Research Article

Language Tagging, Annotation and Segmentation of Multilingual Roman Urdu-English Text

Sajadul Hassan Kumhar[1], **Syed Immamul Ansarullah**[2], **Mudasir Kirmani**[3], **Sami Alshmrany**[4]

1. Sri Satya Sai University of Technology and Medical Sciences (SSSUTM), Sehore, India; 2. Government Degree College Bemina, Srinagar, India; 3. Division of Social Science, FoFy, Sher-E-Kashmir University of Agricultural Sciences and Technology, India; 4. Islamic University of Madinah, Medina, Saudi Arabia

The complexity of multilingual text in Roman Urdu and English necessitates appropriate tagging for language identification. This study collects multilingual text from various social media platforms including Twitter, Facebook, Instagram, and Google+. The Python Application Programming Interfaces are utilized to retrieve the Roman Urdu-English multilingual text. The corpus is prepared through the removal of hashtags, re-tweets, extra spacing, numerals, and the conversion of interfacial area to single spaces. The accompanying information obtained from social media and the web is considered noise and is addressed through web mining to maintain the quality of the Roman Urdu-English multilingual text data. After preprocessing and cleansing, the raw text is tokenized and segmented, resulting in efficient annotation, fragmentation, and segmentation of Roman Urdu and English.

Corresponding authors: Sajadul Hassan Kumhar, sajadulhassan@kwintech-rlabs.org; Syed Immamul Ansarullah, drsyedansar@kwintech-rlabs.org

1. Introduction

Language segmentation and tagging for mixed Roman Urdu-English multilingual corpus is a tedious, challenging, and time-consuming process; however, this research attempts to segment and tag the Roman Urdu-English mixed-language corpus correctly. Language segmentation and tagging are concerned with partitioning input text into segments regarding language. The multilingual text of Roman Urdu and English is complex, so it needs appropriate tagging to make them detachable to identify

doi.org/10.32388/BST5S8

the language of the text. We obtain multilingual text from various social media sites, including Twitter, Facebook, Instagram, and Google+, and use different python APIs to get the multilingual text of Roman Urdu and English. First, the corpus is prepared and cleaned by deleting hashtags and re-tweets, additional spacing leading or trailing the block of text, numerals, and exchanging interfacial areas with single spaces. The extra information obtained alongside the text from social networking sites and the web is noise. In addition, duplicate data and diacritics are optional and contain only altered pronunciations. In the next step of preprocessing and cleansing, the raw text is tokenized, segmented, and POS tagged.

Language tagging is concerned with annotating input text into the respective language tag. The language tagger tags the Sentence of text to their respective language. As a first step towards analyzing and tagging a multilingual code mixed corpus, we develop an effective correlated annotation strategy to represent the information's complexity, heterogeneity, and originality. Input data at social, contextual, linguistic and meta-linguistic levels are required to evaluate mixed multiple languages corpus that executes on distinct sub-parts of the discourse. Language tagging facilitates the classifying of systemic linguistic forms such as POS of labels, cutlets, tokens, phrases, contextual role, and socio-pragmatic concepts such as sentiment analysis, intelligent emotion, polarity, etc.

This research paper discusses the existing approaches relevant to language tagging and segmentation. Furthermore, we prepare the refined reports and results for the proper language segmentation and tagging experiments.

2. Literature Review

In the field of literature, there are numerous methods and tools that tackle the challenges in computer linguistics for language segmentation and tagging. To effectively identify languages within large volumes of text, we have carefully compared and discussed the most widely used, multi-functional, and prominent studies. Previous research efforts have paved the way for advancements in language segmentation and tagging, and our comprehensive analysis highlights the significance of these efforts.

Kumhar et al. carried out a detailed investigation on methods and tools for word embeddings generation. The research work also investigates the techniques used for data segmentation and annotation. However, the model has not given any method for the annotation and segmentation of Roman Urdu and English text.

Teahan 2000 throughout their research, they suggested a prediction by partial match (PPM) framework to forecast the next character in the input pattern utilizing text compression, text categorization, and text classification approach. They preprocess data using the minimum cross-entropy technique, and text correction algorithms assist word fragmentation. The research model lacks understanding of the text analyzed for estimating the next character from input sequence character and anticipates incorrect idiomatic tokenize from text dataset.

Al-Ohali et al. 2003 presented a mechanism for creating a database to identify Arabic Handwritten cheque-book through the pathways of segmentation binarization, data labeling, and gratification of the tagging processes. The suggested investigations model does not include research and reflections on all conditions.

