JETIR.ORG ISSN: 2349-5162 | ESTD Year : 2014 | Monthly Issue

JOURNAL OF EMERGING TECHNOLOGIES AND INNOVATIVE RESEARCH (JETIR)

An International Scholarly Open Access, Peer-reviewed, Refereed Journal

Hybrid Machine Learning Strategies for Sentiment Analysis in Mixed-Language Data

[1]Piyush Rai, [2]Dr. Shobhit Sinha

[1]PhD Scholar, [2]Associate Professor

[1,2]Department of Computer Science & Engg., SRMU, Barabanki, UP, India

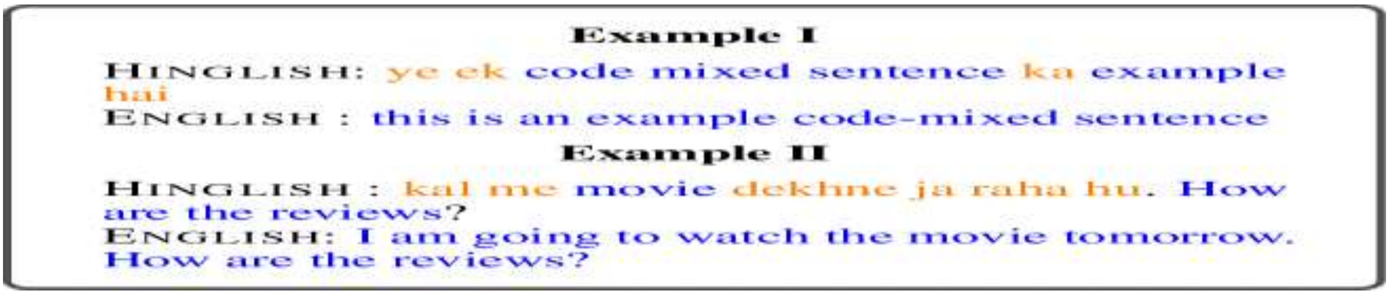
Abstract :

Sentiment analysis, crucial task of Natural Language Processing (NLP), becomes more tedious in the presence of code-mixed slang, where English blends with internet shorthand, colloquial expressions. Traditional monolingual NLP techniques are often inadequate for handling such irregular and informally structured text. The fundamental purpose of sentiment analysis, which is an important part of natural language processing (NLP), is to find and remove the emotional tone of text data. Sentiment analysis now faces additional difficulties as code-mixed languages, like Hinglish, a combination of Hindi and English, become more comm on. Conventional methods mainly handle monolingual data, which leaves code-mixed scenarios unexplored. A comprehensive approach to sentiment analysis on Hinglish is presented in this research article, which addresses problems such as inconsistent transliteration, a lack of standardized grammar, and the dearth of annotated datasets. We illustrate the potential of our method to efficiently analyze sentiments in code-mixed languages by building a solid dataset and utilizing cutting-edge machine learning and deep learning models. Our results make a substantial contribution to the developing field of multilingual natural language processing. Social media, an omnipresent web-based platform, has become a primary forum for discussion and expression, leading to the evolution of a "pseudo-language" in multilingual regions like India. This new linguistic phenomenon often involves code- mixing, where speakers seamlessly interchange languages within utterances, posing significant challenges for Natural Language Processing (NLP) research. This paper focuses on developing methods for sentiment analysis of such code-mixed social media text, specifically involving Indian languages. Sentiment analysis is a well-established area of natural language processing (NLP), but it is hard to do with Indian language texts that are informal and blend code.

This work addresses important issues like inconsistent orthography, non-standard grammar, unpredictable abbreviations, and a lack of annotated resources by presenting a thorough framework for sentiment analysis on slang-rich code-mixed data. In order to capture the linguistic variability typical of code-mixed online communication, we create three new corpora from Twitter and Facebook. In order to achieve sentiment classification, the suggested pipeline combines code-mix complexity analysis, sentence boundary detection, word-level language identification, and part-of-speech tagging. Experiments show that deep learning models perform significantly better than traditional machine learning techniques, especially BiLSTM architectures and transformer-based systems like BERT. The results contribute to the wider development of multilingual and sociolinguistic NLP by demonstrating the efficacy of sophisticated neural models for sentiment processing in highly informal, linguistically hybrid social media text.

IndexTerms - **Sentiment Analysis, Hinglish, Codemixed, BERT, Machine Learning.**

I. Introduction



Example I

HINGLISH: ye ek code mixed sentence ka example hai

ENGLISH : this is an example code-mixed sentence

Example II

HINGLISH : kal me movie dekhne ja raha hu. How are the reviews?

ENGLISH: I am going to watch the movie tomorrow. How are the reviews?

Each word is annotated with the English and Hindi language tag in the above sentence. A language of this kind of mix is popularly known as Hinglish. Code-mixing, also referred to as languages mixing, represents a linguistic phenomenon frequently observed in bilingual or multilingual societies. This phenomenon occurs more frequently in informal communications, which contributes to its popularity in social media platforms such as Twitter, Facebook, and WhatsApp. Code-mixing is the phenomenon where elements such as morphemes, words, phrases, clauses, and sentences from one language are integrated into another language.

When code alternation or switching takes place within an utterance and below the clause level, it is typically termed code-mixing. In contrast, code-switching is a broader concept that usually pertains to inter-clausal code alternation. The speaker is typically shaped by a range of social factors, such as the participants in the discussion, the social context, the social status of those involved, the topics being discussed, the level of formality in language, and the intended purposes and functional uses of language. As language captures cultural footprints, every language also has some culture-specific words which have a particular significance that is not possible to interpret for what it's worth in other languages. Such words play an instrumental role in communication in a specific language. A multilingual person generally uses code-mixed words as there are no such equivalent words in the other language. So, code-mixing really helps in communication by letting the speaker share their thoughts while talking to a multilingual audience. It seems like a lot of the research on social media texts has focused mainly on English. But nowadays, most of these texts are actually in other languages. So, it looks like the Twitter language map shows that Europe and South-East Asia are the most diverse when it comes to languages, especially among the regions that are really active on Twitter right now.

