International Journal on Recent and Innovation Trends in Computing and Communication
ISSN: 2321-8169 Volume: 11 Issue: 11
***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 December 2023*

Sentiment Analysis of Code-mixed Roman Urdu-English Social Media Text

Gazi Imtiyaz Ahmad, Syed Ishfaq Manzoor, Jimmy Singla

ABSTRACT

Evaluation of digital and social landscape continues in unexpected ways. The internet continuously grows as individuals discover new means to consume information. Internet platforms led the charge for Internet usage. Almost 60% of the world's population are active social media users who come to rely on social media platforms to satisfy their daily data needs. Every day, billions of text messages are produced on social media sites like Facebook, Instagram, and Twitter, creating a substantial amount of data from people all around the world. Businesses and organizations use text analysis tool in social media monitoring to learn how the public views their brand, goods, and services. Sentiment analysis is one of the most popular applications of text analysis in social media monitoring. This method classifies text as neutral, negative, or positive using natural language processing (NLP) techniques. Sentiment analysis can be used to monitor changes over time and spot trends in customer sentiment. This can be helpful for tracking the success of marketing activities and for spotting possible problems or opportunities for a business.

The use of code-mixed text on social media platforms serves as a key communication method for multilingual users. It enables the integration of more than one language within a single message, which often mirrors cultural identity, conveys nuanced meanings, and promotes effective communication among diverse linguistic groups, especially in areas where multiple languages are prevalent. However, due to the intricate interactions between multiple languages, such as word meaning ambiguity, grammatical variations, and the scarcity of easily accessible labeled datasets, code-mixed text sentiment analysis presents numerous challenges that make it challenging to accurately assess the sentiment of a text. This paper focuses on leveraging Long Short-Term Memory (LSTM) networks for sentiment analysis of Urdu-English code-mixed text. By addressing the inherent challenges and optimizing LSTM models, this study contributes to developing effective sentiment classification frameworks for this underrepresented language pair.

Key Words: Code-mixed, Urdu-English, Machine Learning, Deep Learning, LSTM

INTRODUCTION

Sentiment analysis, also known as opinion mining, is a natural language processing task used to determine the emotional tone of text, and used in various contexts like social media and customer feedback. In natural language processing, sentiment analysis is a common activity with the objective of classifying the text depending on the mood or attitude conveyed in the text, which can be positive negative, or neutral. Sentiment analysis is the contextual meaning of words that reveal a brand's social sentiment and assist a company in assessing whether or not the product it is producing will be in high demand. Text analysis and natural language processing techniques are used in sentiment analysis to categorize words as neutral, negative, or positive. This enables businesses to get a general idea of how their clients feel about the brand.

Non-English speakers often use phonetic typing, code-mixing, and Anglicisms in social media texts, making automatic language detection challenging. Linguists use terms like 'code-mixing' or 'code-switching' to describe these phenomena, but 'code-mixing' is mainly used for intra-sentential code-switching, where language changes occur within a sentence, and 'code-switching' for inter-sentential phenomena. Different criteria distinguish these phenomena, making automatic language detection challenging. [1] Compared to more formal communications, code-mixing is far more common on social media. Code-mixed text, often found in social media content like Facebook, Twitter, and forums, is a common practice among multilingual speakers and writers who prefer to convey information in their native language. [2]

When it comes to the resources and tools available for sentiment analysis, English is regarded as a rich language. However, there are a number of other languages that are regarded as having inadequate resource availability, and Roman Urdu is regrettably one of them. Urdu is the official language of Pakistan and is spoken and understood in five South Asian countries. Therefore, standard corpora are also necessary for any NLP assignment, but morphologically, it is quite challenging because Urdu and Hindi are relatively similar spoken languages that can be understood by people from both countries, but their writing systems diverge. Roman Urdu, which uses English alphabets to write Urdu script, is hence chosen over Urdu script because Roman Urdu scripts are typically used by South Asian users (such as those from Pakistan, India, and other South Asian countries) to communicate on social media. For instance, " where are you going" is written in Roman Urdu as " kahan ja rahay ho." Because it enables non-Urdu speakers to comprehend the feelings of Urdu speakers behind text, sentiment analysis for the Roman Urdu language is therefore very crucial. Additionally, by comprehending the attitudes of global consumers, businesses can effectively grow their market on a global scale. Since the majority of Urdu speakers use Roman to convey their feelings, ideas, thoughts, and opinions, sentiment research for Roman Urdu is crucial if the South Asian market is to be targeted. [3]

Therefore, Sentiment analysis in Roman Urdu-English is crucial for understanding the emotions of non-Urdu speaking users and expanding international markets. Urdu is spoken in five South Asian regions, and its writing scripts differ from Hindi. Using Roman Urdu scripts on social media platforms like Facebook, Twitter, and YouTube allows for better understanding of emotions and ideas expressed by Urdu speakers. This is especially important for targeting the South Asian market, as most Urdu speakers express their thoughts and opinions using Roman Urdu scripts.

Most of the work on sentiment analysis has been carried out on monolingual and resource rich languages such as English. However, with the increase of multilingual users on the Internet and on social media platforms, it is now a day a common practice of people to post their views, opinions, reviews on products, services, social events and government policies in multiple languages which make Sentiment Analysis process more challenging. People use words or phrases of one language and mix them in another language in one sentence. Also these kind of posts are usually informal in nature.

For a number of reasons, including sociolinguistics, cultural identification, and the unique affordances of online communication, code-mixed and code-switched languages have emerged as the de facto languages on social media. The frequency of code-mixed and code-switched languages on social media illustrates how digital communication in today's globalized world is fluid, dynamic, and hybrid.

Code-mixed texts, in which users combine two or more languages in a single chat, have become a prevalent form of expression with the rise of social media and online communication, particularly in multilingual communities. Sentiment Analysis of such text is an essential field of research for comprehending user sentiment, public opinion, and sociocultural trends.

