INTERNATIONAL JOURNAL OF NOVEL RESEARCH AND DEVELOPMENT (IJNRD)
© 2026 IJNRD | Volume 11, Issue 3, MARCH 2026 | ISSN: 2456-4184 | IJNRD.ORG

SMART READING FOR MULTILINGUAL TEXT

Author: Vijay Kumar Tati
Department of AI & DS, School of Engineering and Technology, Dhanalakshmi Srinivasan University, Samayapuram Campus, Tiruchirappalli, Tamil Nadu – 621112, India
Email: tativijay1@gmail.com

Syed Mansoor
Department of AI & DS, School of Engineering and Technology, Dhanalakshmi Srinivasan University, Samayapuram Campus, Tiruchirappalli, Tamil Nadu – 621112, India
Email: sdmansoormansoorsd29@gmail.com

Kundan Sai Tavva
Department of AI & DS, School of Engineering and Technology, Dhanalakshmi Srinivasan University, Samayapuram Campus, Tiruchirappalli, Tamil Nadu – 621112, India
Email: kundansaitavva@gmail.com

Guided by: Mrs. M. Suguna
Assistant Professor, Department of AI &DS,
School of Engineering and Technology,
Dhanalakshmi Srinivasan University,
Samayapuram Campus, Tiruchirappalli,
Tamil Nadu – 621112, India
Email: suguna15.9@gmail.com

***Abstract*—Sentiment analysis plays a crucial role in extracting insights from customer feedback across digital platforms. Traditional deep learning models process entire text sequences, resulting in increased computational cost and limited interpretability. This paper proposes a Selective Multilingual Text Understanding system that integrates contextual embeddings from XLM-Roberta with a GRU-based selective token mechanism and a CNN classifier. A sparsity-controlled regularization term governed by a tunable lambda parameter enables dynamic token selection while maintaining competitive accuracy. The system is trained on the IMDB sentiment dataset and deployed using a FastAPI-based web interface with real-time inference. Experimental results demonstrate that the selective model achieves comparable accuracy to a baseline model while reducing average token utilization, thereby improving computational efficiency. The proposed system further provides business-oriented insights by extracting strengths and concerns from customer feedback.**

Index Terms — Sentiment Analysis, Selective Reading, Multilingual NLP, GRU, CNN, XLM-Roberta, Business Intelligence.

1. INTRODUCTION

The rapid expansion of e-commerce and digital platforms has led to an exponential growth in user-generated content. Businesses rely heavily on sentiment analysis to interpret customer feedback and improve products and services. However, most deep learning-based sentiment models process entire text sequences without distinguishing between relevant and irrelevant tokens, leading to redundant computation.

To address this limitation, this work introduces a selective reading framework for multilingual sentiment analysis. The proposed system focuses on dynamically selecting important tokens before classification, thereby improving computational efficiency while preserving predictive performance. Additionally, the system provides structured business insights by categorizing extracted keywords into strengths and concerns.

2. LITERATURE REVIEW

Sentiment analysis has progressed from traditional machine learning approaches to deep learning and transformer-based architectures capable of modeling contextual and sequential dependencies in text data.

A. Hybrid Deep Learning Models

Transformer-based models such as BERT have significantly improved sentiment classification performance due to their bidirectional contextual representation. However, standalone transformer models may not fully capture sequential dynamics present in textual data. To address this limitation, hybrid architectures integrating BERT with recurrent networks such as GRU have been proposed.

These hybrid models leverage contextual embeddings generated by transformers and refine them through sequential modeling layers. Experimental results in prior work demonstrate that such architectures outperform traditional machine learning models (e.g., SVM) and standalone transformer-based systems in terms of accuracy, precision, recall, and F1-score. The integration of GRU layers enhances the model's ability to capture sentiment shifts, long-range dependencies, and complex linguistic patterns such as sarcasm and clause-level polarity changes.

Despite performance improvements, these models typically process the entire input sequence, leading to increased computational cost.

B. Selective Token Processing

Recent research has explored selective reading mechanisms to improve computational efficiency. Instead of processing all tokens, reinforcement learning-based policy networks are employed to identify and select only sentiment-relevant tokens. The selected tokens are then passed to a classification module for final prediction.

Such approaches demonstrate that not all words contribute equally to sentiment classification. By reducing the number of processed tokens while maintaining high accuracy, selective reading frameworks improve model efficiency. However, the integration of reinforcement learning introduces additional training complexity and computational overhead.

C. Research Gaps

Although prior research has achieved substantial improvements in sentiment classification accuracy, several limitations remain:

1. Most studies focus on single-language datasets rather than unified multilingual frameworks.
2. Existing systems primarily perform polarity classification without generating structured business insights such as keyword extraction.
3. Limited work addresses scalable deployment for bulk review analysis and real-world business intelligence applications.
4. Cross-lingual transformer architectures remain underexplored in integrated sentiment and keyword extraction systems.

