Improving Sentiment Analysis of Mobile App Reviews through Emoji and Slang Normalization

Tinne Dey
Data Science Research Centre
Tampere University
Tampere, Finland
tinnedey09@gmail.com

Zheying Zhang
Software Engineering Research Center
Tampere University
Tampere, Finland
zheying.zhang@tuni.fi

Kostas Stefanidis
Data Science Research Centre
Tampere University
Tampere, Finland
konstantinos.stefanidis@tuni.fi

***Abstract*—Mobile app reviews provide valuable feedback for developers and policymakers, but their informal and noisy nature poses significant challenges for sentiment analysis systems. In particular, the frequent use of emojis and slang expressions can distort the semantic interpretation of user opinions and negatively impact classification performance. This paper investigates the effect of emoji and slang normalization on BERT-based sentiment analysis of mobile app reviews. Using a dataset of Instagram and WhatsApp reviews labeled through star ratings, we compare a baseline BERT model with enhanced variants incorporating emoji normalization, slang normalization, and their combination. Experimental results show that normalization improves overall classification performance, achieving an increase of approximately 3–4 percentage points in macro F1-score compared to the baseline. The findings suggest that handling informal digital expressions reduces noise and improves model robustness when analyzing user-generated content. This study highlights the importance of preprocessing strategies for more reliable sentiment analysis in real-world app review environments and provides insights for developing more inclusive and adaptable text analytics systems.**

***Index Terms*—Sentiment analysis, BERT, emoji normalization, slang normalization, linguistic bias, inclusion**

I. Introduction

Mobile applications have become an essential part of everyday digital life, supporting communication, entertainment, commerce, and social interaction. User-generated reviews on app stores provide a rich source of feedback for developers, organizations, and policymakers to understand user satisfaction, detect emerging issues, and improve services. Sentiment analysis techniques are widely used to automatically classify such reviews into positive, neutral, or negative categories, enabling large-scale opinion mining [1]. However, the informal, noisy nature of app reviews poses significant challenges for natural language processing (NLP) systems. Unlike formal text, mobile app reviews often include emojis, abbreviations, creative spellings, and slang. These elements carry strong emotional and contextual meaning but are often not adequately handled by standard NLP pipelines. Transformer-based models such as BERT have demonstrated strong performance in sentiment classification tasks [1]; nevertheless, their effectiveness can be limited when confronted with non-standard and informal digital language. Emojis may represent complex affective states [2], [3], and slang terms may encode community-specific meanings that are not explicitly captured by pretrained vocabularies [4]. As a result, models may misinterpret or ignore important sentiment cues.

Beyond performance degradation, overlooking informal digital expressions may introduce a form of linguistic bias. Different demographic and cultural groups vary in their use of emojis and slang. Younger users, multilingual communities, and marginalized groups often rely more heavily on informal or non-standard expressions to convey tone and emotion. If sentiment analysis systems systematically misclassify such expressions, certain user voices may be underrepresented or inaccurately interpreted. In applications where review analytics informs decision-making, this can indirectly affect perceptions of user well-being and satisfaction. Therefore, improving the robustness of sentiment analysis models to informal language is not only a technical challenge, but also a step toward more inclusive and equitable data-driven systems.

In this paper, we investigate the impact of emoji and slang normalization on BERT-based sentiment analysis of mobile app reviews. We construct a dataset of English-language reviews from Instagram and WhatsApp collected from the Google Play Store and automatically label them using star ratings [5]. We compare a baseline BERT model trained on raw text with three enhanced variants: emoji-only normalization, slang-only normalization, and a combined emoji-and-slang normalization approach. The proposed normalization strategies transform informal expressions into standardized textual representations before tokenization, enabling the model to better capture their semantic contribution.

Our experimental results demonstrate that incorporating normalization strategies improves overall classification performance, with notable gains in macro F1-score compared to the baseline. The findings indicate that handling informal digital expressions reduces noise and enhances the robustness of transformer-based sentiment classifiers in real-world app review environments.

The main contributions of this work are as follows:

- We systematically evaluate how emoji and slang usage affects BERT-based sentiment analysis performance on mobile app reviews.
- We design a modular preprocessing pipeline that converts emojis and slang into standardized textual representations

prior to tokenization, enabling controlled experimentation.
- We compare four experimental variants under identical settings and analyze their impact on balanced performance, highlighting how normalization reduces performance gaps associated with informal language use.
- We discuss how improved handling of informal expressions contributes to reducing linguistic bias and strengthening inclusion in data-driven review analytics.

Rather than proposing a new model architecture, this work provides a controlled and reproducible empirical study of how explicit preprocessing of informal language affects transformer-based sentiment analysis in mobile app reviews.

The remainder of this paper is organized as follows. Section II reviews related work on sentiment analysis, informal language processing, and linguistic bias in NLP systems. Section III describes the dataset, preprocessing pipeline, and experimental design. Section IV presents the empirical results and comparative analysis of normalization strategies. Section V discusses the implications of our findings for robustness and inclusion in sentiment analysis. Finally, Section VI concludes the paper and outlines directions for future research.

