ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

"ROMAN URDU AND CODE-MIXED LANGUAGE PROCESSING FOR SOCIAL MEDIA ANALYTICS IN PAKISTAN."

Sabeen Amjad
Department of English ,SZABIST University
sabeen.amjad@szabist.edu.pk
Ijlal Hussain
Department of English,SZABIST University
ijlal.hussain@szabist.edu.pk
Nasir Ullah Khan
Department of English,SZABIST University
nasirullah.khan@szabist.edu.pk

Absrtact
This study examines Roman Urdu and Urdu-English code-mixing. It also examines the problems associated with using online Roman Urdu and code-mixing and digital Urdu. The study also includes the Roman Urdu social media linguistic structure, code-mixing, and Roman Urdu with a focus on Urdu Cricket, Urdu Dramas, and Politics in Urdu. A manual corpus-based research approach combined with other methods was used in this study. This included the collection of 900 posts from YouTube, Facebook, and Twitter (now X). It was found that there is a lot of code-mixing and a lot of Roman Urdu in the social media posts collected for this research. Most of the code-mixing was intra-sentential, as compared to inter-sentential. The study's analysis classified the social media posts sent with an almost equal balance of positive, negative, or neutral sentiments. Other social media posts dealt with several issues of Natural Language Processing, such as a lack of standard corpus, spelling variation, and linguistic uncertainty. The study, for the first time, also explored the necessity of an organized and advanced Natural Language Processing Technology for the multi-lingual digital space of Pakistan.

Keywords: *Roman Urdu, Code-Mixing, Social Media Analytics, Natural Language Processing (NLP), Sentiment Analysis, Computational Linguistics, Low-Resource Languages, Pakistani Digital Communication*

Introduction

Social media platforms such as Twitter (X), Facebook and WhatsApp have grown so fast that they have revolutionised the way Pakistanis communicate. These sites facilitate communication among people, enable them to express their views, and share knowledge, resulting in an immense amount of user-generated textual data (Chung, 2025). The language used on social media is informal, flexible, and linguistically varied, making it difficult for computational analysis. It is common in digital interactions in Pakistan to encounter Roman Urdu, a unique way to communicate in Urdu using the Latin script to enable more informal conversation. Because there are no standard spelling rules in Roman Urdu, one word can have an extensive range of spellings. The absence of standardization makes it vague and difficult to text processing (Han, Cook, & Baldwin, 2013).

There is also an increasing incidence of code-mixing, where speakers switch back and forth between Urdu and English within a single text or conversation. Code mixing is prevalent in digital interactions and is increasing rapidly within sociolinguistic contexts of multilingual communities (Bali, Sharma, Choudhury, & Vyas, 2014; Jawad, Ahmad, Alvi, & Alvi, 2024). These hybrid language forms are nonconforming to standard grammar, making it challenging for standard NLP systems to analyse. These features of the language pose major challenges for Natural Language Processing (NLP). Roman Urdu and code-mixed texts remain largely untouched by most of the existing NLP models, since they were only trained on standard English and Urdu (Nastalik script) texts (Johri, Khatri, Al-Taani, Sabharwal, Suvanov, & Kumar, 2021). Despite the limitations present in this field of data mining, social media data

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

analysis is still important in applications like market analysis, political analysis, and public opinion analysis (Mao, Liu, & Zhang, 2024).

Several challenges remain unaddressed when it comes to the challenges of Roman Urdu and code-mixed language processing. High lexical variability due to the lack of standardization in spelling is the first challenge that complicates the process of normalization and feature extraction. complicating normalization and feature extraction processes (Han, Cook, & Baldwin, 2013). Second, text that is mixed with codes introduces structural and linguistic complexity, which makes tasks like language identification and syntactic parsing much more difficult (Nazir, Bilal, & Shongwe, 2026).

The NLP tools that are available are mainly tailored to English or standard Urdu corpora, and hence, they are not effective with Pakistani social media data. Therefore, important tasks in this regard, including sentiment analysis, topic detection, and text classification, are prone to suboptimal accuracy. In addition, there is a clear dearth in the Pakistani identity of specifically developed solutions to the linguistic realities of Pakistan. This gap highlights the importance of research that will seek to adapt computational approaches to accommodate Means a fair amount: Roman Urdu and code-mixed language.

Research Objectives

The research objectives of the present study are:

- To analyse the linguistic features of Roman Urdu and Urdu-English code-mixed text.
- To identify computational challenges in processing such language.
- To develop or adapt NLP models for improved social media analytics.
- To evaluate the effectiveness of different NLP techniques on Roman Urdu data.

Research Questions

This study is guided by the following research questions:

- What are the structural characteristics of Roman Urdu and code-mixed language?
- What challenges arise in processing this language computationally?
- Which NLP techniques and models perform most effectively for such data?
- How can the accuracy of social media analytics be improved in the Pakistani context?

