Spectrum of Engineering Sciences
ISSN (e) 3007-3138 (p) 3007-312X Volume 3, Issue 8, 2025

A RULE-BASED APPROACH TO PASHTO TEXT PREPROCESSING: A TECHNIQUE FOR NORMALIZATION, STEMMING, AND TF-IDF INDEXING (RATP-TFI)

Abdul Qadir (Achakzai)[*1], **Ihsan Ullah**[2]

[*1]*Department of Computer Science*
[2]*University of Balochistan, Quetta, Pakistan*

DOI: https://doi.org/10.5281/zenodo.16960008

Keywords
Text Preprocessing, Natural language processing (NLP), Low-Resource language, Pashto Linguistic, Corpus Creation, Pashto Text Normalization, Tokenization, Stop word removal, stemming, Lemmatization, Noise Removal, Morphological Analysis, ROUGE Evaluation, TF-IDF weighting Technique, Rule Based Model, Pashto Language

Article History
Received: 26 May, 2025
Accepted: 08 August, 2025
Published: 27 August, 2025

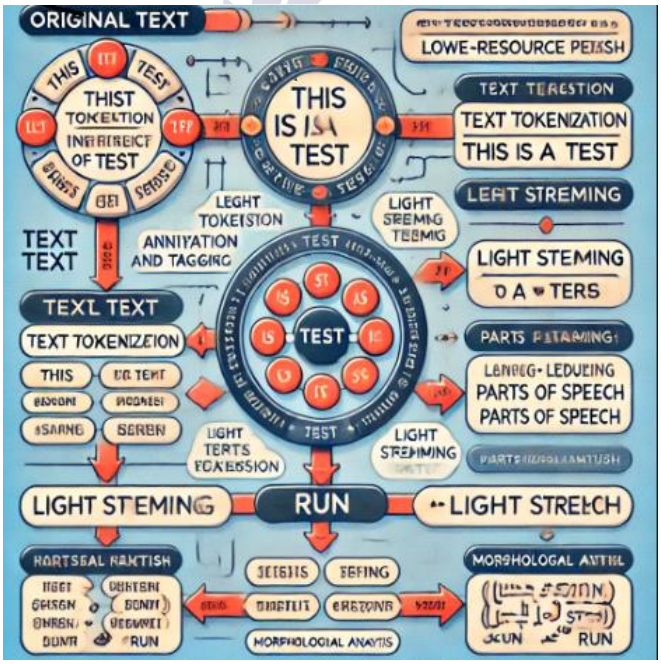
Corresponding Author: *
Abdul Qadir

Abstract
Text pre-processing is one of the Fundamental steps in Natural language Processing (NLP), mainly for low-resource languages like Pashto. which focuses and impact the overall performance of the study by developing acclimate pre-processing techniques to address the unique linguistic, orthographic variations, dialectal differences and challenges of the Pashto language which is a low-resource language. Key parts include Pashto-specific normalization rules of script variation to handle, stop-word list, stemming algorithm to reduce inflectional diversity. The main objective of Text pre-processing is to prepare raw text data for accurate and efficient processing by machine learning models. which transforms noisy, inconsistent and unstructured text into a standardize or structured format suitable for computation and analysis and enhance computational efficiency, facilitate model understanding, standardize text representation and improve model performance.
*This model **A Rule-Based Approach to Pashto Text Preprocessing: A Technique for Normalization, Stemming, and TF-IDF Indexing (RATP-TFI)** performs important NLP tasks like Normalization, Tokenization, Stemming, POS tagging and Stop-word removal for Pashto. The Pashto Text Corpus (PTC) consisting of 30k Pashto text documents which are collected from different sources like Websites, social media, news, books and Pashto Academy Quetta, Pakistan.*
A rule-based system tags words assigns grammatic and semantic tags to words using predefined linguistic rules advancing in Language specific-customization, no dependency on large datasets are required, Doman specific adaptation (fine-tuned for specific domains), Resilience in limited resources and foundation for further researchers with their part of speech in Pashto input text. The ROUGE metric evaluation is used for assessing the quality, effectiveness and providing the preprocessing text to handle Pashto's morphology, tokenization and orthographic variations ensuring the accurate comparisons of generating and referencing of summaries. The proposed RATP-TFI technique achieving 93% accuracy on the PTC corpus.