II. Related Work

Sentiment analysis has been extensively studied in the context of social media and user-generated content, where short and informal texts pose unique challenges. Early approaches relied on lexicon-based and rule-based methods, such as AFINN and LIWC [6], [7], which assign sentiment scores to predefined word lists. While these approaches are interpretable and simple to implement, they struggle to capture contextual meaning, negation, and domain-specific expressions. As machine learning techniques evolved, supervised models using handcrafted features became dominant, followed by deep learning architectures capable of learning contextual representations directly from data.

With the introduction of transformer-based models, particularly BERT (Bidirectional Encoder Representations from Transformers), sentiment classification achieved significant improvements [1], [8]–[10]. BERT leverages bidirectional self-attention mechanisms to model contextual dependencies between words, allowing it to outperform traditional recurrent and convolutional neural networks in many NLP tasks. Several studies have applied BERT and its variants to sentiment analysis in app reviews and social media platforms, demonstrating strong baseline performance. However, pretrained models such as BERT are typically trained on formal corpora (e.g., Wikipedia and BooksCorpus), which may limit their ability to fully capture the semantics of informal digital language. Recent work has also explored sentiment analysis in mobile app maintenance and broader customer feedback analytics. Li et al. propose a sentiment-aware analysis method to elicit user opinions regarding particular app updates by detecting similar topics before and after an update, supporting update-centric review analytics [11]. Puustinen et al. introduce STracker, a framework for identifying sentiment changes in customer feedback by tracking discovered topics and their sentiments over time [12]. Katsarou et al. propose MUTUAL, a multi-domain sentiment classification approach that incorporates uncertainty sampling to improve learning under limited labeled data and domain variation [13]. The role of emojis in sentiment analysis has received increasing attention. Kralj Novak et al. [2] developed an emoji sentiment lexicon that quantifies the polarity of emojis based on large-scale Twitter data, highlighting their significant emotional impact. Felbo et al. [3] introduced DeepMoji, a model pretrained on emoji prediction, showing that emojis can serve as powerful supervision signals for capturing emotional content. These studies demonstrate that emojis are not merely decorative symbols but meaningful carriers of affective information. Nevertheless, many sentiment analysis pipelines either remove emojis during preprocessing or treat them as unknown tokens, potentially losing valuable sentiment cues.

Similarly, slang and informal language present challenges for NLP systems. Social media texts frequently contain abbreviations, creative spellings, and colloquial expressions that may not appear in standard vocabularies. Prior research has explored normalization techniques, including dictionary-based mappings, statistical normalization, and neural approaches for transforming informal text into standardized forms [4]. Urban Dictionary and other crowd-sourced resources have often been used to build slang lexicons. However, the integration of systematic slang normalization within transformer-based sentiment classification pipelines remains relatively underexplored, particularly in the context of mobile app reviews.

Beyond performance considerations, recent work in responsible and inclusive AI highlights the risk of linguistic bias in NLP systems. Models trained on formal or majority-language data may underperform on texts produced by specific demographic or cultural groups that use distinctive linguistic patterns. Informal expressions, dialects, and community-specific slang may be misinterpreted or penalized by automated systems, leading to uneven performance across user groups. Improving robustness to informal language can therefore contribute not only to higher accuracy but also to more equitable treatment of diverse user voices in digital analytics.

In contrast to prior work that examines emojis or slang in isolation, this study provides a systematic empirical comparison of emoji normalization, slang normalization, and their combination within a BERT-based sentiment analysis framework for mobile app reviews. By evaluating multiple preprocessing strategies under identical experimental conditions, we aim to clarify their impact on model performance and robustness in real-world user-generated text.

III. Methodology

This section describes the dataset construction, preprocessing strategies, model architecture, and experimental setup used to evaluate the impact of emoji and slang normalization on sentiment classification performance.

A. Dataset Collection and Statistics

User reviews were collected from the Google Play Store for two widely used mobile applications: Instagram and WhatsApp. The data collection process was conducted using the `google-play-scraper` Python library, which programmatically retrieves publicly available review content through the Google Play interface. The tool extracts review text, star ratings, timestamps, and associated metadata. Only publicly accessible reviews were collected, and no private or restricted information was accessed.

The constructed dataset, referred to as the **IGWA dataset** (Instagram and WhatsApp Reviews), contains a total of 54,544 English-language reviews after filtering and preprocessing. Reviews were labeled into three sentiment classes based on star ratings: ratings of 1–2 were mapped to *Negative*, rating 3 to *Neutral*, and ratings of 4–5 to *Positive*. The dataset was randomly split into training, validation, and test sets using an 80/10/10 ratio to ensure consistent evaluation across all model variants. Table I summarizes the dataset statistics. Sentiment

TABLE I
SUMMARY STATISTICS OF THE IGWA-ONLY DATASET.

