

Investigating the Phenomenon of Code-Switching through Computational Linguistics Techniques

Muhammad Asharib

National College of Business

Administration & Economics,

Email: muhammadasharib879@gmail.com

Memona Naeem Hashmi

National College of Business

Administration & Economics,

Email: memonahashmi007@gmail.com

Ayesha Asim

National College of Business

Administration & Economics,

Email: ayeshasim0803@gmail.com

Irsa Waseem

National College of Business

Administration & Economics,

Email: shahidmalik082@gmail.com

Aqsa Malik

National College of Business

Administration & Economics,

Email: Irsafahad0@gmail.com

Shahid Hameed

National College of Business

Administration & Economics,

Email: aqsamalik0955@gmail.com

ABSTRACT

Code-switching, the use of two or more languages within a single discussion or utterance is a widespread feature of multilingual communication, mainly in digital environments such as social media and messaging platforms. While sociolinguistic studies have provided a valuable insight into the phenomenon, it remains a complex challenge for computational modeling, especially when dealing with informal structures, grammatical variations, and underrepresented language pairs such as Urdu-English and Hindi-English. This study investigates the phenomenon of code-switching using computational techniques to improved analyze and interpret multilingual language behavior in actual contexts. The research applies advanced

Natural Language Processing (NLP) tools including language credentials, part-of-speech tagging, and syntactic parsing on data composed from online interactions. Using annotated corpora and models such as BiLSTM and multilingual BERT (mBERT), the study assesses model correctness and identifies core limitations in processing code-switched text. Furthermore, the study proposes improvements through context-aware strategies that assimilate linguistic insights with machine learning. By connecting computational techniques with language theory, the research contributes to the growth of more inclusive and accurate multilingual NLP tools, with potential applications in sentiment analysis, speech recognition, and machine translation. It also enhances value to the educational and technological landscape of linguistically diverse regions like South Asia.

Keywords: Code-Switching; Language context; Computational linguistic; Digital humanities; Natural language processing.

Introduction

Language serves not only as a tool for communication but also as a social and cognitive marker of identity, community, and intention. One of the most visible expressions of multilingual competence is code-switching the practice of alternating between two or more languages or language varieties within a single interaction. According to Poplack (1980), code-switching is a “systematic and rule-governed phenomenon” that reflects both linguistic competence and social context. It is not merely a matter of linguistic confusion or deficiency, as was once assumed, but rather a sophisticated strategy that serves various communicative functions such as identity signaling, emphasis, topic shift, and discourse alignment (Myers-Scotton, 1993; Gumperz, 1982).

Traditionally, code-switching has been studied within the domain of sociolinguistics, using qualitative and discourse-analytic approaches. However, the rise of digital communication, social media, and multilingual datasets has led to a renewed interest in this phenomenon from a computational perspective. Researchers have begun to recognize the importance of modeling and processing code-switching for real-world applications in Natural Language Processing (NLP) and Machine Learning (ML) (Solorio et al., 2014). These technologies are now being employed to perform language identification, part-of-speech tagging, syntactic parsing, and even machine translation of code-switched texts.

The integration of computational techniques has significantly advanced the study of codeswitching. Barman et al. (2014) emphasize that large-scale computational analysis enables the extraction of patterns from massive multilingual corpora, revealing insights that are not easily accessible through traditional methods. For example, models trained on code-switched corpora can identify frequent switch points, predict switching behavior, and detect grammatical patterns unique to mixed-language utterances. Furthermore, with the emergence of transformer-based models such as BERT and multilingual BERT (Devlin et al., 2019), there is growing potential for improving the semantic and syntactic understanding of code-switched content.

Despite these advances, code-switching poses complex challenges for computational systems. These include language ambiguity, grammatical inconsistency, data sparsity,

and the lack of annotated resources for many language pairs (Molina et al., 2016). The informal and contextsensitive nature of code-switching makes it difficult to generalize across linguistic communities, particularly in under-resourced languages such as Urdu-English or ArabicFrench. As Sridhar and Sridhar (1980) noted, the underlying grammar of code-switching often involves dual linguistic competence, which computational models must be trained to handle.