Kumhar et al. 2021 proposed an efficient model for embedding Roman Urdu and English text on social media. However, the model did not give detailed insight into how the text was annotated, tagged, and segmented.

Durrani and Hussain. 2010 suggested a framework for Urdu word segmentation. They use orthographic and linguistic features to trigger Urdu segmentation and tokenization. This framework also used a composite n-gram method and a rule-based maximum matching heuristic. Bi-gram statistical natural language processing was not included in the research proposal to integrate morphemes for tagging.

Kumhar et al. 2020 proposed a sentiment analysis approach with the help of Recurrent Neural Network for social media text. The collected corpus of Roman Urdu has been represented in vector form using the Word2vec model. They use LSTM with the SoftMax function to polarize the text into positive, negative, and neutral.

Khan et al. 2018 suggested a machine learning approach for Urdu word segmentation using conditional random fields (CRF). The proposed research model lacks tangible word segmentation boundaries, such as space insertions, exclusions, and repetition of borrowed words, and is incapable of producing better precision.

Kumhar et al. 2020 proposed the effectiveness of tools and methods used to collect Roman Urdu and English text from different social media sites. However, the model gives no direction regarding the annotation and segmentation.

Ghosh et al. 2016 suggested a part of the speech tagging model built using Conditional Random Field (CRF). The CRF can mark the tag to the text at code-switching points where the language of text changes.

Bach et al. 2018 suggests an empirical study on part-of-speech tagging for Vietnam social media text. They perform part of speech tagging in supervised and semi-supervised scenarios. However, the model has not done the work on the mixed dataset.

Halgamuge et al. 2021 suggested a hashtag segmentation algorithm that significantly improves computational complexity. The algorithm generated candidate keys of data packets to the corpus comprising Unigram (1) and Bigram (2).

Techniques	Objective (s)	Result	Limitations	References
CRF	Urdu Word Segmentation	73.25%	Not concrete boundaries for space omissions, reduplication for foreign words	Khan S. N. et al.
n-gram model, texcat and Var cat	Tagging language for identification of languages in bilingual text	86.35%	Short words are not distinguishable using Uni-gram	Zampari M. et al.
RNN	Part of Speech tagging	89.25%	Produce low values on standardization of words and are consequently missed by the tagger.	Sharief Z. et al.
SVM, NB	To Segment bilingual text	91.20%	Not extendable for all language set	Rafique, A. et al.
ANN	Language tagging	87.51%	Propagates the errors from dataset to dataset	Khan, S. et al.
Uni-gram, bi-gram	Identify long dependencies	89.11%	It did not produce much better results for long words	Syed et al.
Spectrogram	Identify seven Indic languages and tag them	86.56%	Not extendable for language changing patterns	Mukherjee. et al.

Table 1. Comparison table of various researches taken for language tagging

doi.org/10.32388/BST5S8

3. Materials and Methods

The art of annotation, which encompasses both segmentation and tagging, will be gracefully carried out through the algorithm flowchart depicted in Figure 1. We gather text from social media havens such as Instagram, WhatsApp, Facebook, and Google+ utilizing the Python API. This results in the collection of bilingual text in both Roman Urdu and English, ready for further refinement. Natural language processing methods and tools are then employed to categorize the social media dataset. Subsequently, the social media text undergoes tokenization, breaking it down into small, meaningful words and phrases. However, if remnants of noise or hashtags persist, the tokenized words and segments will be subject to another iteration of the process. The tagged tokens are then processed using Support Vector Machines and a Bi-Directional Neural Network to determine their weights in English. If the token is not in English, it is searched in the Wordnet to verify its existence. The language tagger evaluates the weight of each token. If the weight is less than 30, it is tagged as English, but if it surpasses 30, it is deemed as Roman Urdu. These tagged tokens are then catalogued in a dictionary for future reference. Figure 1 showcases the seamless workflow of the language tagging and segmentation for the multilingual Roman Urdu and English data.