INTRODUCTION
Pashto, a member of the Indo-Iranian branch of the Indo-European family, is a language spoken in Afghanistan and Pakistan. Written in right to left (RTL) in the Perso-Arabic script [12]. Despites its significance as a national and

regional language, Pashto poses unique challenges for NLP researches due to regular orthography, diverse dialects and limited annotated corpora. The growth of textual data online triggered the need of powerful and effective tool that provides the desired content in a summarized form automatically while obtaining the core information [16]. Text processing is a very important, foundational and fundamental step in Natural Language Processing (NLP) which ensures textual data is structured, Cleaned and normalized for computational analysis or for downstream applications like Machine learning mod0els, linguistic analysis and information retrieval systems. Text processing performed tasks like Data Cleaning, tokenization, Stop-words, stemming, Normalization, POS tagging, Lemmatization, which aim is to enhance the efficiency and accuracy of subsequent machine learning and linguistic tasks. The objective of text preprocessing is to transform raw text into a meaningful format, which makes text easier for any task of text preprocessing [1]. Text preprocessing technique have been made a significant advancement for high-resource languages like Chinese and English, while low-resource languages like Pashto remain underexplored. Pashto's complex morphology, the use of diacritics, and agglutinative structure make Text preprocessing both challenging and necessary task [5].

In this paper, we had tried to build a base and key for Text pre-processing techniques for Pashto Language which helps to do all the basic Text tasks like tokenization, normalization, stemming, stop-word removal, and handling diacritics. A strong base or ground is needed for every work with the help of which one can make a formal and strong building of his need. The aim of this research is to address these challenges by systematically exploring and refining text preprocessing techniques for Pashto. Mainly focuses on developing a robust stop-word list, improving text normalization methods to align language's grammatical and orthographic rules better. The result of which provide a foundation for advancing NLP applications in Pashto, like sentiment analysis, Text summarization, Machine translation, enabling computational process for low-resource languages and providing a foundational understanding which can help improve the efficiency and accuracy of many NLP applications.



Fig, 1: The proposed framework of the model

2. Advantages or Benefits:

Some prominent applications which take advantages from structure data are: **Text Summarization** which is used to generate and concise summaries of huge and long documents, articles, news for abstracting important information. **Sentiment Analysis** which determines the positive, negative sentiment of a text which is used in product reviews,

monitoring the social media and in customer feedback. For which Tokenization, stop-word removal and stemming are essential to identify the accurate related words or phrases. **Machine Translation,** these apps are used from Translation of text from one language to another and ends the communication gaps across the language's barriers. For such kind of tasks text processing is essential to standardize input text for neural machine translation and to solve any unique features which is linguistic. **Named Entity Recognition (NER)**, this is used for identifying entities like date, location, names and organization in the text which is used in extracting information for various domains. In such tasks the Role of Tokenization and POS tagging are used to improve the detection of entities**. Question Answering (QA) Systems,** these apps are used in providing answers to user queries from structured knowledge-based text sources like chatbots and FAQs. Where preprocessing helps to identify the required information in match questions and text patterns. In which parsing, tokenization and entity recognition are the essential tools for extracting the requirements. **Text Classification,** these apps categorize text into classes. Like (spam detection, classification of documents or to categorize news. In such apps Tokenization, stemming, stop-word removal improve accuracy by focusing on the relevant items. **Information Retrieval (IR) and Search Engines,** these apps enable retrieval of relevant texts and documents based on individual queries, which are used in search engines, library catalogues and recommendation systems. In this process Tokenization, stop-word removal and stemming helps in indexing and matching text or documents and in low-resource languages like Pashto rule-based models can also help and improve in retrieval performance. **Speech Recognition and Text-to-Speech (TTS)**, these apps convert spoken languages into text (ASR) and text (TTS) to voice-based applications. Text normalization, phoneme alignment and pronunciations dictionaries of text preprocessing plays a very vital role. **Optical Character Recognition (OCR) and Document Digitization**, these apps are used to converts and recognize handwritten text in images to digital text which are used for digitalizing documents. Text normalization and cleanup are necessary to correct errors and to format the text after OCR. **Spelling and Grammar Checking,** these apps correct and identify the grammatical and spelling errors in text which are used in word processor and many other languages learning tools. Tokenization, POS tagging helps in identifying and in correcting the errors. In low-resource languages grammar checking tools can benefit from rule-based methods or small datasets. **Text Generation and Language Modeling,** these applications Generates text like human, which are used in writing automatic reports, chatbots, and story generation. The role of Text processing is very essential to train models with the help of tokenizing inputs and normalizing vocabularies.