THEORETICAL FRAMEWORK

Linguistic Framework

The paper is founded on the sociolinguistic theory and explains how language is used: appreciated differently in various social contexts and its expression of identity, culture, and language. Wardhaugh & Fuller (2021), Labov (1972), and Hymes (1974) describe the different ways of communication. In multilingual societies such as Pakistan, individuals switch between languages, which causes problems in the system through features such as code switching and code-mixing. The code-mixed language of Roman Urdu and Urdu-English is two such languages that is influenced by the digital communication space. The convenience/technological accessibility of Roman Urdu leads users to use it, while codemixing shows bilingual competence and social identity. Tagg (2020) notes that "digital" Communication gives the flexibility in language; people can plan how to strategically mix. Linguistic tools to express meaning and to engage socially.

The code-mixing is specific; in fact, it follows the social/communication function. The mix of languages enables language users to emphasize certain aspects of the message, to convey feelings and emotions. Not only did they bring the best, but they also embodied modernity and urbanity. This can be matched with sociolinguistic views where language is seen as a means of identity formation and social status. Also, because there is no standardisation of spelling in Roman Urdu, it is a sort of linguistic variation. Spelling is not used correctly, but according to phonetics and pronunciation. The essence of difference lies in understanding the multifaceted nature of Roman Urdu online.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

Computational Framework

It also relies on the ideas of computational linguistics and specifically processing informal and low-resource language data. Even though this study does not use complex machine learning models, it uses a simplified Natural Language Processing (NLP) pipeline to organize the analysis.

This method follows the techniques of corpus linguistics, in which linguistic patterns are obtained based on observed data, and not based on predictive modelling (McEnery & Hardie, 2011).

Significance of the Study

This work has its contribution to the study of computational linguistics and especially in low-resource and non-standard languages. It fills a major gap in the literature of NLP, which has generally concentrated on standardized languages, by concentrating on Roman Urdu and code-mixed text. In practice, the research carries significant implications in various fields. Sentiment analysis can be improved to boost customer insights and decision-making in marketing. Social media analytics can be used in politics to aid in the tracking of opinion and evaluation of the sentiments of the masses. Moreover, the research can also used to create NLP tools that are specifically designed to address the linguistic contexts of Pakistan, making them more useful in practice.

Scope and Limitations

This paper targets Roman Urdu and Urdu-English mixed-language text in the social media contexts. The data is gathered on the specific platforms, i.e., Twitter (X), Facebook, and YouTube, which might interfere with its generalizability to other communication channels. The research is inadequate in that it does not Incorporate the various regional accents or variations in the languages used in Pakistan that may impact language trends. In addition, the annotated datasets and computational resources are also limited, and this can affect the model performance and scalability. However the study provides a concrete and feasible strategy The book is about analysing Roman Urdu and code-mixed language as social media analytics.

LITERATURE REVIEW

The Latin script used to write the Urdu language is known as Roman Urdu. It is a widely used informal digital communication in Pakistan. It is closely correlated with the increase in the convenience and speed of social media and mobile technologies, which is where they are developing better. In comparison to linguistic accuracy (Khan, Naseer, Wali, & Tamoor, 2024), it is important to the users. Unlike other Roman Urdu features, there are no standard spelling rules, and thus the same word can be spelled in a variety of ways. This phonetic difference causes ambiguity and is a challenge to computational systems, such as data processing activities like tokenization and normalization. According to a study of Han, Cook, and Baldwin (2013) on the text of social media, non-standard spelling is an important factor that can impair NLP models' performance.

In addition, Roman Urdu is also a low-resource language variant no known big, annotated datasets and linguistic sources. Khan et al. (2024) pointed out that this disadvantage has Implications for formulation of strong NLP systems, and reduction in model generalization. The nature of Urdu, a language which is flexible and informal in its use, has caused difficulties in computational modelling and linguistic.

Code-Mixing and Code-Switching

The mixing and switching of codes are also common activities in multilingual societies, in particular, during online communication. Code-mixing is the use of two or more languages in a sentence, whereas code-switching occurs between sentences (Muysken, 2000; Myers-Scotton, 1994). Urdu-English code-mixing is so common in the Pakistani context because of the high prevalence of bilingualism and the sociocultural impact of English. Tagg (2020)

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

claimed that Users often confuse languages to express themselves in a better way, convey subtle meanings, and identify with their social circle.

Hidayatullah et al. (2022) affirm that the code-mixed language is a pronounced aspect of social Media communication and poses serious challenges in NLP systems. The intermingling of languages in the same text results in unstructured syntax, mixed lexica, and the boundaries of language are blurred, making language recognition and syntactic parsing difficult (Winata, et al., 2021).

In addition, Tagg (2020) and Javed et al. (2026) stated that code-mixing is closely related to sociolinguistic variables like identity, social status, and communicative competence. It mirrors user interaction with language resources in a digital context, in order to build meaning and express identity.

