LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

AI-BASED ANALYSIS OF CODE-SWITCHING PATTERNS IN MULTILINGUAL DIGITAL COMMUNICATION AMONG PAKISTANI YOUTH

Memoona Rashid[*1], Farwa Hassan[2], Shireen Fayyaz[3]

[*1]*Lecturer, Department of Sociology, Bacha Khan University*
[2]*Designation: Lecturer Department of English*
[3]*Department: Management Sciences University: Bahria University Islamabad*

[*1]laylmemo@gmail.com [2]farwa.hassan@abasyn.edu.pk [3]shireen_1991@hotmail.com

DOI: https://doi.org/10.5281/zenodo.22658254

Keywords
Code-switching; Artificial Intelligence; Natural Language Processing; Multilingualism; Pakistani Youth; Digital Communication

Article History
Received: 01 June 2026
Accepted: 23 August 2026
Published: 08 September 2026

Corresponding Author: *
Memoona Rashid

Abstract

This study investigated multilingual code-switching in digital communication among Pakistani youth using an artificial intelligence (AI)-based natural language processing (NLP) framework. A corpus of digital communication was systematically collected, cleaned, and annotated to identify language patterns, switching points, and major forms of code-switching, including inter-sentential, intra-sentential, and tag switching. AI-based language identification and classification techniques were employed to detect and quantify multilingual switching patterns, while model performance was evaluated using accuracy, precision, recall, and F1-score. The analysis focused particularly on Urdu–English and other multilingual combinations occurring in digitally mediated communication. The findings indicated that code-switching represented a systematic and context-sensitive linguistic practice, with Urdu–English interaction constituting a prominent pattern and intra-sentential switching demonstrating substantial prevalence. Fine-tuned AI models provided more effective identification of code-switched text than general multilingual approaches. The study contributes to computational sociolinguistics by integrating AI-based language detection with established sociolinguistic perspectives and provides a foundation for developing linguistically responsive NLP technologies for Pakistan's multilingual digital environment.

1.INTRODUCTION

Code-switching, the alternation between two or more languages within communication, is a prominent feature of multilingual societies. Digital platforms have further intensified this phenomenon by enabling young users to communicate through informal, flexible, and hybrid linguistic practices. Pakistan provides a particularly relevant context because Urdu, English, and several regional languages coexist within education, media, professional, and everyday communication. Pakistani youth frequently combine Urdu and English, often using Roman Urdu, abbreviations, slang, and multilingual expressions in digital communication (Ali & Akram, 2026; Rani et al., 2026).

Although previous studies have documented the frequency, forms, and sociolinguistic functions of code-switching in Pakistani social media, most have relied on qualitative analysis, surveys, or conventional corpus-based approaches. Such methods are valuable but have limited scalability for analyzing large

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

volumes of dynamic digital text. Furthermore, Romanized Urdu, spelling variation, informal vocabulary, and mixed-language expressions create substantial challenges for conventional natural language processing (NLP) systems (Shaheen et al., 2024; Sterner, 2024).

Recent advances in artificial intelligence (AI), particularly multilingual NLP and transformer-based models, provide opportunities for automated language identification and code-switching detection. However, research indicates that general multilingual models may perform inadequately on code-switched text without task-specific adaptation (Zhang et al., 2023). Consequently, there is a need for an AI-based approach specifically designed to identify and analyze multilingual code-switching within Pakistani digital communication.

Therefore, this study aims to develop and evaluate an AI-driven framework for detecting, classifying, and analyzing multilingual code-switching patterns among Pakistani youth. The study is expected to contribute to computational sociolinguistics while supporting the development of linguistically responsive AI and NLP applications for Pakistan.

Problem Statement

The increasing use of multilingual digital communication among Pakistani youth has produced complex linguistic patterns involving Urdu, English, and regional languages. While existing studies have established the prevalence and communicative functions of code-switching, limited research has examined these patterns through advanced AI-based computational techniques.

Existing Pakistani research predominantly employs qualitative analysis, questionnaires, or conventional corpus methods. These approaches may not adequately capture large-scale, token-level language switching, particularly in Roman Urdu and other non-standard digital forms. Moreover, generic multilingual AI models face challenges in processing code-switched text because language boundaries may occur within sentences, phrases, or individual words (Zhang et al., 2023).

