Webology (ISSN: 1735-188X)
Volume 19, Number 4, 2022

Adapting Machine Learning And Deep Learning Approach Towards Language Identification And Sentiment Analysis Of Code-Mixed Urdu-English And Hindi-English Social Media Text

Gazi Imtiyaz Ahmad[1] , Dr. Sruchi Talwani[2] , Dr. Jimmy Singla[3]

[1]Research Scholar LPU Punjab.

[2]Department of CSE LPU Punjab.

[3]Professor, Dept. of CSE CT University Punjab.

Abstract

Large amounts of textual data are produced by social networking sites in the form of posts, comments, reviews, and other user-generated content. This data can be useful in providing insights into public opinions, sentiments and trends and can be analyzed using Natural Language Processing techniques. This type of data is influenced by regional language text expressed in Romanized (Latin script) form. The idea of using Latin alphabet is that it allows text to be represented in a form that can be easily typed and processed by computers. Moreover, in multilingual societies people express their opinions and sentiments in a code-mixed fashion which refers to the practice of using languages or language verities mixed together in a single stretch of discourse. People often communicate with one another on social networking platforms and microblogging sites using multiple languages language varieties, a practice known as "code-mixing.". This can be due to a variety of factors, such as the multilingual nature of many online communities, the desire to reach a wider audience, or the influence of language trends and memes. However, such informal and code-mixed textual data are under resourced in terms of labelled datasets and language models. Therefore, it is difficult to use Natural Language Processing algorithms on this type of textual data. In this work we present machine learning and deep learning approaches for word level Language Identification and Sentiment classification of Urdu-English and Hindi-English code-mixed text. For deep learning models we use character level and word level feature for embeddings and feature based approach for machine learning models. The paper also describes the development of code-mixed Urdu-English dataset from social media. The dataset was annotated for Sentiment classification and word-level Language detection.

Keywords: Code-mixed, Sentiment Analysis, LSTM, Natural Language Processing, Machine Learning, Deep Learning

Introduction

Social media platforms play a significant role in modern communication, networking, and information sharing, and have become an integral part of many people's personal and professional lives. Social media platforms are used by billions of people around the world to communicate with each other, share information, and stay connected. Social media platforms are also used by businesses and organizations to provide customer service, answer questions, and address concerns in real-time. It is used by individuals and organizations to build networks, find and connect with like-minded people, pursue common interests, share and discover new content.

Textual information is produced in great quantities by social networking sites in the form of posts, comments, reviews, and other user-generated content. This data can provide valuable insights into public opinion, sentiment, and trends, and can be analyzed using natural language processing techniques such as sentiment analysis. Social media data can be used to understand how customers perceive a product or service, respond to customer needs and concerns, track and analyze the sentiment of social media posts, comments and reviews and other text related to social and political issues and can be used to identify areas for improvement or to track the effectiveness of marketing campaigns.

Sentiment analysis is a natural language processing task that involves classifying texts or parts of texts into predefined categories such as positive, negative, and neutral. The goal of sentiment analysis is to determine the attitude or emotion of the writer with respect to a particular topic, product, or service. Sentiment analysis can be used to mine opinions and emotions expressed in texts such as social media posts, reviews and customer feedback and can be useful in a variety of application, such as marketing and customer services.

Code-mixing is a common phenomenon on social networking and microblogging sites, where people often use multiple languages or language varieties to communicate with each other. This can be due to a variety of factors, such as the multilingual nature of many online communities, the desire to reach a wider audience, or the influence of language trends and memes. On social media platforms, code-mixing can take many forms, such as mixing languages within a single sentence or post, using multilingual hashtags, or including translations or transliterations of words or phrases. Researchers have studied code-mixing on social media as a way to understand language use and have have focused on the challenges and opportunities that code-mixing presents for NLP tasks, such as machine translation and sentiment analysis. code-mixing is commonly observed in informal communication and occurs when a conversant use two or more languages together especially in written communication. [1]. A number of machine learning and deep learning approaches and methods have been employed to observe and identify code-mixing in a multilingual text. [2] Several code-mixing measures have been widely applied over time to determine and validate the quality of code-mixed text. [3]. Popular codemixing metrics that measure the complexity of the code-mixed text include Code-Mixed Index (CMI) proposed by Das

and Gambäck [4], Gambäck and Das [5] Multilingual Index (M-index) proposed by Barnett et al. [6], Integration-index (I-Index) proposed by Guzmán et al. [7], Burstiness and Memory Goh and Barabási [8].