People from multilingual societies such as India, Southeast Asia and Latin America where code-mixed communication on social media platforms like Twitter, Facebook, YouTube and WhatsApp is a new norm. Code-mixing is also the default way many users communicate online, particularly in informal settings, making it essential to analyze sentiment in such texts. Conventional sentiment analysis algorithms are unable to manage the linguistic complexity of code-mixed texts because they are designed for monolingual text, frequently in English. Ignoring code-mixed nuances can lead to misinterpretation of user sentiment, especially when the emotional tone depends on specific word combinations or cultural expressions. Code-mixed languages effectively convey emotions, humor, and sarcasm, making them useful for businesses to gauge customer opinions. In multilingual regions, this analysis provides accurate insights into customer satisfaction and brand perception. During events like elections, public sentiment is often expressed in code-mixed languages, necessitating effective analysis to understand grassroots opinions. Therefore, Sentiment analysis of code-mixed text is essential in the multilingual, digital world of today. It facilitates precise insights into user sentiment, closes linguistic gaps, and spurs advancements in AI and NLP. In order to ensure inclusivity and relevance in the study of international online communication, researchers and developers can fully utilize sentiment analysis in a variety of culturally rich contexts by tackling the issues of code-mixing.

This research aims to address the growing need for sentiment analysis techniques that can effectively capture sentiments in code-mixed Urdu-English social media text. Urdu and English are two widely spoken languages, and their code-mixing is a common occurrence, especially in regions where both languages hold significance. Furthermore, the sentiment analysis of code-mixed text in these languages has practical

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 December 2023*

applications in fields such as market research, political analysis, customer service, and social media monitoring. The study uses deep learning and multilingual embeddings to analyze code-mixed conversations, constructing a dataset of Urdu-English text from social media platforms and annotating it with sentiment labels. We propose a sentiment analysis model using LSTM based architecture for classification.

LITERATURE REVIEW

Sentiment analysis, a field of computational linguistics, has evolved from rule-based techniques to machine learning techniques. It uses sentiment lexicons and sentiment scores to classify and identify words. “The Automatic Creation of Literature Abstracts”, a 1958 study by H.P. Luhn [4] served as the foundation for the text analysis methods employed in sentiment analysis. As data and computational power grow, supervised and unsupervised machine learning methods become more popular [5]. Social media platforms like Twitter and Facebook have increased the importance of sentiment analysis [6.]. Deep learning, particularly recurrent neural networks and convolutional neural networks, has revolutionized sentiment analysis by learning features from raw text data [7]. The field continues to evolve with advancements in Natural Language Processing and AI technologies. For sentiment analysis, a variety of methodologies and strategies, from rule-based approaches to sophisticated machine learning and deep learning techniques, have been developed. Rule-based sentiment analysis is a simple, straightforward method for calculating text sentiments using predefined rules and NLP techniques like stemming, part-of-speech tagging, parsing, lexicons, and tokenization. However, it often overlooks text and word combinations. [8]. Lexicon-based sentiment analysis uses a dictionary of words with polarity labels to determine sentiment polarity in a document. Sentences are tokenized, matched with words in the model, and a combining function is used to predict the total text component [9]. A machine learning approach uses a labelled dataset to train a classifier, then uses the model it builds to predict sentiment. The standardization of the text input by pre-processing and elimination of any extraneous information are the first steps taken by machine learning algorithms in sentiment analysis. In order to represent the text as numerical characteristics that can be input into machine learning classifiers, feature extraction techniques like Term Frequency-Inverse Document Frequency (TF-IDF) and N-grams are then applied. Unsupervised learning, which identifies hidden patterns or intrinsic structures in input data, and supervised learning, which builds a model using known input and output data to predict future outputs, are the two types of approaches used in machine learning [10]. Also ensemble approaches where multiple sentiment analysis models are combined to improve the performance, hybrid approaches combined multiple techniques to leverage their strengths are also used form sentiment analysis.

Studies have suggested that English is the dominant language on social media platforms and most frequently used languages for web content. Thus, Natural Language processing tools and techniques have mainly focused on English language. Over the years linguistic and AI tools have been developed and implemented on single language predominantly on English. Also Sentiment analysis, a subset of natural language processing (NLP), has seen significant advancements in the analysis of monolingual data. However, it is typical to find posts written in multiple languages due to the necessity of globalization and the growing number of Internet users worldwide. Additionally, people frequently combine different languages in a single sentence while writing unstructured content, such as tweets. Actually, it has become common to use many languages in a single statement. One common linguistic occurrence seen in multilingual society is code-mixing, which is the mixing of two or more languages in a single sentence or discussion.

Sentiment analysis of code-mixed text is essential in the multilingual, digital world of today. It facilitates precise insights into user sentiment, closes linguistic gaps, and spurs advancements in AI and NLP. In order to ensure inclusivity and relevance in the study of international online communication, researchers and developers can fully utilize sentiment analysis in a variety of culturally rich contexts by tackling the issues of code-mixing.

The study of code-mixing is grounded in a number of linguistic theories. Poplack [11] distinguishes three categories of code-switching: inter-sentential, intra-sentential, and tag-switching. Myers-Scotton's [12] Matrix Language Frame Model offers a framework to understand the interactions between a matrix language, the language whose grammatical structure is followed in the utterance, and an embedded language, the language whose words, phrases, idioms are embedded in matrix language, in sentences. Sociolinguists have long been interested in code-mixing. Along with its linguistic structures, studying its functional form from a sociolinguistic standpoint has grown in importance. Politicians employ language and linguistic systems to build relationships with citizens in a multilingual nation like India. Both social and linguistic situations are affected by code-mixing. In multilingual communities, it serves as a social expression of cultural identity and group membership. It offers important insights into the interactions between various languages and dialects from a linguistic perspective.

The Indo-European language family includes the Urdu language, which belongs to the Indo-Aryan group. [13] Nearly 70 million people speak Urdu as their primary language, and over 100 million, mostly in Pakistan and India, speak it as a second language. In addition to being formally recognized, or "scheduled," in the Indian constitution, it is the official state language of Pakistan. The United States, the United Kingdom, and the United Arab Emirates also have sizable speech communities. Roman Urdu is essentially a combination of the Urdu language and the English/Latin alphabets. In other words, Urdu messages are conveyed using English alphabets. Internet users are pleased with the results when they use Roman Urdu for socializing or informational purposes. Several well-known Urdu news websites have developed their own methods for handling Roman Urdu. For people who are not familiar with Arabic or traditional Nastaliq script, Roman Urdu has been a huge help. A number of social media platforms are serving as testing grounds for this new Urdu script. [14]

People in the subcontinent share their views, opinions, reviews, likes and dislikes on products and services, movies and film stars, elections and politicians, social and cultural issues, government policies etc. and more than half of these views expressed on social platforms are code-mixed in nature. Because Urdu-English code-mixing is so common in media, internet communication, and daily conversation among Urdu-speaking groups, it has attracted more and more attention.