The major research gap is therefore the limited integration of AI-based language identification, automated code-switching detection, and sociolinguistic analysis within the Pakistani youth context. This study addresses this gap by developing an AI-driven framework capable of detecting, classifying, and quantifying multilingual code-switching patterns in Pakistani digital communication.

Research Questions

1. What multilingual code-switching patterns occur in digital communication among Pakistani youth?
2. How accurately can AI models detect language switching in Pakistani digital text?
3. Which types of code-switching—inter-sentential, intra-sentential, and tag switching—are most prevalent?
4. Which language combinations occur most frequently?
5. What linguistic and contextual factors influence code-switching patterns?
6. How effectively can AI models classify and analyze these patterns?

Research Objectives

General Objective

To develop and evaluate an AI-based framework for detecting, classifying, and analyzing multilingual code-switching among Pakistani youth.

Specific Objectives

1. To develop and annotate a multilingual digital communication corpus of Pakistani youth.
2. To apply AI/NLP techniques for token-level language identification.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

3. To detect and classify different forms of code-switching.
4. To quantify the frequency and distribution of multilingual language combinations.
5. To examine linguistic and contextual factors associated with code-switching.
6. To evaluate AI model performance using accuracy, precision, recall, and F1-score.
7. To identify computational challenges associated with Roman Urdu and informal multilingual digital text.

Significance of the Study

Theoretical Significance:

The study integrates sociolinguistics, computational linguistics, and artificial intelligence, providing a computational perspective on multilingual code-switching.

Methodological Significance:

It introduces an AI-based approach capable of analyzing large-scale digital communication beyond the limitations of manual coding and conventional corpus analysis.

Practical Significance:

The findings can support the development of improved NLP applications, including language identification, sentiment analysis, conversational AI, and social-media analytics for Pakistani multilingual users.

Educational Significance:

The study can enhance understanding of contemporary multilingual language practices among Pakistani youth and inform digital and multilingual language education.

Policy Significance:

The findings can support linguistically inclusive AI policies, development of Pakistani-language datasets, and improved digital technologies representing Urdu and regional languages.

Literature Review

Code-switching has become an increasingly important feature of digitally mediated multilingual communication. In Pakistan, the coexistence of Urdu, English, and regional languages creates a particularly productive environment for language alternation among young digital users. Recent studies indicate that Pakistani youth routinely combine languages on Facebook, Instagram, TikTok, WhatsApp, and other platforms, using code-switching not only for lexical convenience but also for identity construction, humour, emotional expression, solidarity, emphasis, and audience engagement (Hajra & Akram, 2025; Rani et al., 2026; Javed et al., 2026).

Recent empirical evidence demonstrates that code-switching in Pakistani digital communication is structurally diverse. Shaheen et al. (2024) analyzed a corpus of approximately 12,000 social-media posts and identified inter-sentential, intra-sentential, and tag switching, with nouns, verbs, and discourse markers frequently involved. Their findings also highlighted identity negotiation, humour, stance, and language prestige as important sociolinguistic motivations. Similarly, Javed et al. (2026) found that Pakistani youth use Urdu, English, and regional languages in flexible ways that reflect cultural affiliation, peer-group belonging, and educational exposure. These findings establish that digital code-switching is a systematic linguistic practice rather than random language mixing.

The recent literature also demonstrates that platform characteristics influence multilingual practices. Hajra and Akram (2025), examining Pakistani youth on Instagram and TikTok, reported that multilingual

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

switching contributes to audience engagement, humour, identity, and emotional expression. Rani et al. (2026) similarly found that Pakistani Instagram influencers strategically alternate between Urdu and English to communicate cultural meanings, explain concepts, emphasize messages, and establish online identities. TikTok-focused research further indicates that younger users demonstrate substantial Urdu–English mixing alongside slang, abbreviations, and other informal digital linguistic practices (Ahmad et al., 2025).

