Journal of Information Systems Engineering and Management
2025, 10(23s)
e-ISSN: 2468-4376
https://www.jisem-journal.com/

Research Article

Expanding Research Horizons for Hinglish Text by Tackling Challenges and Research Gaps

Pratibha .[1*], Amandeep Kaur[1†] and Meenu Khurana[2†]
[1*]Chitkara University Institute of Engineering & Technology, Chitkara University, Rajpura,140401, Punjab, India.
[2*]Chitkara School of Engineering & Technology, Chitkara University, Baddi, Himachal Pradesh, India.
*Corresponding author(s). E-mail(s): pratibha@chitkara.edu.in; Contributing authors: amandeep@chitkara.edu.in; meenu.khurana@chitkara.edu.in;

ARTICLE INFO

Received: 23 Dec 2024
Revised: 10 Feb 2025
Accepted: 24 Feb 2025

ABSTRACT

Hinglish, a hybrid of Hindi and English, poses unique challenges for sentiment analysis due to its linguistic complexity, lack of structured grammar, and limited availability of annotated datasets. This research investigates and analyses existing methods for Hinglish and English sentiment analysis and explores the complexities of this code-mixed domain. By analyzing research papers, reports, and white papers, the study identifies key obstacles like language detection, intricate grammar, and limited data availability. It also highlights the difficulties of adapting models to new contexts. The research proposes innovative solutions based on these challenges and emphasizes the need for a multi-faceted approach to achieve accurate sentiment analysis in Hinglish. This paves the way for further innovation in code-mixed language sentiment analysis, and at the same time, it can form the base for building large language models for sentiment analysis as well as llm poorly understood Hinglish in current capacities.

Keywords: Code-mixed, languages, Sentiment analysis, Topic modeling, Language-independent Features, Business, Social Media, Hinglish, Emotion.

1 INTRODUCTION

Code-mixed languages make sentiment analysis difficult to compute. Thus, an automated system for recognising raw emotions or sentiments requires a comprehensive language model[1][2]. Sentiment analysis-based assistance solutions require huge language models. Fig. 1 shows code-mixed sentences in Roman script Hindi, English, and non-code-mixed languages. The specified entity can confuse codemixed and non-codemixed statements.

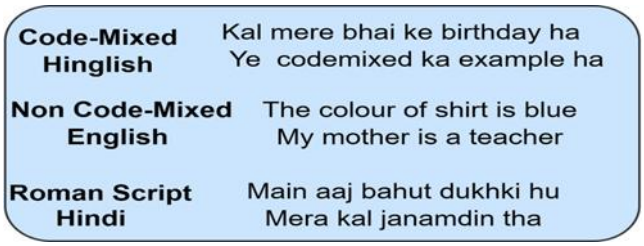

Fig. 1: Code-mixed and Non-Code-mixed Language

Research shows that Hinglish code-mix language morphology is difficult to interpret, and human-labelled data is scarce. Slang, misspellings, and other informal language make processing social media information difficult for sentiment analysis models[3][4]. Because code-mixed language sentiment analysis may misinterpret sarcasm and irony, it is difficult to discern. Negations can also change a statement's tone, making sentiment analysis algorithms less accurate when they're present. Emojis and emoticons are technologically challenging. Emojis and emoticons can express emotions, but sentiment analysis models may have problems distinguishing them. Idiomatic language work is also arduous. Idiomatic statements can express emotions, but sentiment analysis systems may

struggle to understand them [5]-[10].

Textual context infers implicit sentiment. Implicit sentiment is inferred from context, making it hard for sentiment analysis tools to detect. Thus, natural language systems struggle with multilingual and cross-lingual material. Because out-of-vocabulary terms may not be in the model's training data, sentiment analysis algorithms and models may have trouble distinguishing the text's sentiment. Maintaining context and subjectivity is crucial. Sentiment analysis pipelines may have trouble distinguishing subjective from objective assertions. The speaker's perspective may affect a statement's sentiment. Machine learning and deep learning sentiment analysis algorithms may struggle to detect multi-modal input like photos, videos, and sounds. As illustrated above, sentiment analysis is also difficult with long-form content due to the large context window. Thus, consistent work in this sector is needed to design algorithms that benefit humanity's future[11] [12].

In the following part of our research endeavour, we will provide an infographic of the methodology that we use to investigate the potential research gaps that call for the attention of modern scholars. It would be an example of a methodological gap or 'M', for instance, if the majority of the previous research employs one specific kind of machine learning algorithm, but another kind of algorithm would be better appropriate for the issue at hand. The empirical gap is the second kind of research gap that exists nowadays. In this context, the term "gap" refers to the absence of study that addresses a particular research issue by employing particular kinds of data. For instance, this would be considered an empirical gap if the majority of the previous research centred on a specific type of data, such as images, when other types of data, such as text, could have been used to address the issue instead. In this case, images would be an example of the type of data.

In conjunction with this, one of the goals is to identify theoretical gaps ′ T′, which are understood to be a deficiency in existing research that addresses particular theoretical facets of the study subject. For instance, if the majority of previous research focuses on a particular aspect of the issue, such as class classification, but another aspect, such as clustering, is not well addressed, this would be considered a theoretical gap. A serious attempt is being made to uncover application gaps. Application gaps ′ A ′ are gaps that refer to a lack of research that targets specific application areas of the research topic [7]. Subsequently, an effort will also be made to uncover evaluation gaps, which may be defined as a lack of research that makes use of particular evaluation metrics or protocols to evaluate how well-proposed solutions to a research problem work. For example, if the majority of previous research relies on a specific evaluation metric, such as accuracy, but there is a possibility that another metric, such as F1-score, would be more appropriate for the issue at hand, this would be called an evaluation gap. In the framework of sentiment analysis, gaps related to the quantity and quality of the dataset will also be examined. In the next section, we observe the contemporary related work of the problem undertaken for this research. Qualitative analysis by hand is impractical due to the large number of articles. Thus, text mining and clustering find the most promising research subjects. LDA topic modelling and Python programming create distinctive and clear word clouds for academic writing classification. Recognising that a text document may cover multiple themes helps LDA avoid text clustering[12] [13].

1.1 Evolution of Sentiment Analysis and Emotion Detection Techniques

Sentiment analysis is the practice of looking at how people feel and think about a text or piece of information. Fig. 2 illustrates how it changed through time to become multimodal sentiment analysis between 1940 and 2021.In order to comprehend people's attitudes, researchers employed a manual survey method to make the first sentiment analysis finding in 1940. Another finding was made between 1940 and 1947 when people's sentiments were recorded on paper for analysis. Later on, sentiment analysis was developed as a result of numerous significant investigations. Detecting subjective phrases in 1990, recognising emotions from text in 2001, determining polarity (positive or negative), employing lexical information for analysis in 2004, and analysing positive and negative feelings are a few prominent ones.In 2013, scientists developed the first neural network capable of handling tasks using the English language. In terms of text analysis and prediction, this was a huge advancement. Transformers are a sort of model that can comprehend and interpret language in a more sophisticated manner. The notion of transformers was introduced in 2017. Then, in 2018, the BERT model was presented. BERT is a model that has been taught to grasp the context and meaning of words in sentences. It worked well for many different text analysis jobs. In 2019, Google created the Transformer XL model, which gave computers the ability to comprehend context in addition to fixed-length characteristics. As a result, individuals are better able to perceive the context while analysing and interpreting lengthy texts. A trained multimodal

transformer-based model was created in 2021. To better comprehend the whole context, this model can handle many sorts of data, including text, photos, and videos[14]-[16].

