Enhanced Sentiment Analysis for Code-Switched Hinglish and Gen Z Slang in Product Reviews: A Hybrid Lexicon-Based Approach

Aayushi Soni, Ishitaba Umat

Department of Computer Science, Institute of advanced research (IAR)

ABSTRACT

The proliferation of code-mixed languages and evolving internet slang presents significant challenges for traditional sentiment analysis systems. This paper introduces a novel hybrid sentiment analysis framework specifically designed to handle Hinglish (Hindi-English code-switching) and contemporary Gen Z slang in product review contexts. Our approach combines lexicon-based methods with contextual negation handling and intensity modulation to achieve robust sentiment classification. We present a real-time sentiment analyzer that demonstrates superior performance in detecting sentiment nuances across multilingual and generational linguistic boundaries. The system achieves enhanced accuracy through custom-built dictionaries containing over 150 Hinglish and Gen Z slang terms, coupled with sophisticated emotion detection capabilities. Our implementation provides not only sentiment polarity but also confidence metrics, slang density scores, and multi-dimensional emotional profiling, offering comprehensive insights into user-generated content.

Keywords: Sentiment Analysis, Code-Switching, Hinglish, Gen Z Slang, Natural Language Processing, Lexicon-Based Methods, Product Reviews

1. INTRODUCTION

1.1 Background and Motivation

The digital age has fundamentally transformed communication patterns, giving rise to hybrid linguistic phenomena that challenge traditional Natural Language Processing (NLP) systems. Code-switching, particularly between Hindi and English (Hinglish), has become ubiquitous in South Asian digital communities, with over 350 million internet users engaging in bilingual communication patterns. Simultaneously, Gen Z users worldwide have developed distinct lexical patterns characterized by creative slang and rapid semantic evolution.

Traditional sentiment analysis tools, predominantly trained on monolingual English corpora, exhibit significant performance degradation when confronted with code-mixed text. The challenge is compounded by the informal nature of online product reviews, where users freely alternate between languages, employ non-standard orthography, and utilize contemporary slang that may not exist in conventional lexicons.

Product reviews represent a critical source of consumer insights for businesses. A survey of e-commerce platforms in India reveals that over 60% of user-generated content exhibits code-switching behavior, while global platforms increasingly encounter Gen Z linguistic innovations that defy traditional sentiment classification. The inability to accurately process such content results in substantial information loss and misclassification, directly impacting business intelligence and customer understanding.

1.2 Research Contributions

This research makes the following key contributions:

1) Development of comprehensive lexicons containing 150+ Hinglish and Gen Z slang terms with weighted sentiment scores;

2) A hybrid sentiment analysis architecture combining lexicon-based methods with contextual negation detection and intensity modification;

3) Implementation of multi-dimensional emotional profiling alongside traditional polarity classification;

4) A real-time web-based analyzer with interactive visualization capabilities for sentiment trends and confidence metrics;

5) Empirical validation demonstrating improved accuracy in code-mixed and slang-heavy contexts.

2. RELATED WORK

2.1 Code-Switching and Hinglish Analysis

Code-switching has been extensively studied in computational linguistics. Solorio and Liu (2008) pioneered work on automatic language identification in code-switched text, while Vyas et al. (2014) created the first POS-tagged Hinglish corpus. Sharma et al. (2016) developed shallow parsing techniques for code-mixed social media text, achieving modest accuracy improvements over baseline models.

Joshi et al. (2020) proposed a sub-word level composition model for sentiment analysis of Hindi-English code-mixed text, focusing on the challenges of non-standard spelling and morphological variations. However, these approaches often struggle with the rapid evolution of internet slang and the non-standardized nature of Gen Z communication.

2.2 Sentiment Analysis Methodologies

Sentiment analysis has evolved through three primary paradigms: lexicon-based approaches, machine learning methods, and deep learning architectures. Lexicon-based methods, while simpler, offer interpretability and require no training data. Liu (2012) demonstrated that well-crafted lexicons combined with linguistic rules can achieve competitive performance, particularly in domain-specific applications.

Pang and Lee (2008) provided a comprehensive survey of opinion mining techniques, establishing foundational methodologies. More recently, transformer-based models like BERT and GPT have dominated sentiment analysis benchmarks. However, their effectiveness diminishes significantly with out-of-vocabulary terms, code-mixed text, and rapidly evolving slang.

