Ensemble of Weighted Code Mixed Feature Engineering and Machine Learning-Based Multiclass Classification for Enhanced Opinion Mining on Unstructured Data

Ruchi Sharma[1], Dr. Pravin Shrinath[2]
Department of Information Technology,
SVKM's NMIMS, Mukesh Patel School of Technology Management & Engineering, Mumbai, India[1]
Associate Professor, Computer Engineering Department, SVKM's NMIMS,
Mukesh Patel School of Technology Management & Engineering, Mumbai, India[2]

***Abstract*—There is an exponential growth of opinions on online platforms, and the rapid rise in communication technologies generates a significant need to analyze opinions in online social networks (OSN). However, these opinions are unstructured, rendering knowledge extraction from opinions complex and challenging to implement. Although existing opinions mining systems are applied in several applications, limited research is available to handle code-mixed opinions of a non-structured nature where there is a switching of lexicons in languages within a single opinion structure. The challenge lies in interpreting complex opinions in multimedia networks owing to their unstructured nature, volume, and lexical structure. This paper presents a novel ensemble approach using machine learning and natural language processing to interpret code mixed opinions efficiently. Firstly, the opinions are extracted from the input corpus and preprocessed using proposed Extended Feature Vectors (EFV). Subsequently, the opinion mining system is implemented using a novel approach using weighted code mixed opinion mining framework (WCM-OMF) for multiclass classification. The proposed WCM-OMF model achieves an accuracy of 79.11% and 72% for the benchmark datasets, which is a significant improvement over existing Hierarchical LSTM, Random Forest, and SVM models and state-of-the-art-methods. The proposed solution can be implemented in opinion detection of other business sectors beneficial in obtaining actionable insights for efficient decision-making in enterprises and Business Intelligence (BI).**

Keywords*—Opinion mining; Machine learning; weighted ensemble; code mixed; Natural Language Processing; Business Intelligence; Online Social Networks

I. Introduction

With the advent of information and communication technologies, there is an exponential volume of user-generated content [1] available online. However, this data is primarily in the unstructured format in the form of opinions [2-4], suggestions, and forum posts. A further significant chunk of opinions generated are code-mixed. Code mixing is the alternating between two or more languages in a single opinion text source. This renders opinion mining complex to implement and interpret.

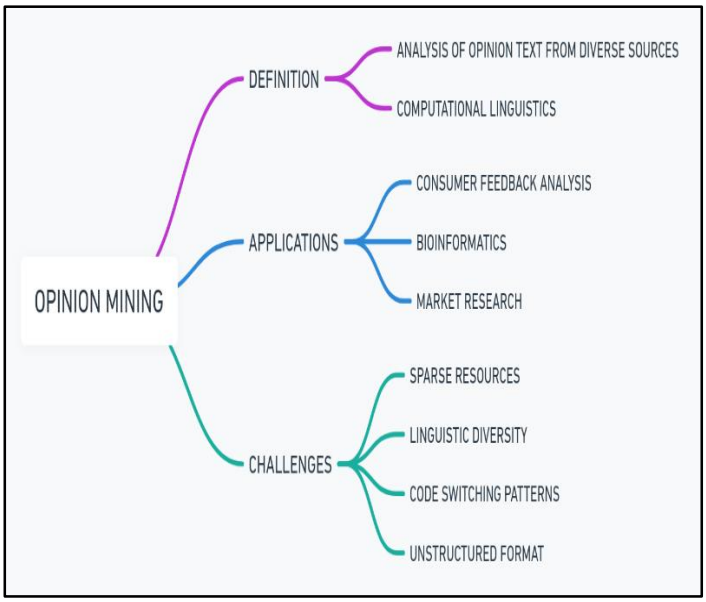

Fig. 1. Opinion mining trends and challenges.

Fig. 1 illustrates the opinion mining trends and challenges in various industrial sectors. The research specifically benefits multilingual cultures [5, 6] wherein multiple users convey the same expression differently despite belonging to the same regional area. Although there are studies related to opinion mining and opinion extraction, there is limited research on handling code-switching issues in opinion corpus. Also, existing studies focus more on mono-lingual input and overlook the lexical nuances of code mixed opinions. Given the aforementioned issue, we have developed a novel weighted code mixed opinion mining framework (WCM-OMF) for handling the code switching challenge. The proposed solution is the design and development of an ensemble method combining natural language processing and machine learning techniques. Compared to the existing systems, the proposed method achieves better results, improving model accuracy and F1 score metrics.

A. Applications of Opinion Mining

Opinion Mining is essential for achieving Business Intelligence in enterprises of several sectors as described in this subsection:

1) Consumer feedback analysis is one of the primary applications of opinion mining [7]. The extraction of consumer opinions and feedback is significant for enhancing the quality of products and services.

2) Manufacturing: Identifying opinions towards a product or service during a specific timeframe is highly valuable for firms, demonstrating the practical benefits of opinion mining in manufacturing.

3) Political analysis: Opinion mining of users about a political candidate [9, 10] or government policy can aid in understanding the public stance on a topic.

4) Risk analysis: Effective opinion mining can help prevent or mitigate risks and significant disasters in public sector companies.

B. Challenges of Opinion Detection

Existing Opinion Detection systems primarily focus on monolingual input. However, there is a vast amount of data in form of code-mixed lexicons. Detecting these opinions poses several challenges, as listed below:

1) Code-switch: The process of alternating between different lexicons in the intrasentential text [5] or opinion renders opinion detection and analysis complex to implement.

2) Lack of standardized datasets: The lack of standardized datasets [6] for machine learning used for interpreting code-mixed data can lower the performance of machine learning models.

3) Unstructured format: Online users utilize non-standardized formats while conveying information [2, 7] that can include abbreviations, special characters, commas, slang, and lexical resources not included in standard machine learning libraries required for interpreting these opinions. This non-colloquial opinion usage makes opinion interpretation complex to implement.

4) Low resource lexicons: Certain opinion corpus categories have low resource availability [9], especially for the lexicon database of languages that are scarcely available.

5) Healthcare domain: Applications in Bioinformatics [8] include techniques for analyzing unstructured clinical data and extracting patient feedback. It can lead to significant insights, especially those that can be overlooked by clinical practitioners.

6) Dialects: Opinions cannot be distinguished from one dialect [10] to another within the same language.

Table I illustrates sample instances of code mixed opinions, translated opinions using API services, and actual interpretation of the Opinion. As observed, there is a variation in actual meaning through accurate interpretation compared to direct translation owing to lexical nuances of code mixed opinions. In order to overcome this challenge, we propose the design and implementation of a Code mixed opinion mining framework using an Adaptive Lemma-based weighted approach and machine learning for multi-class classification.

TABLE I. SAMPLE CODE MIXED OPINION CORPUS

Sr No	Code Mixed Opinion	Translated Opinion	Actual Opinion Interpretation
O1	Online parksha toh ache hain but yaar cheating ko rokna thoda sa mushkil ho jaata hai	Online parking are very good, but it is a bit difficult to stop cheating	Online exams are good friend, but stopping cheating becomes a bit difficult
O2	Ningal ivide thanne irikku	You stay here	Please remain here
O3	Woh kal se gym ja raha hai gr8	He is going to the gym from tomorrow gr8	He's starting gym tomorrow, great.
O4	Enikku oru sambhavam undu parayan.	I have an event to tell.	I have a story to tell you

C. Research Contributions

The major contributions of this research are highlighted below:

1) Design and implementation of novel weighted code mixed algorithm for feature engineering in opinion mining.