However, code-mixed communication is informal in nature and as such there exists several spelling variants of a single word such as the word ‘you’ in roman Urdu can have following different spellings “aap”, “ap”, “app”, “aaap”, “aapp” [15]

Sentiment Analysis is a crucial research area in natural language processing (NLP), with various machine learning and deep learning models proposed. It can be categorized into four domains: subjectivity, word, document, and opinion extraction. It can be applied to various domains, including product reviews, stock prediction, movie reviews, news articles, and political debates. Sentiment analysis can be examined at document, sentence, and aspect levels, with the document level determining the sentiment associated with the target. [16]

Sentiment analysis, a natural language processing (NLP) technique, uses free text data to understand public attitudes and modes towards social, ethical, political, company products, or services. [17]

Sentiment analysis gained prominence with the rise of computer-mediated social networks, where public opinions are expressed freely, often using English script, and researchers have developed resources to calculate these sentiments. [18] [19] [20]

Code-mixed sentiment analysis is a subfield of sentiment analysis that focuses on analyzing text or speech containing a mixture of two or more languages or language varieties. It has gained prominence in recent years due to its importance in understanding sentiment in multilingual and multicultural contexts. Doruöz, A. Seza, et al. in [21] conducted a thorough survey of computational linguists and language technologists regarding the linguistic and social elements of code-switched and code-mixed research with a focus on the multilingual contexts of Europe and India. The survey provides a comprehensive overview of research on code-switching in social media, a crucial aspect for comprehending code-mixed sentiment analysis. Code-mixing is a common social media phenomenon where users switch between languages based on audience, context, and personal preferences. It can express emotions, emphasize words, signal group identity, and accommodate multilingual followers. However, it presents challenges for sentiment analysis models, and understanding these variations is crucial for analysing code-mixed sentiment.

Sentiment analysis is crucial for marketing and branding strategies, enhancing business sales and user experience. Roman Urdu sentiment analysis has received limited research compared to other languages. [22]. Roman Urdu is also a popular social media platform communication language. However, due to

Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 December 2023*

lack of benchmark corpus and other language assess, poses, a challenge crucial for natural language processing tasks [23].

Mehmood et al. [24] conducted sentiment analysis on 779 Roman Urdu reviews using unigram, bigram, and uni-bigram features and five classifiers. They found Naïve Bayes and Logistic Regression outperformed other classifiers in sentiment analysis accuracy, providing a reference point for future studies.

A sentiment analysis system for Roman Urdu/Hindi was presented by Arif et. al [25]. The authors used various supervised learning algorithms on 12000 sentences labeled with positive, negative and neutral sentiments. For feature selection, the authors employed countVectorizer, HashingVectorizer and Term Frequency-Inverse Document Frequency (tf-idf). The experimental results showed that SVM performs better than other machine learning methods. The paper also highlighted the issues and problems of code-mixed data such as spelling inconsistency as well as transliteration issues. Chandio et al. [26] used machine learning techniques to analyze Roman Urdu reviews of e-commerce products, achieving a 92% accuracy rate on a dataset of 26,000 Roman Urdu e-commerce reviews. The authors propose a fine-tuned Support Vector Machine (SVM) powered by RomanUrdu Stemmer, which uses a dictionary-based Roman Urdu stemmer to standardize text and minimize complexity. The model's efficacy is evaluated using various experimental configurations and compared to machine learning and deep learning models. However, they acknowledged potential bias and the issue of informal language.

A self-attention Bidirectional LSTM model is proposed by Manzoor et.al. [27] for sentiment analysis of Roman Urdu social media text of 10000 sentences. A normalized dataset of 3000 sentences also utilized for comparison. The authors developed a deep neural network model for lexical variation and sentiment analysis of Roman Urdu sentences, using the Self-attention Bidirectional LSTM network. The model achieved 69.3% accuracy on sentiment classification, offering insights into advanced neural network architectures.

Code-mixing is a prevalent trend on social media platforms, particularly in multilingual societies, reflecting natural conversational trends and becoming a prominent feature of informal digital communication. Code-mixing on social media is an evolving trend reflecting language's dynamic nature, influenced by context, convenience, and cultural identity, and is expected to significantly shape digital communication norms. However, code-mixed communication on social media platforms presents unique challenges due to the blend of multiple languages, scripts, and cultural contexts, impacting users, platform moderation, and Natural Language Processing tools. Code-mixed text, involving frequent language switching, can complicate language identification, leading to errors in tasks like sentiment analysis or content moderation. Code-mixing, due to its lack of formal rules, leads to inconsistencies in grammar, syntax, and vocabulary, making it challenging to create reliable text processing algorithms or tools. Social media communication is often informal, utilizing slang, abbreviations, and emojis, adding uncertainty and making it challenging for models to interpret context or sentiment. Code-mixed text, containing multiple meanings due to language or cultural context, can cause misinterpretation in automated systems like content moderation or recommendation engines. Code-mixed datasets are limited, particularly for low-resource languages, causing NLP models to struggle to generalize or perform accurately on real-world code-mixed text. Code-mixed text, which often combines languages to alter emotional tone, may be challenging for sentiment analysis models to accurately classify. Code-mixed social media communication challenges require innovative NLP solutions, annotated datasets, and context-aware algorithms, requiring collaboration between linguists, technologists, and companies.

Chandio et al [28] developed a deep recurrent architecture, RU-BiLSTM, using bidirectional LSTM, word embedding, and an attention mechanism for sentiment analysis of Roman Urdu. The model was tested on two Roman Urdu datasets with 26,824 and 10,021 reviews, using three neural word embedding schemes word2Vec, Glove and fast-text. The authors achieved an accuracy of 80.63% and 77.50%. respectively.

Alvi, Muhammd Bux, et al [29] perform sentiment analysis of code-mixed Roman Sinddhi-English textual data of 4500 sentences. The authors developed a lexical Roman Sindhi sentiment dictionary for detecting sentiment orientation. They also created two interfaces to make use of the lexical resources: a Roman Sindhi rule-based sentiment scorer (RBRS3)

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 December 2023*

that rates the sentiment of Roman Sindhi script features, and a Roman Sindhi to English translator (RoSET) that converts a Roman Sindhi feature into an equivalent English term.