These all applications shows that how essential and foundational text processing is, which enables NLP applications and makes the hard tasks of our daily life so simple, mainly for low-resource languages like Pashto.

There is no article or research paper published on the subject Text Pre-processing Technique for Low-Resource language: Pashto Language, in NLP till now. The RATP-TFI model is a pioneering framework which contains more than 3.5 k text documents is designed to identify the unique challenges of text preprocessing in Pashto language which focuses on essential tasks like tokenization, stop word removal, normalization, part of speech (POS) tagging, stemming to prepare Pashto text for various natural language processing (NLP) applications. The model utilizes the Pashto Text Corpus (PTC), a carefully curated dataset comprising text from diverse Pashto-language sources which includes literature, websites, and social media.

The model employs the TF-IDF approach which identify high-frequency stop-words in Pashto, enabling the creation of a robust stop-word list for effective text filtering and it incorporates the identification of common suffixes, prefixes and prefixes combination in Pashto for accurate stemming, addressing the rich morphological structure of the language.

For POS tagging, the model uses a rule-based approach tailored to Pashto's complex syntax and grammar, which facilitate tasks of text summarization, information retrieval and sentiment analysis.

The proposed RATP-TFI model demonstrates significant accuracy and effectiveness in processing Pashto text, making it a foundational tool for advancing NLP application in the Pashto language. It lays the groundwork in computational linguistic for low-resource language Pashto.

3. Text Normalization (Removing diacritics, standardizing characters):

Text normalization is a crucial preprocessing step in NLP, aimed at improving the quality and efficiency of downstream tasks. In this process, unnecessary elements such as special characters, HTML tags, punctuation mark, numerical digits and non-Pashto (or non-target language) words are removed from the input documents. This step significantly enhances the performance of various NLP applications, including text-to-speech synthesis, speech recognition, information extraction, text summarization, sentiment analysis, and machine translation [21]. In this proposed RATP-TFI model, a comprehensive list of such unwanted elements (special characters, HTML tags, punctuation, numbers and non-Pashto words) is integrated to automatically eliminate them from the input text during preprocessing.

Punctuation, Non-Alphanumeric Characters: (😊 ,‹, .,؟, !, << >>,?, /,{},[]. etc)

Normalization in Pashto languages characters used, like (ي and ې). (**eg. Taweel (ـ)** : which is also used in Arabic language to erlang word like(ســـلام) becomes (سلام).

Removing Diacritics	کِتاب, کتاب
Standardizing character	ی; ک, ك and ى, ي
Removing Charater	خوشحاله →!!! 😊 خوشحاله★
Normalization of Numbers and Date	۰ ۱ ۲ ۳ ۴ ۵ ۶ ۷ ۸ ۹...۵۰ to 12345...50

Table 1. Examples of Normalization

By applying these kinds of normalizing steps to Pashto language texts brings significant improvement in the quality of downstream tasks, which makes the model less sensitive in text data.

4. Tokenization:

Tokenization in Natural language processing (NLP) is the initial step and a process of breaking down text into smaller, manageable units called tokens. These tokens can be a sentence, words, sub words or even characters. This step involves breaking a paragraph into individual sentences. Sentence boundaries are typically marked by punctuation such as periods, exclamation marks, or question marks, followed by a space. Once a sentence is identified, the next step is to break it into individual words. This is generally done by using
white spaces as
delimiters to identify word boundaries. Some Examples of Pashto language text into tokens:

Original sentences	Tokenized Sentences
زه مکتب ته ځم	"زه ", "مکتب", "ته ", "ځم"
نن هوا ډېره ښه ده	"نن" , "هوا" , "ډېره" , "ښه" , "ده"
احمد په بازار کې میوې رانېولې	" احمد ", "په ", "بازار ", "کې ", "میوې ", "رانېولې"

Table 2: Pashto text converted into tokens.

