ISSN: 2633-4828 Vol. 6 No.1, January, 2024

International Journal of Applied Engineering & Technology

BUILDING AND ANALYZING A BRAHUI TEXT CORPUS: APPLYING DTM AND TF-IDF TECHNIQUES

Naseer Ahmed[1], Mansoor Ahmed Khuhro[2] and Mazhar Ali Dootio[3]
[1]Department of Computer Science, Sindh Madresatul Islam University, Karachi, Pakistan
[1]Department of Artificial Intelligence and Mathematical Sciences, Sindh Madressatul Islam University, Karachi
[3]Department of computer science and information technology Benazir Bhutto Shaheed University Lyari, Karachi, Sindh Pakistan
[1]naseerbajoi@gmail.com and [2]makhuhro@smiu.edu.pk

ABSTRACT
There are many literary and general writings in the under-resourced language of the Brahui. To build the Brahui text corpus, a variety of books, magazines, and online resources are accessible. No appropriate and useful text corpora are generated and made available online for the study, linguistics analysis, language feature analysis, and information retrieval systems. Current challenges include a lack of resources for computational linguistics research and NLP applications for the Brahui language. However, we developed the Brahui text corpora as text resources for researchers, Natural Language Processing (NLP) specialists, and computational linguists. The Brahui text corpus is developed using books, magazines, and social media platforms. The prosperity of the Brahui text corpus, characterized by its extensive and complex terms, provides a comprehensive foundation for advanced linguistic analysis. The 3-gram technique of the N-gram model is used to build and analyze the Brahui text corpus using the Document Term Matrix and TF-IDF models. These analyses have proved the significance of the corpus for information retrieval. This Corpus may be used for studies on Topic modeling, Word2Vec, sentiment analysis, aspect-based cluster analysis, semantic analysis, and machine translation systems. To fully comprehend and make better use of the Brahui text corpus in NLP and computational linguistics, future research should also investigate more complex language aspects and advanced computational methods.

Keywords: The Brahui language, NLP, Text Corpus, Machine Translation, Computational linguistics, DTM, and TF-IDF.

1. INTRODUCTION

The Brahui language is one of the oldest languages of Pakistan. It comes under the Dravidian group of languages [1]. In the Balochistan province of Pakistan, the Baloch and Brahui tribes mainly speak the Brahui language as a Dravidian language; approximately 2.57 million people talk to Brahui as their mother tongue [2]-[3]. Language is generated from the collection of symbols that apply to spoken and written forms of language. Communication and business transactions are fundamental resources of human society. People use languages to communicate their ideas, values, and resources [4]. Only a few languages have the opportunity to make vast amounts of text corpus available [5]. Around 6,800 active languages are worldwide [6]. Like other human challenges, language is a product of evolution and is a problem for people. Many poor resource languages are still not targeted by the Natural Language Processing community. This is much more challenging to develop a corpus for such languages [7]-[8], and [9]. The Brahui language currently lacks NLP-ready resources for researchers, even despite its expanding online presence across blogs, websites, and social media. On the internet, the Brahui is often used. There are more and more Brahui blogs, literary websites, online news sites, and discussion forums daily. The second most common written and spoken language in Balochistan, a province of Pakistan after Urdu, is the Brahui language. Still, no resources are available for NLP researchers despite its widespread online usage for evaluating Language characteristics and variation analysis; a text corpus is mandatory [10]. Text corpora are essential to natural language processing (NLP) because they offer the baseline information required for developing and evaluating language models, enabling the development of accurate and useful NLP tools. Essential resources like digital lexical databases and linguistic corpora, which are required for language analysis and variation research are unavailable for the Brahui language a significant research gap that this study attempts