This research investigates the application of computational techniques in analyzing, simulating, and processing code-switched conversations. It focuses on existing approaches such as sequence tagging, boundary detection, and contextual language modeling in multilingual contexts (Aguilar et al., 2018). Moreover, it explores the implications for downstream applications including machine translation, sentiment analysis, and speech recognition, all of which require the ability to handle code-switched input effectively. By bridging the theoretical insights of sociolinguistics with the analytical power of computational linguistics, this study contributes to the development of inclusive and contextaware NLP systems. As Joshi (1982) remarked, understanding bilingual behavior like codeswitching not only advances linguistic theory but also enhances the realism and accessibility of language technologies in today's globalized world.

Background of the Study

In today's increasingly interconnected and multilingual world, the ability to move fluidly between languages commonly known as code-switching has become a routine aspect of communication among bilingual and multilingual speakers. This phenomenon involves alternating between two or more languages within a single utterance, conversation, or discourse context. As observed by Gumperz (1982), code-switching is a natural outcome of language contact and serves multiple communicative functions, including clarification, emphasis, and identity construction. It is not merely an act of language mixing but rather a strategic linguistic tool used by speakers to navigate different social roles, cultural affiliations, and pragmatic intentions (Myers-Scotton, 1993).

Historically, code-switching has been studied extensively in the fields of sociolinguistics and psycholinguistics, where the focus has been on its grammatical constraints (Poplack, 1980), sociocultural motivations (Blom & Gumperz, 1972), and cognitive implications (Sridhar & Sridhar, 1980). These studies have provided significant insights into the structural and functional dimensions of bilingual speech. However, with the rise of digital communication platforms such as social media, messaging apps, and online forums, code-switching has become more frequent, spontaneous, and contextually diverse than ever before.

This digital shift has created new challenges and opportunities for computational linguistics. Traditional Natural Language Processing (NLP) tools most of which are designed for monolingual, grammatically consistent input struggle to process code-switched content effectively (Solorio et al., 2014). Code-switching introduces problems such as intra-sentential language shifts, lexical borrowing, and non-standard grammar, all of which complicate tasks like language identification, POS tagging, and machine translation. As Barman et al. (2014) noted, standard NLP pipelines often fail when applied to code-switched data, due to the lack of labeled corpora, language

ambiguity, and resource constraints, particularly in low-resource language pairs like Urdu-English or Bengali-Hindi.

Recent advances in computational methods such as multilingual embeddings, transformer-based models (e.g., mBERT), and neural tagging frameworks have enabled more nuanced analysis and simulation of code-switched data (Aguilar et al., 2018; Devlin et al., 2019). These models have shown promise in recognizing language boundaries, adapting to mixed syntax, and learning from multilingual inputs. However, many of these tools still rely on datasets that are either too small, domain-limited, or biased toward high-resource languages, which restricts their applicability in diverse real-world settings.

Given this context, there is a growing need to explore how computational techniques can be improved and adapted to study code-switching more effectively. This study situates itself at the intersection of sociolinguistics and computational linguistics, aiming to bridge theoretical understanding with practical modeling approaches. By focusing on real-world code-switched conversations, especially in under-represented language contexts, this research seeks to address both the linguistic complexity and technical challenges associated with multilingual data processing.

Statement of the Problem

The increasing prevalence of code-switching in spoken and written communication particularly on digital platforms such as social media, instant messaging, and online forums presents both opportunities and challenges for linguistic analysis and Natural Language Processing (NLP). While code-switching is a natural and meaningful linguistic behavior among multilingual speakers, it poses serious complications for existing computational systems, which are predominantly designed to handle monolingual, grammatically consistent input.

One of the central problems is that current computational tools and language models perform poorly when applied to code-switched data. Standard NLP tasks such as language identification, part-of-speech tagging, syntactic parsing, named entity recognition, and machine translation often fail when encountering intra-sentential language switches, mixed grammar structures, and non-standard vocabulary. This is especially true for low-resource language pairs such as Urdu-English, Hindi-English, or Arabic-French, where the lack of annotated corpora, lexical resources, and pre-trained models further restricts accurate processing (Solorio et al., 2014; Barman et al., 2014).

Moreover, most of the available computational studies on code-switching are domain-specific, language-limited, or based on synthetic datasets that do not reflect real-world usage patterns. These limitations hinder the generalizability and scalability of code-switching models across different linguistic communities and technological applications. As a result, many widely used NLP systems including speech recognition software, translation engines, and chatbots struggle to perform effectively in multilingual environments where codeswitching is common.

