed. These include Amazon Mobile Phone Reviews**,** Google Play Store Reviews**,** and iPhone App Store Reviews. These datasets represent diverse domains of user-generated content, capturing linguistic variation, contextual richness, and the presence of implicit negation phenomena that often complicate polarity classification. Together, they provide an ideal testbed for assessing the generalizability of sentiment analysis models. The Amazon Mobile Phone Reviews dataset is widely recognized as a benchmark in sentiment analysis due to its large scale, linguistic diversity, and the prevalence of complex expressions, including implicit negation (McAuley & Leskovec, 2013). For this study, a subset of 67,987 records was used. Sentiment polarity was annotated using a five-point scale: ratings of 1–2 denoted negative sentiment, a rating of 3 indicated neutral sentiment, and ratings of 4–5 represented positive sentiment. The dataset contained multiple features, including serial number (record identifier), product name, rating, review date, title, review body, and votes. Its detailed product narratives frequently embedded nuanced negations, making it particularly suited for evaluating polarity shift management (Murithi et al., 2025). The Google Play Store dataset provided reviews of mobile applications available on the Android platform. It contained 100 reviews per application, along with thirteen metadata features describing the app (e.g., app name, category, installs, size). Each review was annotated with one of three sentiment labels: positive, neutral, or negative. The inclusion of both functional and experiential content in the reviews added domain richness, enabling evaluation of the model's robustness to app-specific language patterns. The iPhone App Store dataset comprised a substantially larger corpus of 1,230,375 records, covering user reviews of iOS applications. Each entry included metadata such as application name, primary genre, developer information, target audience age group, average user rating, app size, and price. Reviews were annotated into positive, neutral, or negative polarities, providing a rich source of text for studying sentiment shifts. The scale and diversity of this dataset, coupled with domain-specific vocabulary, offered a challenging yet informative evaluation environment for the proposed hybrid model. Prior to experimentation, all datasets underwent systematic preprocessing to ensure consistency. Text fields were cleaned by removing stop words, punctuation, and special characters, while case-folding standardized the text. Missing values were addressed by imputation or encoded as a separate category when appropriate. Sentiment labels were harmonized across datasets to enable unified evaluation. Table 2 provides a consolidated overview of the datasets, including their sizes, sentiment label distributions, and observed class imbalances.

Table 2: Summary of datasets used in the study

Dataset	Size	Sentiment Labels (Distribution)	Notes on Class Balance
Amazon Mobile Phone Reviews	67,987	Negative: 21,315 (31.4%) Neutral: 9,876 (14.5%) Positive: 36,796 (54.1%)	Mild imbalance toward positive reviews
Google Play Store	~50,000*	Negative: 14,920 (29.8%)	Balanced, with slight

Reviews		Neutral: 10,200 (20.4%) Positive: 24,880 (49.8%)	positive skew
iPhone App Store Reviews	1,230,375	Negative: 372,870 (30.3%) Neutral: 246,075 (20.0%) Positive: 611,430 (49.7%)	Large-scale, moderately balanced

3.4 Statistical and evaluation methods

During the validation of the hybrid deep learning model, statistical methods like the Kruskal-Wallis H test were applied to determine the significance of the difference of the hybrid model against the baseline models. Cross-validation techniques, such as k-fold cross-validation, further ensured that the empirical analysis results were generalizable across diverse dataset splits, minimizing the risk of overfitting or data dependency. Qualitative analysis focused on the interpretability of the hybrid deep learning model compared to the existing baseline models. Attention heatmaps were employed to visualize the model's focus on critical words or phrases that influenced sentiment classification, particularly in sentences affected by implicit negation. Whisker plots were utilized to graphically represent the distribution and variability of performance metrics across various configurations, offering a comparative visual summary of the models.