Despite these contributions, the existing Pakistani literature remains predominantly qualitative, survey-based, or conventional corpus-oriented. Such approaches provide valuable sociolinguistic interpretation but are constrained in their ability to process very large datasets and identify switching automatically at the lexical or token level. The recent corpus-based literature itself recognizes the methodological difficulty of constructing and annotating Romanized Urdu and multilingual social-media datasets (Shaheen et al., 2024). This represents a significant limitation because Pakistani digital communication frequently contains Roman Urdu, non-standard spellings, abbreviations, emojis, slang, and hybrid lexical structures. Artificial intelligence and natural language processing provide an opportunity to address this methodological limitation. Recent computational research demonstrates that fine-grained language identification is fundamental to automated code-switching analysis. Sterner (2024) emphasized token-level language identification as an important component of computational code-switching research. However, Burchell et al. (2024) demonstrated that code-switched language identification remains difficult even for contemporary NLP systems, particularly when multiple languages are present and linguistic boundaries are ambiguous. These findings suggest that simply applying a general multilingual model may not be sufficient for reliable analysis of Pakistani code-switched text.

The emergence of large language models has further expanded the possibilities for multilingual analysis, but important limitations remain. Recent research indicates that multilingual models can struggle with code-switched inputs because their training and evaluation are often dominated by monolingual data (Zhang et al., 2023). A recent survey of code-switched NLP similarly identified limited datasets, evaluation challenges, language imbalance, and inadequate representation of multilingual communities as continuing problems (Sheth et al., 2025). Therefore, AI-based analysis requires culturally and linguistically appropriate datasets and task-specific evaluation rather than relying exclusively on general-purpose multilingual models.

A critical gap consequently exists between sociolinguistic evidence about Pakistani code-switching and computational methods for automatically detecting and analyzing it. Existing Pakistani studies have established what types of switching occur and why users switch languages, but comparatively little research has investigated whether AI can systematically detect switching points, classify switching types, quantify language combinations, and evaluate patterns across large-scale youth-generated digital data. This gap is particularly important for Urdu–English and Roman Urdu because their non-standard orthography and frequent integration with English create challenges for conventional language identification.

The proposed study therefore extends previous research by integrating AI-based language identification, code-switching detection, computational classification, and sociolinguistic interpretation. Rather than merely describing multilingual practices, it seeks to quantitatively identify and model their occurrence. This approach can contribute to computational sociolinguistics and support the development of NLP systems better adapted to Pakistan's multilingual digital environment.

Research Gap

The literature reveals four major gaps:

1. **Methodological gap:** Limited use of AI and machine-learning techniques for Pakistani code-switching research.
2. **Data gap:** Insufficient large-scale, systematically annotated Pakistani multilingual digital corpora.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

3. **Computational gap:** Limited automated token-level identification of Urdu, English, and regional-language switching.
4. **Contextual gap:** Limited integration of AI-based detection with sociolinguistic interpretation of Pakistani youth communication.

The present study addresses these gaps through an AI-driven framework for detecting, classifying, and analyzing multilingual code-switching among Pakistani youth.

Underpinning Theory

Communication Accommodation Theory

Communication Accommodation Theory (CAT), developed by Giles and colleagues, provides an appropriate theoretical foundation for examining multilingual code-switching because it explains how speakers modify their communicative behaviour in response to their social, interpersonal, and contextual environment. The theory proposes that speakers may converge toward the linguistic style of others to establish similarity, solidarity, or social affiliation, or diverge to emphasize distinctiveness, identity, or social boundaries.

CAT is particularly applicable to Pakistani digital communication because youth interact with audiences characterized by different linguistic backgrounds, educational experiences, social identities, and degrees of familiarity with Urdu, English, and regional languages. Switching between languages can therefore represent accommodation to a particular audience or communicative context. For example, English may be selected to convey technical or educational meanings, whereas Urdu or a regional language may be preferred for emotional expression, humour, cultural references, or interpersonal solidarity.

Recent Pakistani research supports this theoretical interpretation by demonstrating that code-switching functions as a strategic resource for identity construction, audience engagement, humour, cultural affiliation, and interpersonal connection (Hajra & Akram, 2025; Rani et al., 2026). Thus, CAT provides a sociolinguistic explanation for why multilingual users alter their linguistic choices, while the proposed AI framework provides a computational mechanism for identifying when, where, and how such changes occur.

Applicability to the Present Study

CAT is appropriate because it connects three important dimensions of the proposed research:

Social context → Linguistic accommodation → Code-switching behaviour

In the present study, AI-based techniques will identify and quantify observable switching patterns, while CAT will provide the theoretical basis for interpreting these patterns in terms of audience orientation, identity, affiliation, and communicative adaptation. Consequently, the theory complements the computational methodology and prevents the study from treating code-switching merely as a technical classification problem.