In this context, there is a pressing need for systematic research that not only addresses the technical limitations of existing models but also incorporates linguistic theory and social context to better understand and process code-switched communication. Bridging this gap is essential to building inclusive, multilingual NLP systems capable

of handling the diversity of language use in contemporary digital interactions.

Therefore, this study aims to explore and evaluate computational approaches to analyzing and processing code-switched conversations, with a particular focus on the development and adaptation of models that can handle mixed-language input, support multilingual understanding, and improve performance in practical applications such as speech processing, social media analytics, and cross-lingual information retrieval

Objectives

The primary objective of this study is to investigate how computational techniques can be applied to analyze, simulate, and process code-switched conversations effectively. To achieve this goal, the study will pursue the following specific objectives:

To examine the linguistic patterns and structural characteristics of code-switching in bilingual and multilingual conversations, particularly within real-world communication contexts (e.g., social media, informal speech).

To explore existing computational approaches (such as Natural Language Processing, Machine Learning, and Speech Recognition techniques) used for the detection, classification, and analysis of code-switched data.

To evaluate the performance and limitations of current NLP models in handling codeswitched texts, especially in low-resource language pairs like Urdu-English or HindiEnglish.

To develop or adapt computational models and algorithms that can identify language boundaries, perform syntactic parsing, and support part-of-speech tagging in codeswitched data.

To create or utilize annotated datasets and corpora for training and testing computational models designed for code-switching scenarios.

To assess the practical implications of computational code-switching analysis for downstream applications such as Machine translation, Sentiment analysis, Named entity recognition, Automatic speech recognition (ASR)

To bridge the gap between sociolinguistic theory and computational modeling by incorporating linguistic insights into the design of more context-aware and inclusive multilingual systems

Research Questions

What are the common structural patterns and linguistic features of code-switching in bilingual or multilingual conversations, particularly in informal and digital communication contexts?

Which computational techniques (e.g., NLP, machine learning, speech processing) have been developed or adapted to detect and analyze code-switched language data?

How accurately do existing computational models perform in identifying and processing code-switched texts across different language pairs, such as Urdu-English or Hindi-English?

What are the main challenges and limitations faced by current computational systems when dealing with code-switched data (e.g., language ambiguity, resource scarcity, grammatical inconsistency)?

How can annotated code-switched corpora and other linguistic resources be developed or improved to support more robust model training and evaluation?
What role can linguistic theory play in enhancing the design and performance of computational models used to process code-switching?
How can improved computational models for code-switching benefit real-world applications such as machine translation, sentiment analysis, and automatic speech recognition in multilingual environments?

Significance of the Study

In an increasingly multilingual and digitally connected world, the ability to understand and process code-switched language is becoming vital for both linguistic research and technological development. This study is significant for several reasons, as it lies at the intersection of sociolinguistics, computational linguistics, and artificial intelligence.

First, the study contributes to the theoretical understanding of code-switching by examining its linguistic features through a computational lens. While code-switching has been extensively studied in sociolinguistics, there remains a gap in applying those insights to computational models that can analyze and simulate this phenomenon at scale. By bridging this gap, the study will offer a more comprehensive understanding of how multilingual speakers use language in natural conversation, especially in real-world, informal settings.

Second, the research has practical importance in the field of Natural Language Processing (NLP). Most existing NLP tools and models are designed for monolingual inputs and perform poorly when applied to code-switched data. This study will evaluate and improve the performance of such models, making them more adaptable to mixed-language inputs. This advancement is crucial for enhancing the effectiveness of applications such as:

Machine translation for multilingual users

Speech recognition systems in multilingual regions

Social media monitoring tools that analyze user sentiment or intent

Conversational AI and chatbots that interact with code-switching users

Third, the study addresses the needs of underrepresented and low-resource language communities, such as those that code-switch between Urdu and English, Hindi and English, or Arabic and French. By focusing on these contexts, the research promotes linguistic inclusivity and supports the development of language technologies that serve a broader, more diverse population.

Lastly, the outcomes of this research can guide the development of annotated corpora, language models, and linguistic resources that can be shared with the academic and developer communities. These resources will facilitate further research in computational linguistics and multilingual NLP, helping to close the digital divide between high-resource and low-resource languages.

In sum, this study holds theoretical, technological, and societal relevance. It not only enhances our understanding of bilingual behavior but also contributes to the design of context-aware, inclusive, and intelligent language processing systems in an increasingly globalized linguistic landscape.