Total Reviews	54,544
Train / Val / Test	43,634 / 5,455 / 5,455
Negative	20,767 (38.07%)
Neutral	3,646 (6.68%)
Positive	30,131 (55.24%)
Reviews with Emojis	11.25%
Reviews with Slang	0.29%
Average Length (tokens)	15.44

labels were derived from star ratings and therefore should be interpreted as weak supervision rather than manual semantic annotation. While this labeling strategy is common in large-scale app-review research, rating-text mismatches may occur, especially for neutral, mixed, or context-dependent reviews. We retain this setup because it enables consistent large-scale comparison across model variants, but we acknowledge that label noise may affect the reported results. The relatively small proportion of Neutral samples (6.68%) introduces class imbalance, which contributed to lower class-wise performance for the Neutral category, as discussed in Section IV. This imbalance was not artificially adjusted in order to preserve the natural distribution of user opinions in real-world app reviews. **Dataset Availability:** The anonymized IGWA dataset and preprocessing scripts are publicly available at: https://github.com/tinne2/sentiment-analysis. All personally identifiable information has been removed prior to public release.

B. Preprocessing Pipeline and Normalization Strategy

Mobile app reviews frequently contain informal expressions, emojis, abbreviations, and creative spellings. To systematically evaluate the impact of normalization, a modular preprocessing pipeline was designed. The pipeline consists of the following steps:

1) Basic text cleaning (removal of URLs, excessive whitespace, and non-text artifacts).
2) Optional emoji normalization.
3) Optional slang normalization.
4) Tokenization using the BERT tokenizer.

This modular structure allows controlled experimentation by enabling or disabling specific normalization components. Emojis serve as strong carriers of emotional meaning in digital communication. However, when left unprocessed, they may be treated as unknown or poorly represented tokens by pretrained language models. In the proposed approach, emojis were converted into standardized textual descriptions before tokenization. Each emoji was mapped to a short semantic label (e.g., “smiling_face”, “sad_face”, “angry_face”) based on commonly accepted emoji descriptions. This transformation preserves the emotional signal encoded by emojis while converting them into tokenizable textual units that can be better integrated into the BERT embedding space. Slang expressions and abbreviations are prevalent in mobile app reviews (e.g., “lol”, “omg”, “btw”, “idk”). Such terms may not always be explicitly represented in pretrained vocabularies or may be interpreted inconsistently. A custom slang dictionary was constructed by compiling frequent slang terms observed in the dataset and mapping them to their standardized equivalents (e.g., “lol” $\rightarrow$ “laugh out loud”, “idk” $\rightarrow$ “I do not know”). The dictionary was manually curated and applied as a rule-based replacement step before tokenization. This normalization reduces lexical variability and aligns informal expressions with their standard semantic counterparts.

C. Model Architecture and Training Configuration

We selected BERT-base-uncased as a strong, widely adopted transformer baseline that allows preprocessing effects to be isolated under stable modeling conditions. The purpose of the study is to conduct a controlled comparison of normalization strategies rather than benchmarking a broad range of architectures. For sentiment classification, we employed the BERT-base-uncased model, which consists of 12 transformer encoder layers and approximately 110 million parameters [1]. The model produces contextualized token representations using bidirectional self-attention mechanisms. Each input sequence was truncated or padded to a maximum length of 96 tokens. The final hidden representation corresponding to the special [CLS] token was extracted and passed to a fully connected dense layer with softmax activation to predict one of three sentiment classes: negative, neutral, or positive. To reduce computational complexity and mitigate overfitting, the lower eight transformer layers were frozen during fine-tuning, while the top four layers and the classification head were updated during training. Freezing the lower eight transformer layers serves multiple purposes. First, it reduces computational cost during fine-tuning by limiting the number of trainable parameters. Second, it mitigates overfitting, particularly given the moderate size of the dataset. Lower layers in BERT typically capture general linguistic features learned during pretraining, while higher layers encode more task-specific representations. By fine-tuning only the top four layers and the classification head, the model adapts to the sentiment

classification task while preserving robust pretrained linguistic knowledge. Additionally, freezing layers ensures consistent experimental conditions across all normalization variants, allowing fair comparison of preprocessing strategies without introducing variability due to full-network fine-tuning. We do not claim that freezing eight layers is globally optimal; rather, it provides a consistent and computationally efficient fine-tuning setting for comparing preprocessing variants fairly. As shown in Fig. 1, The input text undergoes emoji and slang normalization before tokenization. The tokenized sequence is then passed through the BERT encoder. A sentiment classifier processes the [CLS] token's representation to predict one of three output classes: Positive, Neutral, or Negative. The model was trained using the Adam optimizer with a learning rate of 2e-5. Training was conducted for up to three epochs, with early stopping based on validation macro F1-score to prevent overfitting. The macro F1-score was selected as the primary evaluation metric due to its balanced treatment of class-wise performance, particularly important in the presence of class imbalance.