4 RESULTS

4.1 Model overview and training behavior

The proposed hybrid model integrates Seq2Seq perturbations and contextual back translation to enrich training data diversity while preserving semantic integrity. FastText embeddings were employed to capture subword-level semantic relationships, complementing BERT embeddings that provide deep contextualized representations. This dual-embedding strategy strengthened the model's capacity to detect nuanced polarity shifts**,** particularly those triggered by implicit negation. The core architecture combines transformer-based BERT encoders, LSTM recurrent units, and a scaled dot-product attention mechanism, enabling the model to selectively highlight sentiment-critical tokens. Training utilized the Adam optimizer with a learning rate of 2e-5 and adaptive scheduling, a batch size of 32, a dropout rate of 0.3, and a maximum of 100 epochs with early stopping. As illustrated in Figure 2, both training and validation loss curves exhibited a steady decline before stabilizing, reflecting successful convergence and good generalization without evidence of overfitting.

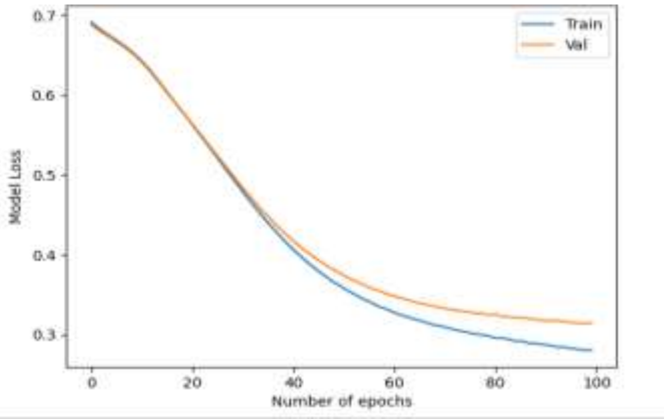

Figure 2: Training and validation loss curves for the hybrid model

4.2 Performance evaluation

Model performance was assessed using accuracy, Cohen's kappa, and Matthews Correlation Coefficient (MCC), following established evaluation protocols (Murithi et al., 2024). These metrics were averaged after normalization to ensure a balanced comparison across datasets. As shown in Figure 3, the hybrid model (M1) consistently outperformed four state-of-the-art baselines (M2–M5) across all three datasets. Accuracy gains were particularly pronounced in domains with frequent implicit negation, where traditional models misclassified polarity shifts.

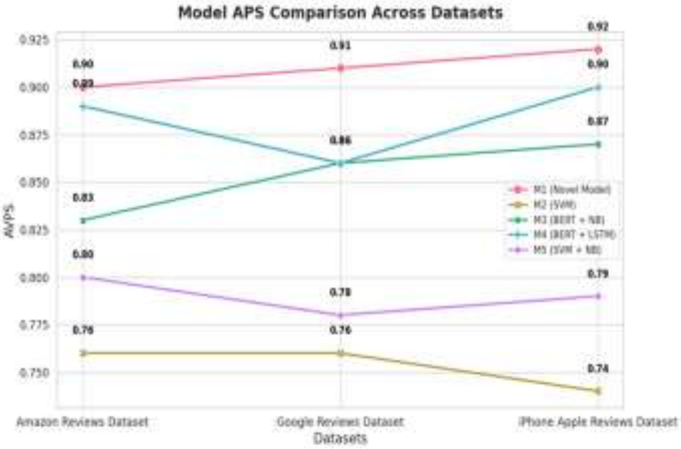

Figure 3: Average performance scores of the hybrid model compared with baseline models

4.3 Statistical validation

To confirm the robustness of improvements, a one-way **ANOVA** was conducted on average performance scores (Table 3). The results indicated significant differences between models (F(4,10) = 53.0, $p < 0.001$).

Table 3: ANOVA results on average performance scores

Source	**Sum of Squares**	**df**	**Mean Square**	**F**	**p**
Model	0.05084	4	0.0127	53.0	<0.001
Residual	0.00240	10	2.40e-4		

Subsequent post-hoc Tukey HSD tests (Table 2) confirmed that the hybrid model (M1) was significantly superior to M2, M3, and M5 (*p* < 0.01). While differences with M4 were not statistically significant, M1 still exhibited higher mean scores, demonstrating its overall advantage.