Literature Review

The phenomenon of code-switching the alternation between two or more languages within a single discourse is well-documented in sociolinguistic and psycholinguistic literature. Early studies laid the groundwork for understanding its structure, motivation, and social functions, while more recent works have explored its implications and challenges for computational processing.

Sociolinguistic Foundations of Code-Switching

The study of code-switching began as part of a broader interest in bilingualism and multilingual communication. Blom and Gumperz (1972) introduced the concepts of situational and metaphorical code-switching, highlighting how speakers switch codes based on social settings or to achieve specific conversational goals. Gumperz (1982) further emphasized the contextualization cues embedded in code-switching, showing its importance for signaling speaker identity, solidarity, or social distance.

Poplack (1980) proposed a syntactic typology of code-switching, identifying inter-sentential, intra-sentential, and tag switching. She argued that fluent bilinguals adhere to grammatical rules of both languages when switching, suggesting that code-switching is systematic and rule-governed. Myers-Scotton (1993) introduced the Markedness Model, proposing that speakers make conscious or unconscious choices to switch codes based on what is socially “unmarked” or expected in a given situation.

Cognitive and Psycholinguistic Perspectives

From a cognitive standpoint, Sridhar and Sridhar (1980) suggested that code-switching reflects the cognitive flexibility of bilinguals who can access and control two linguistic systems. Research in this domain has shown that bilinguals are capable of fast and efficient language switching without loss of fluency or comprehension, especially in informal and emotionally charged contexts.

Emergence of Computational Approaches

The past decade has seen increasing interest in applying computational methods to analyze code-switching, particularly due to the rise of social media, WhatsApp chats, and multilingual voice assistants. Early computational studies focused on language identification, with efforts to label each token (word) in code-switched text with its corresponding language. Solorio et al. (2014) launched the first shared task on language identification in codeswitched data, highlighting the technical difficulties posed by short, informal, and highly mixed language inputs.

Barman et al. (2014) explored code-mixing challenges in the context of Bengali-HindiEnglish tweets. They emphasized the lack of annotated datasets and inconsistencies in script usage (e.g., Romanized vs. native script), which complicate both classification and feature extraction. Their work underlined the need for domain-specific corpora and adaptable algorithms.

NLP Techniques for Code-Switched Texts

Advanced Natural Language Processing (NLP) techniques have been increasingly applied to process code-switched data. Studies have used sequence labeling,

conditional random fields (CRF), and more recently, deep learning models such as LSTMs and transformers (e.g., BERT, mBERT). Aguilar et al. (2018) demonstrated the effectiveness of multi-task learning for Named Entity Recognition (NER) on code-switched social media data. Similarly, Winata et al. (2019) applied neural language models to detect language boundaries in code-switched utterances, showing improved results with multilingual pre-training.

The study by Janjua et al. (2024) explores advancements in contextual understanding for chatbots and NLP, shedding light on techniques that enhance interaction. Rehman et al. (2024) focus on machine learning applications for classifying lung diseases, offering insights into classification systems that could be adapted to multilingual contexts. Janjua, Ali, and Chaudhry (2016) discuss change management within hybrid agile models, providing a framework that can influence the evolution of NLP systems. Almansour et al. (2024) investigate the application of machine learning and rule induction across multiple sectors, a methodology that holds promise for multilingual processing. Abbas et al. (2023) propose an agricultural IoT-based intelligent system for disease forecasting, which integrates various computational tools that may enhance the future of multilingual NLP applications.

Challenges and Gaps in Current Research

Despite significant progress, several gaps persist in the literature:

Lack of annotated code-switched corpora, especially for South Asian and African languages.

Insufficient attention to informal, spontaneous, and oral code-switching, which is more complex than scripted or edited text.

Limited integration of linguistic theory into the design of computational models.

Underrepresentation of low-resource language pairs, which are common in postcolonial multilingual societies.

These limitations point to the need for a more interdisciplinary approach that combines linguistic insight with data-driven modeling, a gap this current study seeks to address.

Methodology

“Code-Switching in Conversations Using Computational Techniques” structured according to academic research standards and aligned with your objectives and research questions.

Research Methodology

Research Design

This study adopts a mixed-methods design combining both qualitative linguistic analysis and quantitative computational modeling. The qualitative component involves identifying the structural and functional patterns of code-switching in conversational data, while the quantitative component employs Natural Language Processing (NLP) and machine learning techniques to detect, analyze, and evaluate code-switched utterances computationally.