Social Media Language Characteristics

The kind of language used on social media sites is quite different from the written language. It is mostly informal and dynamic, and it is characterized by:

- Abbreviations and slang
- Phonetic spellings
- Emojis and symbols of expression.
- Non-standard grammar and sentence structure

Social media is producing massive amounts of user-generated, unorganized, and noisy text, which is valuable and difficult to analyse (Zappavigna, 2022). The users of Roman Urdu are phonetic typers and imaginative forms, which leads to more variability in their Urdu. It affects NLP applications like sentiment analysis, language detection, etc. The studies have recently identified the need for unique approaches to processing of social media text, as it is dynamic and linguistically inconsistent (Nazir, Bilal, & Shongwe, 2026). Such properties make it hard to reach high accuracy with traditional NLP models.

NLP Techniques for Low-Resource Languages

Minor languages like Roman Urdu are not only a challenge for NLP due to the scarcity of data, but also because they vary linguistically. One of the biggest obstacles to the development of effective computational models is the lack of standardized corpora and annotated datasets (Hidayatullah et al., 2022). Naive Bayes and Support Vector Machine are traditional machine learning models that have been extensively applied in text classification due to their efficiency (Joachims, 1998). Nevertheless, they tend to perform poorly when handling extremely variable and code-mixed data.

LSTM models and transformer-based models, based on deep learning, have demonstrated a better ability to capture contextual relationships within text. However, large datasets and computational resources are needed for these models, which can not be achieved in low-resource languages (Nazir et al., 2026). Data augmentation and transfer learning are recent methods that can overcome these issues. Although such methods enhance performance, they do not address the problems of linguistic variability and code-mixing.

Previous Studies

Current studies on Roman Urdu and code-mixed language processing have been mainly conducted on sentiment analysis, language identification, and the development of models. Recent works created Roman Urdu datasets and used deep learning methods to obtain better classification. For example, Wali and Tamoor (2024) introduced a Roman Urdu corpus and showed a better sentiment analysis performance with neural models. Similarly, Jawad et al. (2024) suggested a Roman Urdu sentiment analysis system in which language-specific preprocessing methods are important. Other studies have investigated the use of machine learning and hybrid methods to perform code-mixed text classification (Younas, Nasim, Ali, Wang, & Qi, 2020)

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

Although these improvements have been made, there still exist some constraints:

- Scarcity of good-quality annotated datasets.
- Depending on computational models without linguistic analysis
- Absence of emphasis on the actual Pakistani social media data.

Most of the current studies are focused on the performance of models, without considering the qualitative information on the language patterns and user behavior, which would be crucial to comprehend the context of code-mixing (Hidayatullah et al., 2022; Tagg, 2020).

Research Gap

According to the literature, the following research gaps are identified:

- Scarcity of studies specifically addressing Roman Urdu in real social media situation.
- Lack of integration of qualitative linguistic analysis and quantitative methods.
- Excessive reliance on complicated computational models, even with limited data.
- A shortage of small, manually annotated datasets to study in detail.

The current study will fill these gaps through a corpus-based manual analysis methodology used to examine linguistic patterns, code-mixing behavior, and sentiment trends in Roman Urdu and code-mixed text.

RESEARCH METHODOLOGY

This study utilizes a qualitatively-dominant mixed-method research strategy, merging fundamental quantitative analysis with a qualitative linguistic review. The primary goal is to analyse the language patterns, code-mixing, and sentiment in the dialogue on social media in Roman Urdu. Rather than intricate computational models, a corpus-based manual analysis method is used. This method is suitable for preliminary research for low-resource, non-standard languages, because it enables enhanced contextual comprehension of text (McEnery & Hardie, 2022). Furthermore, sociolinguistics often employs qualitative analysis to study language variability and identity in electronic communication (Tagg, 2023).

Data Collection

Platform Selection

Data was sourced from YouTube comments because there are large quantities of user-generated content that are publicly accessible. Social media captures real-life usage of language and communication practices (Zappavigna, 2024).

Topic Selection

Comments on the following subjects will be harvested to demonstrate the varied use of language:

- Pakistani dramas
- Cricket matches
- Political discussions

These domains are selected as they generate high engagement and varied emotional expression in online Pakistani communities.

Data Collection Method

The data is gathered through manual data mining, which involves extracting comments from the selected videos and recording them in an organized format using Microsoft Excel. In small-scale corpus studies, manual data collection provides control over data relevance and quality (McEnery & Hardie, 2011).

Sample Size

About 500-1500 comments are used as a dataset. This is a good size of qualitative and exploratory research where depth of analysis is the key rather than large-scale automation (Hirose, & Creswell, 2023).

Data Recording

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

All the comments with the following attributes are stored:

- Comment ID
- Text
- Platform
- Topic

Nature of Data

The dataset is composed of:

- Informal Roman Urdu text
- Urdu-English code-mixed language
- Short conversational expressions
- User-generated opinions

This data is indicative of actual linguistic practices in the online world, and is informal, creative, and variable (Tagg, 2020).

Data Cleaning

Limited preprocessing is done to maintain linguistic authenticity. It is carried out in the following steps:

- Removal of URLs
- Optional removal of emojis
- Preservation of original spelling and form

In linguistics, it is crucial to keep the original text in its original form, since overcleaning can eliminate useful variation (Zappavigna, & Ross, 2024).