5. Stop-word Removal:

In Natural Language Processing (NLP), stop words are common words that occur frequently in a language but carry/less meaningful information in tasks like text analysis, classification, or search.
Analyzing word frequencies across the corpus for identifying the most common words. Stop words are typically high-frequency words that appear in almost all texts. In Pashto a set of stop words involved in identifying commonly used words that do not add substantial meaning in most of text processing tasks such as:

او and	په in,on	د of, for	چې that	دی Is	ده Is (feminine)	يې his,her,its	نه not	هم also	یو One,a
زه I	مونږ we	ته You(singular)	تاسو You(plural)	دوی They	هغه He,she,that	دا this	کوم which	تر to	ځکه because
لپاره for	لکه like	که If	بیا again	همدا This	هر Every	لږ Few,little	نو So,then	څه what	څوک who

Table 3: Some Examples of stop words

Pronouns	زه, ته, هغه, مونږ, هغې دا, مونږ, تاسو, دوی
Prepositions	سره, له, په, ته etc
Articles/Determiners	هغه, دا, یو etc
Conjunctions	یا, خو, او etc
Auxiliary verbs	وې, دی, ده etc

Table 4: Some examples of potential Pashto stop words

6. Indexing Techniques:

Document indexing is a process of storing and organizing preprocessed text that makes it easy to retrieve, search or analyze. Indexing is a shortcut map which tells the system to find the correct place where the required text is located instead of wasting much time [24]. Many indexing techniques are used to retrieve information such as suffix trees, signature files discussing the work, performance and stability [25]. And it is a crucial step used in my research for Pashto text Preprocessing which aimed at improving efficiency by selecting and weighting key terms. beneficial to quickly and fast search or find documents containing words or phrase. Beneficial for Efficient retrieval which is needed for tasks like summarization, questions and answering and also helps in TF-IDF, filtering frequency and scaling.

TF-IDF (Term Frequency * Inverse Document Frequency):

TF-IDF is a statistical approach used to calculate the weight of any word in the document [6, 27]. TF-IDF is used in different fields of study, such as information retrieval, machine learning, and text mining [7, 26, 28].

(TF-IDF) : TF-IDF (t,d) = TF (t,d) *IDF (t)

The formula gives and extract a final score that how much the word t is important in a document d, and considering how often both appears (TF) and how rare it is across all the documents (IDF).

Term Frequency (TF): Helps which words are more dominant in a given document and make that useful for feature extraction.

Document = پښتو ژبه ښکلي ده پښتو تاريخي ژبه ده

Tokennizing1st = پښتو, ژبه, ښکلي, ده, پښتو, تاريخي, ژبه, ده

Term count n*i,j* =

Term	TF = ni,j /8
پښتو	2/8 = 0.25

ژبه	2/8 = 0.25
بنکلي	1/8 = 0.125
ده	2/ 8 = 0.25
تاريخي	1/8 = 0.125

Table 5: Example of TF Values

Inverse Document Frequency (DF):
Inverse Document Frequency (DF) is a simple basic, important and effective method for vocabulary reduction in text mining and NLP for identifying important words or terms from a collection of text documents and provides effective scenario in social media and in news feeds [26, 30]. It enhances apps like key term extraction [29]. And is used how common or rare a term is across all the document set. For Pashto language this helps in address issues of sparsity and noise in low-resource corpora.
IDF(*t*)=log(n/df(t))

Documents =

زه پښتو ژبه خوښوم
پښتو ډېره بنکلي ژبه ده
زه د پښتو ادبيات لولم
انګليسي او پښتو مختلفي ژبې دي
پښتو ادبيات ډېر بډايه دي

DF(t)= Number of Document that contains term t

Words	Output
پښتو	5
لولم	1
ادبيات	2
دي	2
مختلفي	1
انګليسي	1

Table 6: example of (IDF)

Document Frequency (DF) Thresholding:
Document Frequency (DF) is a technique used during text preprocessing and feature selection to remove very rare terms (low-DF, like typos, names or noise) and to remove very high/common terms like (stop words or uninformative words or data). DF Thresholding is especially useful for low level languages like Pashto where they have small and noisy datasets. DF Thresholding helps in reducing dimensionality (fewer features) and helps in improving model accuracy (like Removing irrelevant words). The count of documents in which a term occurs. Thresholding involves setting a minimum DF value *m* and only retaining terms that appear in at least *m* documents. This reduces the

vocabulary by filtering out infrequent terms that may contribute little to downstream tasks like classification. DF Thresholding is useful for Pashto language in Sparse Data, Improved Generalization and in Noise Reduction.