Data Collection

Corpus Selection

The research will utilize real-world conversational datasets containing code-switched language, focusing particularly on Urdu-English and Hindi-English language pairs. Possible sources include:

Social media platforms (e.g., Twitter, Facebook)

WhatsApp or SMS conversations (with consent and anonymization)

Publicly available corpora (e.g., ICON 2015, LinCE, SEAME)

Annotation and Preprocessing

The datasets will be manually or semi-automatically annotated at the word level for language identification (LID), part-of-speech (POS) tags, and syntactic structure.

Data will be preprocessed to handle spelling variations, Romanized script, emojis, and abbreviations.

Annotation tools such as ELAN, BRAT, or Prodigy may be used.

Tools and Techniques

Language Identification (LID)

Apply supervised models such as Conditional Random Fields (CRF) or Support Vector Machines (SVM).

Deep learning models like BiLSTM-CRF and transformer-based models (e.g., mBERT, XLM-R) will be trained to identify language at the token level.

Syntactic and Semantic Analysis

Use NLP libraries (e.g., spaCy, Stanza, NLTK) for tokenization, POS tagging, and dependency parsing.

Modify syntactic parsers to accommodate mixed grammar structures.

Named Entity Recognition (NER) and Sentiment Analysis

Apply multi-task learning approaches to perform both LID and NER simultaneously (as demonstrated by Aguilar et al., 2018).

Analyze the effectiveness of existing sentiment models on code-switched data.

Evaluation Metrics

Model performance will be evaluated using:

Accuracy, Precision, Recall, and F1-score for classification tasks (e.g., LID, POS tagging)

BLEU or ROUGE scores for translation tasks (if applicable)

Confusion matrices and error analysis for qualitative insight

Model Training and Testing

Data will be split into training (70%), validation (15%), and test (15%) sets.

Cross-validation will be applied to ensure reliability.

Experiments will be conducted using platforms such as Google Colab, Jupyter Notebook, or PyTorch/TensorFlow environments.

Ethical Considerations

Data collected from personal conversations or social media will be anonymized and used with informed consent, where applicable.

The study complies with ethical research guidelines on data privacy, participant anonymity, and responsible use of public data.

Limitations of Methodology

Availability of high-quality, annotated code-switched corpora is limited, especially for low-resource language pairs.

Romanized and non-standard spellings may reduce model accuracy.

Informal, conversational tone makes syntactic parsing and semantic labeling more challenging.

Data Collection and Analysis Process

The success of this study depends on the availability and quality of data that reflects authentic instances of code-switching. Therefore, the data collection process is designed to source natural, spontaneous, and context-rich code-switched conversations, especially between Urdu-English or Hindi-English language pairs.

Data Sources

Social Media Platforms: Publicly available posts, comments, and replies from platforms like Twitter, Facebook, and YouTube, where code-switching is naturally used.

WhatsApp Chats / Informal Messaging Data: With informed consent and anonymization, conversational data may be collected from bilingual users who frequently switch between languages.

Publicly Available Corpora: Datasets such as LINCE (Linguistic Code-Switching Evaluation Corpus, ICON Code-Mixed Corpora (2015, 2016, SEAME Corpus (for speech-based code-switching)

Data Annotation

Token-level Language Identification: Each word will be labeled with its respective language (e.g., URD for Urdu, ENG for English).

Part-of-Speech Tagging (POS): Using modified taggers trained on bilingual datasets to account for non-standard grammar in code-switched content.

Syntactic/Discourse Features: Additional labels may include switch points, discourse markers, or sentence-level functions of code-switching (e.g., emphasis, clarification).

Ethical Considerations

Only public or consent-based data will be used.

All personal identifiers will be removed to ensure anonymity and confidentiality.

The study follows research ethics guidelines under data protection and responsible digital scholarship.