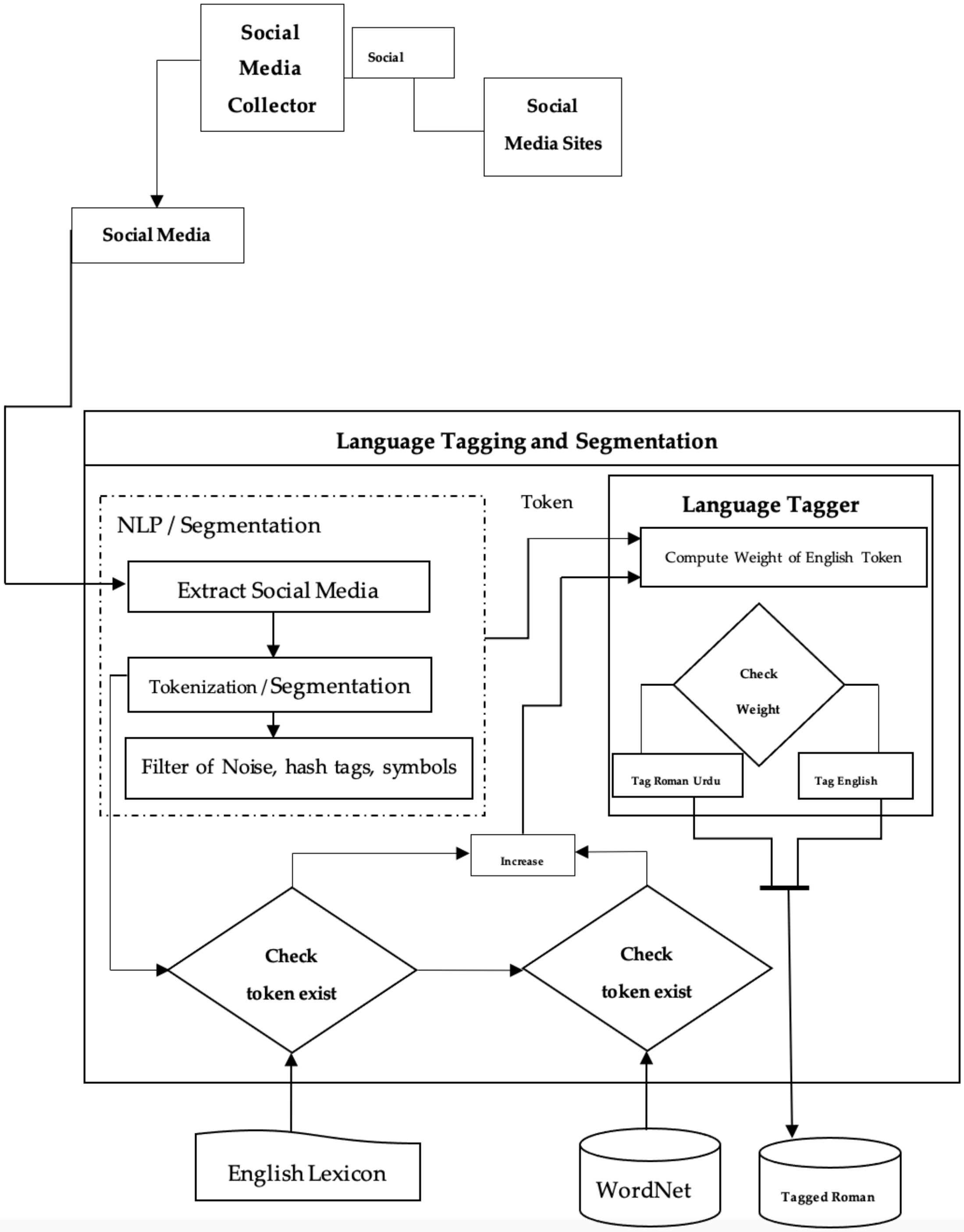

Figure1. Workflow of Language Tagging of Roman Urdu and English Text.

The intermediate fusion level integration is done in multiple steps starting from feature extraction by deep learning models, concatenation, and later passed to the classification layer. Decision level combinations are usually applied to single modalities.

3.1. Language Tagging and Segmentation

Language segmentation and tagging is a challenging and time-consuming process, and in this research, we segment and tag it correctly. Language segmentation and tagging are concerned with partitioning input text into segments relating to language. The segments are the subsequence of instances written in the same language in the research study. e.g.

i. (Movie) En (Kb) Ur (start) En (hogi) Ur

The Urdu equivalent will be:

(Transliteration)ی (Translation) شروع (Transliteration)(Translation) فلم

ii. (Mujay Jana hoga) Ur (There) En

The Urdu equivalent will be:

(Translation) (Transliteration)

Language tagging is concerned with the annotation of input text into the respective language tag. The language tagger tags the Sentence of text to their respective language. A cost-effective correlative annotation arrangement is transformed to obtain the data's uniqueness, diversification, and exclusivity as a first step towards analyzing and tagging multilingual code mixed corpus. The examination of mixed, multilingual corpora that work on different interaction subsets necessitates inputs at the social, situational, linguistic, and meta-linguistic levels. Language tagging facilitates classifying spatial, linguistic forms such as POS of tagging, chunks, tokens, phrases, contextual role, and categorizing sociopragmatic basic ideas such as sentiment classification, intelligent sentiment, and polarization.