to address. The Brahui language processing tools, including extensive computerized lexicons and linguistic corpora, are still in the early stages of development [11]. The scripts used to write the Brahui include Perso-Arabic, Devnagri, and Roman (Sindhi). The Brahui writings are typically written in Perso-Arabic style [12]. There are currently no resources for the Brahui language, such as extensive computational lexicons or linguistic corpora. The lack of this information prevents the development and optimization of algorithms specifically designed for Brahui, which severely impedes NLP research. The problems associated with developing corpora are data collecting, annotation, and the requirement for linguistic and computational skills, which are particularly challenging for languages with limited resources. The absence of these resources pushes the Brahui language closer to being endangered. Numerous academic institutions and people aim to create linguistic corpora for the various Pakistani languages [13]. It becomes more precise and scientific through the process of linguistics. In addition to studying grammar, linguistics is a scientific study of a language that helps people learn its features, improve their communication skills, and comprehensively understand its features and purposes. The theoretical and applied components of a language are addressed by computational linguistics, which uses a variety of computational applications to solve and concentrate on human language problems and academic issues; languages present both cognitive issues and practical difficulties. The advancement of computational linguistics and the natural language processing process has made it easier for people from diverse nations and cultures to understand one another's languages. Recent improvements include text segmentation, text analysis, information retrieval, Machine Translation, syntactic parsing, etc. Natural language processing and computational linguistics research and developments have made them possible [14]. Languages around the world with limited access to technology, like the Brahui language, are inspired by these advances and research projects. The Sindhi language has linguistic issues, which the authors evaluated and analyzed [15]-[16]. In the same way, much more studies have been done on the English language [17]. As a result, there are a lot of NLP materials for English that are inappropriate for other languages, including Brahui, Urdu, and Sindhi. Languages written in the right hand are essential for machine learning, deep learning, and NLP applications. The Brahui language uses an Arabic-Persian script and is the right-hand side of the written language [16]. On the World Wide Web (www), just a few websites, news channels, and social media pages are accessible. We desire to fill the gap in Brahui's digital resources with this corpus and offer a strong basis for various Natural Language Processing (NLP) applications. The development of language corpora has enabled the advancement of NLP activities. Significant progress has been made in natural language processing tasks such as part-of-speech tagging using corpora developed specifically for morphologically rich languages like English. Similarly, researchers and developers may develop powerful natural language processing tools and applications customized to this unique language using a text corpus written in the Brahui language. For Brahui, this new corpus could significantly accelerate the advancement of NLP. Researchers may use this type of corpus for a variety of tasks, including sentiment analysis, topic modeling, aspect-based cluster analysis, sentiment embedding[18], (Word2Vector), and machine translation [14].

2. OBJECTIVES AND SIGNIFICANCE OF THE STUDY

In this way, we aim to support the preservation and restoration of the Brahui language by providing access to this text corpus. By providing this resource, the study bridges the way for natural language processing advancements for other under-resourced languages of the world. This initiative also provides the path and a way forward for advancements, of the Brahui language in future research and a motivation for the other under-resourced languages. This newly developed text corpus could serve as a foundation, for languages facing resource constraints and development advancements in language and technology on a broader scale. Such as information retrieval, Machine Translation, Text summarization, and computational linguistics often suffer from insufficient data for under-resourced languages, in the same way, The Brahui text corpus has been annotated using various NLP techniques for several purposes, such as TF-IDF and DTM.

3. RELATED WORK

This section covers studies on the use of TF-IDF and Document Term Matrix approaches for language analysis with a particular focus on text data sets. Recent developments in machine learning and natural language