Online ISSN

3007-3197

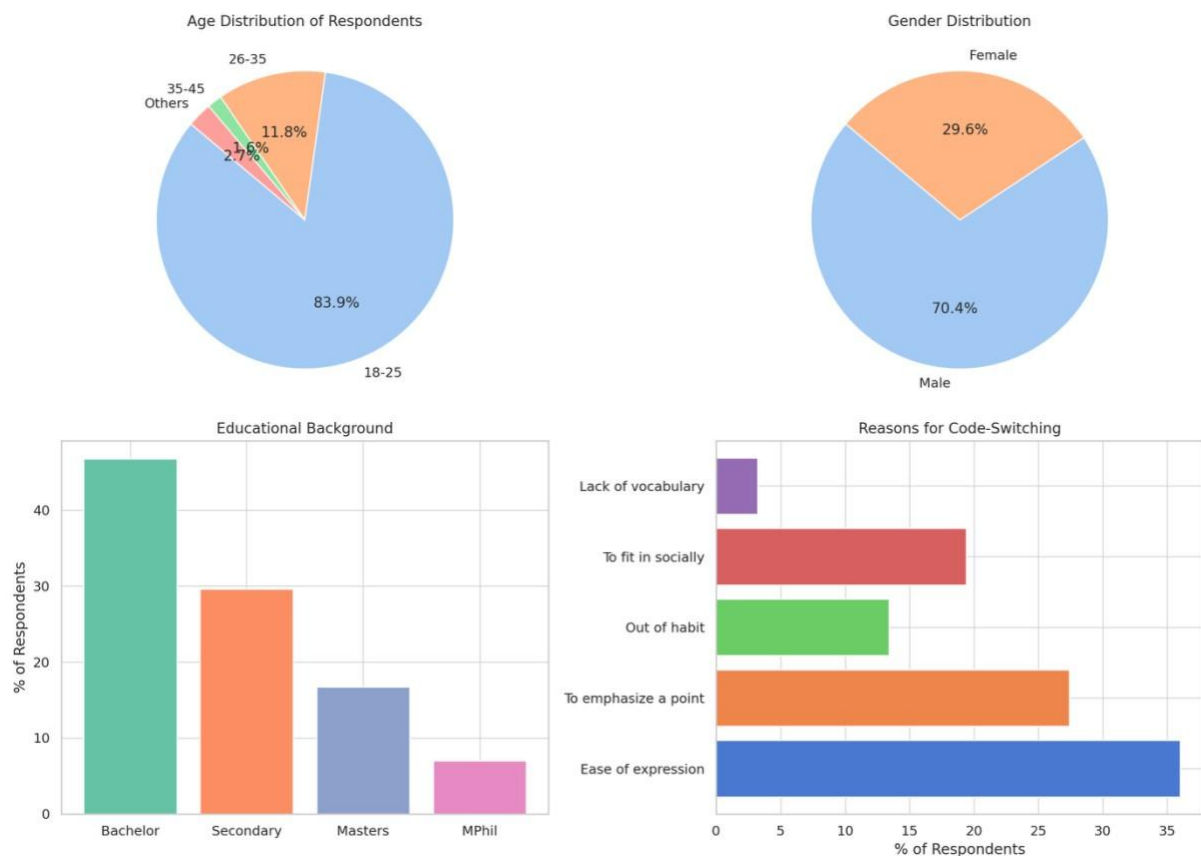
Print ISSN

3007-3189

<http://amresearchreview.com/index.php/Journal/about>

Data Analysis

The analysis process will combine linguistic feature extraction with computational modeling techniques, enabling both qualitative understanding and quantitative evaluation of code switched discourse.



Computational Techniques

Task	Technique / Tool Used
Language Identification	CRF, BiLSTM, or mBERT fine-tuned on token-level code
(LID)	Switched data
POS Tagging	Adapted NLP tools (e.g., spaCy, Stanza, NLTK)

Named Entity Recognition Multi-task learning using transformer models

Syntactic Parsing Dependency parsers fine-tuned for mixed grammar Sentiment

Analysis Pre-trained multilingual sentiment models

Experimental Setup

Data will be split into Training (70%), Validation (15%), and Test (15%) sets. Machine learning models will be trained using Python libraries such as scikit-learn, Keras, PyTorch, or Transformers (HuggingFace). Experiments will be run in Google Colab or Jupyter Notebooks for transparency and reproducibility.

Evaluation Metrics

To evaluate the performance of computational models:

Task	Evaluation Metric
Language Identification	Accuracy, Precision, Recall, F1-Score
POS Tagging & NER	Precision, Recall, F1-Score
Sentiment Classification	Accuracy, Confusion Matrix Parsing/Translation (if applied) BLEU / ROUGE scores

Error Analysis

After model evaluation, the study will include manual error analysis to identify:
Misclassified tokens
Incorrect language switches
Limitations due to informal text features like emojis, abbreviations, and Romanized script

Summary

The combined data collection and analysis process will ensure that:
Authentic, diverse, and multilingual data is used;
Computational techniques are evaluated rigorously in real-world contexts;
The findings contribute both to theory and practical NLP model development for code-switched language processing.