Therefore, it is a difficult issue for researchers studying Natural Language Processing (NLP) to construct a robust code-mixing metric that scales the level of code-mixing and quantifies the text's readability and grammatical accuracy.

The growth of social media in recent years has brought about a number of new opportunities for language technology and information access, as well as a number of new challenges, particularly because these types of coarse texts are characterized by having a high percentage of spelling mistakes and by containing phenomena like creative spelling, word play, abbreviations, Meta tags, and phonetic typing [9].

Urdu is one of the most spoken languages in the world. Almost 170 million people speak Urdu in different parts of the world. Although being the official language of Pakistan, its usage in northern India has grown significantly during the past few decades. One of the Indo-Aryan languages that is spoken widely in South Asia is Urdu, which is written from right to left. [10] However, on social networking sites, people who know how to write and speak Urdu use Latin script (Romanized) to communicate with each other. People may code-mix for a variety of reasons, such as to express their identity or to communicate with a wider audience. On social media sites, code-mixing is often used to reach a wider audience or to connect with others who speak different languages or varieties of language. Code-mixing can also be a way to add flair or personality to a social media post or to create a sense of in-group solidarity among users who share a particular language or cultural background. On social media platforms, a lot of text is produced, much of it informal and cod-mixed. Roman Urdu is a written form of the Urdu language that uses the Roman alphabet rather than the Arabic script. It is commonly used on the Internet and on social networking sites to communicate with a wider audience, as many people may not be familiar with the Arabic script. Roman Urdu is used in a variety of contexts on social media, including personal and professional communication, news and information sharing, and as a tool for marketing and advertising. It is also used in online communities and forums where people discuss a variety of topics in Urdu. Some researchers have studied the use of Roman Urdu on social media as a way to understand language change and the role of social media in shaping language use. Others have focused on the challenges and opportunities that Roman Urdu presents for natural language processing tasks, such as machine translation and sentiment analysis.

It is difficult to analyze the sentiment of code-mixed text, especially for languages with limited resources like Urdu. Although researchers have started to work to develop resources for the code-mixed Urdu-English textual data, but there are many inconsistencies which are hard to decipher when it comes to machine than the humans to understand a particular word or sentence in code-mixed form.

Related Work

Sentiment analysis is a technique used to automatically determine the sentiment or emotion expressed in a piece of text, such as a review, social media post, or news article. The goal of sentiment analysis is to identify the overall sentiment of the text as being positive, negative, or neutral. Sentiment analysis can be used to gather insights about public opinion, identify trends and patterns, and track changes in sentiment over time. It can also be used to help businesses and organizations understand how their products, services, or brand are perceived by customers or the public. People in multilingual societies communicate their feelings and opinions about a good or service in more than one language. Due to ease in typing, people favor use of Latin (Roman) Script instead of native script to show their feelings in language(s) other than English. Sentiment analysis of unilingual language(s) particularly of English language is almost a complete task. A number of language related resources such as annotated datasets, WordNet, SentiWordNets, POS taggers etc. have been developed over the years. However, most of the natural languages such as Urdu have not been extensively explored. In addition, the code-mixed language pairs are often under-resourced because they are not recognized as distinct languages by many language resource development organizations and funding agencies. As a result, there is often a lack of resources, such as language models and lexical resources, that are specifically designed for code-mixed languages. This makes it difficult for researchers and developers to work with code-mixed languages and limit the availability of natural language processing tools and services for these languages. Additionally, since code-mixed languages are often informal in nature and is used for short messages, which further contribute to their under-resourcing.