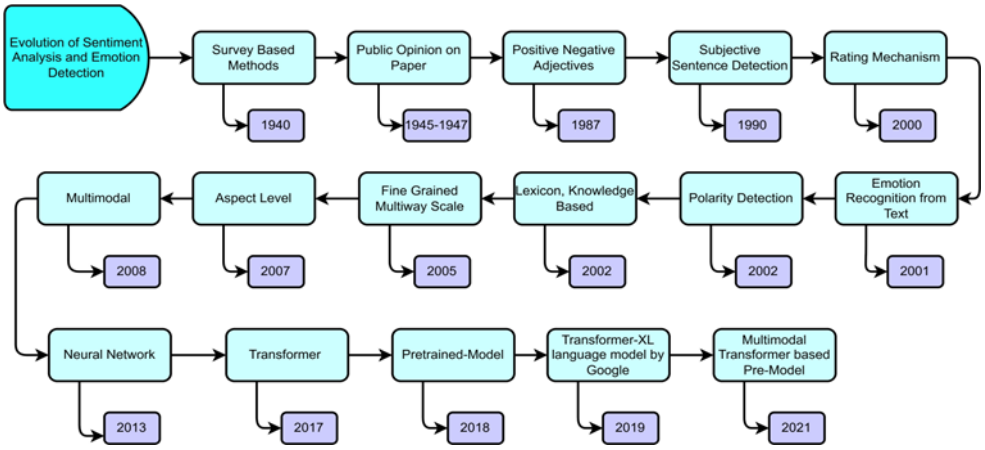

Fig. 2: Evolution of Sentiment and Emotion Detection

As a result of these developments, sentiment analysis and emotion recognition will become more crucial. The goal is to assess the ideas and feelings expressed in large volumes of unstructured data found on social media platforms to determine how individuals think and feel. These studies have been essential for comprehending and researching unstructured social media data, where sentiment analysis is key for examining the ideas and feelings individuals express [82].

1.2 Process of Sentiment and Emotion Detection

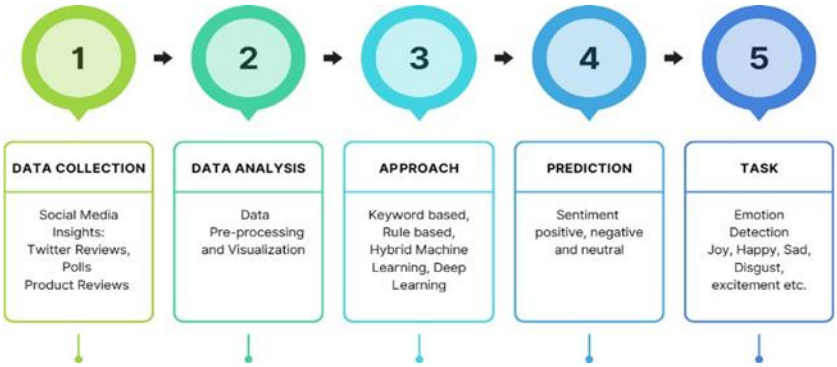

Fig. 3: Workflow of Sentiment or Emotion Detection

1.2.1 Data collection

Collect data from a variety of sources, including social media, customer reviews, emails, and polls[17].

1.2.2 Data Analysis

Data analysis is done to remove noise from the data by deleting extraneous information, stop words, and punctuation marks. Normalize the text by making all of the words lowercase and stemming or lemmatizing them[11][18]. Convert the text to a numerical form, such as word embeddings.

1.2.3 Approach

Choose a suitable approach for identification. In the case of deep learning convolutional Neural Networks, Recurrent Neural Networks are popular approaches, and Transformers and CNNs are excellent at recognizing local patterns in text, but RNNs are excellent at modeling long-term relationships. Transformers- ers are a relatively new concept that has been shown to perform exceptionally well in natural language processing applications. Train the deep learning model on the pre-processed data[19]. Use suitable loss functions for classification or prediction issues, such as cross-entropy loss. Use optimization techniques such as Stochastic Gradient Descent (SGD) or Adam to update the model weights. Evaluate the correctness and efficacy of the

sentiment analysis or emotion detection model using measurements such as precision, recall, F1-score, or confusion matrix [16][20].

1.2.4 Prediction

The last step is to detect the sentiment and emotion in the form of happy, sad, and joy[21].

1.3 Frameworks Spectrum

The overview of various techniques for sentiment detection is presented in Figure 4.