Statistically, multilingual countries like the US, India , Brazil and Indonesia have the highest Facebook audience. So, in the creation of code-mixed data, social media users in multilingual countries play an indispensable role. It's no secret that while English reigns supreme as the go-to language for online communication, there's an undeniable and ever-increasing demand to create technologies that cater to other languages. The world is changing, and so must we!

Therefore, to achieve global scale social media analysis, it is essential to solving the problem of code-mixing language processing. Despite having the code-mixed as an additional challenge, Natural Language Processing (NLP) for social media text has many more other problems to solve – such as having a high percentage of spelling errors and containing phenomena such as creative spellings, word play, abbreviations, meta tags and non-standard phonetic typing. Moreover, technological and cultural advancements give rise to new terminology that essentially represents extra meanings or nuances of existing words, such as "texto" for "SMS error" (compare with "typo" for "spelling error"). The phrase "social media text" shall be employed in this article to denote the manner of communication of these texts, while recognizing that social media does not represent a distinct textual domain. Conversely, a wide array of distinct text kinds is conveyed in this manner, as elaborated by Eisenstein [66] and Androutsopoulos [13]. Both contend that the fundamental characteristic of social media text is not its inherent "noise" or informality, but rather its representation of language in (rapid) evolution, which has substantial ramifications for natural language processing: if researchers develop a system capable of managing a particular variant of social media text today, it will become obsolete by tomorrow. Factors that render the application of machine learning and adaptive approaches to the problem particularly attractive. Given that code-mixing exemplifies a rapidly evolving linguistic phenomenon, it necessitated the initiation of essential NLP research from the outset. Therefore, this thesis started its endeavour with the study of the complex nature of the code-mixed corpora. After that, the research mainly focused on solving fundamental NLP problems of code-mixing, such as - sentence boundary detection; word-level language identification; and parts-of- speech tagging of code-mixed corpora. Finally, an attempt is made to apply such modules towards solving a real-life problem like sentiment analysis techniques code-mixed text. The term code-mixing consists of two individual words i.e. code and mixing. The term code is often surrounded by a whirlwind of confusion, with its meanings dancing across various fields like communication, semiotics, programming, and cryptography, each adding its own flavor to the mix. In the fascinating world of computational linguistics, the term code encompasses not just a myriad of languages, but also the delightful varieties and unique styles that exist within a single language. It's a vibrant tapestry of communication that showcases the richness of human expression!

. The term code denotes a comparatively impartial understanding of linguistic variety, encompassing both speech and dialect. All languages and writing systems can be regarded as 'codes' for anthropological cognition. The term code-mixing, when analyzed linguistically, denotes the amalgamation of distinct lexicons to create a unified entity. Consequently, the phrase code-mixing refers to the amalgamation of two or more languages or language variants in spoken or written discourse. The emergence of social media has allowed users to converse informally, which has contributed towards the development of a new lingual signature that has evolved organically over the times. Such mutations do not only include syntactic/grammatical violations, but rather it also allows to conjoin several languages together to address the multi-lingual society. Although code-mixing is not a new phenomena, it was always present in spoken language conversations, but social media allows it to become a common practice. Therefore, code-mixing has emerged as the new research area. So various NLP scientists, therefore, develop new technologies and techniques to process code-mixed speech and text, to create useful tools for translation and speech recognition and also to construct engaging user interfaces. The goal is to make natural and multilingual interaction in-between computer and human. In every natural language processing or text processing research corpora plays an important role to build NLP system. In general, finding monolingual corpora is more easier than the code- mixed corpora, because monolingual discourse is more prevalent in formal environments and thus is more likely to be preserved. Although the changing of code is a common practice in different communities worldwide, access to code-mixed data is often difficult. Thus, code-mixed data is less likely to be archived and therefore harder to find as training data. Linguistic complexity as well as course nature of the code-mixed text makes the annotation process ambiguous, which makes challenges in different NLP classification problem such as language identification, Parts-of-Speech (POS) tagging etc. So in this research, motivation to discuss the necessary theoretical study on code-mixing including a Code Mixing Index (CMI) which determines the level of mixing between languages and finally, choose sentiment analysis as a test case to solve which is basically a multi-faceted problem that includes several sub- problems such as Sentence Boundary Detection (SBD), POS tagging, language identification. Sentiment analysis, otherwise known as opinion mining or text analysis which refers to the automated process of analysing the textual data to extract, identify, or otherwise characterize the sentiment content of a text unit using NLP, computational linguistics or statistical approach. The fundamental task is to identify the polarity or emotions or intentions of a given text in the document level, sentence or sub-sentence or clause level, feature or aspect level. The type of polarity based on positive, negative, or neutral. Sentiment analysis from social media code-mixed text is extremely important in toady's scenario, as this is the era of social media which demands continuous monitoring to get the wider public opinion or sentiments behind the certain specific topics. Despite several recent advances in NLP, the handling of code-mixed data for sentiment analysis remains a challenge. In the fast-paced world of the last ten years, social media platforms like Facebook, Twitter, and WhatsApp have opened up a treasure trove of opportunities for accessing information and advancing language technology. However, this digital revolution has also brought its fair share of hurdles. The text generated on these platforms tends to be rough around the edges and riddled with noise, showcasing a plethora of spelling blunders and inventive spellings—like using "u" instead of "you" or "b4" for "before."

