IJRAR.ORG E-ISSN: 2348-1269, P-ISSN: 2349-5138

INTERNATIONAL JOURNAL OF RESEARCH AND ANALYTICAL REVIEWS (IJRAR) | IJRAR.ORG

An International Open Access, Peer-reviewed, Refereed Journal

Enhancing Hinglish Sentiment Analysis using Phonetic Normalization and Explainable Transformer Models

Aditi Chillal, Nidhi Chavhan
Software Developer, Software Developer

Abstract:—Hinglish, a code-mixed language combining Hindi and English, presents challenges due to phonetic variability and inconsistent spelling. This paper proposes a hybrid phonetic normalization approach combining rule-based and dictionary based techniques. Using a transformer-based model, we compare baseline performance with normalized text. Experimental results show improvement in accuracy from approximately 56–57% to 59–60% and F1-score from 55–56% to 58–59%. Minor variations of ±1–2% are observed due to training stochasticity. SHAPbased analysis further demonstrates improved interpretability after normalization.

Index Terms—*Hinglish, Sentiment Analysis, Normalization, Transformers, Explainability, SHAP.*

I. INTRODUCTION

Hinglish, a widely used code-mixed language combining Hindi and English, is prevalent on social media platforms. It lacks standardized spelling and grammar, leading to high variability in text representation. For example, the word “nahi” may appear as “nhi” or “nai”, making it difficult for natural language processing (NLP) systems to interpret consistently.

Transformer-based models have shown strong performance in multilingual tasks, but they remain sensitive to noisy inputs. To address this, we propose a phonetic normalization approach to standardize Hinglish text before classification. Additionally, we incorporate SHAP explainability to analyze model behavior.

The main contributions of this work are:

• A hybrid phonetic and dictionary-based normalization method
• Empirical improvement in sentiment classification performance
• Explainability analysis using SHAP

II. METHODOLOGY

A. Dataset

The dataset consists of Hinglish text samples labeled into three sentiment classes: negative, neutral, and positive. Missing values were handled and all text was converted into string format. The dataset consists of approximately 14,000 Hinglish text samples obtained from the Sentimix shared task, annotated into three sentiment classes: positive, negative, and neutral.

B. Normalization

A two-stage normalization pipeline was used:

Phonetic normalization:

• aa → a
• ee → i
• ph → f
• kh → k

Dictionary normalization:

• nhi, nai → nahi
• accha, achha → acha
• bohot, bohut → bahut

C. Model

A transformer-based model (IndicBERT) was fine-tuned for 3-class sentiment classification.

D. Training Setup

• Train-test split: 90:10
• Batch size: 16
• Learning rate: 2 × 10− 5
• Epochs: 2
• Max length: 128

A fixed random seed was used; however, slight variations (±1–2%) were observed across runs.

III. RESULTS

A. Quantitative Results

Table I
Performance Comparison

Model	Accuracy	F1 Score
Baseline	56–57%	55–56%
Normalized	59–60%	58–59%

The normalized model consistently outperforms the baseline across both metrics.

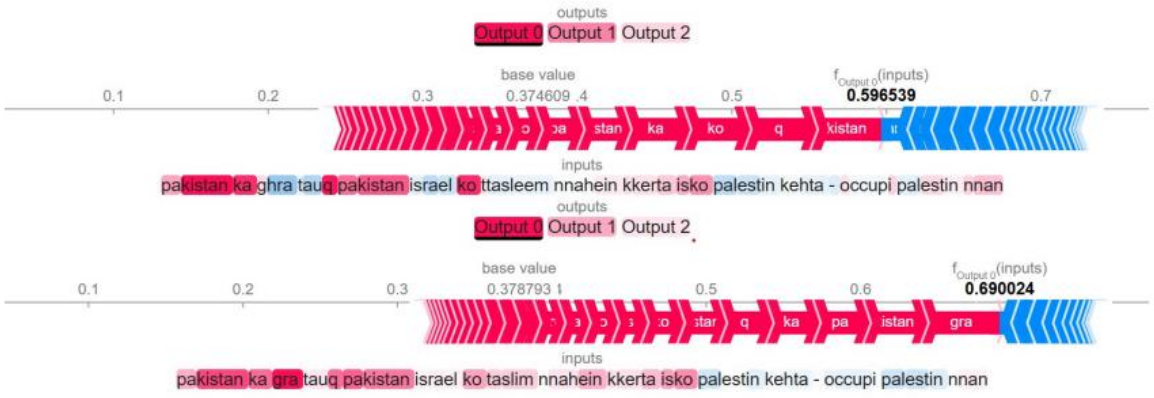
B. Discussion

The improvement is attributed to reduced lexical variability and improved token consistency. Despite minor fluctuations due to stochastic training, normalization consistently improves performance. This work focuses specifically on analyzing the impact of normalization on transformer-based models. Exploring its effectiveness across traditional machine learning models is left for future work. The current results clearly demonstrate that normalization consistently enhances model performance and interpretability

IV. EXPLAINABILITY

We use SHAP to interpret model predictions.

Fig. 1. SHAP explanations before and after normalization



Normalization results in clearer and more stable token-level importance, improving interpretability.

V. CONCLUSION

This paper demonstrates that phonetic normalization improves Hinglish sentiment analysis performance and interpretability. The approach is simple yet effective for handling noisy code-mixed text.

VI. FUTURE WORK

Future work includes incorporating larger datasets, experimenting with advanced multilingual transformer models, and extending normalization techniques using neural phonetic models.

VII. ACKNOWLEDGMENT

The author thanks the open-source community and dataset contributors.

REFERENCES

[1] Vaswani et al., “Attention Is All You Need,” 2017.
[2] Khanuja et al., “IndicBERT,” 2021.
[3] Lundberg and Lee, “SHAP,” 2017.
[4] Devlin et al., “BERT,” 2019.
[5] S. Patwa et al., “SemEval-2020 Task 9: Overview of Sentimix,” 2020.