Jhanwar & Das [11] proposed an ensemble model for SA of Hin-Eng code-mixed data where in sentiment analysis of limited and inconsistent Hin-Eng code-mixed data was performed. The proposed model used the combinational strength of successive patterns from LSTM model and probabilistic n-gram model's keyword polarity for classification of sentiments. The authors used dataset of joshi et. al [12] for experimentation. The dataset consists of 3879 sentences which were collected from popular Facebook pages and achieved best accuracy of 70.8% on the dataset.

Attention-based deep learning architectures are a type of neural network that use attention mechanisms to selectively focus on different parts of the input sequence, improving their performance on tasks such as language modeling, machine translation, and sentiment analysis. Mukherjee, Siddhartha, et al. [13.] proposed a similar architecture for SA of Hin-Eng code-mixed text. The authors created a reliable code-mixed text classifier using both character level and word level embeddings. The accuracy of the suggested classification model was 71.97%.

Mishra et al. [14] have presented machine learning and deep learning algorithms for code-mixed social media text. The Authors used two datasets Hindi-English and Bengali-English released by SAIL-2020. Linear SVM, Random Forests, and Logistic Regression with TF-IDF feature vectors of character n-grams were utilized by the authors as part of ensemble voting model. For Bi-LSTM model, the authors used GloVe for word embeddings. The proposed model achieved F1-score of 0.569 for Hin-Eng dataset and 0.526 for Ben-Eng dataset.

Chakravarthi, Bharathi Raja, et al. [15.] developed a gold standard dataset for sentiment analysis of Malayalam-English code-mixed text. The dataset was assessed in terms of sentiment categorization using machine learning and deep learning methods.

A hybrid system for sentiment analysis combines multiple approaches, such as rule-based and machine learning-based techniques, to improve the overall performance of the sentiment analysis system. Hybrid systems can leverage the strengths of each approach while mitigating their weaknesses, resulting in more accurate and robust sentiment analysis. Kumar and Dhar [16] proposed one such system for SA of Hin-Eng textual data. The data was gathered from well-liked Indian Facebook accounts. For both the overall sentiment of the phrase and for specific words and sub words that convey sentiment, the model used two BiLSTM systems. To operate more effectively, the model applied orthogonal and word embedding features. The proposed system had an F1-score of 0.827 and an accuracy of 83.5%.

Choudhary, Nurendra, et al. in [17], Introduced a special technique for classifying Hin-Eng text as "positive," "negative," or "neutral." The recommended model uses Siamese networks to map sentences in code-mixed and standard languages into a common sentiment space as well as simple clustering-based preprocessing approaches to capture variations of transliterated words. The proposed model had an accuracy of 77.3% and an F-score of 0.759.

Mahmood, Zainab, et. al. [18] created a sentiment analysis system for Roman Urdu utilizing a DL approach of RNN model. The authors collected data from many websites and social networking sites on a range of subjects, including politics, sports, entertainment, food, movies, software, and electrical items. The annotation process of classifying the sentences into ‘positive’, ‘negative’, and ‘neutral’ was done of manually. Two model a rule-based model and an RNN mode were employed to predict the sentiment of text. According to the experimental findings, the RNN model outperforms the rule-based model with an accuracy of 57.2%.

SVM, NB, and LR with Stochastic Gradient Descent are three supervised machine learning algorithms that Rafique, Ayesha, et al. [19] used to provide a SA system for Roman Urdu. The authors created a dataset collected from various websites and microblogging platforms. The experimental findings demonstrated that SVM outperforms the other two supervised ML algorithms, with 87.22% accuracy.

Roman Urdu-English code-mixed text was investigated for sentiment analysis by Younas, Aqsa, et al. [20]. For the system, the authors used the Multilingual BERT and XLM-RoBERTa models. The authors presented the sentiment classification which is independent of lexical normalization, language dictionary and code-transfer indicator. The XLM-R model attained f1 score of 71% with tuned hyper-parameters.