D. Motivation for Proposed Work

To address these limitations, there is a need for a multilingual sentiment intelligence system that combines transformer-based contextual understanding with structured keyword extraction. Such a system should:

- Support multilingual input,
- Perform accurate sentiment classification,
- Extract dominant positive and negative keywords,
- Enable bulk review processing,
- Provide actionable insights for business decision-making.

The proposed work aims to bridge the gap between high-accuracy deep learning models and scalable, multilingual business-oriented sentiment analytics systems.

3. EXISTING SYSTEM

Current sentiment analysis systems are primarily designed to classify textual data into predefined polarity categories such as positive, negative, or neutral. Most existing systems rely on traditional machine learning algorithms (e.g., SVM, Naïve Bayes) or deep learning architectures such as CNNs, LSTMs, GRUs, and transformer-based models like BERT.

Recent transformer-based approaches significantly improve contextual understanding by leveraging bidirectional attention mechanisms. Hybrid models combining transformers with recurrent networks further enhance sequential dependency modelling. These systems achieve high classification accuracy on benchmark datasets.

However, existing systems exhibit several limitations:

1. Single-Task Focus: Most models perform only sentiment polarity classification without extracting structured insights such as dominant positive or negative keywords.
2. Monolingual Orientation: Many implementations are designed for a single language (e.g., English or Arabic), limiting their applicability in multilingual business environments.
3. Full-Sequence Processing: Existing models typically process entire text sequences, increasing computational cost and reducing efficiency for large-scale datasets.
4. Lack of Business Integration: Current research-oriented models are not optimized for practical deployment scenarios such as bulk CSV/XLS review ingestion, automated dashboard generation, or business-oriented analytics.
5. Limited Interpretability: While models provide sentiment labels, they often fail to offer interpretable outputs that explain why a particular sentiment was assigned.

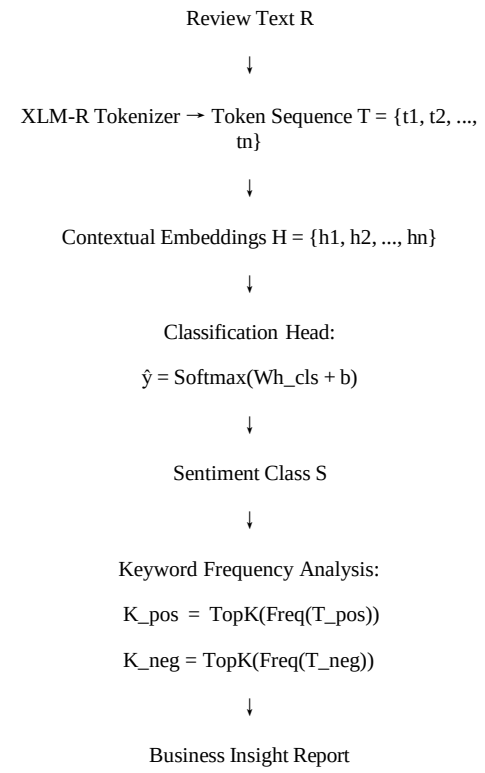
As a result, although existing systems achieve strong performance in controlled research settings, they do not fully address the requirements of scalable, multilingual, insight-driven sentiment intelligence platforms.

4. PROPOSED METHODOLOGY

4.1. System Architecture

The proposed architecture consists of four primary components:

1. **XLM-Roberta Encoder**
   Generates contextual embeddings for multilingual text input.
2. **GRU-Based Selective Module**
   Assigns importance scores to each token and applies sparsity-based filtering.
3. **CNN Classifier**
   Extracts discriminative features from selected token representations for sentiment prediction.
4. **Sparsity Regularization (Lambda Parameter)**
   Controls the trade-off between token selection and classification accuracy.



4.2. Dataset and Preprocessing

The IMDB dataset containing 50,000 labeled movie reviews (positive and negative) is used for training and evaluation. Text is tokenized using the XLM-Roberta tokenizer and truncated to a maximum sequence length of 128 tokens. Inputs are converted into tensor representations for GPU-accelerated training.

4.3. Training Strategy

The model is trained using cross-entropy loss combined with a sparsity penalty term:

where:

- $L_{classification}$ represents cross-entropy loss
- $L_{sparsity}$ penalizes excessive token selection
- λ controls sparsity intensity

Extensive hyperparameter tuning was performed, and lambda optimization required multiple training cycles spanning approximately 6–7 hours to identify an optimal trade-off.

4.4. Experimental Results

The performance of the proposed selective model is compared with a baseline XLM-Roberta classifier.