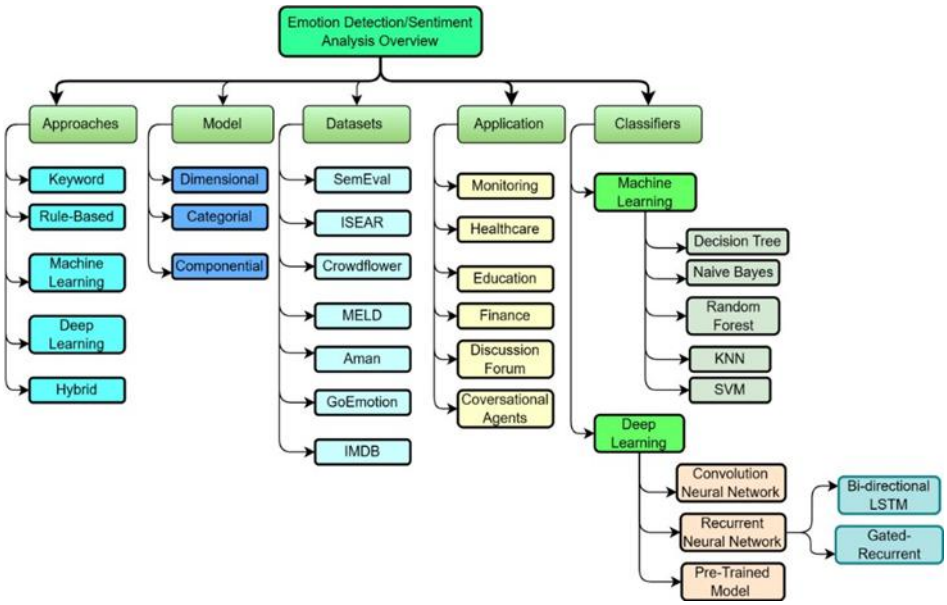

Fig. 4: Overview of Techniques

1.4 Approaches for Sentiment Analysis or Emotion Detection

- Keyword-based approach: It involves finding occurrences of a keyword in a text and matching them with labels stored in the Data dataset. Using this method, the emotion keyword list is derived from standard lexical databases like WordNet-Affect and WordNet. Next, emotion keywords from the text are compared to predefined keyword lists to see which show the most similarities. We analyse keywords to determine keyword intensity and then determine the emotion label[22].
- Rule-based: As the name suggests, to detect the emotion, the rule-based approach is designed with a set of rules and logic. This involves assigning probabilistic affinity or lexical affinity to each word. Following that, linguistic and statistical theories are used to mine rules for emotion recognition from text; after the selection of final rules, emotion can be determined[23].
- Machine Learning: This approach is generally known as supervised or unsu-
- pervised learning. and starts with the text pre-processing step on the emotion dataset with the help of tokenization, stop word removal, lemmatization, and POS tagging. Then, the features are extracted and selected from the text. After that, the system is trained with the given feature set and emotion labels to classify the emotion[24][25].
- Deep Learning: This is an approach that uses unstructured or unlabeled data
- to facilitate unsupervised learning. Learning complex concepts is accomplished by constructing them from simpler ones. The dataset is preprocessed by tokenizing, removing stop words, and lemmatizing. The embeddings are then built after that. These vectors are then given to deep neural network layers with components that are equal to emotion labels via classification, where data forms are found and used to estimate labels[26].
- Hybrid Approach: It is a combination of keyword-based and machine learn-

- ing approaches. Using the emotional keyword, the system first looks for emotion keywords and then the sentence is categorized. The classifier is used if the input phrase lacks emotional keywords[27].

1.5 Dataset

Dataset is a collection of structured data related to a particular piece of work. The data set can be public or private. It can be domain-specific. Table 1 explains the publicly available dataset in English or Hinglish language. The researcher can use publicly available datasets or create their own. Here are some datasets that are used widely in sentiment/emotion detection[23][5].

Table1: Existing Available Dataset with Description

Dataset	Modality	Size	Source	Model	Emotion
ISER[28]	English	7666	Interview	Dimensional	Disgust, Guilt, Fear, Anger, Sadness, Shame
SemEval 2019[17]	English		Tweets, News headlines	Discrete	Joy,Fear,Sadness, Anger Surprise, Disgust
HinGE[29]	Hindi, English	1,976	Humans, rule-based	–	–
MELD [30]	English	13000	Conversation utterances	Discrete	Joy, Fear, Anger,Sadness, Disgust,Surprise Neutral
Emobank [31]	English		Blogs, News, Essay	Discrete	Joy, Fear, Sadness, Anger, Surprise, Disgust
IMDB[32]	English	50,000	Review	–	Positive, Negative
SemEval 2020 Dataset[33]	Hinglish	17000	Twitter	–	Negative,Neutral, Positive
WASSA-2017[34]	English, Spanish, Arabic	7097	Twitter	Russell's Circumplex	Joy,Sadness, Fear, and Anger
OffensEval 2020[35]	Arabic, Danish, English, Greek, Turkish	10,48,027	Twitter	–	–
TwitterUS Airline Sentiment[36	English	14000	Twitter	–	Positive, Negative, Neutral
GoEmotions [37]	English	58k	Reddit	Ekman	Remorse, Grief, Gratitude, Love,Sadness, Pride, Admiration, Disgust, Anger
DailyDialog [38]	English	13118	chat	Discrete	Happiness, sad anger,disgust, fear
Emotion Stimulus [39]	–	1594	FrameNets	–	Happiness, Fear, Surprise,Disgust, Sadness, Anger,Shame.

The Valence and Arousal [40]	English	2895	Facebook	Dimension	Happy, Sad, Love, Hate, Bored
HinGE [29]	Hinglish	10,731	Twitter/'X'		—
PHINC [25]	Hinglish	13,738	Twitter	–	–
[41]	Hinglish	3999 memes	Social Media	–	Negative, Positive, Neutral
Hinglish-TOP[42]	(10K)	human annotated	-	-	
ICON 2017 [43]	Hindi-English Bengali English	18461	Social Media	-	Positive, negative, and Neutral
Emotion Detection[44	Hindi-English	151311	Social Media	-	Happiness, Fear, Sadness, Anger, disgust, Surprise

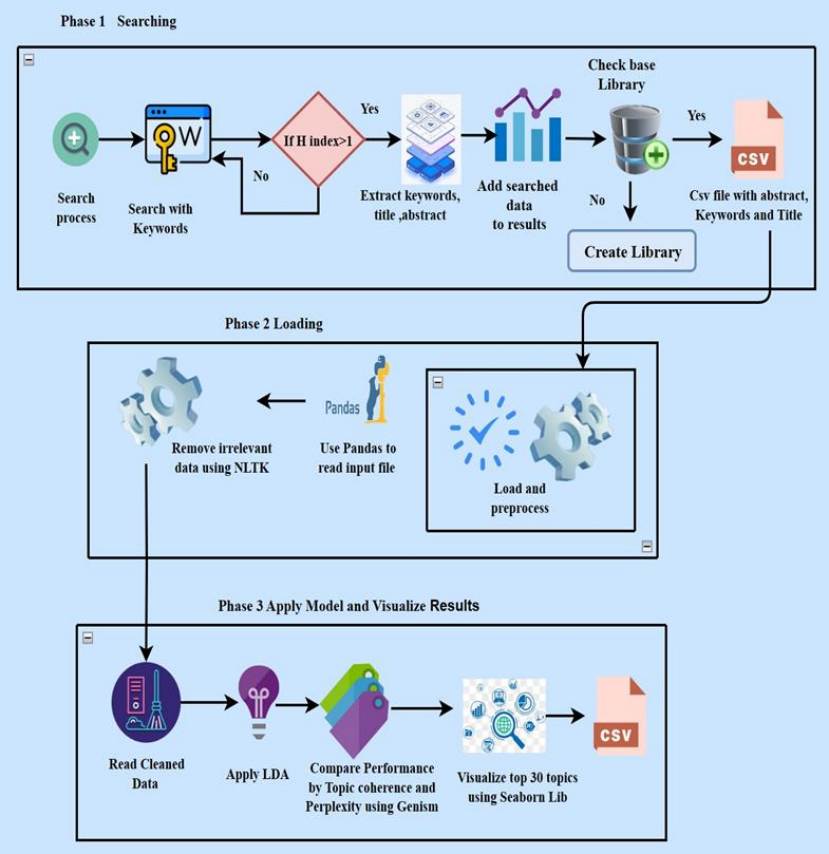

Fig. 5: Workflow of Methodology

1.6 Systematic Workflow of Methodology

Contemporary research points out that quantitative analysis by hand is impractical due to the large number of articles from which issues and challenges can be identified. Hence, in this context, we followed a systematic approach, starting with a few sets of keywords related to Hinglish code-mixed language on the theme of war and conflicts. A three-phase automation script was authored for gathering citations and associated material shown in Fig.5 and the algorithm steps are explained in the next section. A contemporary survey of those datasets and from the trained module hints that limited work has been copied out in the context of Hinglish.

2 Pseudo-Code for Systematic Methodology

The algorithm described below performs several functions, including managing directories and CSV files, performing an advanced search for academic papers using given keywords, sorting outcomes based on a specified h-index threshold, preliminary processing retrieved data by combining and cleaning it, using Latent Dirichlet Allocation (LDA) for topic modelling and visualisation, and finally presenting the results in a structured format.