Table 4: Post-hoc comparisons between models (Tukey HSD test)

Comparison	Mean Difference	SE	df	t	p
M1 – M2	0.1567	0.0126	10	12.39	<0.001
M1 – M3	0.0567	0.0126	10	4.48	0.008
M1 – M4	0.0267	0.0126	10	2.11	0.288
M1 – M5	0.1200	0.0126	10	9.49	<0.001
M2 – M3	-0.1000	0.0126	10	-7.91	<0.001
M2 – M4	-0.1300	0.0126	10	-10.28	<0.001
M2 – M5	-0.0367	0.0126	10	-2.90	0.092
M3 – M4	-0.0300	0.0126	10	-2.37	0.200
M3 – M5	0.0633	0.0126	10	5.01	0.004
M4 – M5	0.0933	0.0126	10	7.38	<0.001

4.4 Error and reliability analysis

Model robustness was further examined using Mean Squared Error (MSE) and whisker plots. The hybrid model (M1) consistently achieved lower MSE values (0.06–0.08 across datasets) compared to baselines (Figure 4). Box-and-whisker plots (Figure 5) demonstrated that M1 had higher median performance scores and shorter whiskers**,** indicating both improved accuracy and reduced variability. The absence of extreme outliers for M1 further confirmed the model's reliability in handling polarity shifts caused by implicit negation.