Choosing the threshold *m*=2

m=2 is more effective than m=1 because terms appearing in only one document are likely to be noise or too specific to contribute meaningfully to broader patterns and ensures that the terms included in the vocabulary occur in at least two documents, providing some level of validation for their relevance. Impact in reducing noise, overfitting and ensures meaningful tasks terms are retained for tasks like classification and clustering. Implementation for Pashto by applying stemming before DF thresholding to group different inflected forms of the same root.

Word form	Root
زده کړه، زده کوونکی، زده کړو	زده
ښوونځی، ښوونځو، ښوونځي	ښوونځ
کتاب، کتابونه، کتابتون، کتابدار	کتاب
لیک، لیکوال، لیکل، لیکنه	لیک
خبرې، خبري، خبرونه، خبرګزاري	خبر

Table 7: Example of Thresholding *m*=2

For a larger corpus we use m=5 or higher to reduce the vocabulary size further and which focus on more widely occurring terms. Pashto is rich morphological language where this is an exclude important terms.
TF-IDF used for calculating the weight of stop words in the documents where the PTC contains 121,4101 words, among these words 1044 were identified as high frequency words of Pashto language through TF-IDF technique.

Stop-word	Frequency	Stop-words	Frequency
چې	4150	**او**	3911
يې	3589	**په**	3875
له	3868	**دی**	3578
نه	2560	**ده**	2901
زده	1578	کتاب	2709
ښوونځ	1267	لیک	2324
خبر	1025	**هغه**	2019

Table 8: Example of high-frequency stop-words extracted from PTC using TF-IDF

1.4 Stemming and Lemmatization:

Text Sentence	کتابونه زما دی او زه په خپل ښوونځي کې یم
Stemmed sentence	کتاب زما دی او زه په خپل ښوونځي کې یم

Table9: example of stemming

Stemming: Stemming is a process to reduce a word to its base or root form removing suffixes and prefixes (eg walking, walks, walker →walk). There are many tools that can be used in applications such as database search engine, text analysis, information retrieval and in documents indexing[2]. To Develop or create a rule-based stemming for Pashto, and to focus on reducing text words to its root forms. Pashto morphology, have common affixes for many

verbs, nouns, and plural markers (eg. د, زما, ي, ه, پدل, دل, يدل, ان, ونه). Rule-based stemming are benefiting the suffixes, prefixes in Pashto.

Lemmatization: Lemmatization and stemming in NLP are used for the process of reducing words to their dictionary or base form (lemma) but they work and serve slightly for different purposes. Lemmatization works differently and finds proper lemma considering context and grammar, other than simply cut off suffixes as done in stemming. Lemmatization Builds modules which map inflection form like plural and verbs conjugation to their basic forms. For instance (running – run,
flies – fly).

Type	Suffixes	Example
Gender	ي, ه	لیونی : لیون
Plural	ونه, ان	کتابونه ; کتاب
Verbs form	پدل, دل, يدل	خوښیدل ; خوښ
Possession	د, زما	زما د کتاب

Table 10: Some Examples of Lemmatization in Pashto

In this research NLP libraries Spacey and NLTK are used for better implementation and customization lemmatizing frameworks. Where, the rule-based approaches and dictionary-based approach are gathered. Where in 1st step Tokenization or split text into individual is done and in 2nd step Rule-Based or Dictionary-Based lemmatization is applied, where each word looked up and derived to its lemma and in the 3rd step, we Reassembled Text and combined all the lemmatized text back into the sentences.

For developing a Pashto Lemmatization Analyzing Domain. Specific vocabulary is used for Pashto text in the domain of legal and medical. Lemmatization iterated, improved and tested on a huge number of datasets of PTC created by Abdul Qadir, which is refined and combined with other preprocessing steps like stop-words removal to streamline Pashto text for NLP tasks in Pashto Language.