Gaurav Singh in [21] implemented various ML and DL techniques for SA on Hindi-English dataset obtained from task 9 on the Semevel-2020. The author performed various preprocessing techniques for cleaning of the dataset. Data transformation techniques included word2vec, doc2vec, count Vectorizer, one hot Vectorizer, tf-idf Vectorizer, and fasttext embeddings. On the cleaned dataset, machine learning models like SVM, KNN, DT, RF, MNB, LR, and ensemble

voting were used. The author reports that an ensemble voting classifier was used to acquire the best F1 score of 69.07.

Proposed Work

Datasets

Two datasets were used for word level Language detection and Sentiment classification.

Hindi-English Dataset

The Hindi-English code-mixed dataset[1] was released by SemEval-2020 ("International Workshop on Semantic Evaluation.") The dataset consists of tweets containing Hindi and English words written in Latin (Roman) script. The polarity of these code-mixed tweets was divided into three categories: 'positive', 'negative', and 'neutral'. Data was given in a tab separated format in the.txt file. Each tweet or statement was tokenized into words, with Hindi (Hin), English (Eng), and other (O) language tags applied to each word. The individual words in the data were initially turned into sentences, with 3000 sentences each for testing and validation and 14,000 sentences for training.

Table 1 presents a full overview of the dataset, and Table 2 provides a description of the language tags.

Categories	Number of sentences		
	Training	Validation	Test
Positive	4634	982	1000
Negative	4102	890	900
Neutral	5264	1128	1100

Table 1: Dataset Description

Language Label	Number of words		
	Training Data	Validation Data	Test Data
Hindi (Hin)	169893	37004	36122
English (Eng)	121412	25615	26358
Other (O)	73713	15959	15779

Table 2: Language Tags

1. https://zenodo.org/record/3974927#.Y-ZWo3ZByUk

Urdu-English

Code-mixed datasets for popular Indian languages such as Hindi-English, Bengali-English and the Dravidian languages like Tamil-English, Telugu-English are available but no resource is available for other Indian languages such as Urdu-English [22]. Therefore, a gold standard corpus for SA of Urdu-English text has been developed using YouTube API to download comments and replies to comments using YouTube search results for different keywords popular in Urdu

language on varying subjects like sports, entertainment, politics etc. The results include username, comment text, date & time, likes, replies, reply count and update date & time. The dataset downloaded consists of 8549 comments which were filtered for monolingual and code-mixed sentences using langdetect library.

1. Read Sentence S
2. Calculate length of S
3. Split S into words W_1, W_2, W_3……W_n
4. For each word W_i =1 to n
5. If langdetect(w_i) equal to English
   Count=count+1
   Else
   break
6. If count value equals to number of words extracted from S
   Label S as English
   Else
   Label S as mixed

Pseudo-code for sentence classification into monolingual and code-mixed sentences.

One of the important steps in social media data analysis is data pre-processing. Pre-processing data is mostly done to get the data set ready for accurate analysis using various computational methods [23]. The dataset was preprocessed using a number of preprocessing techniques. First sentences which were less than 3 words were removed. The emoji's, spaces, hashtags, @user etc. were removed using regular expressions. The sentences were also converted into lowercase. After cleaning, the dataset with the Urdu-English code mixing contained 3805 sentences.

Annotation

The processed dataset was annotated using a data annotation tool which was in-house developed. The tools serve two purposes, viz. word level language identification into Urdu, English and others and sentence level polarity into 'positive', 'negative' and 'neutral'. The data was annotated by two experts who were quire familiar with both Urdu and English languages. The sentences which do not possess inter-annotation agreement were removed. Figure 1 and Table 3 contain a screenshot of the data annotation tool and a description of the dataset respectively.