Input Parameters
Keywords Array: keywords = [k1, k2, ..., kn] Base Directory Path: base_dir Citation Data CSV Filename: citation_csv Cleaned Data CSV Filename: cleaned_csv H-index Threshold: h_index_threshold LDA Parameters: lda_params
1. Initialize Variables
Create an empty list to store results: results = []
2. Enhanced Paper Search with Error Handling
For each keyword *k* **in keywords:** Construct advanced search queries using Boolean operators and domain-specific filters. **Error Handling:** Try scholarly search; if it fails, log the error and continue.
For each paper *p* **in search results:**
If *h-index*(*p*) ≥ *h_index_threshold*, extract: **Title (t) Keyword (k) Abstract (a)** Append (*t, k, a*) to results.
3. Directory and CSV Handling
Ensure base_dir exists; create if not. Create CSV file citation_csv in base_dir. Write data from results to citation_csv.
4. Data Preprocessing with Error Handling
Error Handling: Try loading data; if it fails, log the error and halt execution. Merge **Title, Keywords, and Abstract** into a new column **merged** in the DataFrame (df). Customize the stop word list according to the research domain and remove them from **Merged**. Remove special and illegal characters from **Merged**. Save the cleaned DataFrame as df_clean. Write df_clean to cleaned_csv.
5. LDA Topic Modeling
Apply **Latent Dirichlet Allocation (LDA)** on df_clean using lda_params. Calculate multiple **topic coherence metrics** alongside perplexity. **Visualize the top topics.** Return the DataFrame with LDA results and visualizations.

2.1 M E1 T A E2 D Framework of Identifying Gaps

After a Python script collected all the research papers [see table 1], it was time to review them to identify research gaps. Thus, a framework called [M.E1.T.A.E2.D] was created based on the following definitions of research gaps. [45].

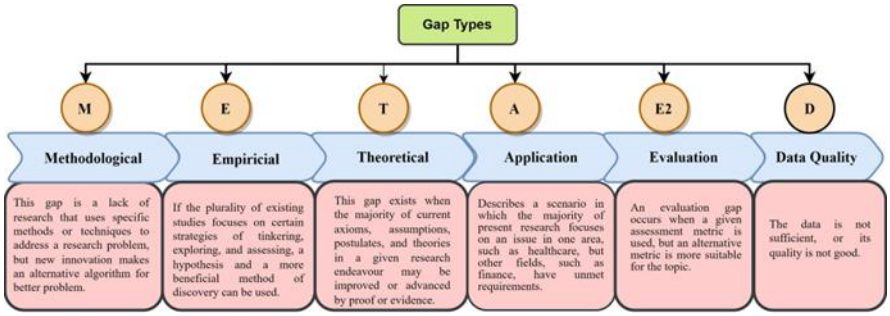

Fig. 6: Framework for Identifying Gaps

M = Methodological gaps occur when existing research uses, for example, a specific machine learning algorithm, but innovations make an alternate algorithm better for the problem.

T = Theoretical gaps If the majority of existing axioms, assumptions, postulates, and theories in particular research work can be improved or enhanced with proof or evidence, this is referred to as a theoretical gap or **T** [45].

A = Application gaps This refers to a situation in which the majority of the existing research addresses a problem in one particular domain, such as healthcare, while other domains, such as finance, do not have their needs adequately addressed.

D = Data Quality Gap In this gap, quality is not good, or data is not sufficient for better results.

E1 = Empirical gaps If the majority of existing research focuses on particular methods of tinkering, experimenting, and evaluating, a hypothesis and a more advantageous method or experiments can be suggested to the work.

E2 = Evaluation gaps If the majority of the previous research relied on a specific evaluation metric, such as accuracy, but a different evaluation metric, such as F1-score, might be more appropriate for the issue at hand.

3 Challenges in Hinglish Code-Mixed Language

In this section, testing of challenges in the domain of hinglish are mentioned**:**

1. **Identifying the Languages Used in the Text and Voice:** In hinglish languages, writing may contain multiple languages or dialects, making language identification a tiresome algorithmic issue. Code-mixed languages often lack clear boundaries between languages, so the text may contain words, phrases, or even whole sentences from other languages. Code-mixed languages may include dialects or geographical variances, complicating language identification e.g awadhi,baua bahut ladayiaa ha. Many code-mixed languages follow specific patterns and laws that are unique to the speakers' language and culture. Speakers' languages and cultures dictate these patterns and conventions. Due to patterns and conventions, statistical language recognition algorithms may have trouble distinguishing text language and culture [46][47].
2. **Intricate Morphological Structures of Languages:** These can make it challenging to correctly identify the language by examining the structural makeup of the words included inside the text. Additionally, there is frequently an absence of labelled data for code-mixed languages, which is true for hinglish data, which makes it challenging to train machine learning models for language identification. Increasing the amount of labelled data for code-mixed languages is one approach that might be used to address this difficulty. This can be accomplished through the use of existing data sources, such as social media and online forums, or by the crowdsourcing of data from native speakers. In addition, strategies such as data augmentation, transfer learning, and multi-task learning can be utilised to make use of the data that is available in order to train models, even when there is just a little quantity of data available[48].Future researchers may develop algorithms that recognise languages with limited data. Unsupervised methods like clustering and inadequate supervision, which uses a little amount of labelled data to train the model, can do this. In the absence of labelled data, language-independent variables like character n-grams or phonetic features can improve language recognition models.

Fig. 7: Identified Challenges in Hinglish Language

3. **Data Availability**: Many codemixed languages lack annotated data (Hinglish is no exception), making sentiment analysis machine learning model training problematic.Thus, code-mixed language data collection is urgently needed to tackle this issue. Second, data augmentation can increase the amount of labelled data available for training by adding tiny random changes to existing data [49]. Transfer Learning is another method that applies pre-trained models to codemixed language jobs. Fine-tuning the model with a target task-specific dataset does this. Multi-task learning trains a model to do related tasks like language identification and sentiment analysis. This technique allows tasks to share representations and features. "Learning via Semi-supervision" cannot develop natural language processing. When labelled data is scarce, this strategy trains the model using both labelled and unlabeled data. Weakly-supervised Learning works well too[50]. In this strategy, the algorithm trains the model using weak labels like hashtags or emoticons, which may be useful when there is no labelled data. Many modern academics have presented algorithms to handle the issue of limited labelled data in addition to these methods. Unsupervised learning methods like clustering or topic modelling can find data patterns to train a sentiment analysis model. These methods can find patterns. In addition, weakly supervised learning methods like distant supervision, in which a model is trained using a large amount of loosely labelled data, could improve sentiment analysis algorithms[51].Lastly, as mentioned, there are a limited number of annotated datasets with the theme of war and conflict. However, some of the datasets in this context that are useful are :

   - Hinglish Social Media Text Dataset: This dataset contains code-mixed social media text in Hinglish. It includes annotations for a variety of linguistic phenomena, including transliteration, phonetic and orthographic changes, and language transfers[14].
   - This [25] gives a parallel corpus of 13,738 code-mixed English-Hindi utterances and their English translation. The annotators do sentence translations by hand. It includes text from social media, Politics, social events, sports, among others.
   - [52] Detect hate speech in code-mixed texts and provide a Hindi-English code-mixed dataset comprised of tweets posted on Twitter. The tweets are annotated along with the language and class at the word level.
   - IIT Bombay Code-Mixed English-Hindi Corpus: The IIT Bombay English-Hindi Code-Mixed Corpus is an exhaustive corpus of code-mixed English and Hindi sentences compiled by researchers at IIT Bombay. It contains social media posts, translations from movies, and news headlines [53].