D. Experimental Variants

To isolate the effect of normalization strategies, four experimental variants were evaluated under identical hyperparameter settings:

1) Baseline Model – Raw text without emoji or slang normalization.
2) Emoji-Only Model – Emoji normalization applied, no slang normalization.
3) Slang-Only Model – Slang normalization applied, no emoji normalization.
4) Enhanced Model – Both emoji and slang normalization applied.

By comparing these variants, we aim to quantify the individual and combined contributions of emoji and slang normalization to sentiment classification performance.

IV. Experimental Results and Analysis

A. Evaluation Metrics and Overall Performance

To evaluate the performance of the proposed models, we report Accuracy, Macro Precision, Macro Recall, and Macro F1-score. While accuracy provides an overall measure of correctness, it can be influenced by class imbalance. Therefore, macro-averaged metrics are particularly important, as they compute performance independently for each class (negative, neutral, positive) and then average the results.

Macro F1-score is used as the primary evaluation metric because it balances precision and recall across all sentiment classes and provides a more reliable assessment of performance when class distributions are uneven.

Table II presents the performance of the four experimental variants under identical hyperparameter settings.

The results demonstrate that normalization strategies consistently improve performance compared to the baseline model trained on raw text. In particular, the slang-only model achieves the highest macro F1-score (0.5718), corresponding to an improvement of approximately 4.4 percentage points over the baseline (0.5277). The emoji-only model and the combined enhanced model also show notable gains of approximately 3–4 percentage points in macro F1-score.

TABLE II
Performance Comparison of Model Variants

Model	Acc	Prec	Rec	F1
Baseline (Raw Text)	0.7705	0.5089	0.5489	0.5277
Emoji-only Norm.	0.7784	0.6096	0.5664	0.5638
Slang-only Norm.	0.7822	0.6232	0.5724	0.5718
Enhanced (E+S)	0.7793	0.6054	0.5642	0.5590

Although improvements in accuracy appear modest, macro-level metrics show clearer benefits, indicating that normalization primarily enhances balanced class-wise performance rather than simply increasing majority-class predictions.

B. Statistical Significance Testing

To verify whether the observed improvements between model variants are statistically significant, we applied McNemar's test on paired predictions over the same test set. This non-parametric test evaluates whether two classifiers differ significantly in terms of their error distributions.

We compared the Baseline and Enhanced models over the IGWA-only test set. The contingency analysis showed that the Enhanced model corrected 228 errors made by the Baseline, while introducing 180 new errors.

McNemar's test yielded $\chi^2 = 5.41$ with $p = 0.020$, indicating that the performance improvement of the Enhanced model over the Baseline is statistically significant at the $\alpha = 0.05$ level.

TABLE III
McNemar's test results on the IGWA-only test set. n_{01}: Baseline correct / Enhanced wrong. n_{10}: Baseline wrong / Enhanced correct.

Comparison	n_{01}	n_{10}	p-value
Baseline vs Enhanced	180	228	0.020

C. Impact of Normalization Strategies

The results suggest that slang normalization has a slightly stronger impact on performance than emoji normalization in this dataset. Although slang expressions appear in only a small portion of the reviews (0.29%), they often occur in short, sentiment-critical contexts where a single token can strongly influence the interpretation of the entire review. Converting informal abbreviations into standardized textual forms enables the model to better capture semantic meaning and reduce ambiguity during tokenization.

Because macro F1-score assigns equal weight to each sentiment class, improvements on a relatively small but challenging subset of examples can still produce noticeable gains in overall performance. In particular, resolving ambiguous slang tokens may help the classifier better interpret difficult or borderline cases, which contributes to improved macro-level metrics.

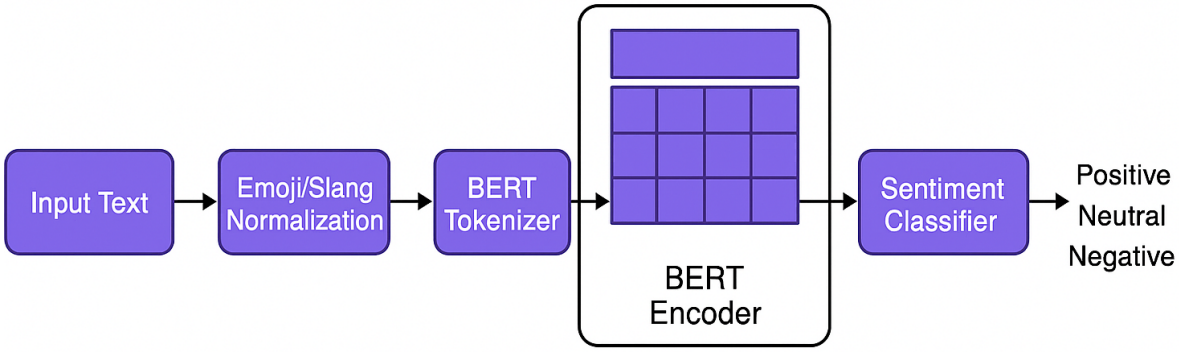

Fig. 1. BERT-based sentiment classification architecture as implemented in this study.