processing have made language analysis through resources more and more common. Linguistic experts have found that corpus-based studies are a great resource for studying languages with limited resources. Research on languages like Cebuano demonstrates how corpora can be used to analyze linguistic features [19]. Studying endangered Uralic languages, like Udmurt shows how the development of a corpus for a language can help describe patterns for revitalization of these languages [20], [21]. To bridge the linguistic gap of computing resources, researchers have developed a text corpus for the Yoruba language using textual data. This approach has enabled the development of tools such, as analyzers and part of speech taggers [22]. They provided a summary of the methods used in analyzing languages, with limited resources. They discussed the application of methods such as TF IDF, in text analysis as well as the challenges and strategies involved in building text datasets [23]. The researchers developed a collection of texts, in the Hausa language. Applied DTM and TF IDF for quantitative linguistic analysis. Their study demonstrated the effectiveness of these methods in extracting insights, from data [24]. They developed a collection of Pashto text. Applied TF IDF to study the frequency of phrases. Their findings show how TF-IDF could enhance the development of language materials, for low-resourced languages [25]. A thorough examination of the applications of (DTM) in linguistics plays an important role, such as topic modeling and text clustering. Their study provided insights into the use of DTM, in linguistic analysis.[26]. The primary goal of the study was to develop a collection of Gujarati texts corpus and the authors used TF IDF to evaluate the significance of terms. Their research highlighted the role of these methods in handling language study[27]. The author's study introduced an approach, to analyzing text data by applying TF IDF and DTM methods. Their results demonstrated these techniques' effectiveness in capturing linguistic patterns [28]. The research focused on the analysis of the language corpus using TF IDF and DTM. Their study illustrated these methods' effectiveness in handling text collections and identifying linguistic characteristics [29]. In this study, a dataset, for the Punjabi language was developed. TF- IDF was used for assessment. The study also highlights the importance of such methods, in researching low-resource languages[30]. In this study, the researchers applied (DTM) to analyze a collection of Korean text corpus. The findings show the capability of DTM to detect the core themes within text datasets[31]. This study shows how collections of texts can assist in preserving languages by developing text corpus. Over the past few decades, electronic repositories of semi-real texts, known as corpora, along with methods, for analyzing them have offered valuable insights into language structures[32]. In recent years, there have been changes in the understanding of language instruction due to the expanding technology and the widespread acceptance of the internet. Using study materials known as corpora a collection of machine-readable authentic texts (including transcripts of spoken data) that is sampled to be representative of a particular language or language variety corpus linguistics is a cutting-edge approach to language analysis [33]. The application of corpus linguistics extends from language study to include learning a language [34]. Users may express and evaluate their opinions using text, which is a popular and effective technique[35]. Appropriate tools and approaches are required for proper corpus development. The processes involve data collection, annotation, and analysis of data [36]. Data gathering, linguistic processing, basic corpus cleaning, and corpus evaluation are the four important processes in producing and studying corpora. Use a multilingual corpus to comprehend the diversity and complexity of the languages. The Corpus can be categorized polarity- and topic-wise [37]. Conducted their research on building and analysis of corpora in the Urdu language. The system for cluster analysis was successfully trained using the K-means machine learning technique as a result [38]. These studies show how important corpora are to the growth of any language and language modeling. The Brahui language has a complex morphology and grammar, just like Arabic and Sindhi. Thus, developing a text corpus and analyzing its contents is to address the Brahui language's computational linguistics issues. The development of the Brahui text corpus is a great attraction for the NLP community to work on the Brahui language development and its analysis and make it safe from being an endangered language. To develop and analyze the Brahui text corpus for linguistic variance. There are several fields of language study where TF-IDF and DTM are applied. The TF-DF is useful for finding typical vocabulary in languages as demonstrated by studies examining the Yorùbá langua ge [39]. As shown in research on Udmurt, DTM visualizations assist in highlighting word co-occurrence patterns, helping language analysis [40]. The application of TF-IDF to text summarization and topic modeling is

investigated in the Cebuano study [41]. while demonstrating the adaptability of TF-IDF and DTM when used for sentiment analysis in Hindi text [42]. Finally, the study conducted by Johnson and Brown concentrated on the usage of DTM and TF-IDF in the analysis of the Yoruba language text collection. Their research shows how useful these techniques are for computational linguistic analysis[43]. Table 1 summarizes the Corpora Project, including its initiation, duration, and accomplishments. It likely delves into details about how the project built its language corpora and how it chose the data it included [44].

Table 1: Development of language corpus for South Asian languages

S#	Language	Started	Closed	Output
1	Urdu, Sindhi, and Kashmiri	1992	1995	3 million words corpus for each language of the group
2	English, Hindi, and Panjabi	-do-	-do-	-do-
3	Marathi and Gujarati	-do-	-do-	-do-
4	Sanskrit	-do-	-do-	-do-
5	Tamil	-do-	-do-	-do-
6	Bangali	-do-	-do-	-do-