2.3 Gen Z Language and Internet Slang

The linguistic patterns of Gen Z have received limited attention in NLP research. McCulloch (2019) documented the sociolinguistic aspects of internet language evolution, noting the rapid pace of semantic shift in online communities. Tagliamonte (2016) analyzed social media language among young adults, identifying key morphological and lexical innovations.

However, computational approaches to Gen Z slang remain scarce. Existing sentiment lexicons like VADER (Hutto and Gilbert, 2014) and SentiWordNet lack coverage of contemporary slang terms such as "bussin," "cap," "mid," and "slaps," necessitating custom lexicon development.

3. METHODOLOGY

3.1 System Architecture

Our sentiment analysis system employs a modular architecture consisting of five primary components: preprocessing, lexicon-based scoring, contextual analysis, emotion detection, and confidence estimation. The system processes text through a pipeline that maintains linguistic context while accommodating orthographic variations common in informal communication.

The architecture is implemented using Python with the TextBlob library for baseline sentiment computation, augmented with custom modules for Hinglish and Gen Z slang processing. The system supports real-time analysis through a Streamlit-based web interface with Plotly visualizations.

3.2 Lexicon Construction

We constructed four specialized lexicons through a combination of corpus analysis and expert annotation:

Hinglish Positive Lexicon: Contains 45 terms including transliterations ("mast," "zabardast," "kamaal") and hybrid expressions

("ekdum best"). Each term is assigned a sentiment score ranging from 0.6 to 1.0 based on intensity.

Hinglish Negative Lexicon: Comprises 35 terms including colloquial expressions ("bekaar," "faltu," "bakwas") with scores from -0.6 to -1.0.

Gen Z Positive Slang: Contains 40 contemporary terms ("bussin," "slaps," "goated," "no cap") with contextual variations.

Gen Z Negative Slang: Includes 30 terms ("mid," "cringe," "L," "ratio") representing disapproval or criticism.

Additionally, we compiled an intensifier lexicon of 20 terms common in Hinglish discourse ("boht," "ekdum," "bilkul"), each with multiplier values ranging from 1.2 to 1.5.

3.3 Sentiment Scoring Algorithm

The sentiment scoring process combines lexicon matching with contextual modulation. For a given text T tokenized into words $w_1, w_2, ..., w_n$, the custom sentiment score is computed as:

$$S_{custom} = (1/N) \times \Sigma (I_i \times L_i \times N_i)$$

where N is the count of matched lexicon terms, I_i is the intensifier multiplier (default 1.0), L_i is the lexicon sentiment score, and N_i is the negation modifier (+1 or contextual adjustment).

3.4 Negation Handling

Negation significantly impacts sentiment polarity. We implement a context-window approach that examines three tokens before and after each sentiment-bearing word. Negation markers include both English ("not," "no," "never") and Hinglish variants ("nahi," "mat," "nai").

When negation is detected, positive sentiment scores are inverted, while negative scores are either inverted or attenuated by 40% depending on the negation context. Special handling is implemented for phrases like "no cap" (Gen Z slang meaning "no lie"), which appears negated but carries positive assertion.

3.5 Hybrid Polarity Calculation

The final sentiment polarity combines custom lexicon scores with TextBlob's baseline polarity using weighted averaging:

$$P_{final} = 0.8 \times P_{custom} + 0.2 \times P_{baseline}$$

This weighting scheme prioritizes our domain-specific lexicons while retaining baseline coverage for standard English sentiment expressions. Classification thresholds are set at ±0.15 to distinguish positive, negative, and neutral sentiments.

3.C Emotion Detection Framework

Beyond binary or ternary sentiment classification, we implement multi-dimensional emotion profiling across four categories: joy, sadness, anger, and excitement. Each category contains 10-15 keyword indicators drawn from both lexicons and emotional expression databases.

Emotion scores are normalized by total emotional content, providing proportional representation rather than absolute counts. This approach reveals nuanced emotional profiles that simple polarity classification cannot capture.

3.7 Confidence and Slang Metrics

System confidence is estimated based on absolute polarity magnitude and lexicon match count:

$$C = min(|P_{final}| \times 100 + match\_bonus, 98)$$

where match_bonus is +10 if custom lexicon matches occur. The slang score quantifies code-mixing intensity as:

$$Slang\_Score = min(matches \times 20, 100)$$

These metrics provide transparency into the analysis process and help identify reviews requiring human verification.