3.2. Annotation

A group of three persons is selected to annotate the corpus. One among the three is from a Computer Science background, and the other two are proficient in English and Urdu languages. We design a simplified annotation framework to assist these annotators in identifying and distinguishing the languages existing in the textual data. To Annotators, we give four fundamental language tags while annotating the mixed Roman Urdu-English multiple languages corpus, including Inclusion, fragment, Sentence, and wlcm (Word-level code-mixing). In addition, English (En), Urdu written in Roman script (Ur), mixed (Mixd), Universal (Uni), and undefined is offered for each of the five characteristics (Undef). We couple the Undefined attribute with signs, sentiments, interpretations, numbers, and generic word phrases like lol, ha-ha.... and so forth. We use the Undef attribute for words and sentences that cannot be allocated or classified with words, sentiments, symbols, or tags. Furthermore, we instruct these experts to annotate each named entity independently. The following is a description of annotation.

a. *Sentence (Sent) tag:* The sentence tag marks tagging to words of inter-sentential mixed text. As the first annotation task, we instructed annotators to identify the language in the inter-sentential Sentence of words such as En, Ur, or Mixd, along with other sentence boundaries attributes such as Uni and Undef, etc. When the language included in the inter-sentential terms of a sentence comprises several languages in the same proportion, the Mixd characteristics are employed. A phrase may contain any word mixings, such as fragment and Inclusion. If a word is not designated as Uni or has no tokens or words classified as Ur, Mixdtags, or En, then Sentence might be marked with the Undef attribute in the statement. Sentence (Sent) tags include the following:

- English – Sentence:

  [Sentence-Language = ″En″] Such a .... tremendous bowling.... but fantastic and nicely played [/Sentence]
- Urdu-Sentence:

  [Sentence-Language = "Ur"] Keahloaorkaro eash[/Sentence]
- Mixed-Sentence:

  [Sentence-Language = "Mixd"] [Fragment-Language = "Ur"] Oiehy .... anrezeemaikhatehainaa [/Fragment] [Fragment-Language = "En"] I adore you...! [/Fragment] [/Sentence]
- Universal-Sentence:

  [Sentence-Language = "Uni"] hahahahaaa....! [/Sentence]
- Undefined-Sentence:

  [Sentence-Language = "Undef"] vevee [/Sentence]

b. *Fragment (Frag) tag:* The groups of foreign words present in a sentence are grammatically related to the Sentence. The fragment tag present in the Sentence conveys that inter-sentential mixed language of words occurred within the sentence boundary. The second annotation is to find these fragments in a sentence. As a result, a phrase with the attribute mixed must have numerous segments with the other particular fragment property. In the example mentioned in serial no third above, the Sentence contains multiple or mixed fragments, including Urdu (Ur) fragment Oyehoye .... anrezeemaikhatehaike, English (En) fragment I love you...! hence the Sentence is Mixd sentence. A sentence can consist of fragments, Inclusion, word or tokens, and languages tags and symbols. e.g.

- Inclusion with Fragment

- [Sentence-Language = “Mixd”] [Fragment-Language = “En”] [Inclusion-Language = “Ur”] Jio... [/Inclusion] amazing prank [/Fragment] [Fragment-Language = “Ur”] haiwoh [/Fragment] [/Sentence]
- Fragment with word-level Inclusion
- [Sentence-Language= “Mixd”] [Fragment-Language= “En”] I would track you down and stay married you. [/Fragment] [Fragment-Language = “Ur”] [wlcm-type= “En-and-Ur-suffix”] typer [/Wlcm] hoiegloitoi!: D [/Fragment] [/Sentence]

c. *Inclusion (Incl):* Inclusion are imported Phrases or words commonly embedded or absorbed in the original language. The identification of inclusion words is made after fragmenting and tagging a sentence by annotation is present in the Sentence denotes the inter-sentential mixing of languages. The Sentence with the inclusion word is as:

- [Sentence-language = “Ur”] mujay [inclusion-language = “En”] seriously [/inclusion] bolo naa [/sentence]

However, languages′ word-level mixing (wlcm) is rear for the Urdu language. The word-level multilingual mixing is the smallest of mixing, and the tag is introduced to capture inter-word mixing that has occurred with a single word. The last task of annotation is identifying the inter-word mixing of languages. The annotators were deployed for annotation and instructed to attribute the word level mixing of language by mentioning the base language + second language.