2.1 The Brahui Text Corpus

The majority of people on the globe are using the Internet, and Industry 4.0, which will completely transform how businesses operate, is rapidly approaching. One of the key performance indicators has been considered to be the ability of tools and applications to transfer data in real-time across the internet. The Internet makes it possible to gather and distribute data to assess the effectiveness and shortcomings of its developers and operators. These data can be, by definition, social media, scientific, biological, or operational, demonstrating the diversity of various datasets. News headlines, tweets, social media posts, blog entries, user comments, news stories, scientific articles, and other sources are some of the many sources that provide textual information [45] The study of language using naturally occurring language samples is known as corpus linguistics; analyses are often performed on computers using particular software. Therefore, corpus linguistics is not a theory of language but a technique for gathering and analyzing data both statistically and qualitatively [46]. Language corpora typically consist of a significant number of representative samples taken from texts that cover a variety of linguistic domains and language variations [47]. Any language's Corpus might have a significant role. A text corpus is a substantial collection of texts or information for linguistic study [48]. As a result, important information offers lexicographers, grammarians, and other individuals interested in languages an excellent description of a language. Information on morphology, syntactic parsing, lexicon structure, semantics, pragmatics, and other language elements is obtained via Corpus analysis. The Corpus can be utilized for many different tasks, such as information retrieval, pattern recognition, speech recognition, machine translation, Word-to-vector analysis, vectorization, dictionary, text-to-speech, feature extraction, synthesis analysis, and speech-to-text recognition, text parsing, text tagging, machine learning, morphological analysis, cluster analysis, classification, and WordNet. Less research has been conducted on the Brahui language and its text corpus. Kanpur an organization was given the task of developing tools for language processing and machine translation from English to Indian languages[49]. The usage of general corpora in language studies has been and continues to be significant. The annotation feature of general corpora significantly increases its usefulness for theoretical linguistic analysis [50]. For under-resourced languages, building language resources like parallel corpora, language models, and linguistic descriptions is a major problem. Building language resources, such as parallel corpora, language models, and linguistic descriptions, is a major problem for languages with low resources. [51]. Since morphologically rich languages have such a huge variety of word forms and sentence structures, their computational complexity rises rapidly [52]. Thus, developing linguistic resources like corpora to support these languages is important. The resources ought to be of the highest standard and show diversity in terms of language. Given the difficulties of developing resources for languages with insufficient resources, this is not an easy task[53]. with word forms and syntax developing and analyzing the corpus requires strong efforts [54].This collection includes books, articles, newspapers, and digital

media materials written in the Brahui scripts. The analysis of the Brahui text corpus involved methods like DTM and TF IDF to identify aspects and patterns, within the corpus. Effective techniques, like DTM and TF IDF, are very best, for analyzing text collections especially when dealing with under-resourced languages[55]. This research has taken a significant step toward the development of the Brahui text corpus, its computational analysis, and the preservation of the Brahui language. When it comes to bridging this gap and focusing on the development and analysis of the Brahui text corpus, the lack of an appropriate Brahui text corpus and research work is a significant drawback of NLP. The Brahui text corpus was developed, and its linguistic variance was examined in this study. The newly developed text corpus is available on https://github.com/Naseerbajoi/the-Brahui-Corpus for the researchers belonging to the NLP community by just sending an official email.

2.2 Development of Brahui Text Corpus

For the development of the Text corpus, several methods are used to generate the text corpus. The texts have been compiled from the Brahui books and magazines. Tokenization of the Brahui text corpus is performed. The construction of the blowdown figure declares the process of building the Brahui text corpus, which starts with problem understanding and ends with the TF, IDF Technique.

Figure 1 Depicts the process of developing the Brahui text Corpus.

The steps involved in developing the Brahui text corpus are shown in Figure 1. First, information is taken from the Brahui Dataset. Continues by preprocessing the data. Word tokenization divides the text into words after that Vectorization is used to transform the data into representations. Subsequently, DTM and TF IDF approaches are used to extract significant features. Finally, word sequences are captured by using N-grams. These steps together play a role, in developing the Brahui text corpus.

2.4 Document Term Matrix (DTM)

The base of text mining applications is the Document Term Matrix (DTM). A set of documents is shown as a matrix, with rows denoting the documents and columns denoting distinct terms [56]. The mathematical formulas related to the development and management of DTMs are explained [57].

2.4.1. DTM Construction

Construction of Document Term Matrix (DTM)

Equation 1: $$|DTM(i,j)| = \left\{ \frac{n(i,j)}{|d_i|_{(TF-Norm)}\ TF(i,j) \times IDF(j)\ _{(TF-IDF)}} \right\}$$

where:

- DTM (i, j) represents the weight of term j in document i within the DTM matrix. n(i, j) represents the number of occurrences of term j in document i.
- |d_i| represents the total number of terms in document i.
- TF (i, j) represents the Term Frequency of term j in document i (n(i, j) / |d_i|).
- IDF(j) represents the Inverse Document Frequency of term j (refer to relevant literature for IDF calculation).