The combined enhanced model (emoji + slang) does not significantly outperform the slang-only model, suggesting that the contributions of normalization strategies are not strictly additive. One possible explanation is that certain sentiment cues overlap, and additional normalization may introduce minor semantic smoothing effects that do not further increase discriminative power.

D. Class-wise Behavior and Neutral Class Challenges

Across all model variants, the neutral class remains the most challenging to classify accurately. Neutral reviews often contain mixed sentiment, ambiguous phrasing, or factual statements without strong emotional polarity. As a result, confusion frequently occurs between neutral and positive or neutral and negative classes.

Normalization improves macro F1-score partly by stabilizing predictions for minority or ambiguous classes. By making informal sentiment cues explicit in textual form, normalization reduces the likelihood that emotionally charged emojis or slang terms are ignored or misinterpreted. However, the difficulty of modeling subtle or context-dependent neutrality persists across all variants.

E. Error Analysis

To better understand the impact of normalization, we qualitatively analyze representative test-set examples comparing the Baseline and Enhanced models. The representative examples were selected based on systematic misclassification patterns observed in the confusion matrices. Specifically, we considered (i) instances misclassified by the Baseline model but correctly predicted by the Enhanced model after normalization, and (ii) instances consistently misclassified across both models. This selection strategy ensures that the presented examples reflect recurring error patterns rather than isolated cases.

a) Enhanced Correcting Baseline Errors.: Review: "Nice app but too many ads (angry emoji)"
True label: Negative
Baseline: Neutral
Enhanced: Negative
The baseline model misclassified this review as Neutral, likely due to the mixed sentiment structure ("Nice app but…"). After emoji normalization, the angry emoji (angry face) strengthened the negative polarity signal, enabling the Enhanced model to correctly predict Negative sentiment.

b) Emoji-Driven Polarity Disambiguation.: Review: "Not working again (sad emoji)"
True label: Negative
Baseline: Neutral
Enhanced: Negative
Here, the baseline model failed to capture dissatisfaction expressed implicitly. Demojization of the sad emoji improved sentiment alignment and contributed to correct classification.

c) Persistent Failure Case (Sarcasm).: Review: "Great. Another update that broke everything."
True label: Negative
Baseline: Positive
Enhanced: Positive
Both models misclassified this sarcastic review. Although lexically positive words are present ("Great"), the actual sentiment is negative. This illustrates a limitation of surface-level normalization and suggests that deeper contextual or pragmatic modeling may be necessary.

d) Discussion of Recurring Patterns.: Across the test set, three recurring patterns emerge. First, sarcasm and implicit criticism remain challenging even after normalization. Second, emojis generally provide strong sentiment cues, and their textual normalization improves polarity detection in short reviews. Third, slang expressions, while relatively infrequent in the dataset, may carry nuanced pragmatic meanings that are not fully captured by literal expansions. These findings indicate that normalization enhances robustness but does not fully address higher-level discourse phenomena. These observations align with our quantitative findings, where normalization yields statistically significant improvements yet leaves room for modeling richer contextual phenomena.

F. Confusion Matrix Analysis

To further analyze class-wise behavior, confusion matrices for the baseline and normalization-based models are presented in Fig. 2, Fig. 3, Fig. 4, and Fig. 5. As shown in Fig. 2, the baseline model exhibits noticeable confusion between neutral and positive classes, as well as between neutral and negative classes. This indicates difficulty in distinguishing moderately expressed or ambiguous sentiment without explicit normalization of informal expressions. The emoji-only normalization model (Fig. 3) reduces certain misclassifications by making emotional cues embedded in emojis more explicit in textual form. This leads to improved recognition of sentiment-bearing tokens that may otherwise be treated as unknown or weakly represented features. The slang-only normalization model (Fig. 4) demonstrates further reduction in class confusion, particularly for reviews containing informal abbreviations. Converting slang expressions into standardized equivalents enables more consistent semantic interpretation during tokenization.

Finally, the enhanced model integrating both emoji and slang normalization (Fig. 5) shows improved stability across classes compared to the baseline. Although the combined approach does not drastically outperform the slang-only model, it contributes to more balanced predictions and reduced variability in class-wise errors. As shown in Fig. 2, (Baseline)

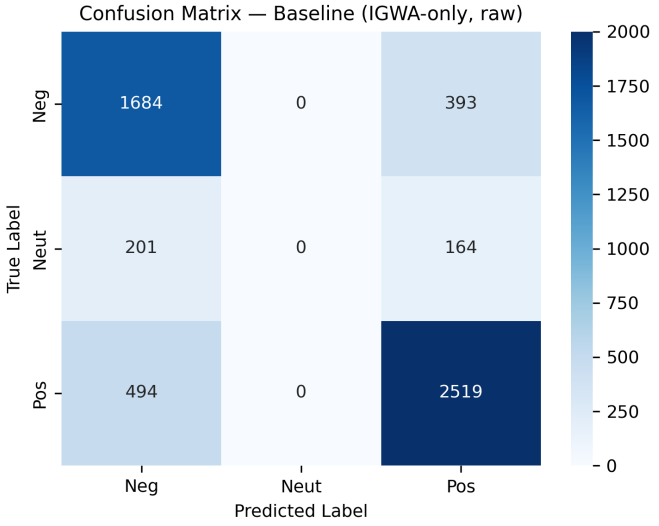

Fig. 2. Confusion matrix of the baseline BERT model without normalization.