3.3. Inter Annotation Agreement

We collect the word-level inter annotation agreement for the randomly selected subset of 100 comments using the Cohen Kappa method on two annotators. The two annotators have agreed on a word that, if annotated jointly, the word will have the same attribute (Ur, En, Undef, Uni) regardless of whether the word is in the Inclusion, Sentence, or fragment. A set of annotations at a point refers to k statistics that measure the inter annotation agreements. More precisely, k measures the statistical agreement between annotation, making the category judgment. For a given set of judgment category of two annotators, the k is calculated as:

$$\mathrm{K} = \mathrm{Prop}(A) - \mathrm{Prop}(E)1 - \mathrm{Prop}(E)$$

Here Prop(A) is the proportion of times an annotator agrees with another annotator, and Prop(E) is the proportion of agreement accepted purely on chance. A baseline is achieved by chance when multiple annotators are coding and annotating the same data. The word-level annotation process is not more confusing after performing experiments and observations. The annotation is validated and more straightforward in a high Inter-annotation agreement (IAA) with Kappa Value 0.884.

3.4. Characteristics of Data

In the dataset proposed above, the intra-sentential and inter-sentential mixed multilingual sentences are more prevalent than the word level inter sentential mixing of languages comparable to U. Barman et al. (2014). The proposed dataset contains all kinds of mixing languages in code mixed Roman Urdu-English multilingual corpus in the research study. It includes the English, Urdu inter and inter sentential, and word level mixing. Few examples which show different typing of language mixing are:

- Inter sentential
- [sent-lang = “Ur”] aapkitnayachayhoo [/sent]
- [sent-lang = “Ur”] Really you are great [/sent]
- Intra-sentential
- [Sentence-Language = “Ur”] [Inclusion-Language = “En”] by the way [/Inclusion] oho [Fragment-Language = “En”] My creeping arms will always be empty….. never cling to you [/Fragment] lyen too badaow [Inclusion-Language = “En”] choozy [/Inclusion] [/Sentence] [sentence-language = “”]
- Word level code-mixing
- [sentence-language=“Ur”] [inclusion-language=“En”] First Year [/inclusion] eo to ei [wlcm-type=“En+Ur Suffix”] tymmer [/wlcm] modhhyesoobarjutee jay.. [/sentence]

4. Results and Discussions

The dataset of Roman Urdu and English was segmented and fragmented into language constituents of POS of tagging, chucking, tokenization of phrases, and segments of sentences. Therefore, we require the annotation of the corpus to fragment and segment the above sentences. We select three persons: computer science proficient, and the other two are experts in Urdu and English languages. We evaluate the word-level inter annotation using the Cohen Kappa method on 100 comments of the two annotators. The inter annotation agreement is calculated as:

$$\kappa = \frac{\mathrm{Prop(A)} - \mathrm{Prop(E)}}{1 - \mathrm{Prop(E)}}$$

κ measures the inter-annotation agreement at a certain point for annotation. Prop(A) is the proportion of time one annotator agrees with another, and Prop(E) is the proportion of agreement accepted purely for chances.

Table 2 shows the language label and tagging statistics of our data. Roman Urdu is 70.8 % of the corpus, followed by 28.3% of English tokens, and reset 0.9 % are unknown words that are universal tokens.

Label	Count	Percentage
Roman Urdu	141600	70.8
English	5660	28.3
Uni	180	0.9

Table 2. Roman Urdu-English Data Set along with Language label distribution of tokens

We also measure the level of ambiguity in the corpus. On inspection, 99.1% of all labels are unambiguous, and 0.9% are ambiguous. Among the ambiguous, the examples include tooh, toh, etc. These ambiguous words are either from vocabulary or produced due to Romanian or Urdu. This dataset is annotated manually using Petrov et al. universal POS tag set (2011). The distribution of the POS tag set is in table 3, and the POS tagging categories with the highest frequency are nouns, verbs, and punctuations. On inspecting, 33.2% were Urdu tokens, 15.7 % were verbs, and 16.1% were punctuations. Among all labels, the least were conjunctions and emotions (X).