4. **Complex Long Sequence Recognition Problem**: There are a lot of code-mixed languages out there, and many of them have long repeating features that are hard to understand, which can make it challenging to correctly detect someone's feelings, raw emotions and sentiments. In hinglish also people do the same. This is because code-mixed languages frequently have intricate recursive and fractal structures, which can vary greatly between different languages. As a result, it is challenging to identify sentiment by analysing the structure of the words contained within a text because code-mixed languages are so prevalent[45]. Many codemixed languages lack labelled data, making sentiment anal- ysis machine learning model training problematic. Codemixed languages are employed in numerous circumstances, making this difficult. Thus, code-mixed language data collection is urgently needed to tackle this issue. Social media, internet forums,

and native speaker crowdsourcing can do this. Second, data augmentation can increase the amount of labelled data available for training by adding tiny random changes to existing data. Transfer Learning is another method that applies pre-trained models to codemixed language jobs. Fine-tuning the model with a target task-specific dataset does this. Multitasking Multi-task learning trains a model to do related tasks like language identification and sentiment analysis. This technique allows tasks to share representations and features. "Learning via semi-supervision" cannot develop natural language processing. When labelled data is scarce, this strategy trains the model using both labelled and unlabelled data. Weakly-supervised Learning works well, too. In this strategy, the algorithm trains the model using weak labels like hashtags or emoticons, which may be useful when there is no labelled data. Many modern academics have presented algorithms to handle the issue of limited labelled data in addition to these methods[54] [45].

5. **Domain Adaptation** Sentiment analysis models that have been trained on one domain may not perform well on another domain due to changes in the way languages are used and the cultural environment in which they are used. Before a solutions is developed to overcome this issue, it is essential to keep in mind that domain adaptation is a difficult problem, and that it may be necessary to use a variety of strategies in conjunction with one another in order to obtain good performance in sentiment analysis across a variety of domains[1]. Hence, here also, there is a need to cover methodological gaps in the current scenario, which include the development of transfer learning, multi-task learning and identification of new types of language-independent algorithms[54].

6. **Language-Dependent Resources** This exploration of contemporary research found that sentiment analysis models mostly use language-dependent resources like specific lexicons and word embeddings. However, not all languages have these resources, making sentiment analysis model training difficult. This is true about Hinglish Cross-lingual Sentiment Analysis is one of several algorithms and methodologies that could solve this problem and fulfil the new advance in this domain. A sentiment analysis model is trained in one language and then transferred to another. Bilingual lexicons or machine translation can help achieve this goal. This technique can leverage cross-language representations and traits. Word Embedding induction methods connect comparable words across languages to create word embeddings for a language without resources. Thus, this approach is needed to fix Hinglish code-mix language sentiment analysis. New algorithms to help sentiment algorithms manage code-switching and language-moving are needed [55][56].

4 RESULTS AND DISCUSSION

Table 3 presents key challenges associated with Hinglish code-mixed language. These challenges span various aspects, including linguistic complexities, data limitations, and the need for adaptable models. The cited studies highlight specific difficulties. According to the table, the issue that appears to require

Table 3: Challenges in Hinglish Code-Mixed Language

Challenges	Study
Intricate Morphological Structures Languages	[3] [28], [57], [48], [45][58]
Data Availability	[3] [59], [54], [12], [48], [60], [61], [62], [63] [58][43]
Domain Adaptation	[64], [3], [65], [66], [4][67][54]
Complex Long Sequence Recognition Problem	[64], [68], [69], [45], [70]
Language-Dependent Resources	[3], [71], [45][58] [43][67]

A key challenge in Hinglish code-mixed language research is the availability of data. Many studies highlight this issue, indicating its widespread recognition as a significant obstacle. While other challenges are important, data

availability stands out due to its frequent mention in research. Addressing this issue can greatly contribute to the advancement of Hinglish language processing. Future efforts should prioritize the creation of high-quality datasets, the development of adaptable models, and the establishment of specialized linguistic resources to enhance performance and scalability in practical applications.

A summary of the main research work shown in Table 4 in the context of identifying challenges in the code-mix domain was constructed with the assistance of the framework, which is abbreviated as [M.E1.T.A.E2.D.]. The following is a discussion of the outcomes of the methodology that is utilised in this section. Multiple research articles highlighted six significant research gaps in the code mix language domain. Code mix languages have these.

Table 4: Summary of Research Gaps in Existing Studies

Study	**Gaps in Study**					
	M	**T**	**A**	**D**	**E1**	**E2**
[3]	✓	✓	✓	✓		✓
[8]		✓	✓	✓		
[18]		✓	✓		✓	
[19]		✓	✓		✓	
[27]	✓			✓		
[45]		✓	✓	✓		
[72]	✓		✓			✓
[73]		✓	✓		✓	
[74]		✓	✓		✓	
[75]	✓			✓		✓
[76]		✓		✓		
[77]	✓		✓		✓	
[78]	✓		✓		✓	
[79]	✓			✓		
[64]				✓	✓	
[23]		✓	✓			✓
[80]	✓		✓			✓
[70]	✓	✓		✓		
[54]	✓	✓		✓		
[56]	✓			✓		
[51]	✓					✓
[71]	✓			✓	✓	
[65]	✓			✓	✓	
[66]			✓	✓		
[48]		✓	✓	✓		
[54]				✓		
[58]	✓			✓		
[81]	✓			✓	✓	
[43]	✓			✓	✓	
[67]	✓			✓	✓	

T = Theoretical gaps, A = Application gaps, D = Data Quality Gap, E1 = Empirical gaps, E2= Evaluation gaps

The analysis reveals that the most significant research gaps in existing studies are methodology and Data Quality of hinglish utterances, respectively.

This indicates a lack of well-defined methodologies and robust algorithms for processing Hinglish code-mixed text. Addressing these gaps requires the development of advanced hybrid models, improved annotation strategies, and the optimization of transformer-based architectures. Additionally, establishing benchmark datasets and standardized evaluation metrics can enhance the effectiveness of Hinglish NLP research, paving the way for more accurate and scalable solutions.

5 MAIN CONTRIBUTIONS

This research work provides a comprehensive review of the challenges and research gaps in sentiment analysis for code-mixed languages, with a specific focus on Hinglish (a mix of Hindi and English). The study highlights several significant contributions in conclusion:

Identification of major challenges: The work identifies six major challenges in Hinglish sentiment analysis, including language identification, complex morphological structures, data availability, long sequence recognition, domain adaptation, and language-dependent resources. These challenges are well-recognized in the field of code-mixed language understanding and are still relevant despite the development of large language models (LLMs). Systematic framework for gap analysis: The study proposes a framework called ”M.E1.T.A.E2.D” to systematically analyze research gaps across different dimensions, such as methodological, empirical, theoretical, application, evaluation, and data quality gaps. This structured approach can be valuable for future research in identifying and addressing gaps in code-mixed language understanding.