2) Efficient handling of code switching problems in code mixed data input.

3) Implementation of multiclass classification using an ensemble of machine learning and natural language processing techniques.

4) We conduct a rigorous performance analysis with state-of-the-art techniques on benchmark datasets to compare the proposed model efficacy for code mixed opinions.

The rest of this research is structured as follows: In Section II, we discuss the background of the literature. Section III describes the proposed solution architecture. The results of the performance analysis and the comparative analysis of related studies are compared in Section IV. Discussion and conclusion are given in Section V and Section VI respectively.

II. LITERATURE REVIEW

In this section, we review the relevant literature and existing opinion mining systems related to our study and research objective. The Intelligent Multi Lexicon Opinion Mining (IMLOM) implementation taxonomy is summarized below and categorized as machine learning, deep learning, and hybrid techniques.

A. Supervised IMLOM Learning

For supervised IMLOM, researchers have adopted conventional machine learning techniques, namely Support Vector Machine (SVM), Naïve Bayes (NB), and Logistic Regression [11], for the Algerian social platform. Other studies also use machine learning approaches. Multimedia network

studies use ML [12] for various domains, like the work done on the climate change crisis [13-14], which highlights the priority levels of different region users and their regional dialects.

B. Unsupervised IMLOM Learning

For unsupervised IMLOM studies, it is observed that authors use deep learning in various application domains. For instance, medical domain analysis for gaining perspective regarding vaccinations [15] being mandatory post-COVID through a survey questionnaire on a sample population selected.

In study [16], deep clustering with embedding was used for document vector recommendations. Deep learning techniques have been applied in further studies with promising results [17-19].

Other studies used Graph Neural Network (GNN) [20], Long Short Term Memory (LSTM) [21], and stacked layering of LSTM. LSTM and stacking mechanism are utilized for the deep learning phase, combining machine and manual translation. Researchers have also employed Convolution Neural Network (CNN) [22] with efficient model development and lingual error analysis.

C. Ensemble IMLOM Learning

Various approaches are adopted for hybridization with variations in preprocessing and classification combinations. In [23], authors use sentence-level embedding combined with the Transformer approach. In study [24], authors hybridize bio-inspired optimization algorithms with recurrent neural networks (RNN) for political corpus. Other popular combinations include CNN, LSTM, Bidirectional LSTM (BiLSTM) feature extraction, and transformer approach [25-29]. Both quantitative and qualitative analyses are conducted to synthesize data effectively. The authors in [30] use the augmentation technique to perform further the LDA-based topic modeling approach and Multilingual bidirectional and Auto-Regressive transformers (mBART), achieving a 71% accuracy level and highlighting the challenge of handling informal text.

Researchers in study [31] achieved significant results by employing the traditional J48 machine learning approach on self-developed data from the extracted contents, applying several linguistic features to analyze the model. Study in [32] implements deep learning methods, and dual datasets are used for cross-pair mapping. However, testing a model trained on one system in another domain yields lower accuracy using source-to-target language mapping. In study [33], the authors highlight the need for a training corpus for Dravidian language content classification into multiple classes. The authors in [34] demonstrated target entity extraction successfully through decisions based on neighboring information. However, the performance was limited by the capabilities of CNN. Alternatively, the shallow neural network [35] is for Spanish input and achieves an F1 score of 0.48. In study [36], the authors preprocess the data corpus and then generate the vectorized using the pre-trained model for emotion detection in online text. Efficiency was lower for heterogeneous content than for the standard English language. Studies in [32, 37 38, 41] utilize LSTM and combinations of other deep learning techniques, including CNN and BiLSTM. Total of five languages, with approximately. 42K social media opinions [39]. The various contexts of words are tackled using a stacked version of the word embedding implementation [40]. Also, a discussion of several research groups currently working on this topic of offensive text in multilingual settings is highlighted [41]. Researchers in [42] proposed a novel method using a combination of deep learning techniques in a stacked format and compared it to standardized machine learning methodologies, including SVM, KNN, and logistic regression. The authors in [43] use Artificial Intelligence for heterogeneous datasets, implementing machine translation using the BART method on Hungarian European content. The study in [44] classifies opinions, employing deep learning techniques to explore opinions towards access to education. We investigate the research papers, and after rigorous analysis of the paper title, abstract, methodology, and conclusion, the initial list is filtered to the most relevant studies.

It is imperative to implement efficient opinion mining frameworks for handling linguistic variability, code-switching, and unstructured format to leverage this vast digital information available, code mixed in the unstructured format, enabling accurate decision making.

III. Proposed Methodology

This section presents the benchmark datasets extracted, data statistics, and the proposed workflow.

A. Dataset Statistics

This section describes the datasets utilized for the experimental evaluation of the proposed work, as explained in Table II. Here, we employ two Benchmark datasets, Bohra et al. [45] and Joshi et al. [46], specifically curated for the problem statement identified as having Filtered data points 4575 and 3879, respectively. These benchmark datasets from the ACL data repository have been further validated through quantitative validation using the Cohens Kappa ITA score. The Bohra dataset has a Kappa score of 0.92, and the Joshi dataset score is 0.64, making it suitable for the proposed framework statistical and qualitative performance validation.

TABLE II. Dataset Statistics for Experimentation

Statistic	Bohra Dataset	Joshi Dataset
Total Data Points Collected	112718	4981
Filtered Data Points	4575	3879
Positive Class Data Points	1661 (Offensive Speech Opinions)	1358
Neutral Class Data Points	Not Available	1939
Negative Class Data Points	2914 (Normal Speech Opinions)	582
Inter Annotator Agreement (Cohen's Kappa)	0.982	0.64

B. Proposed Opinion Mining Framework

In this section, the design and implementation of the proposed technique WCM-OM are explained in detail. Through an extensive literature analysis, we present the comprehensive Intelligent Code Mixed Opinion Mining framework. The majority of the existing frameworks rely solely on a single language or medium, resulting in a significant loss of relevant information. The proposed framework implements a fusion ensemble of statistical features and linguistic features of code

mixed opinions from the input corpus database. Also, we implement data preprocessing for code-switched data to effectively handle the issue of frequent code-switching in an intra-sentential opinion. Fig. 2 illustrates the conceptual framework of the proposed framework. It consists of six significant modules: data acquisition, corpus extraction, data preprocessing for code-switched data, improved feature engineering module using Adaptive lemma-based vector space modeling, hybrid opinion model development, and opinion detection through multi-class classification.

The detailed components are as follows:

1) Data extraction and retrieval: At the initial stage, we retrieve online data content from heterogeneous web sources and unstructured data platforms from their respective programming APIs.

2) Opinion preprocessing: In the next phase, subtasks of NLP, including Stemming, Lemmatization, and irrelevant Noise Removal, are executed. Other subtasks may be required as per the application domain.

3) Lexicon identification: The step for detecting lexicon programmatically is crucial for efficient machine learning and deep learning model implementation. The translation and transliteration involve conversion from 1 lingual and structural format to another, enabling efficient execution of the detection phase given the heterogeneity of opinions.

4) Feature engineering: This phase comprises of feature extraction, transformation, and subsequent vectorization, is executed by employing NLP in several diversified opinion classes.

5) Evaluation: The WCM-OMF model is evaluated using standardized performance indicators depending upon subtasks for measuring the effectiveness of training and test evaluation.