Data Annotation

Annotation Framework

The main element of this research is annotation. All comments are categorized according to:

- Sentiment: Positive, Negative, Neutral
- Language Type: Roman Urdu, English, Mixed
- Code-Mixing Type:
  - Intra-sentential
  - Inter-sentential

These categories are common in the field of code-mixing and sentiment analysis (Das & Gambäck, 2014).

Annotation Process

All the annotations are done manually by the researcher. In low-resource language situations, manual annotation is favoured due to its greater contextual accuracy than automated methods (Fort, Adda, & Cohen, 2011). The annotation process is carried out several times to guarantee consistency and reliability in the dataset.

Data Analysis Methods

Frequency Analysis

Frequency counts are used to get a basic quantitative analysis:

- Percentage use of Roman Urdu vs. English.
- The percentage of mixtures of code language comments.
- Distribution of emotion categories.

Frequency analysis works well in determining trends in corpus linguistic research. (McEnery & Hardie, 2011).

Pattern Analysis

The identification of qualitative analysis is done:

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

- Common lexical patterns
- Common English inserts that are often used (e.g., scene, match).
- Common Roman Urdu phrases.

This method offers clues on the structure and use of language.

Code-Mixing Analysis

The study examines:

- Types of code-mixing
- Contexts where mixing is taking place.
- Potential motivators like convenience, fashion, and identity expression.

Code-mixing is closely associated with sociolinguistic parameters like identity and social situation (Tagg, 2023).

Sentiment Analysis

Sentiment is manually analysed, calculating the percentage of positive, negative, and neutral comments as opposed to automated models. This method guarantees increased contextuality in informal and ambiguous text (Liu, 2020).

Tools

The study uses:

- Microsoft Excel / Google Sheets to organize and analyse the data.
- Microsoft Word to document.

Advanced programming tools are not needed, which makes it easily available and affordable.

Presentation of Results

Results are presented using:

- Tables
- Bar charts
- Pie charts

These visualizations are generated using Excel to clearly represent patterns and distributions.

Validity and Reliability

Validity is achieved based on:

- Real, naturally occurring social media data used.
- Well-defined categories of annotations.

The reliability is ensured by:

- Consistent annotation criteria
- Rechecking the dataset several times

These actions improve the credibility of qualitative findings (Hirose, & Creswell, 2023).

Ethical Considerations

- Only publicly available comments are utilized
- The identities of users are anonymized.
- No personal information is published.

This makes it adhere to ethical guidelines in social media research (Zappavigna, 2015; Townsend & Wallace, 2016).

Data Analysis and Interpretation

This chapter gives a quantitative study on the Roman Urdu and code-mixed language in Pakistani social media. In contrast with the previous discussion, which was more oriented to the conceptual understanding of the nature of language, numerical data of the present chapter is used to determine the obvious patterns and tendencies. The objective is to make the results more organized, quantifiable, and comprehensible.

The chapter starts with the allocation of topics revealing the distribution of the collected data in various spheres including politics, drama, and cricket. This assists in interpreting the

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

context of the data and makes sure that the analysis is carried out on the basis of a balanced set of discussions. The distribution of the topic is significant as the language and mood usually vary with a topic under discussion. Then, the frequency table of the categories of sentiment is given. This analysis categorizes comments as positive, negative and neutral comments. It assists in determining the mood of the data in general. Even though this is an easy way to explain complicated human expressions, it comes in handy in giving a straightforward overall picture of how different people react on social media.

The frequency distribution of types of code-mixing is also in the chapter but it dwells on how users mix languages in their communication. It differentiates between intra-sentential and inter-sentential code-mixing. This will enable the study to recognize general trends in the use of two languages. Nonetheless, although this analysis indicates the frequency of each type, it does not exhaust how users come up with some specific combinations.

Lastly, topic and sentiment cross-tabulation is given to find out the differences in sentiment among different topics. This assists in knowing how some of these topics might give positive, negative, or neutral response. This form of analysis is even deeper as it displays correlations among variables, not individual trends.

This chapter is generally well presented and organized, with simple statistical methods used to present the data. It aids in determining the main trends as well as contributing to the overall goals of the research.

Frequency Distribution of Topics

Topic	Frequency	Percentage (%)
Politics	304	33.80%
Drama	298	33.10%
Cricket	298	33.10%
Total	**900**	**100%**

The topic distribution table above reveals that the data is nearly balanced in the number of politics, drama, and cricket with minimal variations in frequency. The number of comments on politics is slightly more than on the other two topics, whereas the drama and cricket are almost equal. It implies that the data is balanced and not centred on a particular area. Analytically, this even distribution is handy since it minimizes bias riskiness of the outcome. Had there been a topic dominating the dataset, it might have affected the general results particularly in aspects such as sentiment and language use. Rather, the close-balanced representation makes it possible to fairly compare the contexts. Simultaneously, the marginally increased proportion of political content can be a consequence of the deep involvement of users in political debates on social media. This aligns with the fact that political issues tend to bring more interaction and views. Nonetheless, the deviation is not so big that it will cause any major distortion to the analysis. In general, the table shows that the dataset has a solid foundation on which one can base further analysis. It aids the research by guaranteeing that all the chosen subjects are sufficiently covered to make justified comparisons in the subsequent sections.