Phonetic typing adds another layer of complexity, as seen with the Hindi word for "eyes," which can be represented in various creative forms. Word play is rampant, with variations like "goooood" for "good" and "soooo" for "so." Abbreviations are the name of the game, with "LOL" standing for "Laugh out loud!" and "OMG" for "Oh My God!" The blending of words, such as "alot" for "A lot" and "ty" for "Thank You," further complicates matters. Confusion reigns with homonyms, as "Principal/Their" can easily be mistaken for "Principle/There." And let's not forget the presence of meta tags, including URLs and hashtags, that add to the mix. In today's fast-paced world, it's no surprise that technological advancements give rise to a plethora of new terms that breathe fresh life into old concepts. Take, for instance, the term "texto," which has emerged to describe a "SMS error," much like how "typo" has become synonymous with a "spelling error." It's a classic case of language evolving to keep up with the times! In the realm of social media, it's a common sight that non-English speakers, including our friends from India, often steer clear of using Unicode or their native scripts. Instead, they embrace the art of romanized phonetic typing, seamlessly weaving in English elements through code-mixing and anglicisms. This delightful blend of languages creates a vibrant tapestry of expression, but it also turns the task of identifying parts of speech and language in social media text into quite the formidable challenge! In the vibrant world of Indian social media, there are countless writing practices that truly stand out and capture the essence of the digital age. It's a melting pot of creativity and expression, where every post is a unique gem waiting to shine. From witty captions to heartfelt messages, the landscape is rich with diverse voices and perspectives that resonate with audiences far and wide. Each interaction is a testament to the power of connection in this fast-paced online realm.

II. Literature Review

Years down the line from Joshi's groundbreaking work on code-switched text processing [1] and the pioneering strides in language identification [2], the spotlight of linguistic research has predominantly shone on the sociological and conversational factors that fuel code-switching [3]. This involves categorizing switches as either inter- or intra-sentential (between or within sentences), intra-word or tag switching (within words or through inserted phrases), and determining if they serve to express group identity or arise from gaps in language proficiency. Initially, social media studies primarily concentrated on English monolingual texts, but in recent times, there has been an increasing fascination with non-English and mixed-language content. It's clear as day from the numerous dedicated workshops on Computational Approaches to Code Switching [4], [5] and the insightful discussions on information retrieval from code-mixed texts at the FIRE workshops [6]-[10]. When it comes to code-mixed Indian language data, researchers have truly left no stone unturned, delving into a myriad of fascinating facets. Bhattacharja [11] took a deep dive into the fascinating world of complex predicates, skillfully weaving together English words and verbs in a captivating exploration. Ahmed et al. [12] shed light on the fact that code-mixing and abbreviations lead to transliteration errors in Hindi and Telugu, which in turn affects input method editors in significant ways. Mukund and Srihari [13] came up with a groundbreaking tagging method for Urdu-English code-mixing that cleverly utilizes POS categories. Das and Gambäck [14] were pioneers in the field, unveiling the very first social media dataset for Indian code-mixing, featuring the dynamic blend of Hindi and English. Barman and colleagues [15], [16] discovered that character n-grams, parts of speech, and lemmas are incredibly valuable for identifying languages and conducted a dictionary-based classification at the word level. Bali and others highlighted the importance of diving deep into both structural and discourse linguistic analysis when it comes to this kind of code-mixing. Diab and Kamboj [18] explored the exciting world of corpus collection, proposing the innovative idea of crowdsourcing to gather formal English-Hindi code-mixed data. Gupta and colleagues [19] took a bold step by harnessing the power of deep learning to uncover term equivalents in code-mixed text, coining the catchy phrase 'mixed-script information retrieval.' Meanwhile, Maharjan and their team [20] embarked on an exciting journey to develop and annotate code-switching tweets, diving into the vibrant worlds of Spanish-English and Nepali-English. Bohra et al. [21] have crafted a remarkable English-Hindi code-mixed dataset aimed at the ever-important task of hate speech classification. When it comes to the motivations driving code-switching, a plethora of studies indicate that social purposes reign supreme, frequently arising from a deep-seated desire to showcase in-group membership. Sotillo [22] discovered this phenomenon in SMS, where mixing often occurs at the start of messages or as straightforward insertions, a revelation that Bock [23] also highlighted for chat messages in English, Afrikaans, and isiXhosa. In a world where research often echoes itself, Xochitiotzi Zarate [24] found comparable outcomes in English-Spanish SMS, just as Shafie and Nayan [25] did with Facebook comments, and Negrón Goldbarg [26] mirrored these findings in Spanish-English emails. This stands in stark contrast to the vibrant studies conducted in the bustling bilingual hubs of Hong Kong and Macao by Li [27] and San [28], where the driving forces behind Chinese-English code-switching are undeniably more rooted in linguistic motivations than in social dynamics. Furthermore, research delves into the fascinating frequency and diverse forms of code-switching that emerge on social media platforms. In a fascinating exploration of language dynamics, Dewaele [29], [30] connected the dots between code-switching rates and the exhilarating thrill of "strong emotional excitement." Meanwhile, Johar [31] noted a delightful trend of increased positive smileys accompanying the art of code-switching, showcasing the joyful interplay of emotions and language. San [28] observed a clear trend of inter-sentential code-switching in blog posts, shining a light on how it stands out when compared to the spoken Macanese language. Hidayat[32] discovered that Facebook users are all about that inter-sentential switching, with a whopping 59% favoring it over intra-sentential at 33% and tag switching trailing behind at just 8%. The reasons behind this trend are as clear as day, with lexical needs leading the pack at 45%, followed closely by topic discussion at 40%, and a sprinkle of clarification at 5%. On the flip side, Das and Gambäck[33] revealed that fewer than half of the code-switching instances in Facebook messages were intra-sentential, while inter-sentential switching made up just around a third. Finally, the time has come for research to dive into the captivating realm of gender-based differences in code-switching. Kishi Adelia [34] embarked on an enlightening journey through a small dataset of Indonesian tweets, revealing that male students tended to embrace intra-sentential switching to build a sense of camaraderie, while female students chose inter-sentential switching as a heartfelt way to express their emotions and demonstrate their appreciation. Ali and Mahmood Aslam [35] observed in a small SMS sample that, without a shadow of a doubt, Pakistani female students were definitely more likely to pepper their Urdu texts with English words than their male counterparts.