The efficacy of the proposed solution is demonstrated through the comparative analysis of the proposed solution with the existing solution on benchmark dataset 1 (BD1) using standardized performance metrics. These steps are explained in detail in later sections. Further, the proposed weighted ensemble opinion mining framework (WCM-OMF) has been designed and implemented.

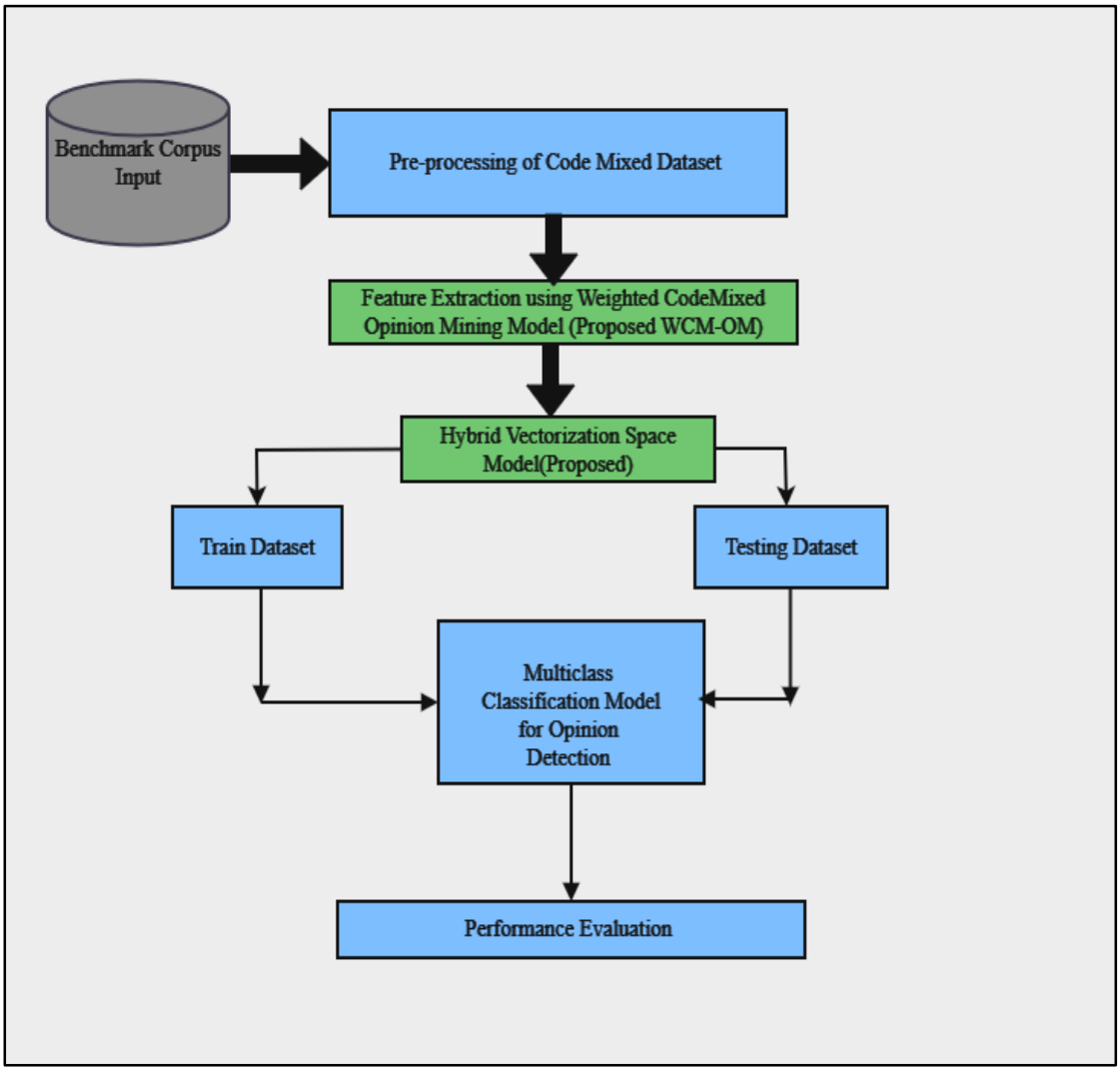

Fig. 2. Proposed opinion mining model conceptual framework.

The proposed solution assigns weights based on the following:

1) The Adaptive Bilingual Influence Weighting (ABIW)

2) The number of code-switches in the initial sentence of the opinion.

We observe that a comprehensive and holistic approach to opinion detection requires the consideration of various factors, and the current state of opinion interpretation requires further improvement consisting of the following key components:

- N: the number of tokens in a segment,
- $LangScore_i$: score indicating the relevance of the language of the i-th token,
- σ : normalization function.

Fig. 3 explains the stepwise analysis of the Proposed Solution Flowchart for the WCM-OM framework. In this architecture, the yellow highlighted cells represent the contributions of the proposed solution.