demonstrates a substantial incidence of off-diagonal mistakes, particularly the inclination to classify Neutral items as Positive, signifying considerable confusion between neutral and positive emotion in the absence of normalization. As shown in Fig. 3, (Emoji-only) demonstrates that the normalization of emojis enhanced the model's ability to differentiate between positive and negative sentiments, thereby decreasing confusion errors, and yielded a modest improvement in recognizing neutral statements, although numerous neutrals remained misclassified with other categories. As shown in Fig. 4, (Slang-

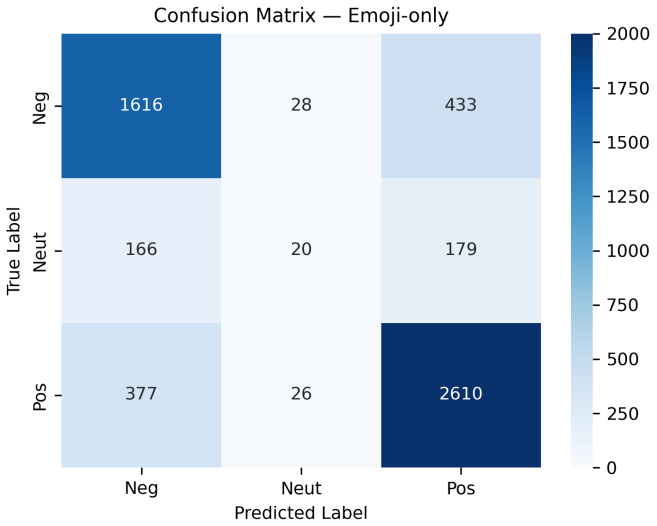

Fig. 3. Confusion matrix of the emoji-only normalization model.

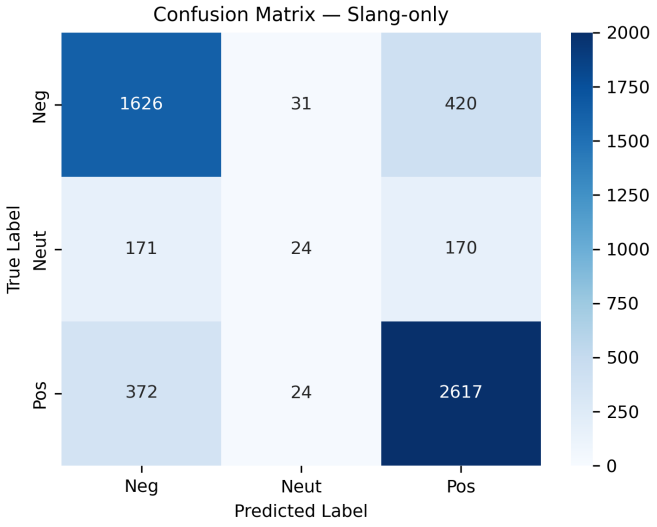

Fig. 4. Confusion matrix of the slang-only normalization model.

only) is a confusion matrix exhibiting a reduction in severe misclassifications: Neutral reviews are far less prone to being confused with Positive or Negative evaluations, and the likelihood of mixing up Negative and Positive reviews is also slightly diminished. The residual errors are less unilateral; for example, a limited quantity of neutral instances may still be categorized as positive if they exhibit genuinely uncertain sentiment, while the bias is significantly reduced compared to the baseline. This visualization highlights that the normalizing of slang significantly enhanced the clarity of Neutral and Negative inputs, resulting in a more distinct separation between the classes in the confusion matrix. As shown in Fig. 5, the Enhanced Model clearly demonstrates a decrease in Neutral to Positive and Neutral to Negative misclassifications, confirming that our normalization method effectively mitigated

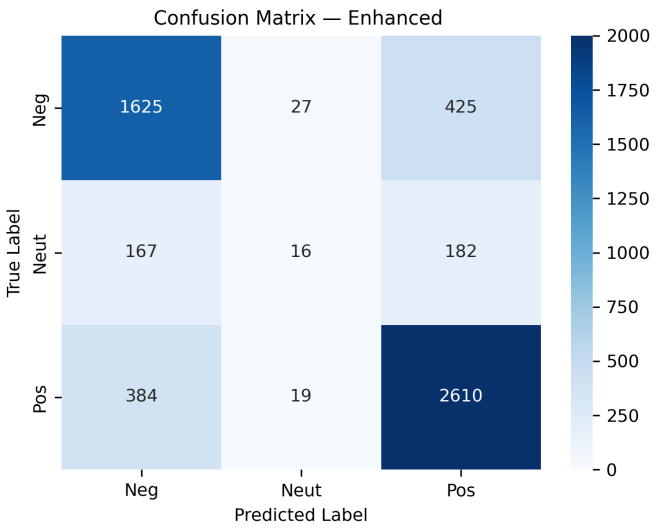

Fig. 5. Confusion matrix of the enhanced emoji and slang normalization model.