III. Proposed Methodology

This study analyzes Three standard classifiers were selected to establish the baselines: Naïve Bayes (NB), Sequential Minimal Optimization (SMO), and Random Forests (RF). The Random Forest classifier is an ensemble of decision trees that has gained popularity for text categorization due to its algorithmic simplicity and superior performance with high-dimensional data. SMO is employed for training a Support Vector Machine (SVM) classifier with a Polynomial kernel; hence, the computation time of SMO is mostly influenced by SVM evaluation, making SMO most efficient for linear SVMs and sparse datasets. The Naïve Bayes classifier utilizes Bayes' Theorem, which is important in computing conditional probabilities, as inverse probabilities are generally more accessible and less subjective than direct probabilities. Naive Bayes is a straightforward and widely utilized technique that merges optimal temporal performance efficiency with acceptable accuracy. The NB classifier posits conditional independence among the characteristics utilized during training. N-grams, as features derived from texts, cannot be regarded as autonomous due to the syntactic and semantic interdependence of the tokens. Tokens exhibit syntactic and semantic interdependence. The investigations are categorized into two approaches: machine learning-based approaches and deep learning-based approaches. In general, I would say the models and techniques were chosen on the basis of 'What performs at the State-of-The- Art for non code-mixed tasks'. In particular for the baseline models – it represents the varied approaches to classification and different learning algorithms. In a way the lower performance of the ML models serves the purpose of showing that the hypothesis in sentiment analysis is highly non linear – and we need a large variety of ML models to add support to this conjecture. Sentiment Analysis is a reasonably complex NLP problem that to when it comes to a mixed language texts. Recent research suggests neural network based techniques are the best performing for the sentiment analysis. Therefore, it is obvious to choose traditional machine learning approaches such as NB, SMO and RF as the baselines which mostly produce linear hypothesis. This gives a clear and differentiable analysis of sentiment analysis performance on code-mixed data using traditional machine learning based approaches Vs. neural network based approaches.To establish a baseline, experiments were run using these popular machine learning algorithms. Stop-words and rare terms were eliminated, and classifiers were trained utilizing characteristics including trigrams and TF-IDF weights. The machine learning studies were conducted using Weka.

Figure 2. Bi LSTM- CNN Learner

A. BiLSTM-CNN learner

The very first deep learning model that was put to the test was a fantastic blend of BiLSTM and convolutional neural networks (CNN), as shown in the illustrious Figure 6.1. In the world of natural language processing, we take a word embedding matrix of a sentence and feed it into a bidirectional LSTM layer. This clever approach allows us to capture those long-range relationships and semantic attributes that are so crucial. By utilizing the bidirectional LSTM, we provide the neural network with both forward and backward contexts, ensuring that it has all the information it needs to understand the nuances of language. The bidirectional LSTM layer takes in 100 time stamps and produces a total of 8 outputs, which is the perfect blend of 4 plus 4 units. The outcome is then fed into a convolutional layer boasting 48 filters, each measuring 9, which glide over the input like a well-oiled machine, akin to a bag of n-grams that expertly pinpoints those all-important word n-grams. A global max pooling layer is used to grab the most important feature from the output of the convolutional layer, ensuring we capture the essence of the data like a pro! The data is then flattened and fed into a first dense layer of size 32, using a rectified linear unit (ReLU) activation. After that, it moves on to a second dense layer featuring a Softmax activation function to predict the probabilities of the three sentiment labels. This model suggests that the BiLSTM will gracefully capture the temporal order in both forward and backward directions, ensuring that the sequence is preserved, ultimately leading to a fresh encoding for the input. Next up, the output from the BiLSTM layer takes a thrilling journey into a convolutional layer, where it will work its magic by extracting those all-important local features! The result of the convolution layer is then brought together into a smaller dimension, ultimately leading to one of three sentiment labels that capture the essence of the analysis. In order to tackle the challenge of overfitting, a range of regularization values from 0.001 to 0.2 and dropout probabilities between 0.2 and 0.5 were thoroughly examined for this model. The sweet spot for optimal performance was found by setting the regularization parameter to a low 0.001 in the convolutional layer, while the dense layer was trained with a dropout probability of 0.2, leading to the best results. Dropout is a fantastic regularization technique that simply involves tossing aside a random selection of layer outputs during the training process, ensuring that your model doesn't get too comfortable and learns to adapt like a pro! When it comes to neuron dropout rates, a range of 20% to 50% is often seen as a solid starting point that really gets the ball rolling. A dropout probability that is too low hardly makes a difference, but if it's too high, it can really hinder the network's ability to learn effectively.

IV. Findings and Insights

Table 6.5 presents a comprehensive overview of the performance of various models on code-mixed social media datasets, showcasing statistics derived from evaluations conducted on the held-out (unseen) test sets. On the flip side, Table 6.6 showcases the results of the ever-popular 10-fold cross-validation performed on the datasets. In the tables, you'll find models that make use of GLoVe embeddings, which have been pre-trained on a whopping 27 billion tweets, clearly marked with a GLoVe suffix.