Comprehensive literature review: The work presents a thorough review of existing literature, analyzing over 50 research papers to identify specific research gaps across various aspects of code-mixed sentiment analysis. This extensive literature review can serve as a valuable resource for researchers in the field.

Novelty in the context of LLMs: Recent developments of large language models (LLMs) have improved natural language understanding; the challenges identified in this work remain relevant for code-mixed languages like Hinglish. LLMs are typically trained on monolingual data and may struggle with the complexities of code-mixing, such as language identification, morpho- logical variations, and domain adaptation. Further, it must be noted that the study was conducted before the recent advancements in LLMs, such as GPT-3, PaLM, and others. These powerful models, combined with techniques like few-shot learning and prompt engineering, may offer new opportunities to address some of the challenges and gaps identified in this work.

Therefore, in essence, this process of reviewing this domain provides a useful contribution to the field of code-mixed language understanding by systematically reviewing the challenges and research gaps in Hinglish sentiment analysis. The complexities of code-mixed languages like Hinglish require continued research and adaptation of LLM models as well to address the specific challenges identified in this work.”

6 CONCLUSION AND FUTURE SCOPE

In this work, the obstacles and concerns that develop when conducting sentiment analysis across code-mixed languages, particularly in the setting of Hinglish, are analysed. Specifically, the focus on the challenges and issues that arise while analysing Hinglish. According to the findings of this study, the difficulties that arise in the context of code-mixed languages can be broken down into several different problem areas. These problem areas include determining the languages that are being used in the text, dealing with complex morphological structures, a lack of labelled data, and the difficulty of training sentiment analysis models for different domains due to changes in language usage and cultural context.

This paper presents an exhaustive investigation. By accumulating linguistic sample data, research papers, scientific reports, and white papers, the research paper offers a deeper understanding of the sentiment analysis problem in the context of Hinglish code-mix languages. This exhaustive study guaranteed a thorough examination of the difficulties and issues associated with sentiment analysis for code-mixed languages. Insights into the complexities of sentiment analysis were pursued with the aid of an automated method (using Python code) for collecting research papers and applying statistical methodologies. The paper emphasised the originality and applicability of a variety of algorithms proposed to surmount the difficulties of code-mixed language

sentiment analysis. These algorithms consist of neural network-based algorithms, hybrid algorithms, rule-based algorithms, transfer learning, multi-task learning, cross-lingual sentiment analysis, multilingual sentiment analysis, language-independent characteristics, and word embedding induction.This domain is multifaceted and fraught with difficulties. The paper acknowledges and addresses the complex and multifaceted challenges that Hinglish code-mixed languages present in sentiment analysis. In recognizing the complexities of code-mixed language sentiment analysis, this paper offers a comprehensive perspective and provides valuable deep information for future scientific inquiry.

This research is incredibly valuable for the general public, particularly those who would like to employ tools for code-mixed language sentiment analysis. In addition, it is valuable for academics interested in examining the issues of code-mixed language sentiment analysis since it provides insight into how machine learning algorithms react to unexpected input or the interactions between multiple languages[49]. This research is also valuable for educators, media writers, and producers, as it allows them to make accessible news reports in colloquial language. The largest beneficiaries of this research will be firms that are creating automated translation technology since they wish to create universally functional automatic translation programmes.

Declarations

- Funding This research received no external funding.
- Ethics approval Not applicable.
- Consent to participate Not applicable.
- Availability of Supporting Data The data generated during the current study are available from the corresponding author on reasonable request.
- Competing Interest Not applicable.
- Author's contributions Pratibha and Amandeep Kaur conceived of the presented idea. Pratibha and Amandeep Kaur wrote the theory. Meenu Khurana and Amandeep Kaur verified the analytical methods. Pratibha and Meenu Khurana made diagrams and did visualization. All authors discussed the gaps and challenges,future scope, and contributed to the final manuscript.
- Acknowledgment: Not applicable.

REFERENCES

[1] Zhu, L., Zhu, Z., Zhang, C., Xu, Y., Kong, X.: Multimodal sentiment analysis based on fusion methods: A survey. Information Fusion **95**, 306– 325 (2023)

[2] Rajeshwari, A.: Shobhaa De Takes a Dig at Indian Olympians: HowTwitter Reacted. https://timesofindia.indiatimes.com/blogs/everything-social/shobhaa-de-takes-a-dig-at-indian-olympians-how-twitter-reacted/

[3] Messaoudi, C., Guessoum, Z., ben Romdhane, L.: A deep learning model for opinion mining in Twitter combining text and emojis. Procedia Computer Science **207**, 2628–2637 (2022)

[4] Wankhade, M., Rao, A.C.S., Kulkarni, C.: A survey on sentiment anal- ysis methods, applications, and challenges. Artificial Intelligence Review **55**(7), 5731–5780 (2022)

[5] Yang, J., Yang, Y., Xiu, L., Yu, G.: Effect of emoji prime on the understanding of emotional words–evidence from erps. Behaviour & information technology **41**(6), 1313–1322 (2022)

[6] Balahur, A., Hermida, J.M., Montoyo, A.: Detecting implicit expressions of emotion in text: A comparative analysis. Decision support systems **53**(4), 742–753 (2012)

[7] Salido Ortega, M.G., Rodr´ıguez, L.-F., Gutierrez-Garcia, J.O.: Towards emotion recognition from contextual information using machine learning. Journal of Ambient Intelligence and Humanized Computing **11**, 3187– 3207 (2020)

[8] Ekbal, A., *et al.*: Quality achhi hai (is good), satisfied! towards aspect based sentiment analysis in code-mixed language. Computer Speech & Language **89**, 101668 (2025)

[9] Lee, K.-P., Song, S.: Developing insights from the collective voice of target users in twitter. Journal of big Data **9**(1), 75 (2022)