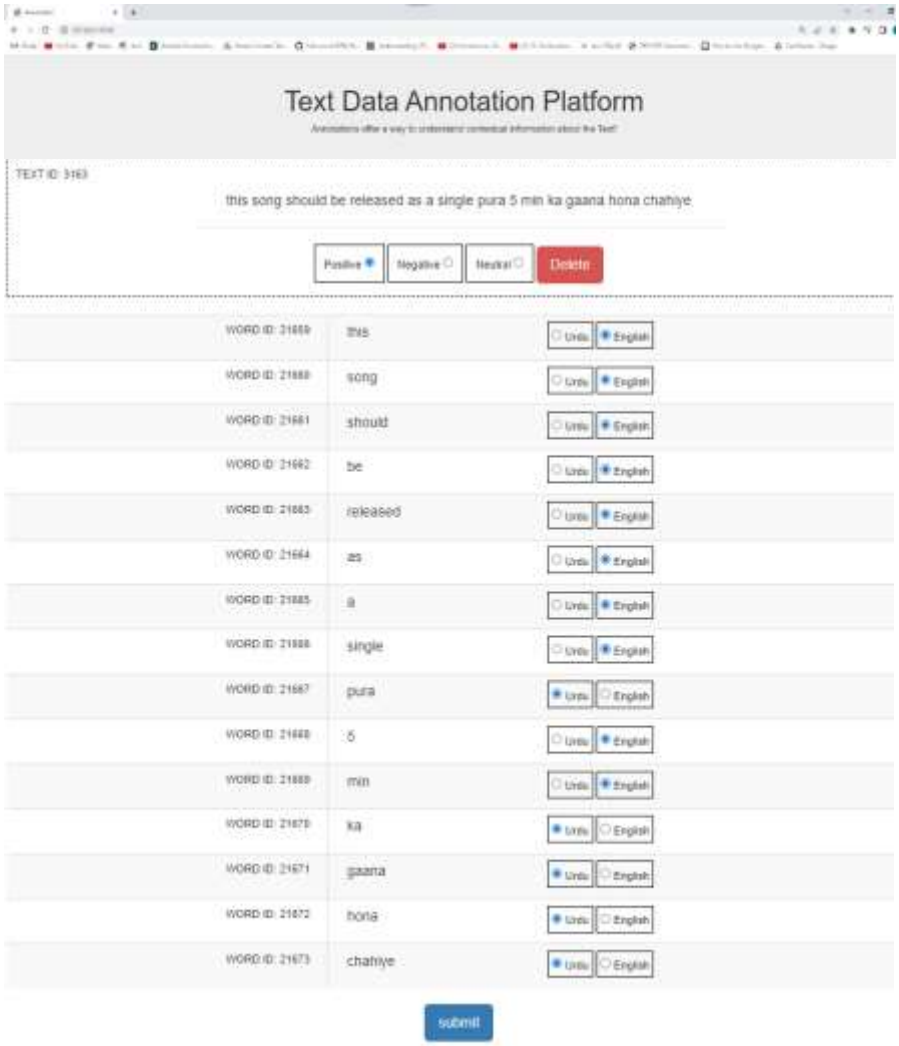

Figure 1: Data Annotation Tool

Number of Sentences		
Positive	1730	45.46%
Negative	1232	32.38%
Neutral	843	22.16%

Number of Words		
English	15562	47.65%
Urdu	14762	45.20%
Other	2335	7.15%

Table 3: Urdu-English Code-mixed Dataset Description

Language Identification Models

Language identification for almost all types of natural languages written in their original script is almost a solved problem. [24] Many tools and applications are available which distinguish between languages in mixed script sentences. Language Identification Tools (LIDs) have been developed over the years for identification of word level languages and are used for many applications such as speech recognition, text summarization, machine translation etc. However, these models fail in

the social media context due to the phenomena like code-mixing, code-switching etc. In code-mixed context, language identification is considered as a classification problem as discussed by [25] [26] [27]. Therefore, we also take language identification as a classification problem. On social networking sites, Multilingual people communicate their feelings, emotions, opinions, and reviews in a mixed phenomenon whereby using words or phrases from more than one language and often write non English text in Latin (Roma) script making the social media communication is informal and noisy in nature. Therefore, language identification in such scenario becomes both a necessary and challenging task to facilitate further processing of such type of text. Since our data has been annotated at word level, therefore we address the task in a fully supervised way. We employ three ML approaches, SVM, MNB and DT for word level language detection on Hindi-English and Urdu-English data. We use word n-grams (n=1 to 3) approach with TF-IDF features which is also followed by many language identification researchers [28.]. We also use a pure neural network with keras for our language identification task. In this model the tokenized words are character level embeddings and the vectors formed by embeddings are made equal length vectors by way of padded with spaces. To reduce the chance of overfitting the model, hyper-parameters are established prior to neural network training and the Adam optimizer is utilized with binary cross entropy as the loss function over fifty epochs. The results of our experimentation are shown in table 4 and table 5.