When we take a closer look at the two tables, it becomes crystal clear that employing a hold-out strategy for partitioning the data set led to outcomes that were just a notch above the rest. While n-fold cross-validation shines with its remarkable statistical accuracy, using a held-out test set can also do wonders for evaluating a model. It ensures that the test set, when used the right way, is made up of fresh data that the model hasn't seen before during training. In addition, cross-validation can be quite the resource hog; thus, hold-out evaluation might be viewed as a "shortcut" to cross-validation, requiring less computational power. This is important because, just like in other studies such as SemEval, similar research has been conducted using a specific test set. As a result, once the trustworthy 10-fold cross-validation results are in, we will be able to draw some meaningful conclusions from them. Table 1 shows that we hit the sweet spot of accuracy without needing any extra information or hand-crafted features. It's a win-win situation! The tables clearly show that deep learning models are head and shoulders above traditional machine learning models, with the attention-based model frequently shining just a bit brighter than the other two network models. Custom word embeddings frequently demonstrate a noticeable boost in performance when compared to the pre-trained GLoVe embeddings derived from a staggering 27 billion tweets. The numbers that show how much we've improved can be found in Table 6.6. This finding really highlights the impressive amount of unknown tokens that come to light when we utilize an embedding sourced from a monolingual corpus. It's truly eye-opening! In light of this, the embeddings obtained from the code-mixed corpus shine brightly, showcasing remarkable performance, even though the pre-training corpus is somewhat limited in size.

Table 1: Result Parameter for EN-HI text analysis

Dataset Name	Classifier used	Accuracy	Precision	Recall	F1 Score
ICON 2017 EN-HI	Naïve Bayes	46.8	32.1	46.8	38.1
	SMO	49.1	53.6	49.1	51.3
	Random Forest	48.9	51.4	48.9	50.1
	BiLSTM-CNN	57.6	57.1	57.6	57.3
	GLoVe BiLSTM-CNN	57.4	56.9	57.4	57.1
	BERT	62.2	62.0	62.2	62.1

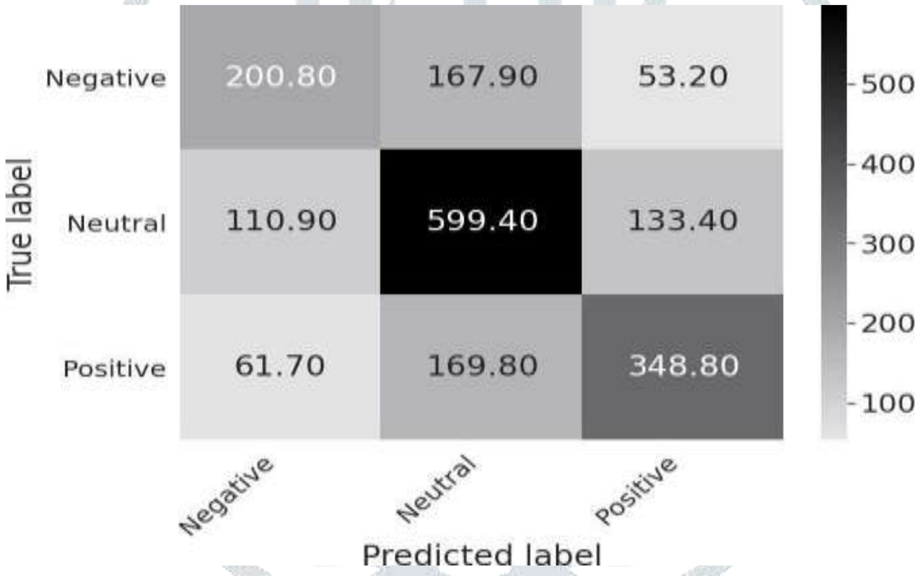


Figure:2 Confusion Matrix of BERT

The custom word embeddings do not significantly outperform the GLoVe word embedding trained on 27 billion tweets, although they somewhat beat it in all dimensions. Given the restricted dataset employed for training the custom embeddings, this can certainly be seen as a success. The improvements in accuracy and F1 score can likely be ascribed to the reduction in the number of unknown tokens due to the use of the custom word embedding. Analysis indicated that 64,072 terms in the lexicon of the custom embeddings were missing from the GLoVe word embeddings, even though the latter was trained on 27 billion tweets. It is expected that models utilizing embeddings trained on a code-mixed corpus will significantly outperform those using embeddings trained on a monolingual corpus, given a more extensive dataset. The exceptional effectiveness of BERT was an unexpected result of the studies. Nevertheless, as shown in Table 1, BERT exhibits comparability to other deep learning models regarding F1 score, providing no significant improvements over them. BERT utilizes WordPiece tokenization techniques that significantly reduce the occurrence of unknown tokens, possibly enhancing its performance. This research definitively demonstrated that deep learning models, particularly the Attention-based model and pre-trained BERT, substantially outperform traditional machine learning techniques (including Naïve Bayes, SMO, and Random Forests) in the sentiment analysis of code-mixed English-Hindi and English social media texts. Despite the challenges posed by little data and the informal nature of social media, the deep learning techniques achieved significantly higher F1 scores. The study highlighted BERT's unforeseen effectiveness despite considerable dataset variability and suggested the potential for domain-specific BERT word embeddings. The attained accuracies are inferior to those of monolingual English sentiment analysis, attributable to the inherent complexity of code-mixed data and the scarcity of comprehensive annotated resources. The diverse corpus collection, including Twitter, Facebook, and WhatsApp, ensured that the proposed methodologies are scalable across other social media platforms and linguistic combinations. Numerous future research opportunities exist in Natural Language Processing (NLP) for code-mixed text. It is essential to enhance language identification to more efficiently handle multilingual code-mixed content, especially considering varied representations like Unicode, transliterated Indian languages, and the English lexicon. Furthermore, examining transliteration and sophisticated language Furthermore, investigating transliteration, sophisticated language modeling, and parsing methodologies for code-mixed text is essential. A promising avenue entails the incorporation of language modeling into code-mixed Part-of-Speech (POS) tagging and examining how code-switched language models can tackle the issue of unknown words in these intricate linguistic contexts.