Frequency Distribution of Sentiments

Sentiment	Frequency	Percentage (%)
Positive	312	34.70%
Negative	295	32.80%
Neutral	293	32.50%
Total	**900**	**100%**

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

The frequency table of the categories of sentiments indicates that the comments are fairly distributed between positive, negative and neutral sentiments. The percentage of positive sentiment is a little higher than those of the other two, whereas negative and neutral are quite close to each other. It means that no one emotional tone is predominant in the dataset. Analytically, such a balanced distribution is significant as it represents a realistic combination of opinions on social media. Users also provide positive as well as negative comments, and neutral notes. This diversity enhances the data set because it enables the research to capture various forms of responses instead of having a single form of sentiment. Nevertheless, the difference among the categories is very low as well, which also indicates that the general mood of the users cannot be described as very polarized. Rather, it indicates that the discussions are comparatively neutral, with no overwhelming bias towards either the positive or negative position. This might be because the contents of the dataset are varied and thus attract varying emotional responses. Generally, the table demonstrates that the dataset is balanced in terms of sentiments. This helps to affirm the validity of the subsequent analysis, particularly that of sentiment and other variables including topic or language use.

Frequency Distribution of Code-Mixing Type

Code-Mixing Type	Frequency	Percentage (%)
Intra-sentential	468	52.00%
Inter-sentential	432	48.00%
Total	**900**	**100%**

The frequency distribution of the types of code-mixing indicates that intra-sentential code-mixing is a little bit more prevalent than inter-sentential code-mixing. A little more than half of the remarks involve intra-sentential mixing and a a little less involves inter-sentential mixing. This means that the users more frequently intermix languages within one sentence as opposed to using languages on a sentence-to-sentence basis. Analytically, this implies that the bilingual users are at ease mixing up the Urdu and English at a more specific level when communicating. It is a natural and fluent application of both languages where there is a combination of words of various languages in one structure. Such mixing is common in the informal digital communication whereby there is no adherence to language rules. The distinction between the two types is however not so huge and it demonstrates that both types of code-mixing are widely used. This underscores the versatility of language usage on social media, whereby users' alternate languages depending on convenience, expression or habit. Simultaneously, this quantitative finding is not an exhaustive explanation of the causes of such mixing, which can depend on the social, cultural or contextual factors. In general, the table shows that the code-mixing is a robust and frequent aspect of the dataset. It helps to prove the point that Roman Urdu and mixed language use are significant peculiarities of online communication in Pakistan which makes them pertinent in the context of further analysis in this paper.

Cross-Tabulation of Topic and Sentiment

Topic	Positive	Negative	Neutral	Total
Politics	102	104	98	304
Drama	106	92	100	298

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Cricket	104	99	95	298
Total	**312**	**295**	**293**	**900**

The cross tabulation of topic and sentiment demonstrates the variation of emotional response on politics, drama and cricket. The table shows that all three topics are relatively equally distributed in terms of their positive, negative, and neutral sentiments with minor disparities between them. This implies that there is no issue that can be deeply connected to a certain kind of sentiment. In politics, the allocation is a little more towards negative sentiment than both positive and neutral. This can be indicative of the sensitive aspect of political discourse on social media, where people tend to show discontent or dissent. Nonetheless, the gap is not so high, which indicates that positive and neutral views are also present in the political discourse. In the case of drama, positive sentiment seems to be a bit more prevalent than negative sentiment, and neutral reactions also take a considerable percentage. This means that the users tend to interact with entertainment material in a more favorable or moderate manner. It implies that conversations related to drama could produce more admiration and less criticism than those related to other aspects. Equally, the sentiment of cricket is displayed with a relatively equal distribution, with a slight positive bias. This could be associated with national interest and emotional attachment with the sport which tends to result in supportive reactions. Nonetheless, the fact that negative and neutral sentiments are also present also indicates that users are also critically involved in consuming cricket-related content. In general, as it can be seen in the table, sentiment is somewhat different across the topics, yet the differences are not drastic. This equal balance hints at the idea that the user attitudes towards social media are varied and situation-specific. It also enhances the credibility of the dataset, given the fact that it has a broad spectrum of emotional reactions in various fields of discussion.

Discussion

The analysis of the present study is centered on the interpretation of the quantitative results in connection with the research goals, especially the analysis of the Roman Urdu and Urdu-English code-mixed language in social media. Prior research on code-mixed language and social media discourse has adopted an analogous approach by integrating quantitative evidence with linguistic interpretation to analyze language trends and user behaviors (Hidayatullah et al., 2022; Tagg, 2020). The findings offer valuable information about the linguistic patterns, distribution of sentiments, and connection between language usage and topic-based conversations. Simultaneously, the results reveal both the advantages and disadvantages of a quantitative method of the analysis of informal and non-standard language.