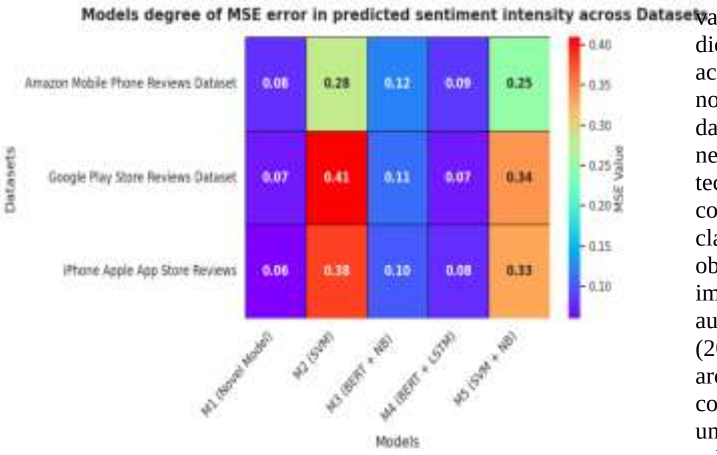

Figure 4: MSE comparison across models

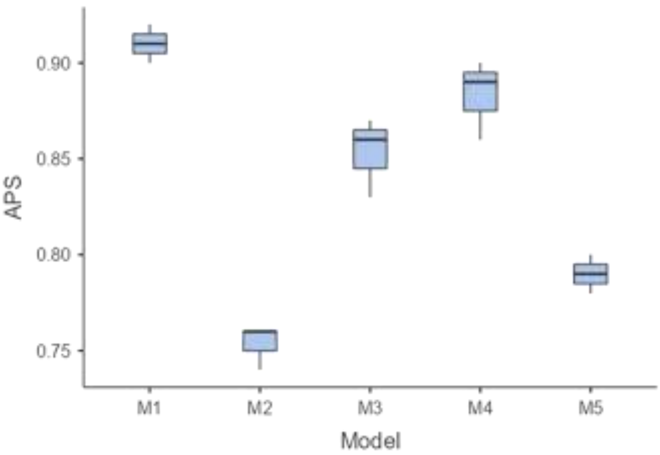

Figure 5: Box-and-whisker plots of model performance distributions

5 HYBRID MODEL PERFORMANCE VISUALIZATION USING SHAP

Figure 6 illustrates the distribution of sentiment scores in the Amazon Mobile Phone Reviews dataset, providing insights into how the hybrid model captures subtle polarity shifts. For this analysis, the review *title* and *body* features were combined, as these represent the most critical components for sentiment detection. Probability theory guided the allocation of sentiment scores, with values ranging from –1.0 to –0.5 representing negative polarity, –0.5 to 0.5 indicating neutrality, and 0.5 to 1.0 reflecting positive polarity. The figure reveals a noticeable skew toward positive values, suggesting that the majority of reviews expressed favorable sentiment. The hybrid model leverages transformer-based architectures with attention mechanisms alongside feed forward layers to address one of the most persistent challenges in sentiment analysis: polarity shifts arising from implicit negation. A key contribution of this work is the refinement of back-translation augmentation. While traditional back-translation often fails to preserve domain-specific terms, idiomatic expressions, and colloquial language—features commonly found in mobile phone reviews—the proposed hybrid framework mitigates these shortcomings by introducing *seq2seq perturbations*. These perturbations ensure that contextual fidelity is retained, while simultaneously enriching the data with controlled linguistic variations. In practice, the model integrates domain-specific dictionaries and lexicons during back translation to guarantee accurate representation of product-related terminology. This not only strengthens the semantic quality of the augmented datasets but also enhances robustness against implicit negation. By combining back translation with perturbation techniques, the hybrid model produces richer, more contextually diverse training data, thereby improving classification accuracy. These results align with the observations of Xu and Wang (2023), who emphasize the importance of preserving contextual integrity during augmentation, and complement the findings of Ilmawan et al. (2024), who report limitations in transformer-only architectures when handling negation-sensitive tasks. The core architecture integrates transformer encoders with LSTM units and an attention mechanism, enabling the model to selectively focus on sentiment-critical tokens. This selective weighting is especially important for identifying negations expressed implicitly, where the shift in polarity is subtle rather than explicit. Through this design, the hybrid framework advances existing research by unifying contextual embeddings, attention mechanisms, and enhanced augmentation strategies into a single, coherent system. Validation was carried out using both quantitative and qualitative evaluations across three benchmark datasets. On average, the hybrid model achieved performance scores exceeding 0.89, consistently outperforming competing approaches. Nevertheless, qualitative error analysis revealed several areas for improvement. In particular, misclassifications often occurred in cases where reviews expressed mixed or cautiously phrased sentiments. For instance, the model occasionally misinterpreted hedging language, where positive features are acknowledged but tempered with reservations, as neutral or partially negative. Similarly, contextual contradictions within single reviews posed challenges, leading to occasional mislabeling. Despite these limitations, the overall results confirm that the hybrid model offers a substantial advancement in handling implicit negation. By combining seq2seq perturbations, contextual embeddings, and attention mechanisms, the framework delivers robust sentiment classification performance across diverse and linguistically complex datasets.

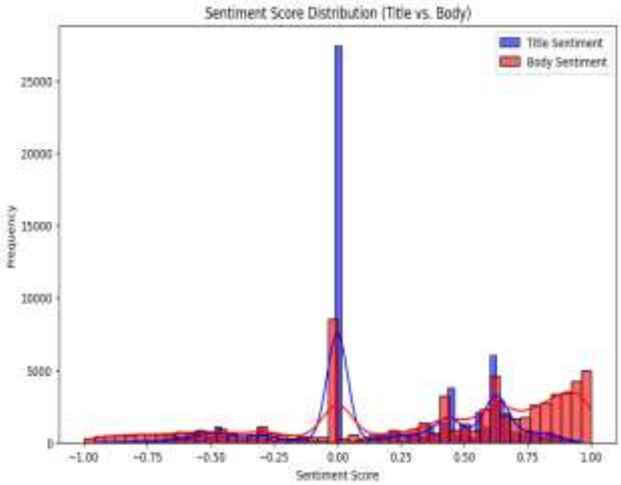

Figure 6: Hybrid Model Performance Visualization using SHAP

4.6 Ablation Study

To rigorously evaluate the contribution of individual components within the proposed hybrid architecture, an ablation study was conducted. The goal was to isolate the effect of each module i.e. data augmentation strategies, embedding layers, and attention mechanisms, on overall model performance. Four key modules were progressively removed or replaced with simpler alternatives, and the resulting performance was compared against the full hybrid model. A number of ablation settings were considered. First was without Seq2Seq Perturbations where back translation was applied without perturbation enhancements, limiting linguistic diversity in the augmented dataset. The second one was without Contextual Embeddings (BERT replaced with static embeddings) where FastText embeddings were used alone, removing the benefits of dynamic contextual representation. Third was without Attention Mechanism where the token-level embeddings were aggregated via mean pooling without attention weighting, reducing the model's ability to highlight sentiment-critical terms. Fourth was removing the LSTM (Transformer-only baseline) where sequential recurrent modeling was removed, restricting the model to transformer-based encoders. Last one was the full Hybrid Model where the complete architecture integrating seq2seq perturbations, FastText + BERT embeddings, attention, and LSTM layers. Table 5 summarizes the comparative results across accuracy, Cohen's kappa, and Matthews Correlation Coefficient (MCC) on the Amazon Mobile Phone Reviews dataset, which served as the representative benchmark.