Nagra, Arfan Ali, et al.[30] proposed a sentiment analysis model on Roman Urdu corpus of 10,021 sentences using faster recurrent convolutional neural network (FRCNN), RCNN, rule-based and N-gram model. The authors achieved an accuracy of 91.73% for binary classification and 89.94% for tertiary classification with faster RCNN model.

Jawad, Kazim, et al [31] developed a comprehensive lexical dictionary for Roman Urdu, a Bilingual Roman Urdu Language Detector, and a Roman Urdu spelling corrector. Data from public reviews on Daraz and GoogleMaps is used to create a Roman Urdu Sentiment Analysis System (RUSAS). A Flask framework-based web app provides the system with a web API, achieving an accuracy of 93.4%.

Early multilingual sentiment analysis focused on individual languages, but as social media and online communication platforms became more prevalent, users began code-mixing in their text and speech. Researchers in the NLP and machine learning communities began working on techniques to analyse sentiment in code-mixed text, using datasets and specialized models incorporating neural networks and deep learning techniques.

The authors in [32] developed a bilingual sentiment analysis system using LSTM and ensemble methods to analyze Hindi-English code-mixed social media content. They developed a bilingual system using baseline models and selected four best performing classifiers. The proposed model achieved an accuracy of 0.73 and f score of 059. Agarwal, Alekh, and Bhattacharyya [33] present a technique for sentiment analysis of movie reviews, incorporating linguistic information without expert intervention. They also propose a Wordnet-based method for improved classification accuracy across a test dataset. Prabhu, Ameya, et al. [34] introduce a Sub-Word Long Short Term Memory model for learning sentiments in a noisy Hindi-English Code Mixed dataset, using sub-word level representations instead of character-level or word-level representations. The authors in [35] used feature extraction with TF-IDF, machine learning methods like Logistic Regression/Random Forest, and transfer learning with the BERT approach for sentiment analysis of Hindi-English code-mixed data, and they were successful in obtaining an F1-score of 0.693. The annotated dataset for sentiment analysis in code-mixed Telugu-English text is introduced by Varma et al. [36]. Sentiment analysis was carried out utilizing unique unsupervised data normalization with an MLP model. On the created dataset, the authors reported an accuracy rate of 80%.

Table 1: Comparative Analysis of Existing Systems

Reference	**Corpus Size**	**Features**	**Methods/ Techniques**	**Results / Performance**
				Accuracy
Mehmood, Khawar, et al (2019)	11000	Word and Character N-gram, TF-IDF	LR, NB, ANN	0.82
Younas et al. (2020)	20,735	Contextual Embeddings	mBERT , XLM-R	0.71
Shakeel et al. (2020)	20, 735	ELMO, ConvNet	McM, LSTM, CNN	0.69
Rizwan et	10, 012	ELMO,	LSTM, BERT, BiLSTM,	0.89

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 December 2023*

al. (2020)		FastText, LASER	CNN, XLM-R	
Ilyas et al. (2023)	20,000	GLOVE, FastText, Word2Vec, Counter Vectorizer	SVM, RF, LR, DT, CNN, LSTM, BERT	0.88
Khan et al. (2022)	20,228	N-gram, TF-IDF, GloVe Word2Vec FastText	RF, LR, SVM, NB, KNN CNN-LSTM	0.90
Ahmad and Singla (2022)	11299	Character Based Embeddings	ANN, LSTM	0.72
Chandio, Bilal Ahmed, et al (2022)	20,824	word2vec, Glove, , FastText	BiLSTM	0.74
Azhar, N., & Latif, S. (2022)	24,000	Word Embeddings	NB, SVM, LR, KNN, ANN, CNN, RNN, ID3 and GB Tree	0.92
Li, Dun, et al (2022)	20229	TF-IDF, word embeddings	NB, SVM, RF, LR, CNN, RNN, LSTM	0.90
Qureshi, Muhammad Aasim, et al (2022)	24,000	TF-IDF, word embeddings	NB, ID3, GB, SVM, LR, KNN, ANN, CNN and RNN	0.92
Javed and Saeed (2023)	89793	Word Embeddings	LSTM, mBERT	0.90
Qureshi et al. (2023)	1990	Rapid-Miner	KNN, NB, DT	: 0.82
Hashmi et. el (2024)	20,735	Contextual Word Embeddings	cm-BERT M-BART, Electra	0.74
Jawad, Kazim, et al (2024)	37,094	POS, and contextual-based sentences	MLAs, DL	0.94

From the literature survey it has been relieved that although monolingual text remains predominant, however, bilingual and multilingual text is also increasing on social media platforms. But the presence of bilingual and multilingual text depends on the regions and user demographics. Therefore, the task of collecting and annotating code-mixed datasets particularly for the resource poor language pairs like Roman Urdu-English datasets is a challenging task. Using several languages or linguistic codes in a discourse is sometimes referred to as code-mixing. It is important to remember, too, that code-mixing can also refer to the use of multiple languages in addition to dialects or variations of the same language. Relatively little research has been done on code-mixing between different dialects of the same language, despite the fact that many studies have concentrated on code-mixing between distinct languages or language codes. Researchers have mostly employed machine learning (ML), followed by lexicon-based and deep learning-based approaches. However, the deep learning approach has drawn interest as a means of improving performance. Sentiment analysis has been done at the sentence level using binary and tertiary sentiment classes. Datasets such as tweets, Facebook comments, YouTube comments, WhatsApp chats were used. These datasets are of varying domains such as product and service reviews, movies reviews, e-commerce reviews and opinions, social and political views etc. But researchers have always used a specific domain for their experimentations. A number of features and methods have been used for sentiment classification. Language models and transformed based learning models have also provide good results.

METHODOLOGY

The methodology of this study is divided into five steps. Data collection, data preprocessing, feature extraction, model development and results & discussion.

Dataset

There does not exists any platform or web resource where people exclusively communicate in code-mixed languages and code-mixed text usually co-exist with other languages. Lack of standardized datasets of code-mixed text particularly for resource poor languages is one of the primarily challenges in sentiment analysis process. Also unavailability of automatic filtering tools hinders in creation of large scale code-mixed corpus [37]. Therefore, due to linguistic and technical issues, collection of code-mixed text data is a complex and challenging task. In this study we have collected data from various sources using API scraping and web crawling. The initial collection of textual data consists of more than 100000 sentences (tweets, comments, views, reviews, opinions etc.).