IV. Conclusion

This research successfully demonstrated that deep learning models, particularly the Attention-based model and pre-trained BERT, substantially outperform traditional machine learning techniques (like Naïve Bayes, SMO, and Random Forests) in sentiment analysis of code-mixed English-Hindi and English- social media texts. Despite the challenges posed by sparse data and the informal nature of social media, the deep learning approaches achieved significantly higher F1 scores. The research also highlighted BERT's surprising effectiveness even with high dataset dissimilarity and showed promise for domain-specific BERT word embeddings. While the obtained accuracies are still lower than those for monolingual English sentiment analysis, this is attributed to the inherent complexity of code-mixed data and the scarcity of large, annotated resources. The diverse corpus collection, spanning Twitter, Facebook, and WhatsApp, ensured the methodologies proposed are scalable across various social media platforms and language combinations. Looking ahead, several avenues exist for further research in Natural Language Processing (NLP) for code-mixed text. There's a need to enhance language identification to better handle multilingual code-mixed text, especially considering various representations like Unicode, transliterated Indian languages, and English words. Additionally, exploring transliteration, advanced language modeling, and parsing techniques for code-mixed content is crucial. A promising direction involves integrating language modeling into code-mixed Part-of-Speech (POS) tagging and investigating how code-switched language models can address the challenge of unknown words in these complex linguistic environments.

References

[1] J. Fong, S Burton, Elecronic Word-of-Mouth, Journal of Interactive Advertising. 6(2)(2006), pp.7-62.https: //api.semanticscholar.org/CorpusID:219544765.

[2] X. Shi, W. Liu, J Zhang, Present and future trends of supply chain management in the presence of COVID-19: a structured literature review, International Journal of Logistics Research Applications.26(2021), 813842.https://doi.org/10.1080/13675567.2021.198890

[3] J. El Baz, S. Ruel, Can supply chain risk management practices mitigate the disruption impacts on supply chains' resilience and robustness? Evidence from an empirical survey in a COVID-19 outbreak era, International Journal of Production Economics.233(2021), p.107972. https://doi.org/10.1016/ j.ijpe.2 020.107972.

[4] J.Hou, S Zhao, HM Wang and E., Bi.Sourcing Decisions with Capacity Reservations under Supply Disruptions, Journal of Management Science and Engineering.2(2)(2019), pp.132-159. https://doi.org/10.3724/SP.J.1 383.202007

[5] T. M. Choi, Risk analysis in logistics systems: A research agenda during and after the COVID-19 pandemic, Tr ansportation Research Part E: Logistics and Transportation Review 145 (2021), pp.102190. https://10.10 16/j.tre.2020.102190.

[6] S.Y. Yang , L.J. Ning, T.F. Jiang, Y.Q. He, Dynamic impacts of COVID-19 pandemic on the regional express logistics: Evidence from China, Transport Policy.111(2021),111-124. https://doi.org/10.1016/j.tranpol.20 21.07.012

[7] N. Zhang, R. Zhang, Z. Pang, X. Li, W. Zhao, Mining express service innovation opportunity from online reviews, Journal of Organizational and End User Computing. 33(6)(2021), pp.1-15. http:/doi.org/10.401 8/JOEUC 20211101.oa3.

[8] K. Li, Research on Customer Satisfaction of Express Service Industry Based on Text Mining.Hefei University of Technology.2021.

[9] L. Zheng, H.W. Duan, L.P. Zhang, D.J. Ergu, F.Y. Liu, The main influencing factors of customer satisfaction and loyalty in city express delivery.FrontPsychol.13(2022),1044032.https://doi.org/10.3389/fpsyg.2022 .1044032.

[10] P. Ding, 2023. Research on Improving Express Service Quality Based on Online Review Data Mining. Shan dong Jiaotong University. https://doi.org/10.1051/matecconf/202032503003.

[11] G. Denktaş-Şakar, E.Sürücü, Stakeholder engagement via social media: an analysis of third-party logistics co mpanies, The Service Industries Journal.40(11-12)(2020),pp.866-889.https://doi.org/10.1080/02642069. 2018.1561874.

[12] W. Hong, C. Zheng, L. Wu, X. Pu, Analyzing the Relationship between Consumer Satisfaction and Fresh E-Commerce Logistics Service Using Text Mining Techniques, Sustainability.11(2019),3570. https://doi.org/10.3390/su11133570.

[13] L. Zheng, Z. He, S. He, An integrated probabilistic graphic model and FMEA approach to identify product defects from social media data, ExpertSystemswith Applications.178(2021), p.115030.https://doi.org/1 0.1016/j.eswa.2021.115030.

[14] J.F. Ding, W.H. Shyu, C.T. Yeh, P.H. Ting, C.T. Ting, C.P. Lin, C.C.Chou, SS Wu, Assessing customer value for express service providers:An empirical study from shippers' perspective in Taiwan, Journal of Air TransportManagement.55(2016),pp.203-211. https://doi.org/10.1016/j.jairtraman.2016.06.004.

[15] The Central People's Government of the People's Republic of China(2008)."People's Republic of China

[16] Postal Industry Standard - Express Service" implementation. https://www.gov.cn/gzdt/2008-01/03/content_849510.htm.2008(accessed 3 January 2008).

[17] A. Vedadi, T.H. Greer, Revisiting How Perceived Uncertainty and Herd Behavior Influence Technology Choice, Journal of Organizational and End User Computing.33(6)(2021),9.http://dx.doi.org/10.4018/JOEUC.20211101.oa1.

[18] S.S.Gu, K.Y. Wang, L.Y. Gao, J. Liu, .Research on express service defect evaluation based on semantic network diagram and SERVQUAL model, Frontiers in Public Health. 10 (2022), p.1056575.https://doi.org/10.3389/fpubh.2022.1056575.

[19]O. Dospinescu, N. Dospinescu, I. Bostan, Determinants of e-commerce satisfaction: a comparative study between Romania and Moldova, Kybernete.51(13)(2022),1-17.https://doi.org/10.1108/K-03-2021-0197.

[20]J. Zhou, S.T. Miao, Performance-only measurement of service quality: An empirical study in Chinese express industry, 2009 6th International Conference on Service Systems and Service Management. 2009, pp.831- 836. http://dx.doi.org/10.1109/ICSSSM.2009.5174996.