4. IMPLEMENTATION

4.1 System Development

The sentiment analyzer is implemented as a web application using Streamlit 1.28.0, providing an accessible interface for real-time analysis. The technology stack includes:

- **Backend:** Python 3.8+ with TextBlob 0.17.1 for NLP operations
- **Data Processing:** Pandas 2.0.3 and NumPy 1.24.3
- **Visualization:** Plotly 5.17.0 for interactive charts
- **Frontend:** Streamlit with custom Tailwind CSS styling

4.2 User Interface Design

The interface employs a glassmorphism design aesthetic with animated gradient backgrounds, providing a modern user experience. Key features include:

Input Section: Multi-line text area with quick-action buttons for loading sample reviews (positive, negative, Gen Z examples).

Results Dashboard: Three-panel display showing sentiment verdict, confidence gauge, and slang score with visual emphasis through color-coded cards and animated emojis.

Advanced Analytics: Radar chart visualizing four-dimensional sentiment profile (polarity, subjectivity, confidence, slang score) and horizontal bar charts for emotion distribution.

Historical Tracking: Sidebar panel displaying recent analyses with timestamp, text preview, and sentiment classification. Users can download session history as CSV for further analysis.

Trend Visualization: Line chart tracking polarity evolution across multiple analyses within a session, revealing sentiment patterns over time.

4.3 Text Preprocessing Pipeline

Input text undergoes cleaning operations while preserving linguistic features essential for code-mixed analysis:

1. URL removal using regex pattern matching
2. Special character filtering while retaining punctuation markers
3. Whitespace normalization
4. Preservation of Unicode characters for Hindi script support

Importantly, we do not perform lowercase normalization during lexicon matching to preserve proper noun capitalization in slang terms (e.g., "W" vs "w").

5. EXPERIMENTAL EVALUATION

5.1 Dataset and Testing

We evaluated the system using a manually curated test set of 200 product reviews collected from Indian e-commerce platforms and global social media. The dataset composition includes:

- 50 pure Hinglish reviews
- 50 Gen Z slang-heavy reviews
- 50 mixed code-switched reviews
- 50 standard English reviews (control group)

Ground truth labels were established through consensus annotation by three independent reviewers familiar with both Hinglish and contemporary internet slang.

5.2 Performance Metrics

The system achieved the following accuracy rates across test subsets:

- **Hinglish Reviews:** 84% accuracy
- **Gen Z Slang:** 82% accuracy
- **Mixed Code-Switched:** 78% accuracy
- **Standard English:** 88% accuracy
- **Overall Accuracy:** 83%

For comparison, TextBlob baseline achieved only 52% accuracy on Hinglish reviews and 48% on Gen Z slang reviews, demonstrating the substantial improvement provided by our custom lexicons and contextual processing.

5.3 Error Analysis

Common error sources include:

Sarcasm Detection: The system struggles with sarcastic expressions that invert apparent sentiment (e.g., "Oh great, another broken product").

Novel Slang: Extremely recent slang terms not captured in lexicons lead to misclassification or neutral defaults.

Complex Negation: Double negatives and implied negation through rhetorical questions occasionally confuse the parser.

Context-Dependent Terms: Words like "sick" or "mad" carry positive sentiment in slang contexts but negative connotations in formal usage.

5.4 Case Studies

Example 1: "Boht mast product hai! Quality ekdum top notch. This slaps fr!"

System correctly identifies strong positive sentiment (polarity: +0.87) with high slang score (60%), detecting intensifier "boht" modifying "mast" and recognizing Gen Z term "slaps."

Example 2: "Total waste of money. Bekaar quality, huge L. Cringe experience."

Accurate negative classification (polarity: -0.91) successfully combining Hinglish "bekaar" with Gen Z terms "L" and "cringe."

Example 3: "Product toh sahi hai but delivery was bakwas"

The system correctly identifies the mixed sentiment, with a neutral overall polarity (0.05) due to the positive "sahi hai" (correct/good) being balanced by the negative "bakwas" (nonsense/ rubbish).

6. DISCUSSION

C.1 Advantages of Hybrid Approach

The lexicon-based methodology offers several advantages over purely data-driven approaches:

Interpretability: Sentiment decisions can be traced to specific lexical matches, enabling system debugging and user trust.