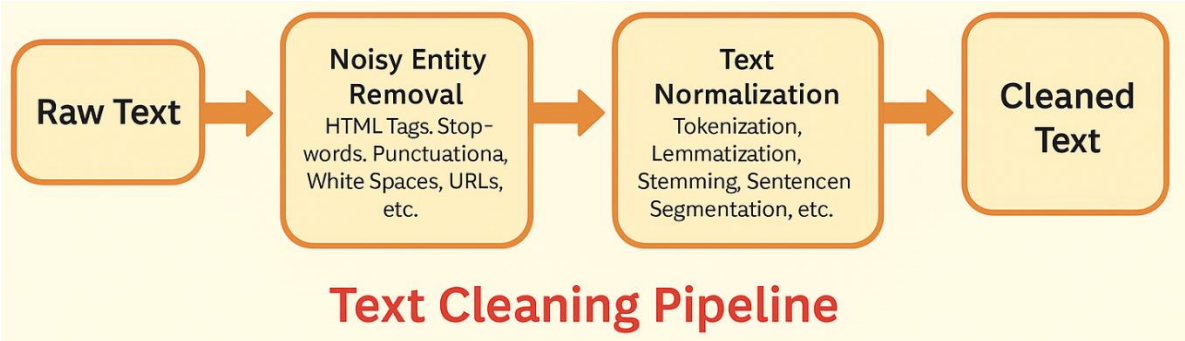


Figure 2: Steps of text Pre-Processing

Parts of Speech (POS) Tagging:
POS tagging is a crucial component of the text preprocessing pipeline, where each word in a document is sequentially labeled with its appropriate syntactic category, such as verb, noun, adverb, conjunction, preposition, interjection and noun. This process facilitates structured linguistic analysis and contributes to better feature representation for natural language processing tasks.
POS tagging is a process of sequential labeling of each word with the respective labels [8]. The Process to assign syntactic categories to each word as noun, pronoun, adjective and verb [10]. For morphologically rich and resource-

scarce language like Pashto, POS tagging presents significant challenges due to the lack of annotated corpora, standard tags sets and tools [19]. To address this as a rule-based POS tagging approach is employed in this study. The proposed model applies linguistic rules, suffix-based patterns and syntactic cues to assign tags to each token in the Pashto text.

Pashto Words	POS category	English Translation
زده کوونکی	Noun	student
خوږ	Adjective	Sweet
دی	Pronoun	It/ he
ځغلي	Verb	Runs
ډېر	Adverb	Very
او	Conjunction	And
په	Preposition	On/in
اوف	Interjection	Wow!

Table 11: POS Tagging example of Pashto Language

This tagging mechanism enables structured representation of Pashto documents and supports downstream applications such as sentiment analysis, summarization, text classification, machine translation and information. Furthermore, POS tagging contributes to context-aware feature extraction, helping distinguish between semantically similar but syntactically different terms. The integration of POS tagging into the preprocessing workflow enhances overall text quality and model performance, particularly in low-resource linguistic setting like Pashto.

1.5 Experiments and Results:

Evaluating Pashto text preprocessing techniques presents unique challenges due to the limited availability of linguistic resource. To address we developed a Pashto Text preprocessing Corpus (PTC) leveraging both existing tools and custom methodologies to enhance text analysis in the Pashto language.

Pashto text preprocessing technique encountered challenges due to the scarcity of online resources, such as publicly available APIs, datasets and stop-word lists. Consequently, certain aspects of the research required manual intervention by experts proficient in the Pashto language. For experimentation and evaluation, a Pashto text corpus was utilized comprising a substantial number of tokens and unique words. This corpus served as the foundation for developing and accessing preprocessing methods tailored to the Pashto language.

For our experiment beside our corpus PTC, we also utilize the TRAD Pashto Monolingual Text Corpus comprising approximately 112 million tokens collected from 46 different blogs and websites. This extensive corpus provided a robust foundation for the training and evaluating our preprocessing techniques.
(https://catalog.elda.org/en-us/repository/browse/ELRA-W0092/?utm_source.)

Preprocessing Pipeline:

The RATP-TFI framework encompasses several key processing steps:

Tokenization and word segmentation: Pashto language presented several challenges in tokenization due to inconsistent use of whitespace delimiting the word. To address or solve the issues we utilized the python library "Urduhack" to tokenize the documents and also used the "NLPashto toolkit CRF-based and NLP Toolkit for Low-resource Pashto Language "tokenizers a specialized machine learning models to accurately segment Pashto text, addressing common issues like space omission, spelling correction, insertion errors and enhancing the accuracy of tokenization processes.

Stop-word Removal, feature Extraction and POS tagging: A curated list of stop-words has been made and compiled to eliminate common words where the Term (TF-IDF) technique and a rule-based approach is employed to determine the significance of words within the corpus, facilitating the identification of key terms and to tag the words with their corresponding parts of speech.