[10] Park, S.-H., Bae, B.-C., Cheong, Y.-G.: Emotion recognition from text stories using an emotion embedding model. In: 2020 IEEE International Conference on Big Data and Smart Computing (BigComp), pp. 579–583 (2020). IEEE
[11] Dhiman, P., & Kaur, A.:Text pre-processing techniques in addressing fake news. In Computational Methods in Science and Technology (pp. 142-146). (2024). CRC Press
[12] Garg, N., Sharma, K.: Annotated corpus creation for sentiment analysis in code-mixed hindi-english (hinglish) social network data. Indian Journal of Science and Technology **13**(40), 4216–4224 (2020)
[13] Farkhod, A., Abdusalomov, A., Makhmudov, F., Cho, Y.I.: Lda-based topic modeling sentiment analysis using topic/document/sentence (tds) model. Applied Sciences **11**(23), 11091 (2021)
[14] Singh, K., Sen, I., Kumaraguru, P.: Language identification and named entity recognition in hinglish code mixed tweets, 52–58 (2018)
[15] Cui, J., Wang, Z., Ho, S.-B., Cambria, E.: Survey on sentiment analysis: evolution of research methods and topics. Artificial Intelligence Review **56**(8), 8469–8510 (2023)
[16] Mozafari, F., Tahayori, H.: Emotion detection by using similarity techniques. In: 2019 7th Iranian Joint Congress on Fuzzy and Intelligent Systems (CFIS), pp. 1–5 (2019). IEEE
[17] Basile, V., Bosco, C., Fersini, E., Nozza, D., Patti, V., Pardo, F.M.R., Rosso, P., Sanguinetti, M.: Semeval-2019 task 5: Multilingual detection of hate speech against immigrants and women in twitter. In: Proceedings of the 13th International Workshop on Semantic Evaluation, pp. 54–63 (2019)
[18] Plaza-del-Arco, F.M., Strapparava, C., Lopez, L.A.U., Mart´ın-Valdivia, M.T.: Emoevent: A multilingual emotion corpus based on different events. In: Proceedings of the 12th Language Resources and Evaluation Conference, pp. 1492–1498 (2020)
[19] Haryadi, D., Kusuma, G.P.: Emotion detection in text using nested long short-term memory. International Journal of Advanced Computer Science and Applications **10**(6) (2019)
[20] Aslam, N., Rustam, F., Lee, E., Washington, P.B., Ashraf, I.: Sentiment analysis and emotion detection on cryptocurrency related tweets using ensemble lstm-gru model. IEEE Access **10**, 39313–39324 (2022)
[21] Saha, T., Saha, S., Bhattacharyya, P.: Tweet act classification: A deep learning based classifier for recognizing speech acts in twitter. In: 2019 International Joint Conference on Neural Networks (IJCNN), pp. 1–8 (2019). IEEE
[22] Choudrie, J., Patil, S., Kotecha, K., Matta, N., Pappas, I.: Applying and understanding an advanced, novel deep learning approach: A covid 19, text based, emotions analysis study. Information Systems Frontiers **23**, 1431–1465 (2021)
[23] Guleria, A., Varshney, K., Pahwa, G., Singhal, S., Sharma, N.: Multimodal sentiment analysis of english and hinglish memes. Multimedia Tools and Applications, 1–26 (2024)
[24] Kratzwald, B., Ili´c, S., Kraus, M., Feuerriegel, S., Prendinger, H.: Deep learning for affective computing: Text-based emotion recognition in decision support. Decision Support Systems **115**, 24–35 (2018)
[25] Srivastava, V., Singh, M.: Phinc: A parallel hinglish social media code- mixed corpus for machine translation. arXiv preprint arXiv:2004.09447 (2020)
[26] Christian, H., Suhartono, D., Chowanda, A., Zamli, K.Z.: Text based personality prediction from multiple social media data sources using pre- trained language model and model averaging. Journal of Big Data **8**(1), 1–20 (2021)
[27] Joshi, R., Joshi, R.: Evaluating input representation for language identifi- cation in hindi-english code mixed text. arXiv preprint arXiv:2011.11263 (2020)
[28] Scherer, K.R.: The role of culture in emotion-antecedent appraisal. Journal of personality and social psychology **73**(5), 902 (1997)
[29] Srivastava, V., Singh, M.: Hinge: A dataset for generation and evaluation of code-mixed hinglish text. arXiv preprint arXiv:2107.03760 (2021)
[30] Poria, S., Hazarika, D., Majumder, N., Naik, G., Cambria, E., Mihalcea, R.: Meld: A multimodal multi-party dataset for emotion recognition in conversations. arXiv preprint arXiv:1810.02508 (2018)
[31] Buechel, S., Hahn, U.: Emobank: Studying the impact of annotation perspective and representation format on dimensional emotion analysis. arXiv preprint arXiv:2205.01996 (2022)
[32] Tan, K.L., Lee, C.P., Lim, K.M.: A survey of sentiment analysis: Approaches, datasets, and future research. Applied Sciences **13**(7), 4550 (2023)
[33] Baroi, S.J., Singh, N., Das, R., Singh, T.D.: Nits-hinglish-sentimix at semeval-2020 task 9: Sentiment analysis

for code-mixed social media text using an ensemble model. arXiv preprint arXiv:2007.12081 (2020)
[34] Mohammad, S.M., Bravo-Marquez, F.: Wassa-2017 shared task on emo- tion intensity. arXiv preprint arXiv:1708.03700 (2017)
[35] Zampieri, M., Nakov, P., Rosenthal, S., Atanasova, P., Karadzhov, G., Mubarak, H., Derczynski, L., Pitenis, Z., Ç ̈oltekin, Ç .: Semeval-2020 task 12: Multilingual offensive language identification in social media (offenseval 2020). arXiv preprint arXiv:2006.07235 (2020)
[36] Rane, A., Kumar, A.: Sentiment classification system of twitter data for us airline service analysis. In: 2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC), vol. 1, pp. 769–773 (2018). IEEE
[37] Russell, J.A., Mehrabian, A.: Evidence for a three-factor theory of emotions. Journal of research in Personality **11**(3), 273–294 (1977)
[38] Li, Y., Su, H., Shen, X., Li, W., Cao, Z., Niu, S.: Dailydialog: A manu- ally labelled multi-turn dialogue dataset. arXiv preprint arXiv:1710.03957 (2017)
[39] Ghazi, D., Inkpen, D., Szpakowicz, S.: Detecting emotion stimuli in emotion-bearing sentences. In: Computational Linguistics and Intelligent Text Processing: 16th International Conference, CICLing 2015, Cairo, Egypt, April 14-20, 2015, Proceedings, Part II 16, pp. 152–165 (2015). Springer
[40] Preo ̧tiuc-Pietro, D., Schwartz, H.A., Park, G., Eichstaedt, J., Kern, M., Ungar, L., Shulman, E.: Modelling valence and arousal in facebook posts. In: Proceedings of the 7th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis, pp. 9–15 (2016)
[41] Guleria, A., Varshney, K., Pahwa, G., Singhal, S., Sharma, N.: Multimodal sentiment analysis of english and hinglish memes. Multimedia Tools and Applications, 1–26 (2024)
[42] Agarwal, A., Gupta, J., Goel, R., Upadhyay, S., Joshi, P., Aravamudhan, R.: Cst5: Data augmentation for code-switched semantic parsing. arXiv preprint arXiv:2211.07514 (2022)
[43] Patra, B.G., Das, D., Das, A.: Sentiment analysis of code-mixed indian languages: An overview of sail code-mixed shared task@ icon-2017. arXiv preprint arXiv:1803.06745 (2018)
[44] Wadhawan, A., Aggarwal, A.: Towards emotion recognition in hindi- english code-mixed data: A transformer based approach. arXiv preprint arXiv:2102.09943 (2021)
[45] Hussein, D.M.E.-D.M.: A survey on sentiment analysis challenges. Journal of King Saud University-Engineering Sciences **30**(4), 330–338 (2018)
[46] Kaur, G., Pratibha, Kaur, A., Khurana, M., *et al.*: A review of opinion mining techniques. ECS Transactions **107**(1), 10125 (2022)
[47] Pratibha, G. Kaur, Kaur, A., Khurana, M., *et al.*: A stem to stern sentiment analysis emotion detection. In: 2022 10th International Confer- ence on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO), pp. 1–5 (2022). IEEE
[48] Pratibha, Kaur, A., Khurana, M., Dama ̆sevi ̆cius, R.: Multimodal hinglish tweet dataset for deep pragmatic analysis. Data **9**(2), 38 (2024)
[49] Choudhary, N., Singh, R., Bindlish, I., Shrivastava, M.: Sentiment anal- ysis of code-mixed languages leveraging resource rich languages, 104–114 (2023). Springer
[50] Shanmugavadivel, K., Sathishkumar, V., Raja, S., Lingaiah, T.B., Nee- lakandan, S., Subramanian, M.: Deep learning based sentiment analysis and offensive language identification on multilingual code-mixed data. Scientific Reports **12**(1), 21557 (2022)
[51] Pratapa, A., Choudhury, M., Sitaram, S.: Word embeddings for code- mixed language processing. In: Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, pp. 3067–3072 (2018)
[52] 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, 36–41 (2018)
[53] Kunchukuttan, A., Mehta, P., Bhattacharyya, P.: The iit bombay english- hindi parallel corpus. arXiv preprint arXiv:1710.02855 (2017)
[54] Ghosh, S., Priyankar, A., Ekbal, A., Bhattacharyya, P.: Multitasking of sentiment detection and emotion recognition in code-mixed hinglish data. Knowledge-Based Systems **260**, 110182 (2023)
[55] Kokab, S.T., Asghar, S., Naz, S.: Transformer-based deep learning models for the sentiment analysis of social media data. Array **14**, 100157 (2022)