[21]S. Gläser, H. Jahnke, N. Strassheim, Opportunities and challenges of crowd logistics on the last mile for courier, express and parcel service providers – a literature review, International Journal of Logistics Research and Applications, 26(8)(2023), pp.1006– 1034.https://doi.org/10.3389/fpubh.2022.1056575.

[22]L. Li, H.F. Hong, Analysis of influencing factors of consumers' purchase intention in the network environment, China Business and Trade. 2(2010),14-15. https://api.semanticscholar.org/CorpusID:270667363.

[23]Y. Xiao, B. Li, Z. Gong, Real-time identification of urban rainstorm waterlogging disasters based on Weibo big data, Natural Hazards. 94(2018), 833-842.https://doi.org/10.1007/s11069-018-3427-4.

[24]D. P. Sakas, D.P. Reklitis, M.C. Terzi, Leading Logistics Firms' Re- Engineering through the Optimization of the Customer's Social Media andWebsiteActivity,Electronics.12(11)(2023), p.2443.https://doi.org/10.3390/electronics12112443.

[25]Y. Zhang, D, Song, P. Zhang, X. Li, P. Wang, A quantum-inspired sentiment representation model for twitter sentiment analysis, Applied Intelligence. 49(2019), pp.3093-3108. https://doi.org/10.1007/s10489- 019-01441-4.

[26]Y. Zhao, S. Cheng, X. Yu, H. Xu, 2020. Chinese Public's Attention to the COVID-19 Epidemic on Social Media: Observational Descriptive Study,J Med Internet Res.22,e18825.https://doi.org/10.2196/18825.

[27]F. A. Yang, L. Yang, Y. Xiaohui, M. Yang, Automatic detection of rumor on sina microblog. In Proceedings of the ACM SIGKDD workshop on mining data semantics.2012, pp.1- 7.https://doi.org/10.1145/2350190.2350203.

[28]S.J. Li, Y.L. Wang, J. Xue, N. Zhao, T.S. Zhu, The Impact of COVID- 19 Epidemic Declaration on Psychologi cal Consequences: A Study on Active Weibo Users, Int J Environ Res Public Health.17(6)(2020). https:// doi.org/10.3390/ijerph17062032.

[29]Y. Guo, S.J. Barnes, Q. Jia, Mining meaning from online ratings and reviews: tourist satisfaction analysis using latent Dirichlet allocation, Tour.Manag.59(2017),pp.467– 483.https://doi.org/10.1016/j.tourman.2016.09.009.

[30]R. Egger, J. Yu, A Topic Modeling Comparison Between LDA, NMF,Top2Vec, and BERTopic to Demystify

[31]Twitter Posts, Front Sociol. 7(2022), p.886498. https://doi.org/10.3389/fsoc.2022.886498.

[32]H. Jelodar, Y. Wang, R. Orji, S. Huang, Deep sentiment classification and topic discovery on novel coronavi rus or COVID-19 online discussions: Nlp using lstm recurrent neural network approach.IEEE Journal of Biomedical and Health Informatics.24(2020), pp.2733- 2742.https://doi.org/10.1109/JBHI.2020.3001216.

[33]F. Hong, C. Lai, H. Guo, E. Shen, X. Yuan, S. Li, FLDA: Latent Dirichlet Allocation Based Unsteady Flow Analysis, IEEE Trans Vis ComputGraph. 20(12)(2014), pp.2545-54. https://doi.org/10.1109/TVCG.2014.2346416.

[34] F. Hong, C. Lai, H. Guo, E. Shen, X. Yuan, S. Li, FLDA: Latent Dirichlet Allocation Based Unsteady Flow Analysis, IEEE Trans Vis ComputGraph.20(12)(2014),pp.2545-54.https://doi.org/10.1109/TVCG.2014.2346416.

[35] M. K. Lim, Y. Li and X. Song, Exploring customer satisfaction in cold chain logistics using a text mining approach, Industrial Management & Data Systems.121(2021), pp.2426-2449.https://doi.org/10.1108/IM DS-05-2021-0283.

[36] B. Sun, M. Kang. S. Zhao, How online reviews with different influencing factors affect the diffusion of new products. International Journal of Consumer Studies. 47(2023), pp.1377-1396. https://doi.org/10.1111/ijcs.12915.

[37] Z. Jingrui, Q.L. Wang, L. Yu, L. Yuan, 2017, A method of optimizing LDA result purity based on semantic similarity. 2017 32nd Youth Academic Annual Conference of Chinese Association of Automation (YAC), IEEE,Hefei,PeoplesRChina.361-365. https://doi.org/10.1109/YAC.2017.7967434

[38]A. McCallum, X. Wang, A. Corrada-Emmanue, Topic and role discovery in social networks with experiments on enron and academic email, Journal of artificial intelligence research. 30(2007), pp.249-272. https://doi.org/10.1613/jair.2229.

[39]J. Xue, J. Chen, R. Hu, C. Chen, C. Zheng, Y. Su, T. Zhu, 2020, Twitter Discussions and Emotions About the COVID-19 Pandemic: Machine Learning Approach, J Med Internet Res. 22(11), e20550. https://doi.org/ 10.2196/20550.

[40]K. Tago, Q. Jin, Influence analysis of emotional behaviors and user relationships based on twitter, Tsinghua Science and Technology. 23(1)(2018), pp.104-113. https://doi.org/10.26599/TST.2018.9010012.

[41]S. Zhang, H. Zhong, C. Wei, D. Zhang, Research on Logistics Service Assessment for Smart City: A Users' Review Sentiment Analysis Approach,Electronics.11(23)(2022), p.4018.https://doi.org/10.3390/electronics11234018.

[42]M. Alkhatib, M. El Barachi, K. Shaalan, An Arabic social media based framework for incidents and events monitoring in smart cities, Journal of Cleaner Production. 220(20)(2019),pp.771-785.https://doi.org/10.1016/j.jclepro.2019.02.063.