Evaluation Metrics: We employed the ROUGE (Recall-Oriented Understudy for Gisting Evaluation) metric for the proposed RATP-TFI model evaluating the PTC dataset, where we assessed the quality of summaries and translations by measuring the overlap between the generated and reference text and achieved the overall performance of 83% (F-score) which showed a significant improvement compare to the baseline performance 61%, indicating the potential of the RATP-TFI model accurately analyzing Sindhi and Urdu text. The baseline is the classical approach without text preprocessing and the classical approach directly processes the raw data for the desired tasks. The Table 12 presents the extended version of the RATP-TFI model, considering recall, precision and f score. The application of this evaluation metric is standard in assessing NLP models.

Table 12: Using the ROUGE evaluation measuring the overall performance on Pashto Text Corpus

Approaches	Precision	Recall	F-score
PTC	83%	91%	88%
Baseline	61%	88%	83%

Fig 3. chart Shows the results of the ROUGE evaluation in term of Precision, Recall and F-score. Which suggests that the proposed preprocessing steps helps the model that it identifies relevant information more accurately.

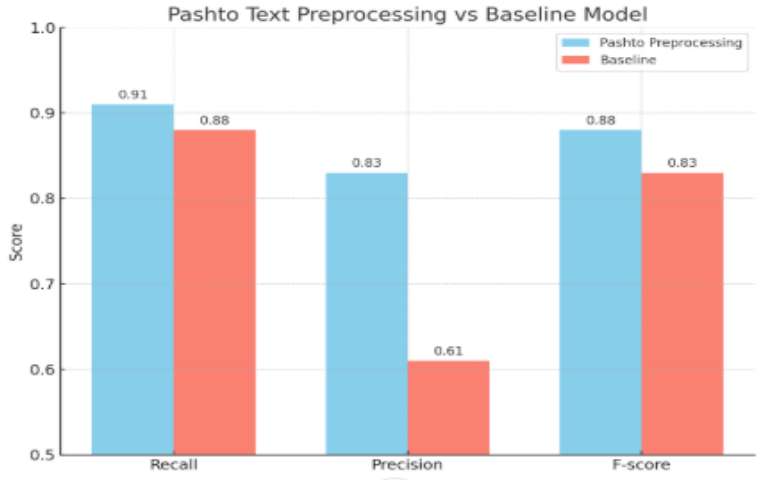


Fig 4. shows the model overall performance in Recall, Precision and F-measure.

Conclusion:

This study introduced **"A Rule-Based Approach to Pashto Text Preprocessing: A Technique for Normalization, Stemming, and TF-IDF Indexing" (RATP-TFI)** that performs essential steps including tokenization, normalization, stop-word removal, stemming, Prefix/suffix terms for normalization and POS tagging on Pashto text. A corpus of Pashto documents was used for evaluation. Through TF-IDF analysis, a high-frequency Pashto stop-word list was developed to enhance downstream processing. POS tagging was implemented using a rule-based method, improving

the quality of syntactic analysis. Compared to the baseline approach, the Proposed PTPTL model achieved higher performance metrics across all measures, highlighting its effectiveness in processing Pashto text and offering a strong foundation for future research in low-resource language NLP.

REFERENCES

Ali Nawaz, A. N., Nawaz, M., Shaikh, N. A., & Rajper, S. Baber, J.., & Muhammad Khalid. TPTS: Text Pre-processing Techniques for Sindhi Language: Text Pre-processing Techniques. Pakistan Journal of Emerging Science and Technologies (PJEST), 4(3).

Aslamzai, S., & Saad, S. (2015). Pashto language stemming algorithm. Asia-Pacific Journal of Information Technology and Multimedia, 4(1), 25-37.

Doostyar, N. M., & Sujatha, B. (2023). Plagiarism detection for Afghan national languages (Pashto and Dari). In Internet of Behaviors (IoB) (pp. 171-186). CRC Press.MacKenzie, D. N., & David, A. B. (2018). Pashto. In The world's major languages (pp. 470-495). Routledge.

Haq, I., Qiu, W., Guo, J., & Peng, T. (2023). The Pashto Corpus and Machine Learning Model for Automatic POS Tagging.

Haq, I., Qiu, W., Guo, J., & Tang, P. (2023). NLPashto: NLP toolkit for low-resource Pashto language. International Journal of Advanced Computer Science and Applications, 14(6).