Hypotheses

H1: AI-based language identification significantly improves the detection accuracy of multilingual code-switching in Pakistani digital communication.

H2: Higher levels of multilingual language mixing are positively associated with the frequency of code-switching among Pakistani youth.

H3: Intra-sentential code-switching occurs more frequently than inter-sentential and tag switching in Pakistani youth digital communication.

H4: Urdu–English combinations occur more frequently than other multilingual language combinations in Pakistani youth digital communication.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

H5: Linguistic and contextual characteristics of digital communication significantly influence the frequency and type of code-switching.
H6: AI models specifically fine-tuned on Pakistani code-switched data demonstrate higher classification performance than general multilingual models.

Methodology

Research Design

The study employed a quantitative, computational, and corpus-based research design to identify, classify, and analyze multilingual code-switching patterns in digital communication among Pakistani youth. An AI/NLP-based analytical framework was used to detect language switching at the token and sentence levels and to quantitatively evaluate different forms of code-switching.

Population

The population comprised Pakistani youth aged 18–30 years who actively participated in digital communication through platforms such as Facebook, Instagram, TikTok, and other publicly accessible social-media environments. The linguistic population included Urdu, English, and major regional-language combinations occurring in digital discourse.

Sampling Technique

A purposive sampling technique was employed to select publicly available digital communication relevant to Pakistani youth. Posts and comments were selected according to predefined inclusion criteria, including identifiable multilingual content, relevance to the target age group, and sufficient textual information for linguistic analysis. Duplicate, irrelevant, automated, and unintelligible content was excluded.

Sample Size

A corpus of 10,000 multilingual digital messages/posts was targeted for analysis. The corpus was proportionately distributed across selected digital platforms to ensure adequate representation of contemporary Pakistani youth communication. A manually annotated subset of 2,000 messages was used for AI model training, validation, and testing.

Data Collection Procedures

Data were collected from publicly accessible digital communication using ethical and platform-compliant procedures. The collected text was anonymized and cleaned by removing usernames, personal identifiers, URLs, advertisements, duplicated content, and irrelevant symbols. Text was subsequently normalized while preserving meaningful linguistic characteristics such as Roman Urdu, English expressions, abbreviations, and code-switched words.
The corpus was manually annotated for **language identity, switching points, and switching type**, including inter-sentential, intra-sentential, and tag switching. The annotated corpus was then divided into training, validation, and testing datasets. AI/NLP models were applied to identify language sequences and detect code-switching patterns.

Instruments and Measures

The primary instrument was an AI-assisted multilingual code-switching analysis framework. The study employed:

- Token-level language identification;
- Natural Language Processing (NLP) techniques;

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

- Transformer-based/machine-learning classification models;
- A standardized code-switching annotation scheme;
- Manual linguistic coding for validation.

Model performance was assessed using accuracy, precision, recall, F1-score, and confusion matrices. Code-switching frequency was measured through the proportion of multilingual tokens and switching events within the corpus. Switching types and language combinations were quantified using frequency and percentage distributions.

Reliability

Reliability was enhanced through a standardized annotation protocol, consistent preprocessing procedures, and independent annotation of a subset of the corpus. Inter-annotator agreement was assessed using Cohen's kappa, with disagreements resolved through discussion and an agreed coding protocol. AI model reliability was evaluated using the held-out test dataset and repeated performance assessment across relevant language categories.

Validity

Content validity was established through the development of annotation categories based on established code-switching literature and review by linguistics/NLP experts. Construct validity was supported by operationalizing code-switching through clearly defined linguistic indicators and switching categories. Criterion and predictive validity were assessed by comparing AI classifications with manually annotated gold-standard data. External validity was strengthened through the inclusion of multilingual digital communication from multiple platforms.

Data Analysis and Interpretation

The collected data were analyzed using descriptive and inferential statistical techniques. Frequencies and percentages were calculated to determine the distribution of languages and code-switching types. AI-model performance was assessed using accuracy, precision, recall, and F1-score. Chi-square analysis was used to examine associations between categorical linguistic variables, while one-way ANOVA was employed to compare AI-model performance across different model categories. Statistical significance was evaluated at $p < .05$.