[43]G. Gautam, D. Yadav, 2014, Sentiment analysis of twitter data using machine learning approaches and semantic analysis, 2014 Seventh international conference on contemporary computing (IC3), IEEE,Noida, INDIA. pp. 437-442. https://doi.org/10.1109/IC3.2014.6897213.

[44]A. Heydari, M. Tavakoli, N. Salim, Detection of fake opinions using time series, Expert Systems with Applications. 58(1)(2016), pp.83-92. https://doi.org/10.1016/j.eswa.2016.03.020.

[45]E. Hirata, T. Matsuda, Uncovering the impact of COVID-19 on shipping and logistics, Maritime Business Review. 7(4)(2017), pp.305-317. https://doi.org/10.1108/MABR-03-2021-0018 .

[46]Y. Bengio, Y. LeCun, Scaling learning algorithms towards AI, Large- scalekernelmachines.34(2007),pp.1-41. https://doi.org/10.7551/mitpress/7496.003.0016.

[47]W.-T. Yih, X. He, C. Meek, Semantic parsing for single-relation question answering. Toutanova, K and Wu, H (Ed.s),Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics,Baltimore,MD.2014,pp.643648d.https://doi.org/10.3115/v1/P14-2105

[48]J.Wang, L.-C. Yu, K. R. Lai, X Zhang, Tree-Structured Regional CNN- LSTM Model for Dimensional Sentiment Analysis, IEEE/ACM Transactions on Audio, Speech and Language Processing. 28(2020), pp. 581-591.https://doi.org/10.1109/TASLP.2019.2959251.

[49]S.Hochreiter, J. Schmidhube, Long short-term memory, Neural computation.9(8)(1997), pp.1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735.

[50]K. Dashtipour, M. Gogate, A. Adeel, H. Larijani, A. Hussain, Sentiment Analysis of Persian Movie Reviews UsingDeepLearning, Entropy. 23(5)(2021),pp.596. https://doi.org/10.3390/e23050596.

[51]P. B. Washington, P. Gali, F. Rustam, I. Ashraf, 2023. Analyzing influence of COVID-19 on crypto and financial markets and sentimentanalysis using deep ensemble model. Plos One.18(9), p. e0286541. https://doi.org/10.1371/journal.pone.0286541.

[52]W. Zhang, L. Li, Y. Zhu, P. Yu, J Wen, CNN-LSTM neural network model for fine-grained negative emotion computing in emergencies, Alexandria Engineering Journal. 61(9)(2022), pp.6755-6767.https:// doi.org/ 10.1016/ j.aej.2021.12.022.

[53]J. Zhao, J. Lin, S. Liang, M. Wang, Sentimental prediction model of personality based on CNN-LSTM in a social media environment, Journal of Intelligent and Fuzzy Systems.40(2)(2021), pp. 3097- 3106.https:// content. iospress.com/doi/10.3233/JIFS-189348.

[54]W. H. Chen, Y. Ca, K.K. Lai, Weibo Mood Towards Stock Market, Database Systems for Advanced Applications.9645(2016), pp. 3-14. https://doi.org/10.1007/978-3-319-32055-7_1.

[55]M. Lu, P. Liu, Denoising Distant Supervision for Relation Extraction with Entropy Weight Method, Chinese Computational Linguistics: 18th China National Conference. 11856(2019), pp.294-305. https://doi.org/10.1007/978-3-030-32381-3_24.

[56]Y.Yang, J.O. Pedersen, A comparative study on feature selection in text categorization, International Conferen ce on Machine Learning.97(1997),pp.412–420.https://dl.acm.org/doi/10.5555/645526.657137.

[57]J. Leskovec, A. Rajaraman, J.D. Ullman, Mining of massive data sets, Cambridge university press,2020.

[58]G.Q. Chao, J.Y. Liu, M. Wang, D.H. Chu, Data augmentation for sentiment classification with semantic preservation and diversity, Knowledge-Based Systems.280(2023),p. 10.https://doi.org/10.1016/j.knosys.2023.111038.

[59]T. Wang, K. Lu, K.P. Chow, Q. Zhu, COVID-19 Sensing: Negative Sentiment Analysis on Social Media in China via BERT Model, IEEE Access.8(2020),pp.138162-138169.https://doi.org/10.1109/ACCESS.2020.3012595.

[60]F. Zhao, X. Ren, S. Yang, Q. Han, P. Zhao, X. Yang, Latent Dirichlet Allocation Model Training With Differential Privacy, IEEE Transactions on Information Forensics and Security. 16(2021), pp.1290- 1305. https://doi.org/10.48550/arXiv.2010.04391.

[61]D. Wu, R. Yang, C. Shen, Sentiment word co-occurrence and knowledge pair feature extraction based LDA short text clustering algorithm, Journal of Intelligent Information Systems.56(1)(2021), pp. 1-23.https:// doi.org/10.1007/s10844-020-00597-7.

[62]T.Y. Kim, S.B. Cho, Predicting residential energy consumption using CNN-LSTM neural networks, Energy.182(2019), pp. 72-81. https://doi.org/10.1016/j.energy.2019.05.230.

[63] M. Schuster, K.K. Paliwal, Bidirectional recurrent neural networks, IEEE transactions on Signal Processing, 45(1997), pp.-2681. https://ieeexplore.ieee.org/document/650093.

[64]W. Jiang, K. Zhou, C. Xiong, G. Du, C. Ou, J. Zhang, KSCB: A novel unsupervised method for text sentiment analysis, Applied Intelligence.53(1)(2023), pp.301-311.https://doi.org/10.1109/TVCG.2014.2346416.

[65]L. Egozi, N. Reiss-Hevlin, R. Dallasheh, A. Pardo. Couriers' safety and health risks before and during the COVID-19 pandemic. International archives of occupational and environmental health, 2022,pp.1-10.https://doi.org/10.1007/s00420-021-01795-8.