Model

	Accuracy	Precision	Recall	F1-score
Multinomial Naïve Bayes	73.04	0.82	0.52	0.53
Decision Tree	71.18	0.64	0.79	0.63
SVM	**75.11**	0.84	0.56	0.55
Neural Network	**75.01**	0.78	0.80	0.79

Table 4: Results of methods proposed for the task of identifying the language of data with Hindi-English code mixing.

Model

	Accuracy	Precision	Recall	F1-score
Multinomial Naïve Bayes	89.62	0.98	0.81	0.89
Decision Tree	**89.82**	0.96	0.82	0.89
SVM	89.28	0.98	0.79	0.88
Neural Network	86.23	0.87	0.86	0.86

Table 5: Results of methods proposed for the task of identifying the language of data with Urdu and English code mixing.

Sentiment Analysis Models

Sentiment analysis also known as appraisal extraction and opinion mining is a computation task of obtaining and evaluating people's opinions, feelings, attitudes and perceptions etc. towards different entities like products, services, social and political events etc. [29]. Sentiment Analysis is a prevailing tool that is used by researchers, business and government organization to extract and evaluate public opinion and perception to gain business insights and improve decision making. [30]. Researchers have employed various machine learning and dictionary based approaches for the task. Deep learning-based approaches, however, have also gained popularity recently since they outperform conventional approaches in terms of performance. Sentiment analysis of monolingual text especially English and other major languages is almost a completed task. This is due to the fact that various natural language processing resources such as datasets, lexical and semantic resources, WordNets and SentiWordNets, POS taggers and parsers are available for these languages. Modern models and methods are only designed for monolingual languages. However, such types of resources and NLP models and approaches are rarely available for code-mixed languages. Therefore, to harness and process huge volume of social media textual data, it is important to build models and application for code-mixed languages [31]. Researchers from the recent past are quite fascinating to extract and decipher useful information from this code-mixed text. various of ML and DL approaches have been developed, but the idea is still new and demands more attention.

In this work we have applied Machine Learning and Deep Learning approaches for sentiment analysis of Hindi-English and Urdu-English code-mixed data. The datasets, preprocessing and annotation has already been discussed. Here we will focus on feature extraction and sentiment analysis process using machine learning and deep learning models. For machine learning models such as SVM, MNB, DT and LR. The class weight was set to balanced and random state set to zero while applying these models on code-mixed data. We have used Count Vectorizer, One Hot Binarizer and TF-IDF Vectorizer with word n-grams (n=1 to 3) for feature extraction and selection process.

We have also used an LSTM model for sentiment classification. We first tokenize the sentences into words and then using word embedding, each word is assigned a unique number. To make the vectors of equal length we pad them with zeroes. Therefore, a sentence is converted into a sequence of equal length vectors where each vector in a sentence corresponds to the individual words on the sentence. Hyper-parameters were specified prior to training, and the data is divided into train and test sets in an 85:15 ratios. We utilize the Adam optimizer with categorical cross entropy for the loss function, with a batch size of 32 across 10 epochs, the results of these models on both the data sets are shown in table 6 and table 7.

Model		Accuracy	Precision	Recall	F1-score
Multinomial Naïve Bayes	One Hot Binarizer	58.21	0.57	0.60	0.57
	Count Vectorizer	59.25	0.58	0.60	0.58
	TF-IDF Vectorizer	60.15	0.62	0.59	0.60
Decision Tree	One Hot Binarizer	53.42	0.53	0.53	0.53
	Count Vectorizer	53.11	0.53	0.53	0.53
	TF-IDF Vectorizer	51.25	0.51	0.52	0.51
SVM	One Hot Binarizer	60.00	0.61	0.59	0.60
	Count Vectorizer	60.00	0.61	0.59	0.60
	TF-IDF Vectorizer	61.22	0.61	0.61	0.61
Logistic Regression	One Hot Binarizer	60.19	0.60	0.60	0.60
	Count Vectorizer	60.10	0.60	0.60	0.60
	TF-IDF Vectorizer	61.24	0.60	0.61	0.61
LSTM	Word Embeddings	**80.12**	0.80	0.80	0.80

Table 6: Results of various approaches proposed for the task of Sentiment Analysis of Hindi-English code-mixed data.