The most important conclusions of the research are the equal representation of the topics in politics, drama, and cricket. This implies that there is a great variety of real-life conversations that are prevalent in Pakistani social media represented in the dataset. In practical terms, such diversity enhances the validity of the study, since it does not over-represent a given domain. In social media and code-mixed language research studies, similarly domain-balanced strategies have been employed to enhance the representativeness and reliability of language analysis (Hidayatullah et al., 2022; Tagg, 2020). It also demonstrates that Roman Urdu and code-mixed language is not tied to a single type of conversation but are applied to various social and cultural environments. Nonetheless, although it is evenly distributed, it is confined to three topics only, and this is why there are other vital areas, including education, business or even religion, that are omitted. This limits the generalization of the results.

The sentiment categories analysis indicates that positive, negative and neutral sentiments are nearly equally distributed in the dataset. This is indicative of the ambivalence of online communication as users use all sorts of opinions and emotions. Previous studies regarding sentiment analysis and discourse on social media show similar findings and have

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

reported that online venues are home to both constructive and negative engagements depending upon the context (Liu, 2012; Zappavigna, 2022). The positive sentiment is a little bit higher, which indicates that social media is not just a place where people criticize but also appreciate and engage. Meanwhile, the fact that there is a considerable volume of negative sentiment confirms the hypothesis that users are actively engaged to spread dissatisfaction, particularly in the context of discussion involving political and social issues. This is supported by earlier findings by (Tagg, 2020) which showed that users demonstrate dissatisfaction and criticism, particularly with respect to politics and social issues. The most important aspect to note at this point is that the classification of sentiment was performed manually which enhances the sensitivity of the context but can result in subjectivity and enables superior understanding of informal and code-mixed languages in contrast to automated methods (Fort et al., 2011). The same comment may have different meanings to different researchers and this may lead to inconsistency.

The results that pertain to code-mixing reveal that intra-sentential mixing is more prevalent as compared to inter-sentential mixing. This shows that users feel free to mix languages in the same sentence as a manifestation of a natural style of bilingual communication. While most other studies have also investigated frameworks where intra-sentential mixing occurs, and have largely classified such practices as normative in online discourse (Das & Gambäck, 2014), the results of the current study contrast those studies where frequent inter-sentential switching tends to occur in multilingual communication as a result of underlying structural simplicity and a lack of grammatical complexity (Muysken, 2000). It also helps to prove the hypothesis that in Pakistan code-mixing is not a random process but certain informal patterns exist. On the linguistic side, this points to Roman Urdu as a versatile means of expression. Nevertheless, frequency analysis is able to reveal the frequency of code-mixing but it does not give an exhaustive explanation of why it takes place. This study does not directly measure factors like identity, social influence, or ease of expression, which limits a deeper understanding.

The cross-tabulation of the topic and sentiment is more informative as it shows the variations in emotional reactions in various regions of discussion. The findings reveal that political discourse is more negatively inclined, whereas drama and cricket exhibits a more neutral or positive course. This correlates with the overall discussion of social media behaviour where political content tends to attract argument and critique, and entertainment and sport content tend to attract more positive interactions (Tagg, 2020). Nevertheless, in the present study the differences are not that significant, and it indicates that sentiment is determined by various factors and can not be described by the topic itself. This shows the intricacy of human expression that may not be brought out completely by mere categorical analysis.

On the whole, the paper illustrates that it is possible to apply quantitative approaches to the investigation of general trends in Roman Urdu and code-mixed language. Frequency distributions and cross-tabulation have given clear and structured findings and the findings are easy to interpret. Concurrently, it is also revealed by the study that pure quantitative analysis has limitations to work with informal and creative language. Other crucial factors like tone, sarcasm, cultural allusions, and intent of user are hard to quantify.

To sum up, the findings confirm the notion that the Roman Urdu and code-mixed language are valuable characteristics of digital communication in Pakistan. They are broadly employed in various subjects and have a flexible and dynamic nature of expression. Although the quantitative method enhances the understandability and validity of the findings, some future study may be improved by integrating it with the qualitative analysis in order to get a more in-depth insight into linguistic behaviour. This would give a clearer picture of how and why users speak in this form of unique language.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

Conclusion and Recommendations

The concluding chapter of the given work summarizes the main conclusions and gives a clear conclusion and recommendations as well as implications. This chapter aims to think about the extent, to which the research objectives were met, and to point out how the results could be used in future research and in practice. The research aimed at investigating the Roman Urdu and the Urdu-English code-mixed language in social media through a quantitative methodology, with specific focus on the topics, sentiments, and the patterns of code-mixing.

The conclusion of the findings demonstrates that the data sample was quite balanced in relation to the three chosen subjects: politics, dra ma, and cricket. The number of comments was similar in each topic, and this aspect guaranteed that the analysis was not biased against a particular area. This balance helps in the dependability of the results and makes some meaningful comparisons in various contexts. The results also ratify the fact that Roman Urdu and mixtures of languages are highly utilized in all these issues and indicate that this type of communication is very popular in the context of different kinds of online discussions in Pakistan.