A Document Term Matrix (DTM) is a matrix with dimensions |D| x |V|, where |D| is the total number of documents in the collection and |V| is the total number of unique words across all documents. Each cell in the DTM, DTM (i, j), represents the weight of the term j in document i. [58]. Common weighting schemes include Term Frequency (TF), Normalized Term Frequency (TF-Norm), [59] and Term Frequency-Inverse Document Frequency (TF-IDF) [60].

2.4.2. DTM Operations:

The DTMs can be tested with several mathematical operations that analyze term distributions and document relationships. Document Similarity Based on term vectors in the DTM, cosine similarity is a commonly used metric to assess how similar two documents are to one another. Cosine Similarity (d1, d2) = (DTM (d1, :) * DTM (d2, :)) / ‖DTM (d1, :)‖ ‖DTM (d2, :)‖. where: d1 and d2 represent document indices. * Denotes element-wise multiplication. ‖ . ‖ denotes the L2 norm (Euclidean norm) of a vector. Term Frequency-Inverse Document Frequency (TF-IDF) Matrix: As mentioned earlier, the TF-IDF weighting scheme incorporates both TF and IDF to capture term frequency and document-level importance. The formula for calculating the TF-IDF matrix was provided previously [60]. Principal Component Analysis (PCA) is one technique that can be used to reduce the dimensionality of DTMs without missing important information. This could increase the effectiveness of computing for future studies [61]. Using the TF-IDF technique, the Corpus's characteristics and weight demonstrate the term significance [62]. Text corpus analysis is a vital area of the natural language process since numerous organizations concentrate on the text corpora of various languages for various purposes. There are many places where the Brahui language is spoken, read, and written. On social media platforms, the Brahui people express their ideas on numerous things, people, and issues, and several blogs are developed in the Brahui language. As a result, the Brahui text corpus is increasingly significant for numerous businesses, linguistics, and NLP research. Term Frequency is used to count the number of words in a document [63]. Not every phrase must be used in actual text corpora documents. The direction of all token occurrences is applied to the text corpus document, considered a multivariate sample. The technique of turning text corpus documents into numerical feature vectors is referred to as vectorization. The DTM of the Brahui text corpus is a matrix with two dimensions of N (14082) rows in the dataset and C (1) columns for text. The rows display the number of times each distinct term appears in each column of documents. In DTM, each row shows a text corpus's document. Numbers like 0, 1, 2, 3, ...and n are displayed on the documents. The DTM displays the frequency and availability of unique words in each cell. In contrast, the Matrix's rows are the Matrix's documents. N-gram is a tokenization technique that divides words according to the kind of token being used. Three different n-gram token types unigram, bigram, and trigram were used in this research. The three different n-gram kinds will be used to type each sentence. A unigram is a word breakdown with n = 1 or a single term in a review sentence. The n-word answer for the review sentence with n = 2 is bigram. While in the review sentence with n = 3, the trigram is an n-word solution [64]. The N-gram model is utilized to develop language models, analyze language features, and perform other computational linguistics tasks. While the n-grams identify the following objects of n items from the provided text Corpus. Moreover, the grams are text items. Tri-gram, bi-gram, and uni-gram N-grams are all possible. N-grams are shown as having one gram, two grams, and three grams in uni-grams, bi-grams, and tri-grams, respectively. As a result, the Brahui text's n-grams determine the order of Brahui words present in the Corpus. As an illustration, a Brahui sentence's (براہوئ بُولی دنیا نا متکونو بولی سے) (the Brahui language is one of the world's oldest language) Below figure 2 represents the uni-gram, the bi-gram, and the tri-gram of the text that the

Brahui language is one of the world's oldest language. This Brahui text portion is divided into the uni-, the bi-, and the tri-gram, as reflected in Figure 2.