Model		Accuracy	Precision	Recall	F1-score
Multinomial Naïve Bayes	One Hot Binarizer	59.35	0.55	0.55	0.54
	Count Vectorizer	59.45	0.55	0.56	0.55
	TF-IDF Vectorizer	52.36	0.38	0.45	0.38
Decision Tree	One Hot Binarizer	48.24	0.46	0.46	0.46
	Count Vectorizer	48.48	0.46	0.46	0.46
	TF-IDF Vectorizer	48.24	0.46	0.46	0.46
SVM	One Hot Binarizer	60.12	0.59	0.56	0.55
	Count Vectorizer	60.23	0.59	0.56	0.54
	TF-IDF Vectorizer	60.41	0.58	0.57	0.56
Logistic Regression	One Hot Binarizer	60.22	0.57	0.56	0.55
	Count Vectorizer	59.00	0.56	0.55	0.54
	TF-IDF Vectorizer	59.00	0.56	0.55	0.53
LSTM	Word Embeddings	**82.49**	0.82	0.82	0.82

Table 7: Results of various approaches proposed for the task of Sentiment Analysis of Urdu-English code-mixed data.

Conclusion

Code-mixing languages has no predefined set of rules as word mixtures are a highly observed occurrence. Also the mixing of languages is informal in nature, therefore, people tend to mix sentence structures on the matrix language. Thus building a model in code-mixed language is a challenging task for NLP researchers due to its linguistic phenomenon. Also linguistic resources especially code-mixed datasets are seldom available. Therefore, in this work we have developed an Urdu-English code-mixed dataset for word level language detection and Sentiment classification. The dataset consists of 3805 sentences and 32659 individual words tagged with corresponding language tag. This dataset shall be published as its first version and shall be updated after adding more and more sentences. In code-mixed Hindi-English and Urdu-English text, we have used machine learning and deep learning models for word level language recognition and sentiment classification. While deep learning models outperform machine learning models for sentiment analysis, machine learning models performed better for language identification. The results are promising and the work shall be carried in future to enhance the performance of the models.

References

1. Sutrisno, B., & Ariesta, Y. (2019). Beyond the use of code mixing by social media influencers in instagram. Advances in Language and Literary Studies, 10(6), 143-151
2. Hussain, A., & Arshad, M. U. (2021). An Attention Based Neural Network for Code Switching Detection: English & Roman Urdu. arXiv preprint arXiv:2103.02252
3. Srivastava, V., & Singh, M. (2021). Challenges and limitations with the metrics measuring the complexity of code-mixed text. arXiv preprint arXiv:2106.10123
4. Amitava Das and Björn Gambäck. 2014. Identifying languages at the word level in code-mixed indian social media text. In Proceedings of the 11th International Conference on Natural Language Processing, pages 378–387.
5. Björn Gambäck and Amitava Das. 2016. Comparing the level of code-switching in corpora. In Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC'16), pages 1850–1855.
6. Parry, J. J., Burnett, I. S., & Chicharo, J. F. (2000). Language-specific phonetic structure and the quantisation of the spectral envelope of speech. Speech communication, 32(4), 229-250.
7. Gualberto A Guzmán, Joseph Ricard, Jacqueline Serigos, Barbara E Bullock, and Almeida Jacqueline Toribio. 2017. Metrics for modeling code-switching across corpora. In INTERSPEECH, pages 67–71
8. K-I Goh and A-L Barabási. 2008. Burstiness and memory in complex systems. EPL (Europhysics Letters), 81(4):48002