The sentiment results also indicated that the positive, negative and neutral categories were nearly equal. The percentage of positive sentiment was a bit higher, but the difference was not that great. It means that one emotion does not dominate the discussion in social media that much. Instead, there are a mixture of opinions expressed by users such as support, criticism and neutral. The resulting balanced distribution of sentiments indicates the heterogeneity of the user views and enhances the validity of the dataset in general.

The code-mixing types were analyzed and intra-sentential code-mixing was a little more common than inter-sentential code-mixing. This shows that users tend to combine Urdu and English in the same sentence instead of changing the sentences. This observation underscores the natural and spontaneous application of bilingual communication in less formal online spaces. It is also an indication that Roman Urdu offers a loose system through which users can easily mix languages.

The cross-tabulation of topic and sentiment also presented further understanding of the variation of emotions in various fields of discussion. Negative sentiment was a little more prevalent in political content, which is indicative of the critical character of political discourse. Conversely, drama and cricket talk exhibited more balanced or even positive trend, implying greater involvement and interest in these. The differences were not so extreme, however, and this can be taken to imply that sentiment is not affected by the topic alone but rather a combination of factors.

According to these results, some recommendations can be offered. To begin with, other areas like education, business and social issues need to be increased in future studies. This would enhance external validity of the findings and give a deeper insight on the use of language in Pakistani social media. Second, more data ought to be employed in order to increase the power of quantitative analysis. Although this research is valuable, a larger sample could have enabled more sophisticated statistical procedures and credible findings.

The second suggestion is the application of the automated tools and machine learning models to classify language and sentiment analysis. Although manual annotation guarantees an improved contextual insight, it is time-consuming and can be subjective. Efficiency and consistency can be enhanced with automated methods that are adequately trained on Roman Urdu and code-mixed data. These models should however be developed with care to address the peculiarities of non-standard language, like spelling variation and colloquialisms.

Moreover, in the future, quantitative and qualitative methods should be considered in one study. Although quantitative analysis can give out clear patterns and trends, qualitative analysis may give more information on why users act the way they do. As an illustration,

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

linguistic and social analysis would have to be more specific to grasp the motives of code-mixing or emotional expression. The mixed-method would thus give a more comprehensive view of the research problem.

This study has both academic and practical implications. Academically, the research is applicable to current body of natural language processing and sociolinguistics by considering a low-resource language setting. It emphasizes the need to create studies that are oriented to the local linguistic reality and not based on standard models of English or Urdu. Another significant finding of the research is that even the simplest forms of quantitative research could be useful when they are used prudently.

Practically, the results can be applicable in businesses, policymakers, and the development of technology. As an illustration, a better sentiment analysis of Roman Urdu data can assist firms to have a more accurate view of the opinions of the customers. In the same vein, such data can be employed by political analysts to monitor the opinion and involvement of the people. To developers, the paper points out the necessity to design the NLP tools to be able to deal with code-mixed and non-standard language, which is prevalent in most multilingual communities.

To sum up, this research has managed to meet its goals by offering a quantitative study of Roman Urdu and Urdu-English code-mixed language in the social media. The results indicate that the use of language is dynamic, flexible and is highly dispersed on a variety of topics. Although the quantitative method is clear and structured, it also lacks in the ability to represent deeper meanings and motivations to use language. Future studies ought to thus expand on this research through the use of larger datasets, sophisticated methodologies and mixed methods. On the whole, the research adds to the improved interpretation of digital communication in Pakistan and the necessity to localize the approach in accordance with the linguistic conditions.

References

- Fort, K., Adda, G., & Cohen, K. B. (2011). Last words: Amazon Mechanical Turk: Gold mine or coal mine?. *Computational Linguistics*, *37*(2), 413-420.
- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Tagg, C. (2020). *Exploring digital communication: Language in action*. Routledge.
- Labov, W. (1972). Sociolinguistic patterns (university of pennsylvania, philadelphia).
- Hymes, D. (2013). *Foundations in sociolinguistics: An ethnographic approach*. Routledge.
- Wardhaugh, R., & Fuller, J. M. (2021). *An introduction to sociolinguistics*. John Wiley & Sons.
- Younas, A., Nasim, R., Ali, S., Wang, G., & Qi, F. (2020, December). Sentiment analysis of code-mixed Roman Urdu-English social media text using deep learning approaches. In *2020 IEEE 23rd international conference on computational science and engineering (CSE)* (pp. 66-71). IEEE.
- Han, B., Cook, P., & Baldwin, T. (2013). Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology, 4*(1), 1–27.
- Hidayatullah, A. F., Qazi, A., Lai, D. T. C., & Apong, R. A. (2022). Language identification of code-mixed text: A systematic review. *IEEE Access, 10*, 114740–114760.
- Jawad, K., Ahmad, M., Alvi, M., & Alvi, M. B. (2024). RUSAS: Roman Urdu sentiment analysis system. *Computers, Materials & Continua, 79*(1), 1–18.
- Nazir, M. K., Bilal, M., & Shongwe, S. C. (2026). Sentiment analysis for code-mixed low-resource languages: a systematic review of approaches, techniques, applications, challenges, and future directions. *Social Network Analysis and Mining*, *16*(1), 47.
- Tagg, C. (2020). *Exploring digital communication: Language in action*. Routledge.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL
JALT