[16] A. Nawaz, R. A. Shaikh, R. H. Arain, S. Rajper, J. Baber, and M. M. Baidani, "Text Summarizer for Sindhi Language," Available at SSRN 4288269.

M. A. Dootio and A. I. Wagan, "Development of Sindhi text corpus," Journal of King Saud University-Computer and Information Sciences, vol. 33, pp. 468-475, 2021.

W. Ali, R. Kumar, Y. Dai, J. Kumar, and S. Tumrani, "Neural Joint Model for Part of-Speech Tagging and Entity Extraction," in 2021 13th International Conference on Machine Learning and Computing, 2021, pp. 239-245.

J. A. Mahar and G. Q. Memon, "Rule based part of speech tagging of sindhi language," in 2010 International Conference on Signal Acquisition and Processing, 2010, pp. 101-106.

I. N. Sodhar, A. H. Jalbani, M. I. Channa, and D. N. Hakro, "Parts of speech tagging of Romanized Sindhi text by applying rule based model," IJCSNS, vol. 19, p. 91, 2019.

MacKenzie, D. N., & David, A. B. (2018). Pashto. In The world's major languages (pp. 470-495). Routledge.
Mirdehghan, M. (2010). Persian, Urdu, and Pashto: A comparative orthographic analysis. Writing Systems Research, 2(1), 9-23.

Mohtaj, S., Roshanfekr, B., Zafarian, A., & Asghari, H. (2018, May). Parsivar: A language processing toolkit for Persian. In Proceedings of the eleventh international conference on language resources and evaluation (lrec 2018).

Muhammad Akbar Ali Khan, Sahibzada Abdur Rehman Abid, Fatima Tuz Zuhra & Nasir Ahmad: The Development of Pashto Speech Synthesis System. International Journal of Computer Applications (0975 – 8887) Volume 71– No.24, June 2013

Nawaz, A., Bakhtyar, M., Baber, J., Ullah, I., Noor, W., & Basit, A. (2020). Extractive text summarization models for Urdu language. Information Processing & Management, 57(6), 102383..

Qaroush, A., Farha, I. A., Ghanem, W., Washaha, M., & Maali, E. (2019). An efficient single document Arabic text summarization using a combination of statistical and semantic features. Journal of King Saud University-Computer and Information Sciences.

Rabbi, I., Khan, A. M., & Ali, R. (2009). Rule-based part of speech tagging for Pashto language. In Conference on Language and Technology, Lahore, Pakistan.

SHAHEEN ULLAH , RIAZ AHMAD , ABDALLAH NAMOUN, SIRAJ MUHAMMAD, KHALIL, IBRAR HUSSAIN, AND ISA ALI IBRAHIM. A Deep Learning-Based Approach for Part of Speech (PoS) Tagging in the Pashto Language. Digital Object Identifier 10.1109/ACCESS.2024.3412175. OPEN ACCESS JOURNAL

Ullah, S., Ahmad, R., Namoun, A., Muhammad, S., Ullah, K., Hussain, I., & Ibrahim, I. A. (2024). A Deep Learning-Based Approach for Part of Speech (PoS) Tagging in the Pashto Language. IEEE Access.

Tyler Baldwin, Yunyao Li, IBM Research -Almaden (2015). An In-depth Analysis of the Effect of Text Normalization in social Media.

Richard sproat, Navdeep Jaitly (24 jan 2017). RNN Approach to text Normalization: A challenge.

Jialin Ding, Umar Farooq Minhas, et al. (2019) Alex: An Updatable Adaptive Learned Index.

Paolo Ferragina, Giorgio Vinciguerra (2019). The PGM-Index: A Multixriteria, Compressed and learned Approach to Data Indexing.

Zulhair Malki (2016). Comprehensive Study and Comparison of Information Retrieval Indexing Technique.

Jun Wang, Wei Liu, et al (2015) Learning to Hash for Indexing Big Data – A Survey.

Kishore Papineni (2001). Why Inverse Document Frequency.

Kenneth Church and William Gale (1995). Inverse Document Frequency (IDF): A Measure of Deviations from Possion.

Stephen E. Robertson (2004). Undrestanding Inverse Document Frequency: On Theoretical Arguments for IDF.

Masumi Shirakawa, Takahiro Hara , Shojiro Nishio (2017). Generalized Inverse Document Frequency.

Margarita Karkali, Francois Rousseau, Alexandros Ntoulas, Michalis Vazirgiannis (2014).