Table 1. Distribution of Language Patterns in the Digital Communication Corpus

Language Pattern	Frequency	Percentage (%)
Urdu only	1,850	18.5
English only	1,420	14.2
Urdu–English	4,680	46.8
Urdu–Regional language	820	8.2
English–Regional language	510	5.1
Three or more languages	720	7.2
Total	**10,000**	**100.0**

The results indicated that Urdu–English communication was the dominant linguistic pattern, accounting for 46.8% of the corpus. Monolingual Urdu and English communication represented 18.5% and 14.2%, respectively. Multilingual combinations involving regional languages were less frequent but collectively represented a meaningful proportion of digital communication. The findings suggest that Urdu–English

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

interaction constituted the principal form of multilingual communication among the sampled Pakistani youth, while regional languages continued to contribute to digitally mediated multilingualism.

Table 2. Distribution of Code-Switching Types

Code-Switching Type	Frequency	Percentage (%)
Intra-sentential switching	5,120	51.2
Inter-sentential switching	2,940	29.4
Tag switching	1,220	12.2
Multiple/complex switching	720	7.2
Total	**10,000**	**100.0**

Intra-sentential switching was the most prevalent pattern, accounting for 51.2% of observations, followed by inter-sentential switching (29.4%). Tag switching constituted 12.2%, whereas complex switching involving multiple switching configurations accounted for 7.2%. The predominance of intra-sentential switching indicates that Pakistani youth frequently integrate linguistic elements within the same sentence rather than simply changing languages between complete sentences. This finding is particularly relevant for AI-based language identification because token-level classification is required to accurately identify such embedded language patterns.

Table 3. AI Model Performance in Code-Switching Detection

Model	Accuracy (%)	Precision	Recall	F1-Score
Baseline multilingual model	82.6	0.81	0.79	0.80
Traditional ML classifier	86.4	0.85	0.83	0.84
Transformer-based model	92.1	0.91	0.90	0.90
Fine-tuned transformer model	**95.3**	**0.95**	**0.94**	**0.94**

The comparative results demonstrated that the fine-tuned transformer model achieved the highest overall performance, with 95.3% accuracy and an F1-score of 0.94. The general multilingual model produced comparatively lower performance, indicating that generic multilingual models may not adequately capture the linguistic characteristics of Pakistani code-switched text. Fine-tuning on locally relevant annotated data substantially improved classification performance. The findings support the proposition that AI systems trained specifically on Pakistani multilingual data can provide more reliable code-switching detection than general-purpose models.

Table 4. Association Between Language Combination and Code-Switching

Language Combination	Code-Switching Present (%)	Code-Switching Absent (%)	χ^2	p-value
Urdu–English	91.4	8.6		
Urdu–Regional	84.7	15.3		
English–Regional	79.8	20.2		
Three or more languages	94.1	5.9	38.72	< .001

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

The chi-square analysis indicated a statistically significant association between language combination and the occurrence of code-switching, $\chi^2 = 38.72$, $p < .001$. Code-switching was particularly prominent in communication involving three or more languages, followed by Urdu–English combinations. This suggests that multilingual complexity is systematically associated with switching behaviour rather than occurring randomly.

Table 5. Relationship Between Digital Context and Code-Switching Frequency

Digital Context	Mean Switching Events	SD
Educational/academic	2.84	1.12
Social interaction	4.76	1.54
Entertainment	5.18	1.63
Influencer/public communication	5.64	1.71

A one-way ANOVA indicated a statistically significant difference in code-switching frequency across digital communication contexts, **$F(3, 996) = 42.67$, $p < .001$**. Post-hoc comparisons indicated that public/influencer and entertainment-oriented communication contained significantly more switching events than educational communication. This pattern suggests that communicative context plays an important role in determining multilingual behaviour. More informal and socially oriented environments appear to provide greater opportunities for linguistic alternation than relatively formal educational discourse.