2.5 The (TF-IDF)

Due to the massive rise in data, processing structured or semi-structured data in every organization is becoming more difficult [33]-[34]. The mathematical framework known as Term frequency-inverse document frequency, or TF-IDF, is important for text mining, machine learning, and information retrieval activities [67]. It is common practice in data mining and information retrieval to assess word significance inside documents. Among the basic concepts is Term Frequency (TF), which quantifies the frequency with which a term (t) appears in a text (d):

Equation No 2: $TF(t,d) = \frac{n(t,d)}{|d|}$

The phrase Term Frequency (TF) of a term, in a text is calculated using Equation 2. To calculate this the total number of words, in the document (|d|) is divided by the frequency of the term 't' within that document (n(t, d)). This standardization makes it simpler to compare term frequencies among papers of lengths.

Equation No 3: $IDF(t,D) = \log\left(\frac{|D|}{df(t)}\right)$

Equation 3 illustrates how the Inverse Document Frequency (IDF) of the word 't' is calculated across a collection of documents 'D'. The equation incorporates a function, log (|D| / df(t)) where df(t) represents the number of documents containing the term and |D| is the number of documents, in the collection. Words that are mentioned in documents receive IDF scores indicating their importance, within the collection. These techniques are utilized by TF IDF to found a weighting system.

Equation No 4: $TF - IDF(t,d,D) = TF(t,d) \times IDF(t,D)$

Equation 4 calculates the score of a term 't', in the document 'd' using TF IDF (Term Frequency Inverse Document Frequency). In collection 'D,' The result of multiplying the term's Term Frequency (TF) within the document by its Inverse Document Frequency (IDF) throughout the collection is this score. Terms that are common inside a text but uncommon throughout the corpus are identified by TF-IDF, which makes them highly indicative of the content of that particular document. Although there are alternative data processing techniques or algorithms, TF-IDF is the main subject of this research. A numerical statistic called TF-IDF can be used to determine the relevance of a keyword to a given document, or it can be used to provide the keywords necessary to identify or categorize a given document. For example, a blogger who has hundreds of contributors to his blog just obtained an intern whose primary responsibility is to produce fresh blog entries every day. It has been noted that interns frequently neglect to maintain tags, which results in a large number of blog articles being uncategorized. This is one of the best situations in which to use the TF-IDF algorithm, which enables bloggers to automatically identify their tags. Bloggers and interns will save a lot of time because they will not be concerned about tags [68], [64]. The Brahui text corpus's TF-IDF document terms are displayed in Fig. 5. Table 2 results of the top 6 keywords from all documents.

1 RESULTS AND DISCUSSION

Figure 2 illustrates the process of analyzing the same collection of words at varying degrees of complexity to identify patterns of language use. Bigrams provide often occurring word pairings, trigrams provide information on more intricate word combinations, and unigrams give fundamental word frequencies. This data may be used for sentiment analysis, machine translation, and text synthesis, among other language-processing activities. Consequently, The DTM development is utilized to determine the occurrence and difference of Brahui terms across various texts in a corpus. This demonstrates the characteristics and significance of the Brahui lexicons and language. Using the n-gram model where n = 3, the DTM for the Brahui text corpus is developed; as a result, the frequency of words corresponds to the documents present in the text corpus based on n-gram words. The extraction of 3 grams demonstrates the Brahui language's complexity. The use of compound terms in a number of the corpus documents is an essential aspect of the Brahui language text corpus. Uni-gram terms may be often

used. However, 3-gram words are not frequently used. The 3-gram keywords correlation to documents demonstrates the value of the Brahui text corpora for analysis and text mining. Due to the identification of distinct terms in the Brahui text corpus, DTM displays linguistic variation in Table 2 and Figure 3 as well as TF-IDF in Table 3 and Figure 4.

```
1-grams for specific words:
[('براہوی',), ('بولی',), ('دنیا',), ('نا',), ('متکونو',), ('بولی',), ('سے',)]
2-grams for specific words:
[('بولی' ,'براہوی'), ('دنیا' ,'بولی'), ('نا' ,'دنیا'), ('متکونو' ,'نا'), ('بولی' ,'متکونو'), (',بولی
سے'')]
3-grams for specific words:
[('دنیا' ,'بولی' ,'براہوی'), ('نا' ,'دنیا' ,'بولی'), ('متکونو' ,'نا' ,'دنیا'), ('بولی' ,'متکونو' ,'نا'),
('سے' ,'بولی' ,'متکونو')]
```

Figure 2 Uni-Gram Model, Bi-Gram Model, and Tri-Gram model

Table 2: The frequency of each Brahui word in the text corpus document is presented by the DTM

Frequency	Brahui Words	Frequency	Brahui Words
1554	براهوی	10376	نا
21	بولی	174	متکونو
0	دنیا	21	بولی
0	:	0	سے

The frequency with which particular Brahui terms occur in a text or set of texts is shown in Table 2. It shows that the most common term is "براهوی" is the most frequent, followed by "نا". Some words like "متکونو" are less common, while others like "دنیا" and "سے" don't appear at all in the analyzed text. Identifying the language's usage patterns and prioritizing language learning efforts can both benefit from this knowledge.