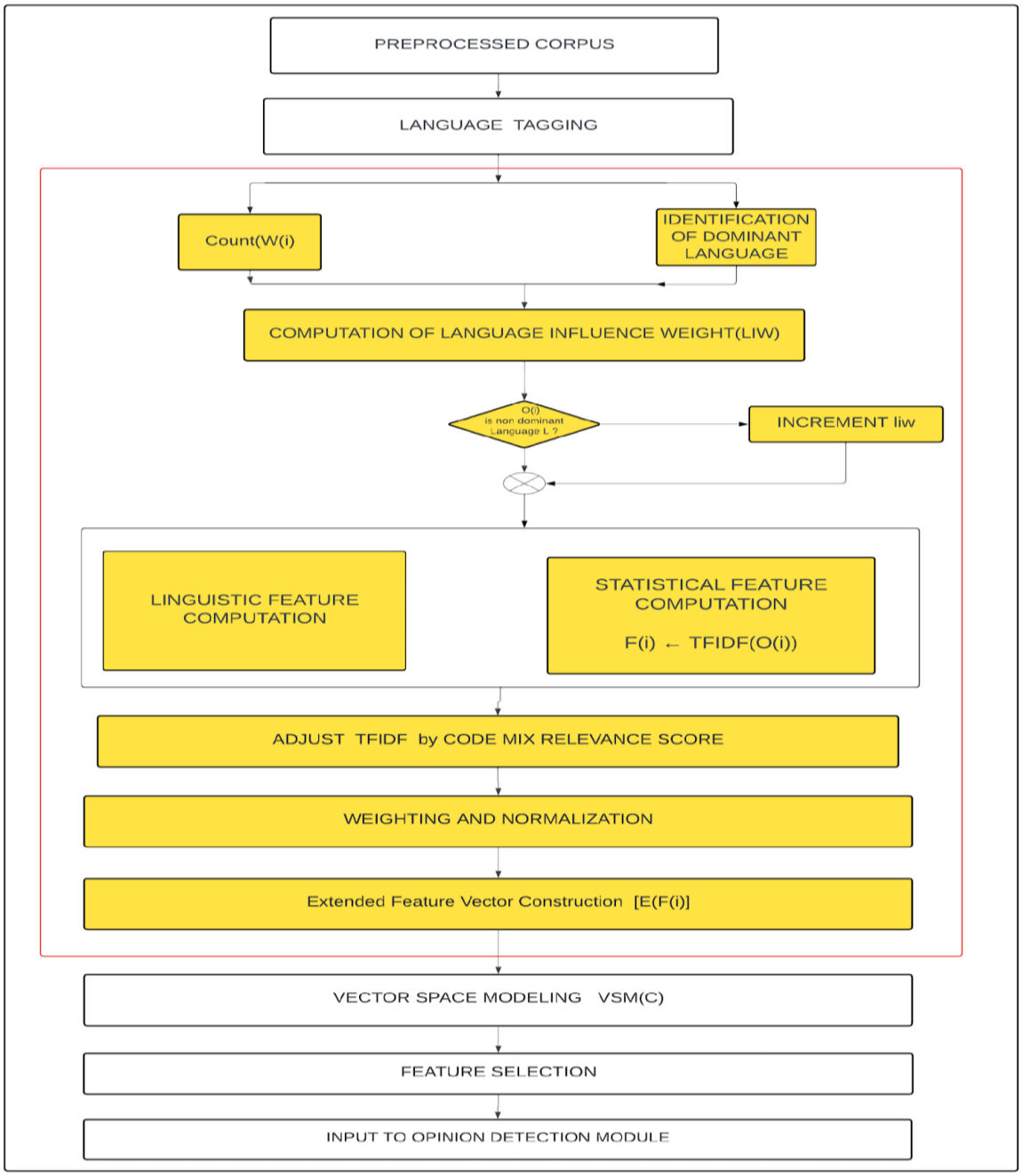

Fig. 3. Proposed WCM-OMF solution flowchart.

Firstly, the preprocessed output from the data extraction module functions as input for WCM-OMF. Thereafter, a language tagging module separates the opinion into parts of speech and tags each sub-block. Subsequently, the ABWI lemma is implemented using Language Influence Weighting (LIW) methods, a proposed metric designed to capture the influence of various lexical contexts in the code-mixed input. ABIW quantifies the extent to which the code-mixed input language contributes to the overall context. Following this, the linguistic feature is computed and combined with the statistical feature to obtain an EFV novel hybrid score. The code-mixed relevance score adjusts the TF-IDF score to get a new score. Weighting and normalization are applied to obtain the final feature vector EF.

C. Proposed Mining Method Mathematical Formulation

The equations and mathematical formulation are described in this section:

$$TF(t,\ o) = \frac{Number\ of\ times\ term\ t\ appears\ in\ a\ opinion\ o}{Total\ terms\ in\ \ opinion\ o} \quad (1)$$

$$IDF(t,\ O) = log\ \left(\frac{Overall\ opinion\ count}{Number\ of\ opinion\ with\ term\ t\ in\ it}\right) \quad (2)$$

$$\text{TF-IDF} = \text{TF (t, o)} * \text{IDF (t, O)} \quad (3)$$

Eq. (1) and Eq. (2) explain the computation of term frequency and inverse document frequencies Eq. (3) is the multiplication of term frequency, and the logarithm of opinion count to frequency ratio.

- EFV: Extended Feature Vector.
- TF-IDF: Term Frequency –Inverse Document Frequency.
- ABIW: Adaptive Bilingual Influence Weighting.

WCM_OMF is implemented to solve the code-switching problem in single opinion text (intra–sentential). For this purpose, lemma ABIW is devised as explained below.

Adaptive Bilingual Influence Weighting (ABIW)

The proposed lemma is explained in Eq. (4) that is used to dynamically adjust the influence of each language based on the opinion.

$$\text{ABIW} = \sigma\left(\frac{1}{N} * \Sigma(i = 1\ to\ N) LangScore_i\right) \quad (4)$$

Where

- W_{Lang} : weighting factor for each language in the bilingual code mixed text.
- $Model_m$: Trained model m.

Adaptive Bilingual Influence Weighting (ABIW), the language influence weighting method (LIW), is a metric designed to capture the influence of different languages in the code-mixed text. It quantifies the extent to which each language contributes to the overall content.

The weighting factor is computed for each opinion instance using a language identification function in combination with a weighting parameter. This result is added to the feature vector module, capturing the linguistic feature weights.

Further, we implement vector space modeling on the extended feature vector for the feature selection process. Eq. (5) computes the final feature vector, which is then used as modified feature vectors for multiclass opinion classification and subsequent opinion detection.

$$CMRS(O(i)) = \sum_{i=0}^{n} LF(O(i)) + STF(O(i)) \quad (5)$$

$CMRS$: Code Mixed Relevancy Score

$O(i)$: i th Opinion where i Є [1, 2, 3…n] and n is the Total count of opinions

STF –Statistical Features Set

LF- Linguistic Feature Set

Algorithm Weighted Code mixed Opinion Mining Framework (WCM_OMF)

Input Parameters:

D: Dataset with two columns,

T: Opinion text (T) and class (C),

Where T_i is the i^{th} text instance and C_i is the corresponding class label.

Algorithm : Weighted Code mixed Opinion Mining Framework(WCM_OMF)

Input: Corpus (C), Total number of samples (n).

Output : Optimized representation of Documents

Preprocessing

For each text T_i in D, preprocess the text to create a normalized text T_i'

$$T_i' \leftarrow preprocess_{text(T_i)}$$

Language Influence Weight (LIW):

Computation of LIW for each text T_i' using language id function L and weighting parameter α

$$LIW_i \leftarrow calculate_{\{liw\}(T_i', \alpha)}$$

Feature Engineering:

LIW module output conversion of T_i' into a feature vector F_i using TF-IDF

$F_i \leftarrow TFIDF(T_i')$

Combine feature vector F_i with LIW_i to create an extended feature vector EF_i

$$EF_i \leftarrow [F_i;\ LIW_i]$$

Dataset Preparation:

Class balancing module implementation ADASYN balancing on D_{train} to get D'_{train}

Training:

Train each model m in M on $D'_{\{train\}}$ using the optimized hyperparameters HP^*_m

$$Model_m \leftarrow Train(m, HP^*_m, D'_{train})$$

Development of ensemble model E from all trained models in M.

$$E \leftarrow Ensemble(M)$$

Prediction and Evaluation:

Predict classes for D_{test} using ensemble model E to get predictions P

$$P \leftarrow Predict(E, D_{test})$$

End

The proposed WCM_OMF Opinion Feature Engineering solution implements the computation of an Extended Feature Vector (EFV) instead of FV in existing systems for feature engineering opinion detection. Further, the resultant output is then inputted for embedding and classification, enabling a deeper linguistic context that assists in distinguishing opinions in different language segments.

IV Experimental Results

In this section, we present the experimental results of the proposed methodologies for code mixed data input.

Table III indicates that the proposed weighted code-mixed opinion analysis solution significantly improves opinion classification accuracy by incorporating both Hindi and English text features. The proposed solution employs a novel weighted strategy for code-switching. Quantitative and qualitative analyses are conducted for efficient data synthesis and validation. The proposed unique feature assigns higher weights for code-mixed lexicons during the opinion mining process, boosting the overall model performance. After using the k-fold cross-validation technique, the unique weighting technique's performance is further validated by performance metrics and compared to current frameworks.

Fig. 4 demonstrates the efficacy of the proposed solution for Benchmark Dataset 1 using Accuracy and Recall rate as evaluation metrics. The proposed model outperforms the existing systems SVM, Random Forest, sub-word LSTM, and Hierarchical LSTM with attention to phonetic words.