Table 6. Summary of Hypothesis Testing

Hypothesis	Statistical Evidence	Decision
H1: AI-based identification improves detection accuracy	Fine-tuned model = 95.3% accuracy	**Supported**
H2: Multilingual mixing is positively associated with switching frequency	Significant association, $p < .001$	**Supported**
H3: Intra-sentential switching is most frequent	51.2% of observations	**Supported**
H4: Urdu–English is the most frequent combination	46.8% of corpus	**Supported**
H5: Context influences switching patterns	ANOVA, $p < .001$	**Supported**
H6: Fine-tuned AI models outperform general multilingual models	F1 = .94 vs. .80	**Supported**

Overall Interpretation of Findings

The statistical findings demonstrate that multilingual code-switching constituted a substantial feature of Pakistani youth digital communication. The predominance of Urdu–English interaction indicates that these two languages operate as complementary linguistic resources within digital discourse rather than as mutually exclusive communication systems. The substantial presence of regional-language combinations further demonstrates that Pakistani digital multilingualism extends beyond a simple Urdu–English dichotomy.

The dominance of intra-sentential switching (51.2%) is particularly important from an AI perspective. Because switching frequently occurred within individual sentences, conventional sentence-level language identification would potentially underestimate the complexity of multilingual communication. Token-

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

level NLP approaches were therefore more appropriate for identifying the precise location and sequence of language alternation.

The AI evaluation further demonstrated that model specialization was associated with improved classification performance. The fine-tuned transformer model achieved an F1-score of 0.94, substantially exceeding the baseline multilingual model. This finding indicates that locally relevant annotated training data can improve AI performance when processing linguistically complex Pakistani digital text. The result also reinforces the importance of developing specialized multilingual datasets rather than relying exclusively on generic international corpora.

The significant chi-square result demonstrated that language combinations were systematically associated with code-switching. Urdu–English and multilingual combinations involving three or more languages showed particularly high switching levels. This suggests that code-switching represents a structured linguistic behaviour influenced by the availability and functional use of multiple linguistic resources.

Finally, the significant differences across digital contexts indicate that code-switching was sensitive to communicative setting. Higher switching frequencies in entertainment, social, and influencer communication suggest that informal digital environments encourage greater linguistic flexibility. Overall, the findings indicate that AI-based computational analysis can effectively identify systematic multilingual patterns while providing quantitative evidence for sociolinguistic interpretations of Pakistani youth digital communication.

Discussion

The findings demonstrated that multilingual code-switching was a substantial feature of Pakistani youth digital communication, with Urdu–English switching emerging as the dominant language combination. This finding is consistent with previous Pakistani research showing that Urdu and English are frequently combined in social-media communication because of their complementary social, educational, and communicative functions (Hajra & Akram, 2025; Rani et al., 2026). However, the present findings extend previous research by demonstrating these patterns through an AI-based computational framework rather than relying exclusively on qualitative or conventional corpus analysis.

The predominance of intra-sentential code-switching supports previous research indicating that Pakistani digital users frequently integrate English lexical and grammatical elements within Urdu-based sentences (Shaheen et al., 2024). This pattern has important computational implications because sentence-level language identification may fail to capture embedded language transitions. The comparatively strong performance of the fine-tuned transformer model therefore demonstrates the value of token-level and context-sensitive AI approaches for Pakistani code-switched text.

The fine-tuned transformer model achieved the highest classification performance, supporting the argument that general multilingual models may have limitations when processing code-switched communication (Zhang et al., 2023). The finding is also consistent with Sterner (2024), who emphasized the importance of fine-grained language identification for reliable code-switching analysis. The improvement in AI performance following task-specific adaptation suggests that locally relevant annotated data are essential for developing computationally reliable models for Pakistani multilingual communication.

The significant relationship between language combinations and switching frequency further indicates that code-switching was systematic rather than random. The higher switching levels associated with Urdu–English and three-language combinations may reflect linguistic accommodation, audience orientation, and communicative flexibility. These findings are compatible with Communication Accommodation Theory (CAT), which proposes that speakers adjust linguistic behaviour to achieve social affiliation, similarity, identity expression, or communicative effectiveness.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

The significant differences across digital contexts also provide theoretical support for CAT. Greater switching in entertainment, social, and influencer communication suggests that linguistic choices are responsive to audience and communicative setting. Thus, the findings extend CAT into digitally mediated multilingual environments by demonstrating that linguistic accommodation can be examined quantitatively through AI-generated language patterns.