Label	Count	Label	Count
ADJ	1658	NUM	172
ADP	1231	PRON	2245
ADV	575	PRT	1143
CONJ	637	PUNCT	3268
DET	789	VERB	4156
NOUN	7491	X	492

Table 3. Distribution of POS label set in Roman Urdu and English dataset.

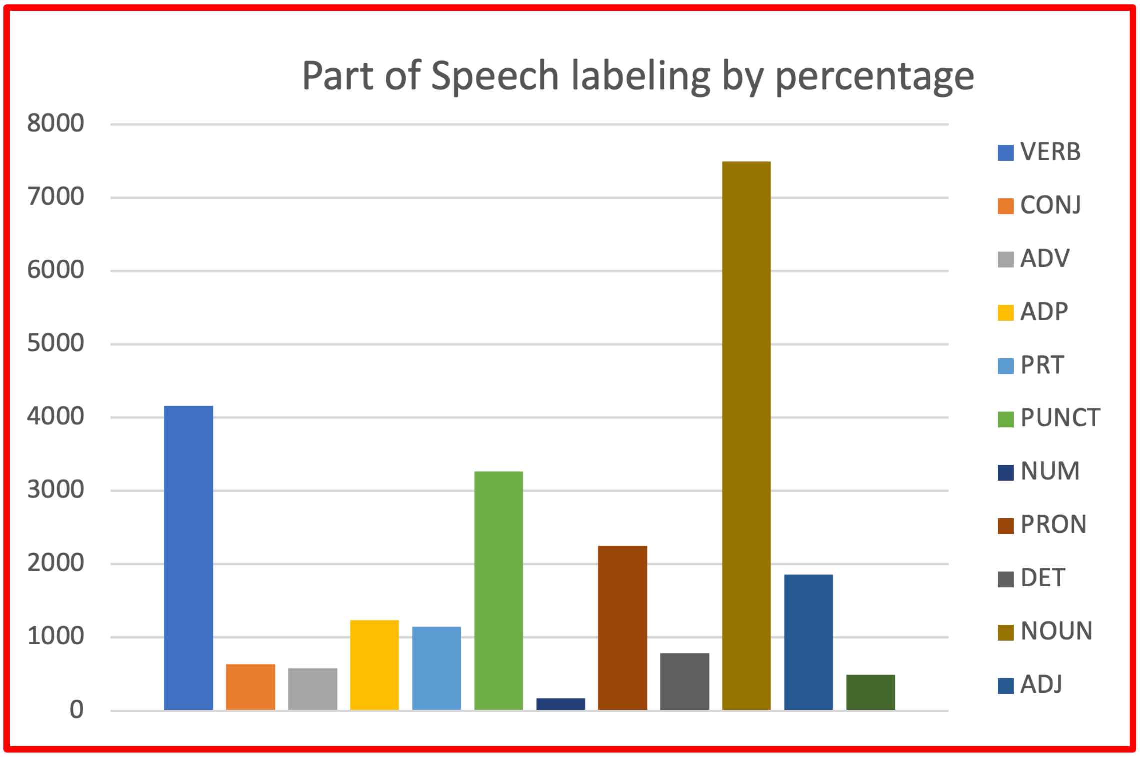

Figure 2. Percentage of POS Labels in the Roman Urdu and English dataset

P \ L	Ur	En	Nu	Uni	Total
ADJ	1201	558	0	1	1760
ADP	715	329	0	3	1047
ADV	348	173	0	9	530
CONJ	491	321	0	13	825
DET	478	512	0	0	990
NOUN	3498	2689	675	89	6951
NUM	16	7	0	111	134
PRON	**1672**	**436**	**0**	**4**	**2112**
PRT	1067	78	0	5	1150
PUNCT	**0**	**0**	**0**	**3973**	**3973**
VERB	2796	972	0	5	3773
X	**3**	**4**	**0**	**387**	**394**