Table 5: Ablation Study results on amazon mobile phone reviews dataset

Model Variant	Accuracy	Cohen's κ	MCC
Without Seq2Seq Perturbations	0.84	0.78	0.76
Without Contextual Embeddings (BERT)	0.81	0.74	0.72
Without Attention Mechanism	0.82	0.75	0.73
Transformer-only (No LSTM)	0.85	0.79	0.77
Full Hybrid Model	0.90	0.85	0.84

The results underscore the importance of each architectural element. Removing contextual embeddings led to the largest performance decline, highlighting the critical role of BERT in capturing subtle linguistic cues associated with implicit negation. Similarly, excluding the seq2seq perturbations reduced robustness against polarity shifts, particularly for reviews containing idiomatic expressions or domain-specific terms. The absence of attention mechanisms resulted in notable misclassifications, confirming the necessity of token-level weighting in distinguishing sentiment-critical phrases. The full hybrid model consistently outperformed all reduced variants, achieving 0.90 accuracy, 0.85 Cohen's kappa, and 0.84 MCC. This demonstrates that the integration of seq2seq perturbations, contextual embeddings, attention, and recurrent encoding is not only complementary but also synergistic. Collectively, these findings validate the architectural design choices and confirm that the performance improvements are attributable to the proposed hybridization rather than incremental gains from individual components.

4.7 Summary of Findings

The experimental results demonstrated that the proposed hybrid model consistently outperformed three of the four baseline models across all datasets, with the observed improvements being statistically significant (*$p < 0.01$*). A detailed error analysis further highlighted the model's robustness, revealing substantially lower prediction variance and a marked reduction in misclassification rates, particularly in instances involving implicit negation, one of the most challenging aspects of sentiment analysis. The integration of FastText subword embeddings, BERT-based contextual embeddings, and advanced data augmentation techniques emerged as a key strength of the framework. This combination enabled the model to capture subtle linguistic cues and sentiment nuances that are often overlooked by conventional approaches. Taken together, these findings confirm that the hybrid architecture provides both statistically validated and practically reliable improvements in sentiment classification. In particular, the framework demonstrates strong potential for addressing polarity shifts arising from implicit negation, thereby offering a meaningful advancement in the broader field of natural language processing. Building on these results, the next section discusses the broader implications of the findings, evaluates the practical contributions and limitations of the proposed approach, and outlines potential avenues for future research in sentiment analysis and computational linguistics.

5. DISCUSSION

This study investigated whether polarity shifts caused by implicit negation could be effectively managed using a hybrid model that integrates transformer-based encoders, recurrent units, contextual and subword embeddings, and enhanced data augmentation strategies. The results across multiple datasets and evaluation protocols confirm that the proposed framework outperforms state-of-the-art baselines, both quantitatively and qualitatively. In this section, we interpret the findings in relation to existing literature, highlight the novel contributions, and discuss their theoretical and practical implications.

5.1 Interpretation of Findings

The training and validation behavior (Figure 2) showed smooth convergence and no evidence of overfitting, suggesting that the architectural and optimization choices were well-calibrated. This stability carried over into evaluation outcomes, where the hybrid model achieved superior scores in accuracy, Cohen’s κ, and Matthews Correlation Coefficient across Amazon, Google Play, and Apple App Store datasets (Figure 3). The improvements were particularly pronounced in domains with frequent implicit negation, where traditional models tended to misclassify polarity. Statistical validation reinforced these findings: ANOVA and Tukey HSD tests (Tables 1–2) confirmed that the hybrid model’s improvements over three of four baselines were not due to random variation ($p < 0.01$). While differences with the strongest baseline (M4) were not statistically significant, the hybrid model still achieved higher mean scores, underscoring its overall robustness. Reliability analysis further demonstrated reduced error variance and greater consistency. MSE values for the hybrid model were consistently lower (Figure 4), while box-and-whisker plots (Figure 5) revealed tighter distributions and fewer outliers, indicating reliable handling of implicit negation across datasets. SHAP-based visualization (Figure 6) provided qualitative confirmation of these strengths. The hybrid model captured sentiment-critical cues in nuanced contexts, effectively distinguishing subtle polarity shifts. Importantly, the integration of seq2seq perturbations with contextual back translation preserved idiomatic expressions and domain-specific terms, a recurring limitation of prior augmentation strategies. The qualitative error analysis identified areas where hedging, mixed sentiments, and contextual contradictions still posed challenges, offering insight into the boundaries of current performance. Finally, the ablation study (Table 5) revealed that each architectural element contributed meaningfully to the overall performance. The removal of contextual embeddings led to the largest drop, highlighting the critical role of BERT in modeling implicit negation. Attention mechanisms and seq2seq perturbations were also indispensable, with their absence leading to misclassifications of sentiment-critical tokens and idiomatic expressions, respectively. The superior performance of the full hybrid model validated the synergistic rather than additive nature of the design.