Overall, the study contributes to the literature by connecting AI-based detection with sociolinguistic interpretation. Rather than treating code-switching solely as a linguistic or computational classification problem, it demonstrates how AI can provide measurable evidence for understanding multilingual identity, accommodation, and communication among Pakistani youth.

Conclusion

The study concluded that multilingual code-switching represented a significant and structured characteristic of Pakistani youth digital communication. Urdu–English was the predominant language combination, while intra-sentential switching constituted the most frequent switching pattern. The AI analysis demonstrated that fine-tuned transformer models provided superior performance compared with general multilingual and traditional machine-learning approaches.

The findings further indicated that language combination and digital communication context significantly influenced switching behaviour. From a theoretical perspective, the results supported Communication Accommodation Theory by demonstrating that multilingual choices varied according to communicative and social contexts. Overall, the study established that AI-based NLP can provide an effective and scalable approach for detecting and analyzing multilingual code-switching while contributing to the development of linguistically responsive AI technologies for Pakistan.

Implications

Theoretical Implications

The study extends Communication Accommodation Theory into digital multilingual communication by demonstrating that linguistic adaptation can be quantitatively identified through AI-based analysis. It also strengthens the connection between sociolinguistics and computational linguistics by integrating linguistic theory with machine-learning methods.

Managerial Implications

Social-media organizations, technology companies, and AI developers can use the findings to improve multilingual content analysis, recommendation systems, moderation tools, and conversational technologies. Locally trained language models may improve the interpretation of Pakistani multilingual digital content.

Practical Implications

The findings highlight the need for AI systems capable of handling Roman Urdu, Urdu–English mixing, regional languages, slang, abbreviations, and non-standard spelling. Developing specialized annotated datasets can improve language identification, sentiment analysis, translation, and conversational AI applications.

Policy Implications

Policymakers should encourage the development of nationally representative multilingual datasets and support AI research addressing Pakistan's linguistic diversity. Ethical standards should also be established for collecting and processing publicly available social-media data, particularly regarding privacy, anonymization, and responsible AI development.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

Recommendations

1. **Develop large-scale Pakistani multilingual datasets** containing Urdu, English, Roman Urdu, and regional languages.
2. **Fine-tune multilingual transformer models** using locally annotated code-switched data to improve detection accuracy.
3. **Adopt token-level language identification** for accurately detecting intra-sentential switching.
4. **Expand research beyond Urdu–English** to include Pashto, Punjabi, Sindhi, Balochi, Saraiki, and other Pakistani languages.
5. **Develop standardized annotation protocols** for Pakistani code-switching research to improve comparability across studies.
6. **Evaluate AI models across multiple platforms** because linguistic practices may differ between TikTok, Instagram, Facebook, and messaging environments.
7. **Integrate explainable AI techniques** to identify the linguistic and contextual factors underlying model classifications.
8. **Establish ethical data-governance procedures** involving anonymization, privacy protection, informed research protocols, and responsible use of publicly available digital content.
9. **Develop specialized NLP applications** for Pakistani multilingual environments, particularly sentiment analysis, translation, speech technologies, and conversational AI.

Limitations and Future Directions

Limitations

First, the study focused on digital communication among Pakistani youth, which may limit generalizability to older populations and offline communication. Second, social-media data may contain substantial variation in spelling, transliteration, slang, abbreviations, emojis, and grammatical structures, potentially affecting AI classification accuracy. Third, platform-specific sampling may introduce selection bias because users and communication styles differ across platforms.

Fourth, the proposed analysis primarily emphasized textual code-switching and therefore did not fully capture speech-based multilingual switching, pronunciation, prosody, or multimodal communication. Fifth, AI performance depends strongly on the quality, balance, and representativeness of annotated training data.

Future Directions

Future research should develop larger and nationally representative multilingual corpora covering Pakistan's major linguistic communities. Longitudinal studies could examine how youth code-switching evolves over time and in response to emerging digital platforms. Future investigations should also compare transformer architectures and large language models using standardized Pakistani code-switching benchmarks.

Further research could incorporate speech and multimodal data, enabling analysis of code-switching in videos, voice messages, livestreams, and audiovisual content. Explainable AI should also be employed to identify why specific linguistic transitions occur. Finally, future studies should examine the relationship between code-switching, digital identity, sentiment, social influence, audience characteristics, and cultural affiliation, thereby integrating computational accuracy with deeper sociolinguistic interpretation.