the model's tendency towards categorizing all instances as positive or negative. The data reveal a reduction in Negative-Positive flip mistakes, suggesting that the presence of both slang and emoji cues diminishes the model's propensity to conflate opposing mood polarities. The study of the confusion matrix corroborates the previous quantitative results: each normalization mitigated distinct error types, collectively producing the most robust classifier with minimal cross-class confusion.

V. Discussion and Implications for Inclusion

The experimental results demonstrate that incorporating emoji and slang normalization consistently improves macro-level performance in sentiment classification of mobile app reviews. While the numerical gains in macro F1-score (approximately 3–4 percentage points) may appear moderate, their implications extend beyond pure accuracy improvements. The findings highlight the importance of modeling informal digital expressions as meaningful linguistic signals rather than treating them as noise.

A. Informal Language and Linguistic Representation

User-generated content on digital platforms reflects diverse communication styles shaped by age, culture, region, and community practices. Emojis and slang are not merely stylistic elements; they are central components of how users express affect, identity, and stance. When NLP systems fail to interpret such expressions correctly, they risk systematically misrepresenting certain groups whose communication patterns deviate from formal written norms.

By normalizing emojis and slang into semantically interpretable textual representations, the proposed approach reduces the gap between informal digital language and pre-trained language model vocabularies. This transformation improves the model's ability to capture sentiment cues embedded in non-standard expressions. In doing so, the system becomes more robust to linguistic diversity present in real-world app reviews.

B. Reducing Linguistic Bias in Sentiment Analysis

Linguistic bias in NLP systems can arise when models are primarily trained on formal corpora and majority-language usage patterns [14]. Informal expressions, dialectal variants, or community-specific slang may be underrepresented in training data. As a result, models may exhibit lower performance for texts produced by specific demographic groups. Although this study does not directly measure fairness, the findings suggest that improved handling of informal expressions may reduce the risk that certain communication styles are systematically underinterpreted by sentiment models. This should be viewed as a hypothesis supported indirectly by performance improvements, rather than as direct evidence of reduced demographic bias.

In practical applications, sentiment analysis results may influence product decisions, feature prioritization, or perceptions of user satisfaction. Ensuring that informal or minority linguistic expressions are not systematically misclassified is therefore relevant to inclusive digital governance and responsible data-driven decision-making.

C. Implications for Well-Being-Oriented Analytics

Mobile app reviews often contain expressions of frustration, dissatisfaction, or appreciation that directly relate to user well-being. Misinterpreting emotionally rich informal expressions may lead to underestimation of negative user experiences or misclassification of nuanced feedback. By improving robustness to informal language, normalization strategies enhance the reliability of automated review analytics.

More reliable sentiment signals can support developers and organizations in identifying problematic updates, addressing user complaints, and improving user experience. From a broader perspective, transparent and robust text analytics pipelines contribute to building trust in algorithmic systems used to interpret public opinion.

D. Limitations and Ethical Considerations

Despite the improvement observed, several limitations remain. First, normalization relies on predefined emoji descriptions and a manually curated slang dictionary, which may not capture evolving linguistic trends or context-dependent meanings. Second, the absence of demographic metadata prevents direct measurement of fairness or group-wise performance differences. Third, the approach does not address deeper semantic challenges such as sarcasm, irony, or pragmatic ambiguity.

Future research could integrate dynamic slang learning mechanisms, contextual emoji embeddings, or fairness-aware evaluation frameworks to further investigate the relationship between informal language handling and equitable model performance.

VI. Conclusion

This paper investigated the impact of emoji and slang normalization on BERT-based sentiment analysis of mobile app reviews. While transformer-based models such as BERT achieve strong performance on formal text, their effectiveness can be limited when confronted with informal digital expressions commonly found in user-generated content. Emojis and slang terms carry significant emotional and contextual meaning, yet they are often underrepresented or inconsistently modeled in standard preprocessing pipelines.

To address this challenge, we designed a modular normalization framework that converts emojis into standardized textual descriptions and expands common slang expressions prior to tokenization. We evaluated four experimental variants—baseline, emoji-only, slang-only, and combined normalization—under identical hyperparameter settings on a dataset of Instagram and WhatsApp reviews labeled using star ratings.

The experimental results demonstrate that normalization consistently improves balanced performance, with macro F1-score increasing by approximately 3–4 percentage points compared to the baseline model. Slang normalization showed the strongest individual contribution, while the combined approach also improved robustness over raw text processing. These findings indicate that explicitly modeling informal digital expressions enhances sentiment classification reliability in real-world app review environments.