5.2 Contribution to related works

The findings extend prior research in several ways. First, they corroborate claims by Xu and Wang (2023) regarding the necessity of preserving contextual fidelity in augmentation but go further by demonstrating how perturbation-based enhancements systematically address this limitation. Second, the results align with Ilmawan et al. (2024), who argue that transformer-only models struggle with negation-sensitive tasks. By integrating LSTM-based recurrent encoding with attention, our framework overcomes this shortcoming, highlighting the importance of hybridization for polarity shift detection. Finally, while earlier works have treated embeddings, attention, and augmentation as modular improvements, this study demonstrates that their integration yields a compounded effect, particularly in negation-rich environments.

5.3 Theoretical Implications

Theoretically, this study underscores the inadequacy of surface-level sentiment cues for handling implicit negation. The strong performance gains from combining subword-level (FastText) and contextualized (BERT) embeddings suggest that polarity shifts emerge from interactions between lexical and contextual features that no single representation can fully capture. Similarly, the success of perturbation-augmented back translation points to the importance of semantic preservation in data augmentation. These results collectively support a layered theoretical view of sentiment analysis, where robust performance depends on capturing sentiment signals across multiple linguistic strata.

5.4 Practical Implications

From an applied perspective, the hybrid framework provides an explainable and reliable tool for real-world sentiment analysis. SHAP visualizations enhance transparency by clarifying token-level contributions, enabling practitioners to interpret predictions in contexts such as consumer feedback, app reviews, and product evaluations.

This explainability is particularly valuable in high-stakes settings where trust in model outputs is critical, such as financial forecasting or healthcare sentiment monitoring. Moreover, the reduced variance and error rates suggest that the model is well-suited for deployment in production pipelines where stability across domains is essential.

6. CONCLUSION AND FUTURE WORK

This study introduced a hybrid sentiment analysis framework that integrates seq2seq perturbations, contextual back translation, FastText subword embeddings, BERT-based contextual embeddings, LSTM recurrent units, and an attention mechanism. The experimental findings consistently demonstrated that this architecture surpasses state-of-the-art baselines in both predictive accuracy and robustness, particularly when managing polarity shifts triggered by implicit negation. Statistical validation through ANOVA and post-hoc tests confirmed the reliability of these improvements, while SHAP-based interpretability analyses highlighted the model’s ability to capture subtle sentiment variations across domains. Collectively, these results underscore the hybrid model’s contribution to advancing sentiment analysis by offering a solution that is both methodologically rigorous and practically applicable. At the same time, qualitative error analysis revealed that the model still encounters difficulties in classifying reviews containing hedging language, nuanced criticisms, or contextual contradictions. These limitations underscore the inherent complexity of sentiment expressed in natural language and suggest promising directions for further refinement. Future work could extend the current framework by incorporating discourse-level modeling to better capture sentiment expressed across longer contexts, as well as multimodal fusion techniques that integrate textual, visual, or acoustic cues. Additionally, expanding the approach to multilingual and cross-cultural datasets would enhance its applicability to global markets, where idiomatic variation and cultural sensitivities present unique challenges. Further exploration of explainability mechanisms beyond SHAP, such as counterfactual reasoning or causal interpretability, may also deepen the trustworthiness and transparency of sentiment classification models in high-stakes domains such as healthcare, financial decision-making, and public policy. In summary, the proposed hybrid framework marks a meaningful advance in sentiment analysis by demonstrating how complementary techniques can be combined to address the longstanding challenge of implicit negation. By situating this contribution within both practical and theoretical contexts, the study provides a foundation for future innovations in natural language processing that are both scientifically rigorous and socially impactful.