Ethical Considerations

Ethical integrity is essential when conducting research involving human language data, especially when that data is sourced from personal communication, social media, or multilingual communities. This study carefully addresses all relevant ethical concerns to ensure that data collection, processing, and analysis are carried out responsibly, transparently, and with respect for individual privacy and digital rights.

Informed Consent and Data Usage

For any data collected from private messaging platforms such as WhatsApp or SMS, explicit written consent will be obtained from participants before data collection. Participants will be informed about the purpose of the research, how their data will be used, stored, and protected. Participants will retain the right to withdraw their data at any time without penalty.

Use of Publicly Available Data

Data obtained from public social media platforms (e.g., Twitter, Facebook, YouTube) will be used only when it is freely accessible without login or password protection. Usernames, profile pictures, and other identifying metadata will be removed or anonymized before analysis.

Only content posted publicly and not protected by privacy settings will be included.

Data Anonymization and Privacy

All datasets will be anonymized before storage and analysis to protect the identity of individuals.

No personally identifiable information (PII) such as names, phone numbers, or email addresses will be included in the final dataset.

Any quotes or examples used in reporting will be paraphrased or masked to avoid back-tracing to the original source.

Data Storage and Security

Collected data will be stored on secure, password-protected servers.

Access to raw data will be limited to the research team only.

All backup copies will be encrypted, and data will be deleted responsibly after the completion of the research, in accordance with institutional and national guidelines.

Ethical Approval

If conducted under a university or institutional affiliation, the study will be submitted for review and approval by the appropriate Ethical Review Committee or Institutional Review Board (IRB) before data collection begins.

Avoiding Bias and Misrepresentation

The research will ensure that all language communities are represented respectfully and accurately.

The study will not reinforce stereotypes or stigmatize code-switching behaviors as inferior or uneducated.

Any computational model developed will be evaluated for language bias and fairness across dialects and demographic groups.

Responsible AI and Open Research

The study follows the principles of responsible AI development by ensuring fairness, accountability, and transparency.

Any models or annotated datasets developed will be shared for academic use with clear licensing and ethical use statements, if permitted by participants and data sources.

Conclusion

This study explored the phenomenon of code-switching in multilingual digital discourse, with a particular focus on Urdu-English and Hindi-English language pairs, using computational techniques grounded in Natural Language Processing (NLP). By integrating linguistic theory with machine learning approaches such as BiLSTM and

mBERT, the research successfully demonstrated the strengths and limitations of current computational models in identifying and analyzing code-switched language data. The use of annotated corpora, language identification, POS tagging, and syntactic parsing provided valuable insights into both structural and functional aspects of multilingual communication.

The findings revealed that while transformer-based models and deep learning techniques show promise in handling code-switched inputs, significant challenges remain especially regarding grammatical inconsistencies, Romanized scripts, and limited annotated datasets for low-resource languages. Furthermore, the research emphasized the importance of context-aware systems that go beyond surface-level tagging to incorporate sociolinguistic features of code-switching. Ultimately, this study contributes to the development of more inclusive and linguistically informed NLP systems, with practical implications for improving sentiment analysis, machine translation, automatic speech recognition, and educational technologies in multilingual regions like South Asia. Bridging the gap between computational methods and sociolinguistic insights, the research encourages future work that continues to refine multilingual processing tools while respecting the complex dynamics of language use in real-world contexts.