- Khan, M., Naseer, A., Wali, A., & Tamoor, M. (2024). A roman Urdu corpus for sentiment analysis. *The Computer Journal*, *67*(9), 2864-2876.
- Zappavigna, M. (2022). *Searchable talk: The linguistic functions of hashtags*. Bloomsbury Academic.
- Muysken, P. (2000). Bilingual speech: A typology of code-mixing.
- Myers-Scotton, C. (1994). Social motivations for codeswitching. Evidence from Africa. *Multilingua-Journal of Interlanguage Communication*, *13*(4), 387-424.
- Winata, G. I., Madotto, A., Lin, Z., Liu, R., Yosinski, J., & Fung, P. (2021, November). Language models are few-shot multilingual learners. In *Proceedings of the 1st Workshop on Multilingual Representation Learning* (pp. 1-15).
- Javed, K., Azam, I., & Khuram, M. (2026). CODE-MIXING ON FACEBOOK POSTS: A CORPUS-BASED ANALYSIS OF LINGUISTIC PRACTICES IN DIGITAL COMMUNICATION AMONG PAKISTANI YOUTH. *Journal of Applied Linguistics and TESOL (JALT)*, *9*(1), 185-195.
- Joachims, T. (1998, April). Text categorization with support vector machines: Learning with many relevant features. In *European conference on machine learning* (pp. 137-142). Berlin, Heidelberg: Springer Berlin Heidelberg.
- Chung, D. T. (2025). How user-generated content on social media platform can shape consumers' purchase behavior? An empirical study from the theory of consumption values perspective. *Cogent Business & Management*, *12*(1), 2471528.
- Bali, K., Sharma, J., Choudhury, M., & Vyas, Y. (2014, October). "i am borrowing ya mixing?" an analysis of english-hindi code mixing in facebook. In *Proceedings of the first workshop on computational approaches to code switching* (pp. 116-126).
- Johri, P., Khatri, S. K., Al-Taani, A. T., Sabharwal, M., Suvanov, S., & Kumar, A. (2021, March). Natural language processing: History, evolution, application, and future work. In *Proceedings of 3rd International Conference on Computing Informatics and Networks: ICCIN 2020* (pp. 365-375). Singapore: Springer Singapore.
- Mao, Y., Liu, Q., & Zhang, Y. (2024). Sentiment analysis methods, applications, and challenges: A systematic literature review. *Journal of King Saud University-Computer and Information Sciences*, *36*(4), 102048.
- Han, B., Cook, P., & Baldwin, T. (2013). Lexical normalization for social media text. *ACM Transactions on Intelligent Systems and Technology (TIST)*, *4*(1), 1-27.
- Nazir, M. K., Bilal, M., & Shongwe, S. C. (2026). Sentiment analysis for code-mixed low-resource languages: a systematic review of approaches, techniques, applications, challenges, and future directions. *Social Network Analysis and Mining*, *16*(1), 47.
- Jawad, K., Ahmad, M., Alvi, M., & Alvi, M. (2024). Rusas: Roman urdu sentiment analysis system. *Computers, Materials, & Continua*, *79*(1), 1463.
- Hirose, M., & Creswell, J. W. (2023). Applying core quality criteria of mixed methods research to an empirical study. *Journal of Mixed Methods Research*, *17*(1), 12-28.
- Das, A., & Gambäck, B. (2014, December). Identifying languages at the word level in code-mixed indian social media text. In *Proceedings of the 11th International Conference on natural language processing* (pp. 378-387).
- Fort, K., Adda, G., & Cohen, K. B. (2011). Last words: Amazon Mechanical Turk: Gold mine or coal mine?. *Computational Linguistics*, *37*(2), 413-420.
- Liu, B. (2020). *Sentiment analysis: Mining opinions, sentiments, and emotions*. Cambridge university press.
- Townsend, L., & Wallace, C. (2016). Social media research: A guide to ethics. *University of Aberdeen*, *1*(16), 1-16.

ISSN E: 2709-8273
ISSN P:2709-8265
JOURNAL OF APPLIED LINGUISTICS AND TESOL

- McEnery, T., & Hardie, A. (2011). *Corpus linguistics: Method, theory and practice*. Cambridge University Press.
- Tagg, C. (2023). Digital language and communication. In *The Routledge handbook of applied linguistics* (pp. 68-80). Routledge.
- Zappavigna, M. (2015). Searchable talk: The linguistic functions of hashtags. *Social semiotics*, *25*(3), 274-291.
- Zappavigna, M., & Ross, A. S. (2024). *Innovations and challenges in social media discourse analysis*. Routledge.