Table 4. Distribution of POS and language labeling individually

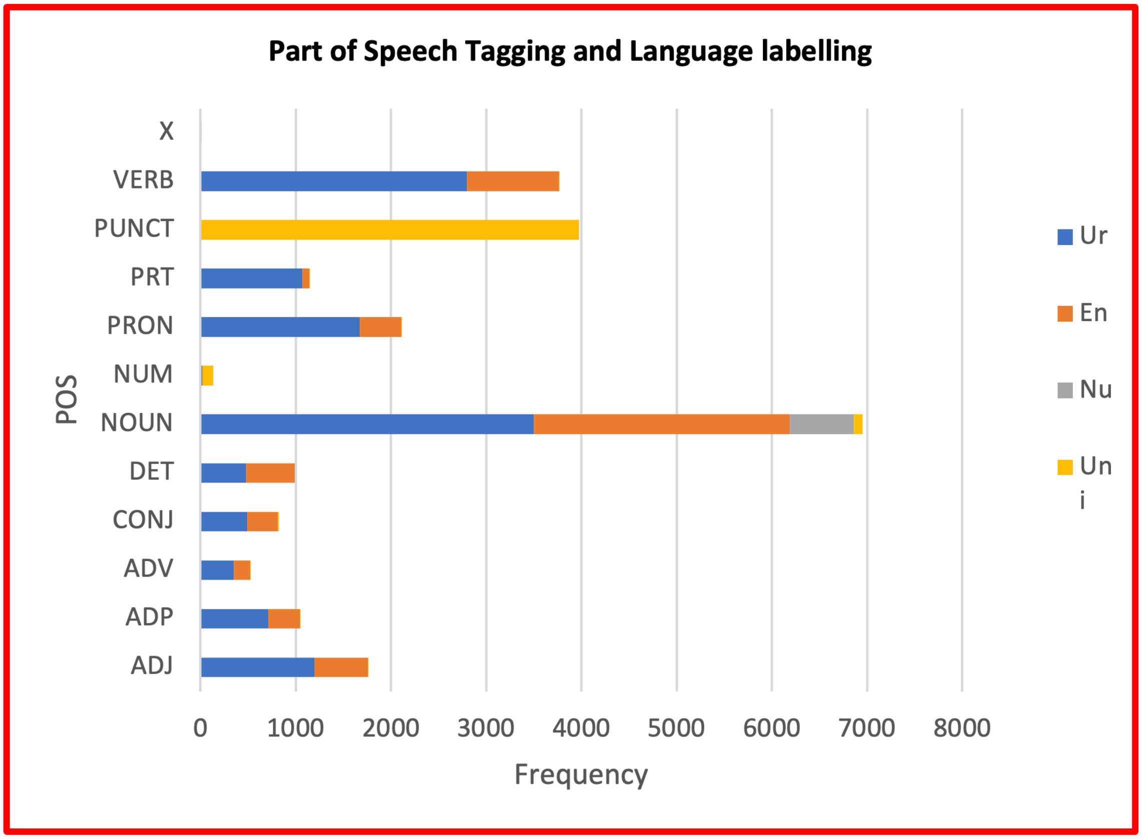

Figure 3. POS distribution, Language Labeling across column visualization from our Roman Urdu and English Dataset

The POS languages labels are distributed depicted in table 4. and Figure 3. Roman Urdu is the dominant language across all POS labeling except for NUM. Only 16 numbers can be seen in the Roman Urdu and English corpus. We observe only 1.59 % of Uniliteral labels are NOUN tagged.

6. Conclusion and Future Work

In conclusion, the language tagging and segmentation of multilingual Roman Urdu and English data is a crucial aspect in the field of computer linguistics. The method described in this study provides a systematic approach to address the complexities of multilingual text. Through the utilization of Python APIs, natural language processing methods and tools, and advanced algorithms such as Support Vector Machines and Bi-Directional Neural Networks, this study has demonstrated the effectiveness of language tagging and segmentation. This groundbreaking research has bridged the gap by collecting multilingual Roman and English texts, making the Roman Urdu and English corpus available for further refinement.

The corpus has been manually tagged and annotated, and subsequently fed into the vocabulary, providing a foundation for future advancements. The results of this study are nothing short of remarkable, demonstrating that multilingual Roman Urdu and English text can be annotated, segmented, and tagged with remarkable efficiency. The text can be further annotated using the dictionary method of Part of Speech Tagging and Segmentation, opening up new avenues for exploration. This research holds tremendous potential for future advancements in the field of computer linguistics and NLP.

Statements and Declarations

Author Contributions: All authors contributed equally to the conceptualization, formal analysis, investigation, methodology, and writing and editing of the original draft. All authors have read and agreed to the published version of the manuscript.

Institutional Review Board Statement: Not Applicable.

Informed Consent Statement: Not Applicable.