References

- Aguilar, G., Maharjan, S., & Solorio, T. (2018). Named entity recognition on codeswitched data: Combining data sets and avoiding annotation bias. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 74–80). <https://doi.org/10.18653/v1/N18-2012>
- Almansour, B. Y., Almansour, A. Y., Janjua, J. I., Zahid, M., & Abbas, T. (2024). Application of machine learning and rule induction in various sectors. In *Proceedings of the 2024 International Conference on Decision Aid Sciences and Applications (DASA)* (pp. 1–8). <https://doi.org/10.1109/DASA63652.2024.10836265>
- Androutsopoulos, J. (2007). Bilingualism in the mass media and on the internet. In P. Auer (Ed.), *Style and social identities: Alternative approaches to linguistic heterogeneity* (pp. 207–232). De Gruyter.
- Barman, U., Das, A., Wagner, J., & Foster, J. (2014). Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching (EMNLP Workshop)* (pp. 13–23). Association for Computational Linguistics.
- Bautista, M. L. S. (2004). Tagalog-English code-switching as a mode of discourse. *Asia Pacific Education Review*, 5(2), 226–233. <https://doi.org/10.1007/BF03024958>
- Blom, J. P., & Gumperz, J. J. (1972). Social meaning in linguistic structure: Codeswitching in Norway. In J. J. Gumperz & D. Hymes (Eds.), *Directions in sociolinguistics: The ethnography of communication* (pp. 407–434). Holt, Rinehart & Winston.
- Devlin, J., Chang, M.-W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of

- deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)* (pp. 4171–4186).
- Duong, H., Cahyawijaya, S., Purwarianti, A., Winata, G. I., & Fung, P. (2013). SEAME: South East Asia Mandarin-English code-switching speech corpus. In *Proceedings of INTERSPEECH 2013*.
- Gumperz, J. J. (1982). *Discourse strategies*. Cambridge University Press.
- ICON 2015 Shared Task. (2015). NLP tools contest on code-mixed Indian social media text. In *Proceedings of the 12th International Conference on Natural Language Processing (ICON-2015)*.
- Iqbal, J., Ali, A., & Chaudhry, M. U. (2016). OPENTCQ: Towards change management in hybrid agile model. In *Proceedings of the 7th International Conference on Computing Communication and Networking Technologies (ICCCNT '16)* (Article 38, 1–8). Association for Computing Machinery. <https://doi.org/10.1145/2967878.2967920>
- Janjua, J. I., Irfan, M., Abbas, T., Ihsan, A., & Ali, B. (2024). Enhancing contextual understanding in chatbots and NLP. In *Proceedings of the 2024 International Conference on TVET Excellence & Development (ICTeD)* (pp. 244–249). <https://doi.org/10.1109/ICTeD62334.2024.10844601>
- Joshi, A. K. (1982). Processing of sentences with intra-sentential code-switching. In *Proceedings of the Conference on Theoretical Issues in Natural Language Processing* (pp. 190–206).
- LinCE: Linguistic Code-switching Evaluation Dataset. (2019). Retrieved from <https://ritual.uh.edu/lince/>
- Myers-Scotton, C. (1993). *Social motivations for codeswitching: Evidence from Africa*. Oxford University Press.
- Poplack, S. (1980). Sometimes I'll start a sentence in Spanish y termino en español: Toward a typology of code-switching. *Linguistics*, 18(7–8), 581–618.
- Rehman, A., Noor, F., Janjua, J. I., Ihsan, A., Saeed, A. Q., & Abbas, T. (2024). Classification of lung diseases using machine learning technique. In *Proceedings of the 2024 International Conference on Decision Aid Sciences and Applications (DASA)* (pp. 1–7). <https://doi.org/10.1109/DASA63652.2024.10836302>
- Sebba, M. (2012). Multilingualism in written discourse: An approach to the analysis of multilingual texts. *International Journal of Bilingualism*, 17(1), 97–118.
- Solorio, T., Blair, E., Maharjan, S., Bethard, S., Diab, M., Hirschberg, J., ... & Chang, A. (2014). Overview for the first shared task on language identification in codeswitched data. In *Proceedings of the First Workshop on Computational Approaches to Code Switching* (pp. 62–72). Association for Computational Linguistics.
- Sridhar, S. N., & Sridhar, K. K. (1980). The syntax and psycholinguistics of bilingual code-mixing. *Canadian Journal of Psychology*, 34(4), 407–416.
- T. Abbas, J. I. Janjua, & M. Irfan. (2023). Proposed agricultural Internet of Things (AIoT) based intelligent system of disease forecaster for agri-domain. In *Proceedings of the 2023 International Conference on Computer and Applications (ICCA)*, Cairo, Egypt, 2023, pp. 1–6.

<https://doi.org/10.1109/ICCA59364.2023.10401794>.

Winata, G. I., Cahyawijaya, S., & Fung, P. (2019). Learning fast code-switching language identification from low-resource multilingual data. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL): Student Research Workshop* (pp. 80–87).