Resource Efficiency: No requirement for large annotated training corpora or GPU-intensive model training, making deployment accessible.

Rapid Adaptation: New slang terms can be incorporated through simple lexicon updates without model retraining.

Multilingual Flexibility: The approach naturally accommodates code-switching without requiring parallel corpora or multilingual embeddings.

C.2 Limitations and Challenges

Despite strong performance, several limitations warrant acknowledgment:

Lexicon Maintenance: Internet slang evolves rapidly, requiring ongoing curation to maintain effectiveness. Terms can shift semantics or fall out of usage within months.

Sarcasm Blindness: Lexicon-based methods inherently struggle with sarcasm and irony, which require pragmatic reasoning beyond surface lexical content.

Compositional Semantics: The bag-of-words approach misses complex compositional effects where sentiment emerges from phrase structure rather than individual word meanings.

Domain Specificity: Lexicons tuned for product reviews may not transfer effectively to other domains like political discourse or entertainment reviews.

C.3 Future Enhancements

Several avenues for improvement merit exploration:

Neural Augmentation: Integrating contextual embeddings from multilingual models could improve handling of unseen terms while retaining lexicon interpretability.

Sarcasm Detection: Incorporating linguistic features like punctuation patterns, capitalization, and emoji usage could enable basic sarcasm identification.

Dynamic Lexicon Updates: Automated corpus analysis could identify emerging slang terms for semi-supervised lexicon expansion.

Aspect-Based Analysis: Extending the system to identify sentiment toward specific product aspects (quality, price, delivery) would provide richer insights.

Multimodal Integration: Product reviews often include images and ratings; integrating these signals could improve overall accuracy.

7. CONCLUSION

This research demonstrates that carefully constructed lexicon-based methods remain viable and effective for sentiment analysis in challenging linguistic contexts. By focusing on Hinglish code-switching and Gen Z slang, we address a critical gap in existing sentiment analysis tools that predominantly handle monolingual, formal text.

Our hybrid approach achieves 83% overall accuracy across diverse review types, substantially outperforming baseline methods on code-mixed content while maintaining computational efficiency. The system's interpretability, ease of deployment, and adaptability make it particularly suitable for practical applications in e-commerce and social media monitoring.

The work highlights the continued importance of linguistic knowledge and domain expertise in NLP system design. While deep learning has dominated recent research, hybrid approaches that combine classical methods with contemporary linguistic insights offer practical solutions for real-world challenges.

As digital communication continues evolving, sentiment analysis systems must adapt to accommodate linguistic innovation. This research provides a foundation for such adaptation, demonstrating that thoughtful engineering of lexical resources combined with contextual processing can effectively bridge the gap between traditional NLP and the linguistic realities of modern internet discourse.

8. REFERENCES

Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014). "I am borrowing ya mixing?" An analysis of English-Hindi code mixing in Facebook. *Proceedings of the First Workshop on Computational Approaches to Code Switching*, 116-126.

Hutto, C. J., & Gilbert, E. (2014). VADER: A parsimonious rule-based model for sentiment analysis of social media text. *Proceedings of the International AAAI Conference on Web and Social Media*, 8(1), 216-225.

Joshi, A., Prabhu, A., Shrivastava, M., & Varma, V. (2020). Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text. *arXiv preprint arXiv:2008.03943*.

Liu, B. (2012). *Sentiment analysis and opinion mining*. Synthesis Lectures on Human Language Technologies, 5(1), 1-167.

McCulloch, G. (2019). *Because Internet: Understanding the New Rules of Language*. Riverhead Books.

Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(1-2), 1-135.

Sharma, A., Gupta, S., Motlani, R., Banerjee, P., Srivastava, M., Mamidi, R., & Sharma, D. M. (2016). Shallow parsing pipeline for Hindi-English code-mixed social media text. *Proceedings of NAACL-HLT*, 1340-1345.

Solorio, T., & Liu, Y. (2008). Learning to predict code-switching points. *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 973-981.

Tagliamonte, S. A. (2016). *Teen Talk: The Language of Adolescents*. Cambridge University Press.

Vyas, Y., Gella, S., Sharma, J., Bali, K., & Choudhury, M. (2014). POS tagging of English-Hindi code-mixed social media content. *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing*, 974-979.