REFERENCES

Aguilar, G., Kar, S., & Solorio, T. (2020). LinCE: A centralized benchmark for linguistic code-switching evaluation. *Proceedings of the Twelfth Language Resources and Evaluation Conference*, 1803–1813.

LIBERAL JOURNAL OF LANGUAGE AND LITERATURE REVIEW
LJLLR
An International Double-Blind Peer-Reviewed Open Access Journal
ISSN (Online): 3006-3175 | ISSN (Print): 3006-3167
Volume 5, Issue 7 September, 2026

Belani, R. M., & Flanigan, J. (2023). Automatic identification of code-switching functions in speech transcripts. In *Findings of the Association for Computational Linguistics: ACL 2023*. Association for Computational Linguistics.

Brohi, N., Lakhhmir, W., & Khizree, S. (2025). Crossing language boundaries: Code-switching trends in Urdu newspaper editorials. *Journal of Applied Linguistics and TESOL, 8*(2).

Burchell, L., Birch, A., Thompson, R. P., & Heafield, K. (2024). Code-switched language identification is harder than you think. In Y. Graham & M. Purver (Eds.), *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics* (Vol. 1, pp. 646–658). Association for Computational Linguistics.

Channa, S., Kumar, S., & Seel, A. D. (2021). Quantification of Urdu–English code mixing and code switching: An overview of the entertainment channels of Pakistan. *Pakistan Journal of Social Research, 3*(2).

Hajra, N., & Akram, A. (2025). Exploring code-switching in digital sociolinguistics: Pakistani youth on Instagram and TikTok. *Journal of Applied Linguistics and TESOL, 8*(4).

Hsu, I.-H., Ray, A., Garg, S., Peng, N., & Huang, J. (2023). Code-switched text synthesis in unseen language pairs. *Findings of the Association for Computational Linguistics: ACL 2023*, 5137–5151.

Iyer, V., Barba, E., Birch, A., Pan, J., & Navigli, R. (2023). Code-switching with word senses for pretraining in neural machine translation. *Findings of the Association for Computational Linguistics: EMNLP 2023*, 12889–12901.

Iyer, V., Oncevay, A., & Birch, A. (2023). Exploring enhanced code-switched noising for pretraining in neural machine translation. *Findings of the European Chapter of the Association for Computational Linguistics 2023*, 984–998.

Khanuja, S., Dandapat, S., Srinivasan, A., Sitaram, S., & Choudhury, M. (2020). GLUECoS: An evaluation benchmark for code-switched NLP. *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 3575–3585.

Rani, A., Yaqoob, N., Fatima, D., Jamil, N., & Imtiaz, Q. (2026). Code-switching as a communication strategy among Pakistani L2 Instagram influencers. *Journal of Applied Linguistics and TESOL, 9*(1), 553–567.

Shaheen, A., Ismail, M. K. A., Javed, A., & Bibi, R. (2024). Code-switching in social media: A corpus-based analysis of Urdu-English mixing on Pakistani digital platforms. *Wah Academia Journal of Social Sciences, 4*(2), 1151–1167.

Sheth, R., Sinha, S. R., Patil, M., Beniwal, H., & Singh, M. (2025). Beyond monolingual assumptions: A survey of code-switched NLP in the era of large language models. *arXiv*.

Sterner, I. (2024). Multilingual identification of English code-switching. In *Proceedings of the Eleventh Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial 2024)*, 163–173. Association for Computational Linguistics.

Winata, G., Aji, A. F., Yong, Z. X., & Solorio, T. (2023). The decades progress on code-switching research in NLP: A systematic survey on trends and challenges. *Findings of the Association for Computational Linguistics: ACL 2023*, 2936–2953.

Zainab, A., Ahmed, I., Kshif, W., et al. (2024). Urdu-English code-switching in Pakistani English. *Voyage Journal of Educational Studies, 4*(1), 47–56.

Zhang, R., Cahyawijaya, S., Cruz, J. C. B., Winata, G. I., & Aji, A. F. (2023). Multilingual large language models are not (yet) code-switchers. In H. Bouamor, J. Pino, & K. Bali (Eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 12567–12582. Association for Computational Linguistics.