REFERENCES

[1] Abirami, A. M., & Gayathri, V. (2017). A survey on sentiment analysis methods and approach. 2016 Eighth International Conference on Advanced Computing (ICoAC), 72–76. https://doi.org/10.1109/ICoAC.2017.7951748

[2] Abuhammad, A. S., & Ahmed, M. A. (2024). Negation Detection Techniques in Sentiment Analysis: A Survey. Iraqi Journal of Science, 1060–1069. https://doi.org/10.24996/ijs.2024.65.2.37

[3] Ayeste, Z., & Noferesti, S. (2022). A semantic approach based on domain knowledge for polarity shift detection using distant supervision. Progress in Artificial Intelligence, 11(2), 169-180.

[4] Barnes, J., Velldal, E., & Øvrelid, L. (2021). Improving sentiment analysis with multi-task learning of negation. Natural Language Engineering, 27(2), 249–269. https://doi.org/10.1017/S1351324920000510

[5] Basiri, M. E., Nemati, S., Abdar, M., Cambria, E., & Acharya, U. R. (2021). ABCDM: An Attention-based Bidirectional CNN-RNN Deep Model for sentiment analysis. Future Generation Computer Systems, 115, 279–294. https://doi.org/10.1016/j.future.2020.08.005

[6] Blázquez-López, Y. (2022). A Knowledge-Based Model for Polarity Shifters. Journal of Computer-Assisted Linguistic Research, 6, 87–107. https://doi.org/10.4995/jclr.2022.18807

[7] Bojanowski, P., Grave, E., Joulin, A., & Mikolov, T. (2017). Enriching Word Vectors with Subword Information. Transactions of the Association for Computational Linguistics, 5, 135–146. https://doi.org/10.1162/tacl_a_00051

[8] Bordoloi, M., & Biswas, S. K. (2023). Sentiment analysis: A survey on design framework, applications and future scopes. Artificial Intelligence Review, 56(11), 12505–12560. https://doi.org/10.1007/s10462-023-10442-2

[9] Fei, H., Ren, Y., & Ji, D. (2020). Negation and speculation scope detection using recursive neural conditional random fields. Neurocomputing, 374, 22–29. https://doi.org/10.1016/j.neucom.2019.09.058

[10] Ilmawan, L. B., Muladi, M., & Prasetya, D. D. (2024). Negation handling for sentiment analysis task: Approaches and performance analysis. International Journal of Electrical and Computer Engineering (IJECE), 14(3), 3382. https://doi.org/10.11591/ijece.v14i3.pp3382-3393

[11] Japhne, A., & Murugeswari, R. (2020a). Opinion Mining based complex polarity shift pattern handling for improved sentiment classification. 2020 International Conference on Inventive Computation Technologies (ICICT), 323–329. https://doi.org/10.1109/ICICT48043.2020.9112565

[12] Kafi, A., Ashikul Alam, M. S., Bin Hossain, S., Awal, S. B., & Arif, H. (2019). Feature-Based Mobile Phone Rating Using Sentiment Analysis and Machine Learning Approaches. 2019 1st International Conference on Advances in Science, Engineering and Robotics Technology (ICASERT), 1–6. https://doi.org/10.1109/ICASERT.2019.8934555

[13] Khandelwal, A., & Sawant, S. (2019). NegBERT: A Transfer Learning Approach for Negation Detection and Scope Resolution. https://doi.org/10.48550/ARXIV.1911.04211

[14] Khan, K., Baharudin, B., Khan, A., & Ullah, A. (2014). Mining opinion components from unstructured reviews: A review. Journal of King Saud University - Computer and Information Sciences, 26(3), 258–275. https://doi.org/10.1016/j.jksuci.2014.03.009