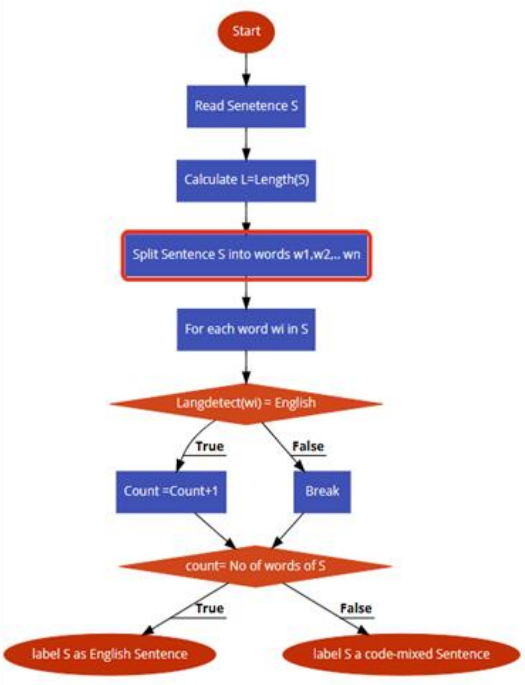


Figure 1. Sentence classification into monolingual and code-mixed sentences.

Data Preprocessing

An essential step in any Natural Language Processing (NLP) is text preprocessing which involves cleaning of raw text data and preparing it for NLP task. Alam and Yao in [38] studies the impact of preprocessing steps on the accuracy of machine learning algorithms for sentiment analysis. The dataset used in this study underwent comprehensive preprocessing to enhance the quality of the input data and improve model performance. First the sentences were filtered for monolingual (English Only) and code-mixed (English and Roman Urdu) sentences. The sentence was treated as code-mixed if it contains one or more words that does not belong to the English vocabulary. The process of separation of sentences into monolingual and code-mixed is given in figure 1.

The raw text data was cleaned using a custom text cleaning function. This function systematically removed user mentions (e.g., @username), URLs, and special characters, retaining only alphanumeric content to maintain meaningful information. After cleaning, all text was converted to lowercase to eliminate discrepancies caused by case sensitivity. Next step involves labelling of sentences as annotation is crucial for building effective sentiment analysis models for Urdu-English code-mixed data, ensuring high-quality training and evaluation of machine learning and deep learning models. For this an in-house annotation tool was developed for the annotation of text. The annotation process was done by experts who were familiar with both English and Urdu languages. The annotation process performs two types of tasks (1) sentence level sentiment classification into “positive”, “negative” and “neutral” and (2) word level language tagging into “English”, “Urdu” and “Others”. The inter annotator agreement using Cohen’s kappa statistic and as explained in [39] was more than 90%.

To further refine the dataset, common stop words that add little semantic value were removed, reducing noise and focusing on informative content.

Table 1: Dataset Description

Sentiment Class		**Training**	**Test**	**Validation**
Positive	25229 **(43%)**	20183	4037	1009
Negative	17349 **(30%)**	13879	2776	694
Neutral	15948 **(27%)**	12758	2552	638

Table 2: Example of Roman Urdu-English Sentences with English translation and polarity

Roman Urdu-English Sentence	**English Translation**	**Polarity**
is mobile ki qeemat bohat munasib or achi hain	The price of this mobile is very reasonable and good	Positive
app istemaal karne aur offers haasil karne ke liye amlay ki behtareen rahnumai	Great guidance from the staff on using the app and availing the offers	Positive
Ilm insaan ka behtareen sathi hai	Knowledge is a person's best companion	Neutral
assalam alikum mujhy QS5 ki charging strip chahye agar koi sale karna chahta plz contact	Assalam aleikum I need charging strip for QS5 if anyone wants to sell please contact	Neutral
khanay ki miqdaar tasweeron mein dukhaay jane walay muqablay mein intehai kam hai	The amount of food is extremely small compared to what is shown in the pictures	Negative
lekin baaz auqaat yeh bohat mehanga par jata hai	But sometimes it turns out to be very costly	Negative

Feature Extraction

Word2Vec given by Mikolov, Tomas, et al [40] is a popular NLP algorithm for creating word embeddings, which capture word meanings, relationships, and contexts in a continuous vector space. Word2Vec is a neural network-based system that learns vector representations of words using pre-processed sentence data. It builds vocabulary using input data and then learns each word's vector representation, resulting in a vector representation of the word. The cleaned text was tokenized, transforming words into sequences of integers, which were subsequently padded to a fixed length of 300 words to ensure uniform input size for the

Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 December 2023*

neural network. The target labels representing sentiment classes (Negative, Neutral, Positive) were encoded into integers and converted into one-hot encoded vectors, making them suitable for multi-class classification with a softmax output layer. A Word2Vec model was trained over 30 epochs to generate word embeddings that captured the semantic relationships between words, and these embeddings were used to initialize the model's embedding layer. We have classified sentiment using an LSTM model. The sentences are first tokenized into words, and each word is then given a distinct number by word embedding. We pad the vectors with zeroes to make them all the same length. As a result, a sentence is transformed into a series of vectors of identical length, where each vector represents a word in the text. The data is split into train and test sets in an 80:20 ratios with 20% of test data for validation, and hyper-parameters were defined before training. For the loss function, we employ the Adam optimizer with category cross entropy over 30 epochs.

Model Development

An The task of sentiment analysis is performed for ternary classification on the model.. After the completion of all requisite preprocessing and labelling process, a total of 58526 sentences including 25229 sentences of positive class, 17349 of negative class and 15948 sentences of neutral class were used for the development of model. The dataset has a balanced distribution of sentiment classes.