Model	Accuracy	Avg. Token Usage
Baseline	≈ 87%	100%
Selective Model	≈ 86–88%	20–30% Reduction

The selective model maintains competitive accuracy while reducing average token utilization. Confidence scores and token selection percentages are displayed through a real-time web interface.

5. DEPLOYMENT AND BUSINESS INSIGHT MODULE

The trained model is deployed using FastAPI with GPU-enabled inference. A web interface allows users to input customer feedback and receive:

- Sentiment prediction
- Confidence score
- Selective token usage percentage
- Business-oriented keyword categorization

Extracted keywords are categorized into:

- **Strengths** (positive indicators)
- **Concerns** (negative indicators)

This extension enhances practical usability for business analytics applications.

6. CHALLENGES

Key challenges encountered during development include:

- Sensitivity of the lambda parameter
- Extensive hyperparameter tuning
- GPU environment configuration
- Model-to-backend integration
- Balancing sparsity with predictive accuracy

7. CONCLUSION AND FUTURE WORK

This paper presents a selective multilingual sentiment analysis framework integrating contextual embeddings with token-level filtering. The system achieves competitive accuracy while reducing computational overhead. Additionally, it bridges research and practical application by generating structured business insights.

Future work includes:

- Multilingual dataset expansion
- Attention-based explainability
- Large-scale batch review analytics
- Ablation studies on selector components
- Optimization for real-time deployment at scale

REFERENCES

[1] M. Zouidine and M. Khalil, “Selective Reading for Arabic Sentiment Analysis,”
IEEE Access, vol. 13, pp. 59157–59169, 2025,
[2] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.
[3] A. Vaswani et al., “Attention is all you need,” in *Proc. NeurIPS*, 2017, pp. 5998–6008.
[4] M. Zouidine, H. Hammouchi, and M. Khalil, “Deep reinforcement learning-based early prediction for Arabic sentiment analysis,” in *Proc. ISIVC*, Marrakech, Morocco, 2024, pp. 1–5.
[5] M. Abdul-Mageed, A. Elmadany, and E. M. B. Nagoudi, “ARBERT & MARBERT: Deep bidirectional transformers for Arabic,” in *Proc. ACL*, 2021, pp. 7088–7105.
[6] W. Antoun, F. Baly, and H. Hajj, “AraBERT: Transformer-based model for Arabic language understanding,” 2020, arXiv:2003.00104.
[7] G. Inoue et al., “The interplay of variant, size, and task type in Arabic pre-trained language models,” in *Proc. WANLP*, 2021, pp. 92–104.
[8] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre- training of deep bidirectional transformers for language understanding,” in *Proc. NAACL-HLT*, Minneapolis, MN, USA, 2019, pp. 4171–4186.
[9] K. Cho et al., “Learning phrase representations using RNN encoder– decoder for statistical machine translation,” in *Proc. EMNLP*, Doha, Qatar, 2014, pp. 1724–1734.
[10] S. Hochreiter and J. Schmidhuber, “Long short-term memory,” *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, 1997.
[11] A. Vaswani et al., “Attention is all you need,” in *Proc. NeurIPS*, 2017, pp. 5998–6008.

AUTHOR BIOGRAPHIES

- Vijay Kumar Tati is the author of this paper, currently pursuing Engineering B-tech in department of Artificial intelligence and Data Science 4th year at Dhanalakshmi Srinivasan University, Trichy, Tamil Nadu. He is currently researching in Multilingual natural Processing and intelligent analytics systems. His research interests include transformer-based models, Sentiment analysis, and cross-lingual representation learning.

- Syed Mansoor is co-author of this paper who is pursuing Engineering B-tech in department of Artificial intelligence and Data Science 4th year at Dhanalakshmi Srinivasan University, Trichy, Tamil Nadu. He is interested in NLP Transformers and Other Machine Learning Projects.

- Kundan Sai Tavva is co-author of this paper who is pursuing Engineering B-tech in department of Artificial intelligence and Data Science 4th year at Dhanalakshmi Srinivasan University, Trichy, Tamil Nadu. He Interested in Data Science and Machine Learning also did research on Multimodal Data Fusion and Graph Based algorithms.

- Guided by Ms. M. Suguna, Assistant Professor in the Department of Artificial Intelligence and Data Science, Dhanalakshmi Srinivasan University, Samayapuram. She began her professional career as a Software Developer before entering the academic field, bringing valuable industry experience into teaching and curriculum development. With more than 3 years of experience in Teaching, she has guided multiple UG and PG research projects in emerging technologies. She previously served as Assistant Professor at Dhanalakshmi Srinivasan College of Arts and Science for Women (Perambalur) and Chidambaram Pillai College for Women., she has published research papers in reputed international journals, including Scopus-indexed publications.