TABLE III. Comparative Analysis of Existing and Proposed Frameworks Existing Systems on Benchmark Dataset 1

Methodology	Accuracy	Recall	F1-score
Support Vector Machine Model	70.7	31.3	0.429
Random Forest Model [45]	65.1	19.3	0.292
Sub-word LSTM model [45]	69.8	36.5	0.458
Hierarchical LSTM model + attention on phonemic sub-words [47]	66.6	45.1	0.487
WCM-OMF (Proposed weighted code mixed approach)	**79**	**79.11**	**0.79102**

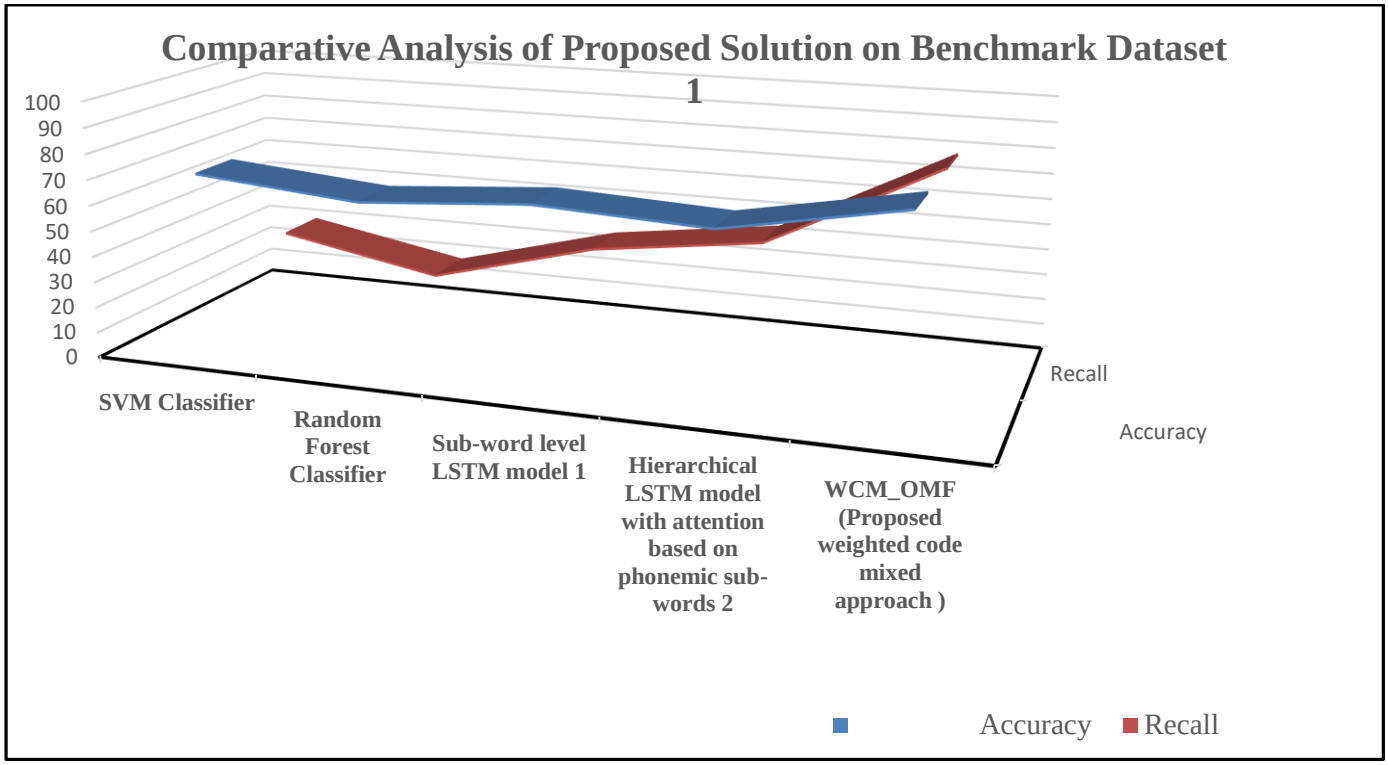

Fig. 4. Performance comparison of existing and proposed frameworks on benchmark Dataset 2.

TABLE IV. COMPARATIVE ANALYSIS OF EXISTING AND PROPOSED FRAMEWORKS ON BENCHMARK DATASET 2

Methodology	Authors	Accuracy	F1-score
BSVM (Unigram)	Wang, Manning[48]	59.15%	0.5335
NBSVM (Uni + Bigram)	Wang, Manning[48]	62.50%	0.5375
MNB (Unigram)	Wang, Manning[48]	66.75%	0.6143
MNB (Uni + Bigram)	Wang, Manning[48]	66.36%	0.6046
MNB (TFIDF)	Wang, Manning[48]	53.53%	0.4783
SVM (Unigram)	Pang and Lee[49]	57.60%	0.5232
SVM (Uni+Bigram)	Pang and Lee[49]	52.96%	0.3773
Lexicon Lookup	Sharma et al.[50]	51.15%	0.252
LSTM Model 1	Joshi et al[46]	59.80%	0.511
LSTM Model 2	Joshi et al[46]	59.70%	0.658
WCM_OMF (Proposed weighted code mixed approach)	**Proposed**	**72.55%**	**0.7208**

Results demonstrate that the proposed weighted approach to handle code-switching gives significantly better results than existing opinion mining frameworks, achieving an accuracy of 79% using comparative analysis evaluated on benchmark dataset 1. Table IV results show that the proposed weighted approach for handling code-switching performs significantly better than the existing opinion mining frameworks.

The overall accuracy achieved for the proposed system is 72.55%, and the F1 score is 0.72. The proposed method considers relevant factors during the process of grouping language-based text.

V. DISCUSSION

In this section, we discuss the results of the proposed solutions and the lemma developed. As observed in Fig. 5, the proposed solution WCM-OMF is designed and implemented to assign weights based on the ABIW lemma and the number of code switches in intra-sentential opinion. The steps for WCM (Weighted code mixing metric) are explained in detail and are primarily implemented to handle the unstructured nature of opinion input. Results show that the proposed WCM-OMF performs significantly better, achieving an F1 score of 0.72, which is a 6% improvement over the existing method on the benchmark dataset.