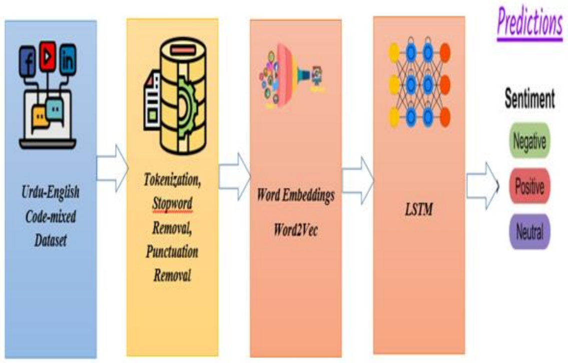


Fig 2. Workflow of the proposed model

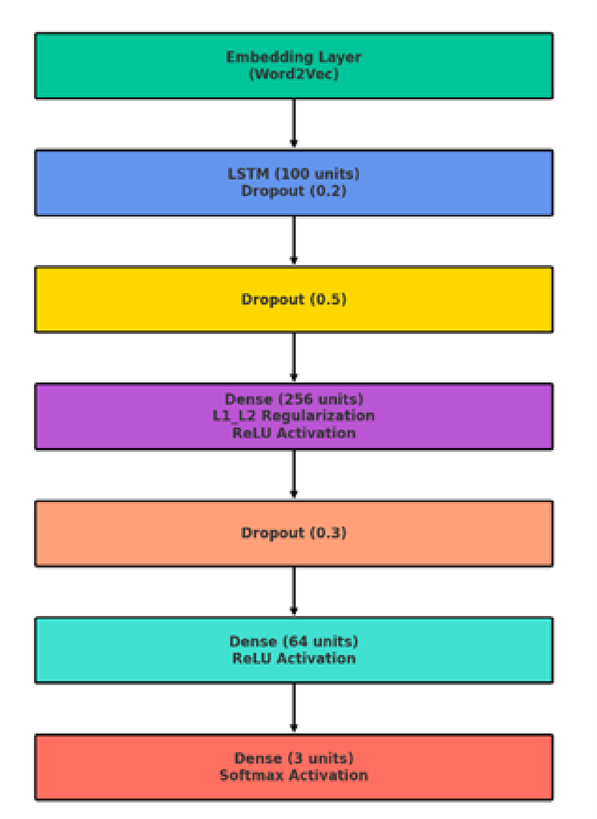


Fig. 3 Layers in the proposed LSTM

The model architecture was designed using a sequential neural network optimized for sentiment analysis. The first layer was an embedding layer initialized with the pre-trained Word2Vec embeddings, which transformed input words into dense vector representations, capturing contextual relationships. To mitigate overfitting, a dropout layer with a dropout rate of 0.5 was added immediately after the embedding layer, randomly disabling half of the neurons during each training step. Following this, a Long Short-Term Memory (LSTM) layer with 100 units was implemented to effectively learn long-range dependencies in the text data. This LSTM layer incorporated both dropout and recurrent dropout rates of 0.2 to further reduce the risk of overfitting. A dense layer with 256 units was added next, utilizing L1 and L2 regularization to penalize complex models and promote generalization. This layer was activated using the ReLU function to introduce non-linearity. Another dropout layer with a dropout rate of 0.3 was included to enhance regularization. Subsequently, a dense layer with 64 units and ReLU activation was employed for additional feature extraction. Finally, the output layer consisted of three units with a softmax activation function, enabling the

Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 December 2023*

model to classify inputs into one of the three sentiment categories: Negative, Neutral, or Positive.

Results and Discussion

The performance evaluation of the model demonstrated its effectiveness in sentiment classification tasks. The model achieved a high test accuracy of **95.06%** with a test loss of **0.1738**, indicating strong generalization on unseen data. The classification report provided detailed performance metrics across all three sentiment classes. For the Negative class, the model achieved a precision of **0.96**, recall of **0.93**, and an F1-score of **0.94**. For the Neutral class, it obtained a precision of **0.91**, recall of **0.95**, and an F1-score of **0.93**. The Positive class showed the highest performance with a precision, recall, and F1-score all around **0.97**. The macro and weighted averages for precision, recall, and F1-score were consistently **0.95**, confirming the model's balanced performance across all classes.

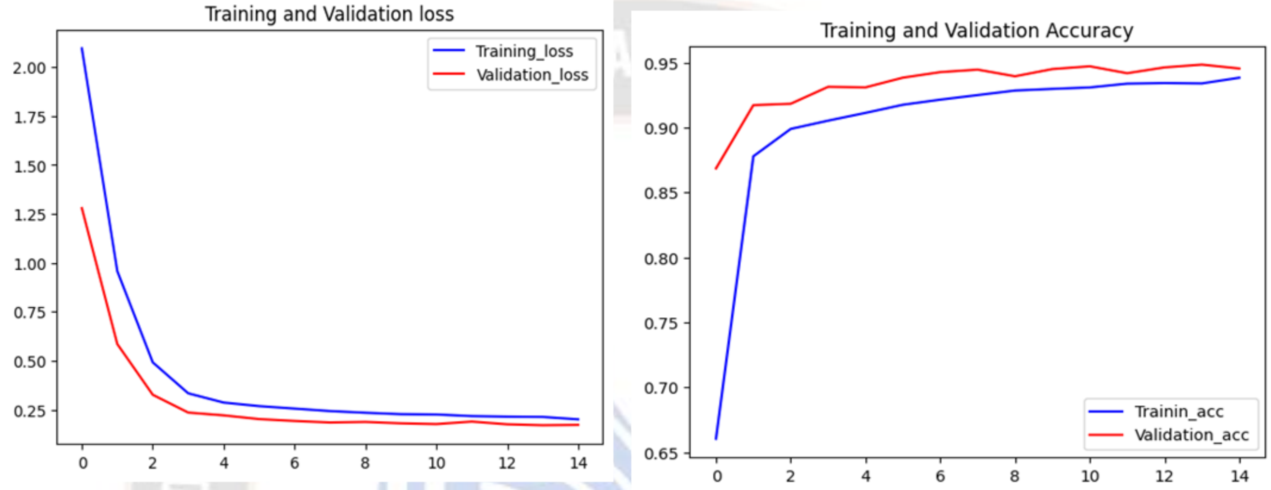
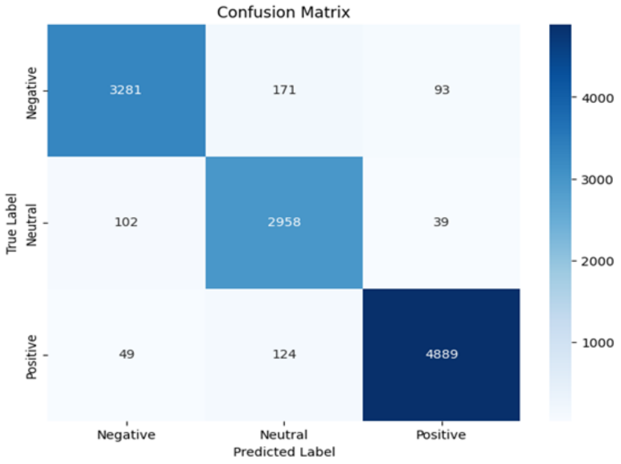
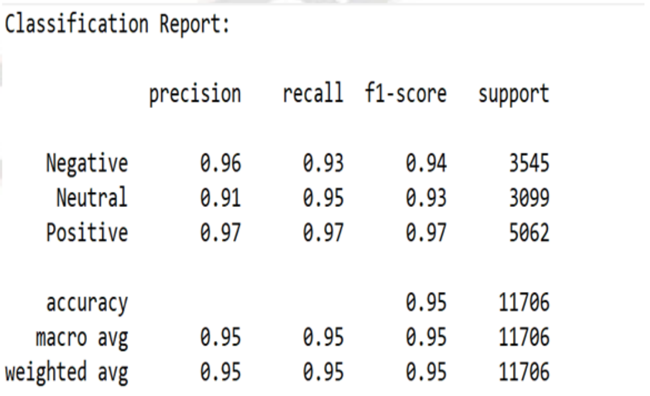