9. Jamatia, A., Swamy, S. D., Gambäck, B., Das, A., & Debbarma, S. (2020). Deep learning based sentiment analysis in a code-mixed English-Hindi and English-Bengali social media corpus. International journal on artificial intelligence tools, 29(05), 2050014
10. 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
11. Jhanwar, M. G., & Das, A. (2018). An ensemble model for sentiment analysis of Hindi-English code-mixed data. arXiv preprint arXiv:1806.04450
12. Aditya Joshi, Ameya Prabhu, Manish Shrivastava, and Vasudeva Varma. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. In Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pages 2482–2491, 2016.
13. Mukherjee, S., Prasan, V., Nediyanchath, A., Shah, M., & Kumar, N. (2019, December). Robust deep learning based sentiment classification of code-mixed text. In Proceedings of the 16th International Conference on Natural Language Processing (pp. 124-129)
14. Mishra, P., Danda, P., & Dhakras, P. (2018). Code-mixed sentiment analysis using machine learning and neural network approaches. arXiv preprint arXiv:1808.03299
15. Chakravarthi, B. R., Jose, N., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2020). A sentiment analysis dataset for code-mixed Malayalam-English. arXiv preprint arXiv:2006.00210
16. Kumar, V., & Dhar, M. (2018). Looking beyond the obvious: Code-mixed sentiment analysis (cmsa).
17. Choudhary, N., Singh, R., Bindlish, I., &Shrivastava, M. (2018). Sentiment analysis of code-mixed languages leveraging resource rich languages. arXiv preprint arXiv:1804.00806
18. Mahmood, Z., Safder, I., Nawab, R. M. A., Bukhari, F., Nawaz, R., Alfakeeh, A. S., ...& Hassan, S. U. (2020). Deep sentiments in Roman Urdu text using recurrent convolutional neural network model. Information Processing & Management, 57(4), 102233
19. Rafique, A., Malik, M. K., Nawaz, Z., Bukhari, F., &Jalbani, A. H. (2019). Sentiment analysis for roman urdu. Mehran University Research Journal of Engineering & Technology, 38(2), 463
20. 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
21. Singh, G. (2021). Sentiment Analysis of Code-Mixed Social Media Text (Hinglish). arXiv preprint arXiv:2102.12149
22. Raja Chakravarthi, B., Jose, N., Suryawanshi, S., Sherly, E., & McCrae, J. P. (2020). A Sentiment Analysis Dataset for Code-Mixed Malayalam-English. arXiv e-prints, arXiv-2006

23. Shanmugavadivel, K., Sampath, S. H., Nandhakumar, P., Mahalingam, P., Subramanian, M., Kumaresan, P. K., & Priyadharshini, R. (2022). An analysis of machine learning models for sentiment analysis of Tamil code-mixed data. Computer Speech & Language, 76, 101407
24. Das, A., & Gambäck, B. (2015). Code-mixing in social media text: the last language identification frontier?
25. Gundapu, S., & Mamidi, R. (2020). Word level language identification in english telugu code mixed data. arXiv preprint arXiv:2010.04482
26. Shekhar, S., Sharma, D. K., & Beg, M. S. (2020). Language identification framework in code-mixed social media text based on quantum LSTM—the word belongs to which language? Modern Physics Letters B, 34(06), 2050086
27. 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)
28. Bhaskaran, S., Paul, G., Gupta, D., & Amudha, J. (2021). Indian language identification for short text. In Advances in Computational Intelligence and Communication Technology: Proceedings of CICT 2019 (pp. 47-58). Springer Singapore
29. Birjali, M., Kasri, M., & Beni-Hssane, A. (2021). A comprehensive survey on sentiment analysis: Approaches, challenges and trends. Knowledge-Based Systems, 226, 107134.
30. Yadav, A., & Vishwakarma, D. K. (2020). Sentiment analysis using deep learning architectures: a review. Artificial Intelligence Review, 53(6), 4335-4385
31. Santy, S., Srinivasan, A., & Choudhury, M. (2021, April). Bertologicomix: How does code-mixing interact with multilingual bert?. In Proceedings of the Second Workshop on Domain Adaptation for NLP (pp. 111-121).