From an inclusion perspective, improving robustness to informal language reduces the risk that certain communication styles are systematically misinterpreted. Although this study does not directly measure demographic fairness, handling linguistic variability contributes to more equitable representation of user sentiment in automated analytics systems. Reliable sentiment interpretation is particularly relevant in contexts where review analysis informs product development, user satisfaction assessment, and digital well-being initiatives.

Several limitations remain. The normalization strategies rely on predefined dictionaries and do not adapt dynamically to evolving slang or context-dependent emoji meanings. Additionally, the dataset is limited to English-language reviews from two applications and does not include demographic annotations for fairness analysis. Future work may explore adaptive slang learning, contextual emoji embeddings, multilingual evaluation, and fairness-aware assessment frameworks to further strengthen inclusion-oriented sentiment analysis systems. Overall, the results indicate that explicit normalization of emojis and slang can improve the robustness of sentiment classification on informal app-review text. At the same time, the findings should be interpreted in light of several constraints, including weak labels derived from star ratings, limited domain coverage, and the absence of subgroup-level fairness evaluation.

AI-Generated Content Acknowledgement

AI-based language assistance tools were used during the writing process to support grammar refinement and improve clarity of expression. All research design, data collection, model development, experimental analysis, and conclusions presented in this paper were conducted and verified by the authors.

References

[1] J. Devlin, M.-W. Chang, K. Lee, and 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. [Online]. Available: https://aclanthology.org

[2] P. Kralj Novak, J. Smailović, B. Sluban, and I. Mozetič, "Sentiment of emojis," *PloS one*, vol. 10, no. 12, p. e0144296, 2015. [Online]. Available: https://journals.plos.org/plosone/article?id=10.1371/journal.pone.0144296

[3] B. Felbo, A. Mislove, A. Søgaard, I. Rahwan, and S. Lehmann, "Using millions of emoji occurrences to learn any-domain representations for detecting sentiment, emotion and sarcasm," in *Proceedings of the 2017 conference on empirical methods in natural language processing*, 2017, pp. 1615–1625.

[4] H. Zhang, R. Sproat, A. H. Ng, F. Stahlberg, X. Peng, K. Gorman, and B. Roark, "Neural models of text normalization for speech applications," *Computational Linguistics*, vol. 45, no. 2, pp. 293–337, 2019. [Online]. Available: https://aclanthology.org/J19-2004/

[5] W. Ullah, Z. Zhang, and K. Stefanidis, "Sentiment analysis of mobile apps using bert," in *International Conference on Industrial, Engineering and Other Applications of Applied Intelligent Systems*. Springer, 2023, pp. 66–78. [Online]. Available: https://link.springer.com/chapter/10.1007/978-3-031-36822-6_6

[6] F. Å. Nielsen, "A new anew: Evaluation of a word list for sentiment analysis in microblogs," *arXiv preprint arXiv:1103.2903*, 2011.

[7] J. W. Pennebaker, M. E. Francis, R. J. Booth *et al.*, "Linguistic inquiry and word count: Liwc 2001," *Mahway: Lawrence Erlbaum Associates*, vol. 71, no. 2001, p. 2001, 2001. [Online]. Available: http://downloads.liwc.net.s3.amazonaws.com/LIWC2015_OperatorManual.pdf

[8] Y. Liu, M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov, "Roberta: A robustly optimized bert pretraining approach," *arXiv preprint arXiv:1907.11692*, 2019. [Online]. Available: https://arxiv.org/abs/1907.11692

[9] Z. Lan, M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut, "Albert: A lite bert for self-supervised learning of language representations," in *8th International Conference on Learning Representations (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=H1eA7AEtvS

[10] K. Clark, M.-T. Luong, Q. V. Le, and C. D. Manning, "Electra: Pre-training text encoders as discriminators rather than generators," in *8th International Conference on Learning Representations (ICLR)*, 2020. [Online]. Available: https://openreview.net/forum?id=r1xMH1BtvB

[11] X. Li, Z. Zhang, and K. Stefanidis, "Sentiment-aware analysis of mobile apps user reviews regarding particular updates," *ICSEA 2018*, p. 109, 2018.

[12] P. Puustinen, M. Stratigi, and K. Stefanidis, "Stracker: A framework for identifying sentiment changes in customer feedbacks," *Information Systems*, vol. 128, p. 102491, 2025.

[13] K. Katsarou, R. Jeney, and K. Stefanidis, "Mutual: Multi-domain sentiment classification via uncertainty sampling," in *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing, SAC 2023*. ACM, 2023, pp. 331–339.

[14] H. Miller, J. Thebault-Spieker, S. Chang, I. Johnson, L. Terveen, and B. Hecht, ""blissfully happy" or "ready tofight": Varying interpretations of emoji," in *Proceedings of the international AAAI conference on web and social media*, vol. 10, no. 1, 2016, pp. 259–268. [Online]. Available: https://ojs.aaai.org/index.php/ICWSM/article/view/14757