Fig 4: Training & Validation (a) loss (b) accuracy

Performance was derived from the accuracy and loss curves. The training and validation accuracy plots revealed that the model successfully learned meaningful patterns without significant overfitting, as evidenced by the convergence of the training and validation accuracy. The loss curves similarly indicated steady declines in both training and validation loss, reflecting consistent learning and robust performance. Additionally, the confusion matrix analysis highlighted the model's strong classification capabilities, with minimal misclassifications between sentiment categories. In particular, the model showed exceptional performance in identifying positive sentiments, achieving high precision and recall in this category.





```
Classification Report:

              precision    recall  f1-score   support

    Negative       0.96      0.93      0.94      3545
     Neutral       0.91      0.95      0.93      3099
    Positive       0.97      0.97      0.97      5062

    accuracy                           0.95     11706
   macro avg       0.95      0.95      0.95     11706
weighted avg       0.95      0.95      0.95     11706
```

Fig. Confusion Matrix and classification report

Article Received: *25 July 2023* ***Revised:*** *12 September 2023* ***Accepted:*** *30 December 2023*

The combination of thorough data preprocessing, effective use of Word2Vec embeddings, and a carefully designed LSTM-based neural network resulted in a highly accurate and reliable sentiment classification model. The inclusion of dropout layers and regularization techniques contributed to the model's ability to generalize well to unseen data. Overall, the proposed model demonstrates significant potential for accurate and efficient sentiment analysis, achieving balanced and robust performance across all sentiment categories.

Table 3: Precision, Recall chart and f1-score for the dataset along with accuracy

Class	Precision	Recall	F1-Score
Positive	0.97	0.97	0.97
Negative	0.96	0.93	0.94
Neutral	0.91	0.95	0.93
Average	0.95	0.95	0.95
Accuracy			**0.95**

Our approach takes into account contextual and language-specific information in addition to word-level sentiment cues to overcome the difficulties caused by code-switching. In order to determine whether certain language combinations and switching techniques are representative of particular sentiment, we also look into code-mixing patterns. Our experimental findings show how well our method works to classify sentiment in Urdu-English code-mixed text with state-of-the-art accuracy.

CONCLUSION AND FUTURE WORK

Code-mixing languages are a challenging task for Natural Language Processing (NLP) researchers due to their informal nature and limited linguistic resources. we explore innovative approach to sentiment analysis of code-mixed Urdu-English social media text that that takes into consideration the unique linguistic challenges exhibited by code-mixed text. We make use of NLP approaches of deep learning and multilingual embeddings to create a model that can successfully identify and classify sentiments. To enhance the accuracy and granularity, the model also takes into account code-mixing patterns and contextual cues. We achieved an accuracy of 95% with weighted average f1-score of 0.95. In our future work we aim to develop a parallel model that can identify language at word level, identify and correct spelling variations as well as automatic identification and removal of roman Urdu stop words before actual classification of sentences.

REFERENCES

1. Das, A., & Gambäck, B. (2013). Code-mixing in social media text. *Traitement Automatique des Langues*, *54*(3), 41-64
2. Mahadzir, N. H. (2021). Sentiment analysis of code-mixed text: a review. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, *12*(3), 2469-2478.
3. Nagra, A. A., Alissa, K., Ghazal, T. M., Kukunuru, S., Asif, M. M., & Fawad, M. (2022). Deep sentiments analysis for roman urdu dataset using faster recurrent convolutional neural network model. *Applied Artificial Intelligence*, *36*(1), 2123094.
4. Luhn, H. P. (1958). The automatic creation of literature abstracts. *IBM Journal of research and development*, *2*(2), 159-165
5. Pang, B., & Lee, L. (2004). A sentimental education: Sentiment analysis using subjectivity summarization based on minimum cuts. *arXiv preprint cs/0409058*.
6. Go, A., Bhayani, R., & Huang, L. (2009). Twitter sentiment classification using distant supervision. *CS224N project report, Stanford*, *1*(12), 2009
7. Kim, Y. (2014). Convolutional neural networks for sentence classification. *arXiv preprint arXiv:1408.5882*
8. Hasan, K. S., & Ng, V. (2014, June). Automatic keyphrase extraction: A survey of the state of the art. In *Proceedings of the 52nd Annual Meeting of the Association for*

Computational Linguistics (Volume 1: Long Papers) (pp. 1262-1273)

9. Wilson, T., Wiebe, J., & Hoffmann, P. (2005, October). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing* (pp. 347-354)

10. Pang, B., Lee, L., & Vaithyanathan, S. (2002). Thumbs up? Sentiment classification using machine learning techniques. *arXiv preprint cs/0205070*

11. Poplack, S. (1980). Sometimes i'll start a sentence in spanish y termino en espanol: toward a typology of code-switching1.

12. Myers-Scotton, C. (1993). Social Motivations for Codeswitching: Evidence from Africa

13. Noor, F., Bakhtyar, M., & Baber, J. (2019). Sentiment analysis in e-commerce using svm on roman urdu text. In *Emerging Technologies in Computing: Second International Conference, iCETiC 2019, London, UK, August 19–20, 2019, Proceedings 2* (pp. 213-222). Springer International Publishing.