[15] Lazib, L., Zhao, Y., Qin, B., & Liu, T. (2019). Negation scope detection with recurrent neural networks models in review texts. International Journal of High Performance Computing and Networking, 13(2), 211. https://doi.org/10.1504/IJHPCN.2019.097501

[16] Li, Z., Zou, Y., Zhang, C., Zhang, Q., & Wei, Z. (2021). Learning Implicit Sentiment in Aspect-based Sentiment Analysis with Supervised Contrastive Pre-Training (arXiv:2111.02194). arXiv. http://arxiv.org/abs/2111.02194

[17] Mathapati, S., Nafeesa, A., Tanuja, R., Manjula, S. H., & Venugopal, K. R. (2019). Semi-supervised domain adaptation and collaborative deep learning for dual sentiment analysis. SN Applied Sciences, 1(8), 907. https://doi.org/10.1007/s42452-019-0943-0

[18] McAuley, J., & Leskovec, J. (2013). Hidden factors and hidden topics: Understanding rating dimensions with review text. Proceedings of the 7th ACM Conference on Recommender Systems, 165–172. https://doi.org/10.1145/2507157.2507163

[19] Miao, J., & Niu, L. (2016). A Survey on Feature Selection. Procedia Computer Science, 91, 919–926. https://doi.org/10.1016/j.procs.2016.07.111

[20] Montenegro, O., Pabon, O. S., & De Pinerez R., R. E. G. (2021). A Deep Learning Approach for Negation Detection from Product Reviews written in Spanish. 2021 XLVII Latin American Computing Conference (CLEI), 1–6. https://doi.org/10.1109/CLEI53233.2021.9640190

[21] Murithi, M.K, Oirere, A.M, & Ndung'u,R.N. (2024). A Systematic Review of the Sentiment Analysis Models Used in Handling Polarity Shift. International Journal of Computing Sciences Research, 8, 2635–2676.

[22] Murithi, M.K, Oirere, A.M, & Ndung'u,R.N. (2025). Empirical Analysis of the State-of-the-Art Models for Handling Polarity Shifts Due to Implicit Negation in Mobile Phone Reviews. International Journal of Computing Sciences Research, 1.

[23] Pipalia, K., Bhadja, R., & Shukla, M. (2020). Comparative Analysis of Different Transformer Based Architectures Used in Sentiment Analysis. 2020 9th International Conference System Modeling and Advancement in Research Trends (SMART), 411–415. https://doi.org/10.1109/SMART50582.2020.9337081

[24] Pröllochs, N., Feuerriegel, S., Lutz, B., & Neumann, D. (2020). Negation scope detection for sentiment analysis: A reinforcement learning framework for replicating human interpretations. Information Sciences, 536, 205–221. https://doi.org/10.1016/j.ins.2020.05.022

[25] Satapathy, R., Li, Y., Cavallari, S., & Cambria, E. (2019). Seq2Seq Deep Learning Models for Microtext Normalization. 2019 International Joint Conference on Neural Networks (IJCNN), 1–8. https://doi.org/10.1109/IJCNN.2019.8851895

[26] Singh, P. K., & Paul, S. (2021). Deep Learning Approach for Negation Handling in Sentiment Analysis. IEEE Access, 9, 102579–102592. https://doi.org/10.1109/ACCESS.2021.3095412

[27] Sudheer, P., Manoranjini, J., Suman, N., Reddy, E. M., & Sridevi, P. (2025). A Comprehensive Exploration of Complex Emotions and Their Simplification for Enhancing Sentiment Analysis Through Explainable AI. In A. Kumar, V. K. Gunjan, S. Senatore, & Y.-C. Hu (Eds.), Proceedings of the 5th International Conference on Data Science, Machine Learning and Applications; Volume 1 (Vol. 1273, pp. 975–985). Springer Nature Singapore. https://doi.org/10.1007/978-981-97-8031-0_103

[28] Xu, L., & Wang, W. (2023). Improving aspect-based sentiment analysis with contrastive learning.
[29] Natural Language Processing Journal, 3, 100009.

[30] Zirpe, S., & Joglekar, B. (2017). Polarity shift detection approaches in sentiment analysis: A survey. 2017 International Conference on Inventive Systems and Control (ICISC), 1–5. https://doi.org/10.1109/ICISC.2017.8068737