Data Availability Statement: Not Applicable.

Conflicts of Interest: The authors declare no conflict of interest.

References

- Al-Ohali, Y., Cheriet, M., and Suen, C. (2003). "Databases for recognition of handwritten Arabic cheques." Pattern Recognition, 36(1), 111-121.
- Bach, N. X., Linh, N. D., and Phuong, T. M. (2018). "An empirical study on POS tagging for Vietnamese social media text." Computer Speech & Language, 50, 1-15.
- Barman, U., Das, A., Wagner, J., and Foster, J. (2014). "Code mixing: A challenge for language identification in the language of social media." In: Proceedings of the First Workshop on Computational Approaches to Code-Switching, 13-23.
- David, Y., Grace, N., and Richard, W. (2001). "Inducing multilingual text analysis tools via robust projection across aligned corpora." In: Proceedings of the First International Conference on Human Language Technology Research, 1-8.
- Durrani, N., and Hussain, S. (2010). "Urdu word segmentation." In: Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, 528-536.

- Ghosh, S., Ghosh, S., and Das, D. (2016). "Part-of-speech tagging of code-mixed social media text." In: Proceedings of the Second Workshop on Computational Approaches to Code-Switching, 90-97.
- Halgamuge, M. N., Caliskan, H., and Mohammad, A. (2021). "Computation Time Optimization on Hashtag Segmentation for Social Media Data." In: 2021 IEEE Wireless Communications and Networking Conference (WCNC), 1-6.
- Hassan, S. M., Ali, F., Wasi, S., Javeed, S., Hussain, I., and Ashraf, S. N. (2019). "Roman-Urdu News Headline Classification with IR Models using Machine Learning Algorithms." Indian Journal of Science and Technology, 12(35), 1-9.
- Jauhiainen, T., Lui, M., Zampieri, M., Baldwin, T., and Lindén, K. (2019). "Automatic language identification in texts: A survey." Journal of Artificial Intelligence Research, 65, 675-782.
- Khan, S. N., Khan, K., Khan, A., Khan, A., Khan, A. U., and Ullah, B. (2018). "Urdu word segmentation using machine learning approaches." International Journal of Advanced Computer Science and Applications, 9(6), 193-200.
- Khan, S. N., and Usman, I. (2019). "A model for English to Urdu and Hindi machine translation system using translation rules and artificial neural network." Int. Arab J. Inf. Technol., 16(1), 125-131.
- Kumhar, S. H., Kirmani, M. M., Sheetlani, J., and Hassan, M. (2020). "Word Embedding Generation Methods and Tools: A Critical Review."
- Kumhar, S. H., Kirmani, M. M., Sheetlani, J., and Hassan, M. (2021). "Word Embedding Generation for Urdu Language using Word2vec model." Materials Today: Proceedings.
- Kumhar, S.H., Kirmani, M.M., Sheetlani, J., and Hassan, M. (2020). "Sentiment Analysis of Urdu Language on Different Social Media Platforms Using Word2vec and LSTM". Turkish Journal of Computer and Mathematics Education (TURCOMAT), 11(3), pp. 1439-1447.
- Kumhar, S.H., Qadir, W., Kirmani, M.M., Bashir, H., and Hassan, M. (2021). "Effectiveness of Methods and Tools Used for Collection of Roman Urdu and English Multilingual Corpus". Design Engineering, pp. 13428-13438.
- Rafique, A., Malik, M.K., Nawaz, Z., Bukhari, F., and Jalbani, A.H. (2019). "Sentiment Analysis for Roman Urdu". Mehran University Research Journal of Engineering & Technology, 38(2), pp. 463-470.
- Roy, R.S., Padmakumar, A., Jeganathan, G.P., and Kumaraguru, P. (2015). "Automated Linguistic Personalization of Targeted Marketing Messages Mining User-Generated Text on Social Media". In: International Conference on Intelligent Text Processing and Computational Linguistics. Springer, Cham, pp. 203-224.

- Sharf, Z., and Rahman, S.U. (2018). "Performing Natural Language Processing on Roman Urdu Datasets". International Journal of Computer Science and Network Security, 18(1), pp. 141-148.
- Teahan, W.J. (2000). "Text Classification and Segmentation Using Minimum Cross-Entropy". In: Content-Based Multimedia Information Access-Volume 2. pp. 943-961.

Declarations

Funding: No specific funding was received for this work.

Potential competing interests: No potential competing interests to declare.