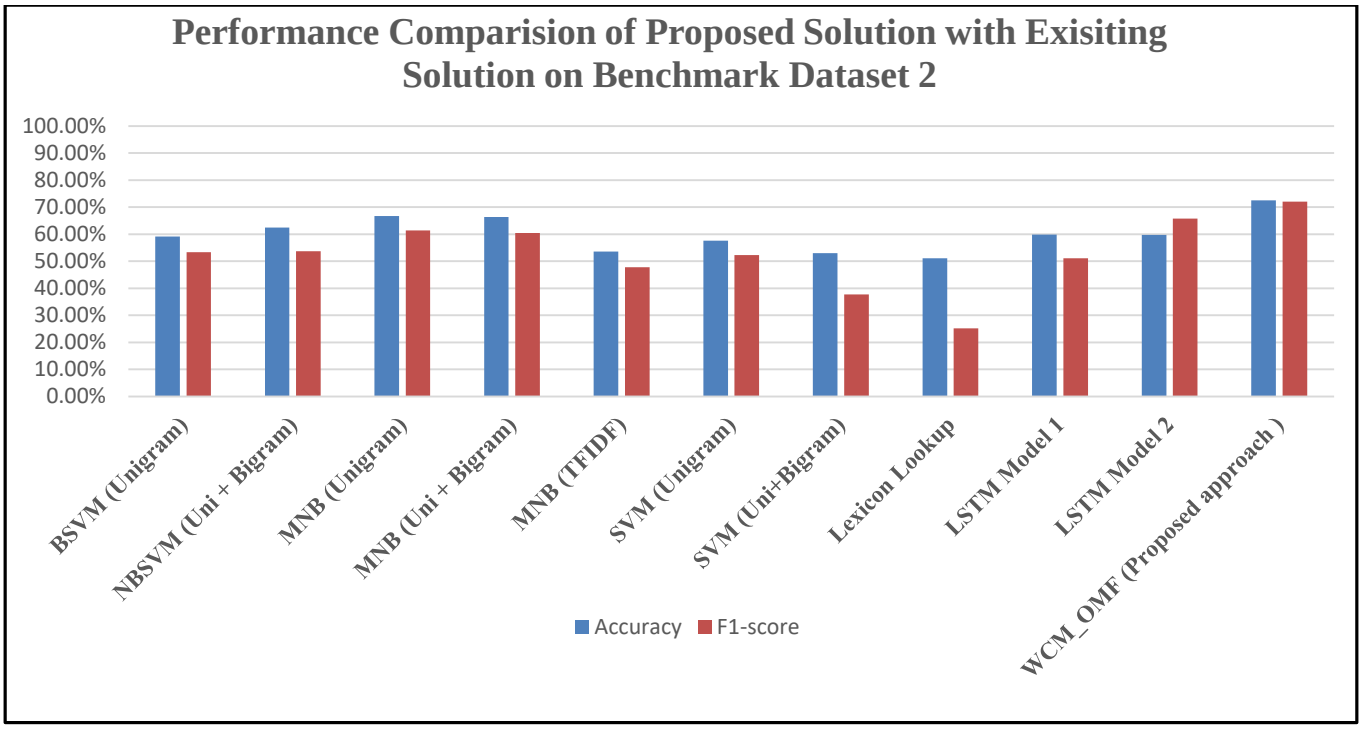

Fig. 5. Performance comparison of the proposed solution on benchmark Dataset 2.

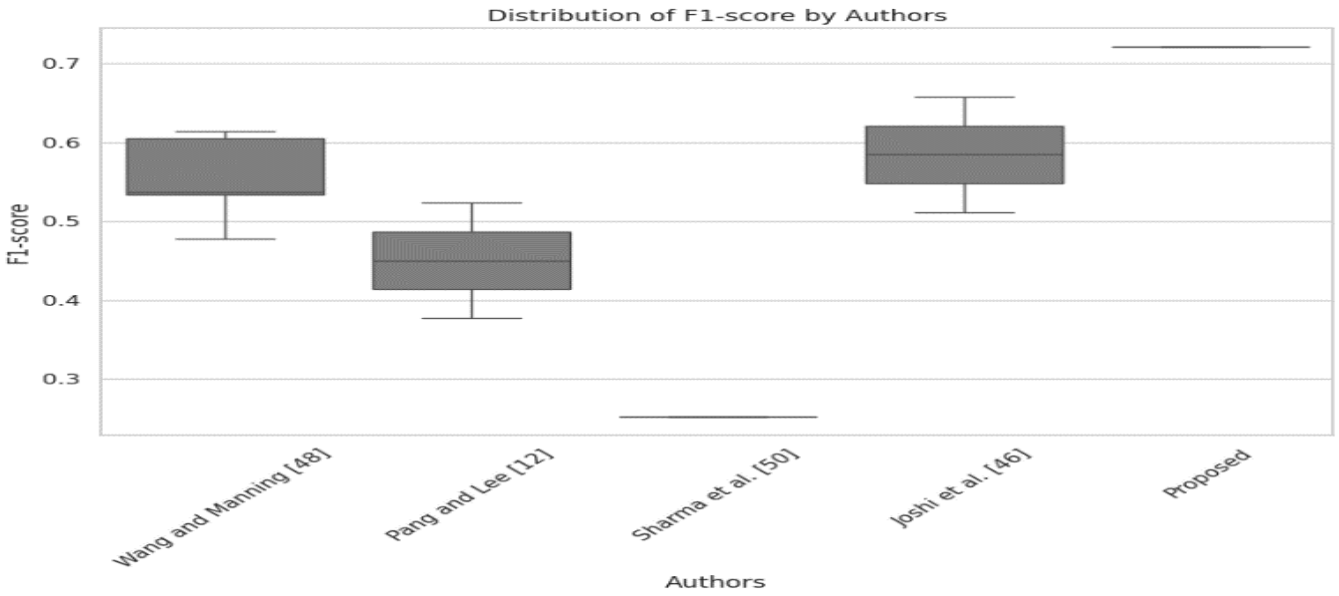

Fig. 6. Box Plot representation of F1 score distribution of proposed and existing systems.

As demonstrated in Fig. 6, the box plot represents the accuracy and F1 score distribution across several existing state-of-the-art models. As observed, the proposed model demonstrates a superior performance, as evidenced by the grouping of median data points. Opinion mining helps in the detection of patterns, improves decisions, and customizes strategy-based data extracted [51], leading to Business Intelligence [52].

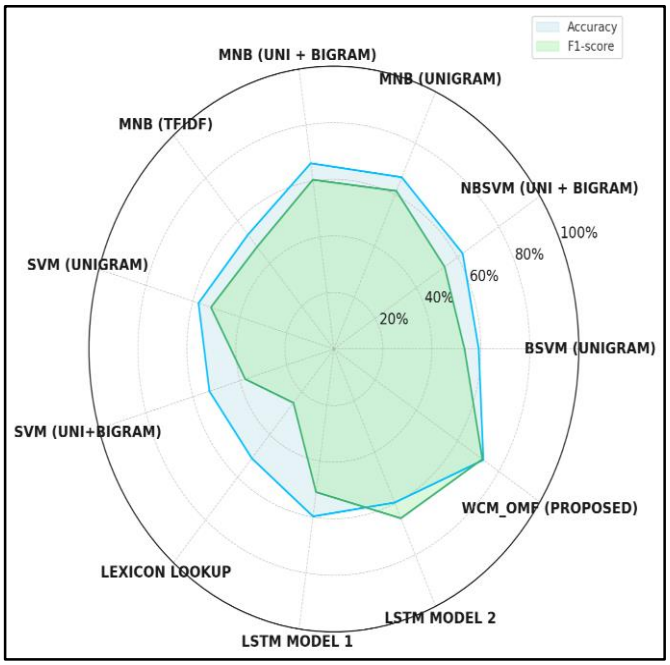

Fig. 7. Radar chart of accuracy and F1 score.

Fig. 7 visualizes the radar chart of accuracy and F1 score that demonstrates the proposed model efficiency across multiple metrics compared to the existing MNB BSVM-Unigram and BSVM- Bigram approaches.

We further validate the proposed results in Table V by implementing multiple Comparison of Means - Tukey HSD (Honestly Significant Difference) test to identify significant differences between pairs of group means with 95% confidence and p-value less than 0.05 mean difference of 9.86 for confidence interval [9.548, 10.171] and 13.21 for confidence interval [12.898, 13.521] indicating proposed method outperforms the existing approaches and differences are statistically significant.