14. Chandio, B., Shaikh, A., Bakhtyar, M., Alrizq, M., Baber, J., Sulaiman, A., ... & Noor, W. (2022). Sentiment analysis of roman Urdu on e-commerce reviews using machine learning. *CMES-Comput. Model. Eng. Sci*, *131*(3), 1263-1287

15. Soomro, M. A., Memon, R. N., Chandio, A. A., Leghari, M., & Soomro, M. H. (2024). A dataset of Roman Urdu text with spelling variations for sentence level sentiment analysis. *Data in Brief*, *57*, 111170

16. Nankani, H., Dutta, H., Shrivastava, H., Rama Krishna, P. V. N. S., Mahata, D., & Shah, R. R. (2020). Multilingual sentiment analysis. Deep learning-based approaches for sentiment analysis, 193-236

17. Ali, Z., Razzaq, A., Ali, S., Qadri, S., & Zia, A. (2021). Improving sentiment analysis efficacy through feature synchronization. Multimedia Tools and Applications, 80, 13325-13338.

18. Baccianella, S., Esuli, A., & Sebastiani, F. (2010, May). Sentiwordnet 3.0: an enhanced lexical resource for sentiment analysis and opinion mining. In Lrec (Vol. 10, No. 2010, pp. 2200-2204).]

19. [Hutto, C., & Gilbert, E. (2014, May). Vader: A parsimonious rule-based model for sentiment analysis of social media text. In Proceedings of the international AAAI conference on web and social media (Vol. 8, No. 1, pp. 216-225).

20. Loria, S., Keen, P., Honnibal, M., Yankovsky, R., Karesh, D., Dempsey, E., ... & Ragnarsson, L. (2019). TextBlob: simplified text processing; 2018. Online: https://textblob. readthedocs. io/en/dev/Accessed, 08-02

21. Doğruöz, A. S., Sitaram, S., Bullock, B. E., & Toribio, A. J. (2023). A survey of code-switching: Linguistic and social perspectives for language technologies. *arXiv preprint arXiv:2301.01967*

22. Mehmood, F., Ghani, M. U., Ibrahim, M. A., Shahzadi, R., Mahmood, W., & Asim, M. N. (2020). A precisely xtreme-multi channel hybrid approach for roman urdu sentiment analysis. *IEEE Access*, *8*, 192740-192759.

23. Majeed, A., Beg, M. O., Arshad, U., & Mujtaba, H. (2022). Deep-EmoRU: mining emotions from roman urdu text using deep learning ensemble. Multimedia Tools and Applications, 81(30), 43163-43188

24. Mehmood, K., Essam, D., & Shafi, K. (2019). Sentiment analysis system for Roman Urdu. In *Intelligent Computing: Proceedings of the 2018 Computing Conference, Volume 1* (pp. 29-42). Springer International Publishing

25. Arif, H., Munir, K., Danyal, A. S., Salman, A., & Fraz, M. M. (2016). Sentiment analysis of roman urdu/hindi using supervised methods. *Proc. ICICC*, *8*, 48-53.

26. Chandio, B., Shaikh, A., Bakhtyar, M., Alrizq, M., Baber, J., Sulaiman, A., ... & Noor, W. (2022). Sentiment analysis of roman Urdu on e-commerce reviews using machine learning. *CMES-Comput. Model. Eng. Sci*, *131*(3), 1263-1287

***Article Received:** 25 July 2023 **Revised:** 12 September 2023 **Accepted:** 30 December 2023*

27. Manzoor, M. A., Mamoon, S., Tao, S. K., Ali, Z., Adil, M., & Lu, J. (2020). Lexical variation and sentiment analysis of Roman Urdu sentences with deep neural networks. *International Journal of Advanced Computer Science and Applications*, *11*(2).

28. Chandio, B. A., Imran, A. S., Bakhtyar, M., Daudpota, S. M., & Baber, J. (2022). Attention-based RU-BiLSTM sentiment analysis model for roman Urdu. Applied Sciences, 12(7), 3641

29. Alvi, M. B., Mahoto, N. A., Reshan, M. S. A., Unar, M., Elmagzoub, M. A., & Shaikh, A. (2023). Count me too: Sentiment analysis of roman sindhi script. SAGE Open, 13(3), 21582440231197452

30. Nagra, A. A., Alissa, K., Ghazal, T. M., Kukunuru, S., Asif, M. M., & Fawad, M. (2022). Deep sentiments analysis for roman urdu dataset using faster recurrent convolutional neural network model. Applied Artificial Intelligence, 36(1), 2123094

31. Yadav, K., Lamba, A., Gupta, D., Gupta, A., Karmakar, P., & Saini, S. (2020, December). Bi-LSTM and ensemble based bilingual sentiment analysis for a code-mixed Hindi-English social media text. In *2020 IEEE 17th India Council International Conference (INDICON)* (pp. 1-6). IEEE

32. Agarwal, A., & Bhattacharyya, P. (2005, December). Sentiment analysis: A new approach for effective use of linguistic knowledge and exploiting similarities in a set of documents to be classified. In *Proceedings of the International Conference on Natural Language Processing (ICON)* (Vol. 22)

33. Prabhu, A., Joshi, A., Shrivastava, M., & Varma, V. (2016). Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. *arXiv preprint arXiv:1611.00472*

34. Parikh, A., Bisht, A. S., & Majumder, P. (2020, December). IRLab_DAIICT at SemEval-2020 task 9: Machine learning and deep learning methods for sentiment analysis of code-mixed tweets. In *Proceedings of the Fourteenth Workshop on Semantic Evaluation* (pp. 1265-1269)

35. Subrahamanyam, S., Sathineni, P. R., & Mamidi, R. (2021). Sentiment analysis in code-mixed telugu-english text with unsupervised data normalization. *Recent advances in natural language process (RANLP)*

36. Srivastava, V., & Singh, M. (2021). Challenges and considerations with code-mixed nlp for multilingual societies. arXiv preprint arXiv:2106.07823

37. *Alam, S., & Yao, N. (2019). The impact of preprocessing steps on the accuracy of machine learning algorithms in sentiment analysis. Computational and Mathematical Organization Theory, 25, 319-335.*

38. *Viera, A. J., & Garrett, J. M. (2005). Understanding interobserver agreement: the kappa statistic. Fam med, 37(5), 360-363*

39. *Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013). Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781.*