[56] Pratapa, A., Bhat, G., Choudhury, M., Sitaram, S., Dandapat, S., Bali, K.: Language modeling for code-mixing: The role of linguistic theory based synthetic data. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 1543–1553 (2018)
[57] Singh, L.G., Singh, S.R.: Empirical study of sentiment analysis tools and techniques on societal topics. Journal of Intelligent Information Systems **56**, 379–407 (2021)
[58] Mangla, A., Bansal, R.K., Bansal, S.: Code-mixing and code-switching on social media text: A brief survey. In: 2023 IEEE International Conference on Computer Vision and Machine Intelligence (CVMI), pp. 1–5 (2023). IEEE
[59] Thara, S., Poornachandran, P.: Transformer-based language identification for Malayalam-english code-mixed text. IEEE Access **9**, 118837–118850 (2021)
[60] Suresh, T., Sengupta, A., Akhtar, M.S., Chakraborty, T.: A comprehensive understanding of code-mixed language semantics using hierarchical transformer. IEEE Transactions on Computational Social Systems (2024)
[61] Perera, A., Caldera, A.: Sentiment analysis of code-mixed text: a comprehensive review. Journal of Universal Computer Science **30**(2), 242 (2024)
[62] Thara, S., Poornachandran, P.: Social media text analytics of malayalam– english code-mixed using deep learning. Journal of big Data **9**(1), 45 (2022)
[63] Singh, N.K., Madal, W., Devi, C.N., Pangsatabam, H., Chanu, Y.J.: Leveraging synthetic data for improved manipuri-english code-switched asr. IEEE Access (2025)
[64] Pota, M., Ventura, M., Fujita, H., Esposito, M.: Multilingual evaluation of pre-processing for bert-based sentiment analysis of tweets. Expert Systems
with Applications **181**, 115119 (2021)
[65] Agarwal, V., Rao, P., Jayagopi, D.B.: Towards code-mixed hinglish dialogue generation. In: Proceedings of the 3rd Workshop on Natural Language Processing for Conversational AI, pp. 271–280 (2021)
[66] Srivastava, V., Singh, M.: Phinc: A parallel hinglish social media code- mixed corpus for machine translation. arXiv preprint arXiv:2004.09447 (2020)
[67] Mundra, S., Mittal, N.: Fa-net: fused attention-based network for hindi english code-mixed offensive text classification. Social Network Analysis and Mining **12**(1), 100 (2022)
[68] Chowanda, A., Sutoyo, R., Tanachutiwat, S., *et al.*: Exploring text- based emotions recognition machine learning techniques on social media conversation. Procedia Computer Science **179**, 821–828 (2021)
[69] Mahmud, T., Ptaszynski, M., Masui, F.: Exhaustive study into machine learning and deep learning methods for multilingual cyberbullying detection in bangla and Chittagonian texts. Electronics **13**(9), 1677 (2024)
[70] Kokab, S.T., Asghar, S., Naz, S.: Transformer-based deep learning models for the sentiment analysis of social media data. Array **14**, 100157 (2022)
[71] Singh, K., Sen, I., Kumaraguru, P.: Language identification and named entity recognition in hinglish code mixed tweets. In: Proceedings of ACL 2018, Student Research Workshop, pp. 52–58 (2018)
[72] Shim, J.-S., Lee, Y., Ahn, H.: A link2vec-based fake news detection model using web search results. Expert Systems with Applications **184**, 115491 (2021)
[73] Seal, D., Roy, U.K., Basak, R.: Sentence-level emotion detection from text based on semantic rules. In: Information and Communication Technology for Sustainable Development: Proceedings of ICT4SD 2018, pp. 423–430 (2020). Springer
[74] Maity, K., Saha, S., Bhattacharyya, P.: Emoji, sentiment and emo tion aided cyberbullying detection in hinglish. IEEE Transactions on Computational Social Systems (2022)
[75] Elhenawy, I.M.M.: Bert-cnn: A deep learning model for detecting emotions from text. Tech Science Press **71**, 2943–2961 (2021)
[76] Zeng, X., Chen, Q., Chen, S., Zuo, J.: Emotion label enhancement via emotion wheel and lexicon. Mathematical Problems in Engineering **2021**,
Towards Effective Hinglish NLP: Addressing Challenges and Research Gaps 27
1–11 (2021)
[77] Agarwal, B., Mittal, N., Bansal, P., Garg, S.: Sentiment analysis using common-sense and context information. Computational intelligence and neuroscience **2015**, 30–30 (2015)
[78] Tripathy, J.K., Chakkaravarthy, S.S., Satapathy, S.C., Sahoo, M., Vaidehi, V.: Albert-based fine-tuning model for cyberbullying analysis. Multimedia Systems **28**(6), 1941–1949 (2022)

[79] Seewann, L., Verwiebe, R., Buder, C., Fritsch, N.-S.: “broadcast your gender.” a comparison of four text-based classification methods of german youtube channels. Frontiers in big Data, 83 (2022)

[80] Zhang, X., Zhang, L.: Topics extraction in incremental short texts based on lstm. Social Network Analysis and Mining **10**(1), 83 (2020)

[81] Patil, A., Patwardhan, V., Phaltankar, A., Takawane, G., Joshi, R.: Comparative study of pre-trained bert models for code-mixed hindi-english data. In: 2023 IEEE 8th International Conference for Convergence in Technology (I2CT), pp. 1–7 (2023). IEEE