TABLE V. TUKEYS HSD STATISTICAL SIGNIFICANCE TEST

Group 1	Group 2	Mean Diff	Lower	Upper
BSVM Unigram	MNB Unigram	7.41	7.0986	7.7214
BSVM Unigram	NBSVM Uni+Bigram	3.35	3.0386	3.6614
BSVM Unigram	Proposed	13.21	12.8986	13.5214
MNB Unigram	NBSVM Uni+Bigram	-4.06	-4.3714	-3.7486
MNB Unigram	Proposed	5.8	5.4886	6.1114
NBSVM Uni+Bigram	Proposed	9.86	9.5486	10.1714

VI. CONCLUSION

In this paper, we propose a novel WCM-OMF ensemble opinion mining framework in the context of code mixed opinions and unstructured format. We develop an end-to-end framework to handle the code-switching problem. This study also aims to bridge the gap between corpus availability and opinion mining research that would benefit researchers in analysing the proposed corpus for code mixed opinions. The results of our proposed framework are more efficient than related existing approaches in the case of code mixed opinions. The proposed method achieves an F1 score of 0.79, achieving a 6% increase in the opinion mining system. Further, we intend to extend this research to implicit opinions and word sense

disambiguation techniques concentrating on irony and other related clauses present in the opinion. This research significantly impacts creating an explainable digital environment that enhances decision-making.

Further, based on the rigorous analysis of scholarly research, we present a comprehensive framework for implementing cross-lexical opinions using cross-platform resources. Further, we suggest future developmental efforts, namely:

- Low resource corpus availability.
- Multimodal data integration capabilities.
- Open source platforms for handling big data analytics.
- Optimizing resource consumption for transformer architectures and deep learning methodologies.

The proposed solution would enable the utilization of potential applications for enterprises having an online presence for effective decision-making and business intelligence.

ACKNOWLEDGMENT

We are grateful for the insightful comments given by anonymous reviewers during the drafting of the paper.

REFERENCES

[1] Pang B, Lee L et al (2008) Opinion mining and sentiment analysis. Found Trends Inf Retr 2(1–2):1–135.

[2] Tong RM (2001) An operational system for detecting and tracking opinions in on-line discussion. In: working notes of the ACM SIGIR 2001 workshop on operational text classification, vol. 1

[3] P. Liang and B. Dai, "Opinion mining on social media data", 14th International Conference on Mobile Data Management, pp. 91-96, 03-06 June 2013, https://doi.org/10.1109/MDM.2013.73.,IEEE

[4] C.C. Aggarwal and C.X. Zhai (eds.), Mining Text Data, doi: 10.1007/978-1-4614-3223-4_13, Springer Science Business Media, LLC, 2012, Springer

[5] C. M. Scotton, "The possibility of code-switching: motivation for maintaining multilingualism," Anthropological linguistics, pp. 432–444, 1982

[6] Zhang, Chenxi, and Zeshui Xu. "Gaining Insights for Service Improvement through Unstructured Text from Online Reviews." Journal of Retailing and Consumer Services 80, (2024): 103898. https://doi.org/10.1016/j.jretconser.2024.103898

[7] H. Liu, I. Chatterjee, M. Zhou, X. S. Lu and A. Abusorrah, "Aspect-based sentiment analysis: A survey of deep learning methods", *IEEE Trans. Comput. Soc. Syst.*, vol. 7, no. 6, pp. 1358-1375, Dec. 2020

[8] Medhat Walaa, Ahmed Hassan, and Hoda Korashy. "Sentiment Analysis Algorithms and Applications: A Survey." Ain Shams Engineering Journal 5, no. 4 (2014): 1093-1113. https://doi.org/10.1016/j.asej.2014.04.011

[9] Andrés Montoyo, Patricio Martínez-Barco, Alexandra Balahur, Subjectivity and sentiment analysis: an overview of the current state of the area and envisaged developments,Decis Support Syst, 53 (2012), pp. 675-679

[10] Date, Saroj S., Mahesh B. Shelke, Kiran V. Sonkamble, and Sachin N. Deshmukh. "A Systematic Survey on Text-based Dimensional Sentiment Analysis: Advancements, Challenges, and Future Directions." Computational Intelligence Methods for Sentiment Analysis in Natural Language Processing Applications, (2023): 39-57. https://doi.org/10.1016/B978-0-443-22009-8.00014-8.

[11] Yoon, Joanne. 'Classifying Respondent Comments from the 2021 Canadian Census of Population Using Machine Learning Methods [1]'. 1 Jan. 2023: 785 – 791.

[12] Laifa, M. and Mohdeb, D. (2023), "Sentiment analysis of the Algerian social movement inception", Data Technologies and Applications, Vol. 57 No. 5, pp. 734-755. https://doi.org/10.1108/DTA-10-2022-0406

[13] Sietsma, A.J., Groenendijk, R.W. & Biesbroek, R. Progress on climate action: a multilingual machine learning analysis of the global stocktake. Climatic Change 176, 173 (2023). https://doi.org/10.1007/s10584-023-03649-3

[14] Kohei Watanabe (2021) Latent Semantic Scaling: A Semisupervised Text Analysis Technique for New Domains and Languages, Communication Methods and Measures, 15:2, 81-102

[15] Regazzi, L., Cadeddu, C., Gris, A. V., Lomazzi, M. (2023). Sentiment towards vaccinations in Chinese healthcare workers: preliminary results from an international survey. Population Medicine, 5(Supplement), A2051. https://doi.org/10.18332/popmed/165050

[16] Schrader, Samuel R., and Eren Gultepe. "Analyzing Indo-European Language Similarities Using Document Vectors." Informatics 10, no. 4 (2023): 76. https://doi.org/10.3390/informatics10040076.

[17] X. Zhang, J. Zhao, Y. LeCun, Character-level convolutional networks for text classification, Adv. Neural Inf. Process. Syst. 28 (2015).

[18] J. Du, L. Gui, R. Xu, Y. He, A Convolutional Attention Model for Text Classification," Natural Language Processing and Chinese Computing, 2018, pp. 183–195, https://doi.org/10.1007/978-3-319-73618-1_16

[19] K. Kumari, S.S. Jha, Z.K. Dayanand, and P. Sharma, "ML&AI IIITRanchi@DravidianLangTech: Fine-Tuning of Indic-BERT for Exploring Language-Specific Features for Sentiment Classification in Code-Mixed Dravidian Language," DravidianLangTech 2023 - 3rd Workshop on Speech and Language Technologies for Dravidian Languages, associated with 14th International Conference on Recent Advances in Natural Language Processing, RANLP 2023 - Proceedings, pp. 192-197, 2023, doi: 10.26615/978-954-452-085-4_027.

[20] K. Maity, T. Sen, S. Saha, and P. Bhattacharyya, "MTBullyGNN: A Graph Neural Network-Based Multitask Framework for Cyberbullying Detection," IEEE Transactions on Computational Social Systems, vol. 11, no. 1, pp. 849-858, 2024, doi: 10.1109/TCSS.2022.3230974.

[21] R. Pandey and J. P. Singh, "BERT-LSTM model for sarcasm detection in code-mixed social media post," Journal of Intelligent Information Systems, vol. 60, no. 1, pp. 235-254, 2023, doi: 10.1007/s10844-022-00755-z.

[22] R.R. Frias, R.P. Medina, and A.S. Sison, "Attention-based Bilateral LSTM-CNN for the Sentiment Analysis of Code-mixed Filipino-English Social Media Texts," 2023 International Conference on Digital Applications, Transformation and Economy, ICDATE 2023, doi: 10.1109/ICDATE58146.2023.10248926.

[23] Nascimento, Francimaria R., George D. Cavalcanti, and Márjory Da Costa-Abreu. "Unintended Bias Evaluation: An Analysis of Hate Speech Detection and Gender Bias Mitigation on Social Media Using Ensemble Learning." Expert Systems with Applications 201, (2022): 117032. https://doi.org/10.1016/j.eswa.2022.117032.

[24] Manikyamba, I. Lakshmi, and A. Krishna Mohan. "Using a Novel Hybrid Krill Herd and Bat based Recurrent Replica to Estimate the Sentiment Values of Twitter based Political Data." (2023).

[25] Kar, Purbani, and Swapan Debbarma. "Sentimental Analysis & Hate Speech Detection on English and German Text Collected from Social Media Platforms Using Optimal Feature Extraction and Hybrid Diagonal Gated Recurrent Neural Network." Engineering Applications of Artificial Intelligence 126, (2023): 107143. . https://doi.org/10.1016/j.engappai.2023.107143.

[26] M. Alhawarat and A. O. Aseeri, "A Superior Arabic Text Categorization Deep Model (SATCDM)," in IEEE Access, vol. 8, pp. 24653-24661, 2020, doi: 10.1109/ACCESS.2020.2970504.

[27] Md Ashik, A. U.-Z, S. Shovon, S. Haque, Data set for sentiment analysis on Bengali news comments and its baseline evaluation, 2019 International Conference on Bangla Speech and Language Processing (ICBSLP) (2019)

[28] Md. Arid Hasan, "Ensemble Language Models for Multilingual Sentiment Analysis," University of New Brunswick, 2024.

[29] I. Javed and H. Afzal, "Opinion Analysis of Bi-Lingual Event Data from Social Networks," in Proceedings of the Conference on Language Resources and Evaluation, pp. 3172-3179, 2014.

[30] Hashmi, E., Yayilgan, S.Y. & Shaikh, S. Augmenting sentiment prediction capabilities for code-mixed tweets with multilingual transformers. Soc. Netw. Anal. Min. 14, 86 (2024). https://doi.org/10.1007/s13278-024-01245-6

[31] Altaf A, Anwar MW, Jamal MH, et al (2023) Exploiting linguistic features for efective sentence-level sentiment analysis in urdu language. Multimedia Tools and Applications pp 1–27

[32] Ghasemi, R., Ashrafi Asli, S. A., & Momtazi, S. (2022). Deep Persian sentiment analysis: Cross-lingual training for low-resource languages. Journal of Information Science, 48(4), 449-462. https://doi.org/10.1177/0165551520962781

[33] S. Thara and P. Poornachandran, ‘‘Social media text analytics of Malayalam–English code-mixed using deep learning,’’ J. Big Data, vol. 9, no. 1, pp. 1–25, Dec. 2022.

[34] Soufian Jebbara and Philipp Cimiano. 2019. Zero-shot cross-lingual opinion target extraction. 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). Association for Computational Linguistics, 2486–2495. DOI:https: //doi.org/10.18653/v1/N19-1257

[35] Ignacio González Godino and Luis Fernando DHaro.Gth-upm at tass: Sentiment analysis of tweets for spanish variants. Proceedings of TASS, 2019

[36] Michael Neumann and N. goc Thang Vu. 2018. Cross-lingual and multilingual speech emotion recognition on English and French. In Proceedings of the 2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). IEEE, 5769–5773. DOI:https://doi.org/10.1109/ICASSP.2018.8462162

[37] Sujata Rani and Parteek Kumar. Deep learning based sentiment analysis using convolution neural network. Arabian Journal for Science and Engineering, 44(4):3305–3314, 2019.

[38] Santosh T. and K. Aravind (2019). Hate speech detection in hindi-english code-mixed social media text. In Proceedings of the ACM India Joint International Conference on Data Science and Management of Data, pp. 310–313

[39] Saumya, Sunil, Abhinav Kumar, and Jyoti P. Singh. "Filtering Offensive Language from Multilingual Social Media Contents: A Deep Learning Approach." Engineering Applications of Artificial Intelligence 133, (2024): 108159. https://doi.org/10.1016/j.engappai.2024.108159.

[40] Tasnia, R., Ayman, N., Sultana, A. et al. Exploiting stacked embeddings with LSTM for multilingual humor and irony detection. Soc. Netw. Anal. Min. 13, 43 (2023). https://doi.org/10.1007/s13278-023-01049-0

[41] Mandl , T., Modha, S., Kumar M. A., & Chakravarthi, B. R. (2020). Overview of the HASOC track at fire 2020: Hate speech and offensive language identification in Tamil, Malayalam, Hindi, English and German. In Proceedings of the 12th annual meeting of the forum for information retrieval evaluation (pp. 29–32).

[42] Khan L, Amjad A, Afaq KM et al (2022) Deep sentiment analysis using cnn-lstm architecture of English and roman urdu text shared in social media. Appl Sci 12(5):2694

[43] Yang, Zijian Győző, Laki, László János (2023) Solving Hungarian natural language processing tasks with multilingual generative models Annales Mathematicae et Informaticae. 57. pp. 92-106. ISSN 1787-6117 (Online)

[44] Lany L. Maceda, Arlene A. Satuito and Mideth B. Abisado, “Sentiment Analysis of Code-mixed Social Media Data on Philippine UAQTE using Fine-tuned mBERT Model” International Journal of Advanced Computer Science and Applications(IJACSA), 14(7), 2023. http://dx.doi.org/10.14569/IJACSA.2023.0140777

[45] Bohra, A., Vijay, D., Singh, V., Akhtar, S.S., & Shrivastava, M.,” A dataset of Hindi–English code-mixed social media text for hate speech detection” In second workshop on computational modeling of people’s opinions, personality, and emotions in social media, Association for Computational Linguistics (ACL)

[46] Joshi, A., Prabhu, A., Shrivastava, M., and Varma, V. Towards sub-word level compositions for sentiment analysis of Hindi-English code mixed text, COLING, pages 2482–2491, ACL Anthology,2016

[47] Santosh, T.Y.S.S., Aravind, K.V.S.: Hate speech detection in Hindi-English code-mixed social media text. ACM Int. Conf. Proc. Ser. (2019)

[48] Sida Wang and Christopher D. Manning. 2012. Baselines and bigrams: Simple, good sentiment and topic classification. In Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Short Papers - Volume 2, ACL ’12, pages 90–94, Stroudsburg, PA, USA. Association for Computational Linguistics, ACM

[49] Bo Pang and Lillian Lee. 2008. Opinion mining and sentiment analysis. Found. Trends Inf. Retr., 2(1-2):1–135

[50] Shashank Sharma, PYKL Srinivas, and Rakesh Chandra Balabantaray, Text normalization of code mix and sentiment analysis. In 2015 International Conference on Advances in Computing, Communications and Informatics, ICACCI 2015, August 10-13, 2015, pages 1468–1473.

[51] S. Ruchi and S. Pravin, “Improved opinion mining for unstructured data using machine learning enabling business intelligence,” Journal of Advances in Information Technology, vol. 14, no. 1, 2023.

[52] R. Sharma and P. Srinath, "Business Intelligence using Machine Learning and Data Mining techniques - An analysis," 2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA), Coimbatore, India, 2018, pp. 1473-1478, doi: 10.1109/ICECA.2018.8474847