Figure 3 below displays the relative frequencies of several Brahui words within a certain context. The word "نا" (no/not) appears the most, followed closely by "براهوی" (Brahui) and "متکونو" (oldest). The word "بولی" (language) appears twice in the graph, likely representing two different occurrences or variations of the word with similar frequencies. The words "دنیا" (world) and "سے" (from) are the least frequent, hardly appearing at all. This implies that the text or dataset analyzed may have been centered on rejecting ("نا") aspects of Brahui language or culture, with minimal reference to the outside world or other sources.

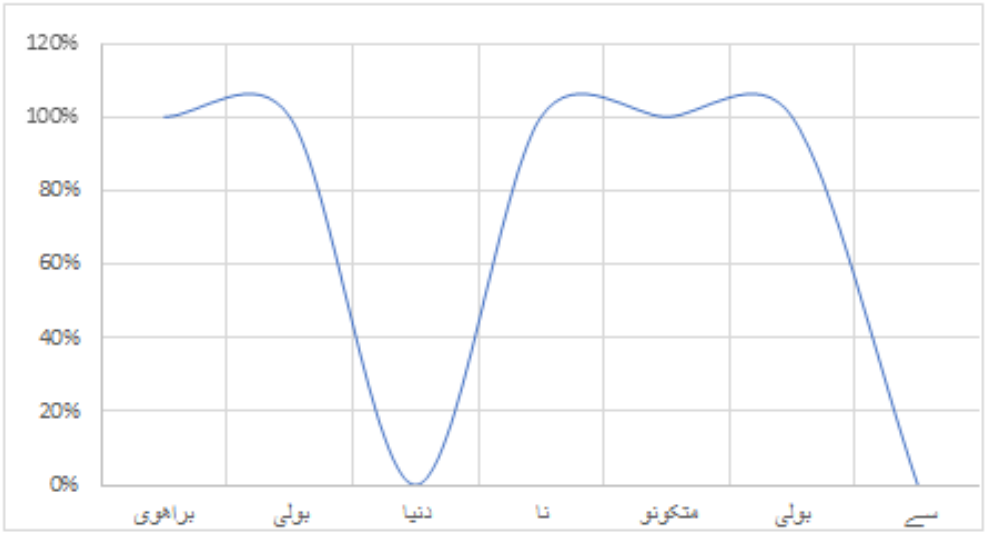

Figure 3: The frequency of each Brahui phrase in the text corpus document is presented by the DTM.

Table 3: TF-IDF of the Brahui words and their association with the documents

Keywords	TF-IDT	نا	0.988883
براھوی	0.148096	متکونو	0.016528
بولی	0.002001	بولی	0.002001
دنیا	0	سے	0

The significance of particular Brahui terms in a given text is ranked in Table 3 . "نا" is the most significant, occurring frequently, and exclusive to this text. while "براھوی" is also important but less different. "متکونو" and "بولی" are less significant, appearing less often and possibly being more common in other texts. "دنیا" and "سے" don't appear at all in the analyzed text. The words that are most important to the text's content and purpose are made easier for us by this analysis.

Figure 4 below shows the TF-IDF (Term Frequency-Inverse Document Frequency) scores for several Brahui terms are shown in this graph. The most significant word is "نا" (no/not), indicating its singularity and significance to the text under analysis. "براھوی" (Brahui) and "متکونو" (oldest) also carry some weight, suggesting a joining to the Brahui language or a specific idea inside it. "بولی" (language) appears double with dissimilar scores, representative potential variations in practice or context. "دنیا" (world) and "سے" (from/with) possess the lowest scores, indicating that the text under analysis does not include them or that they are not as relevant. In general, this analysis helps in identifying the most important terms for understanding the main ideas and content of the text.

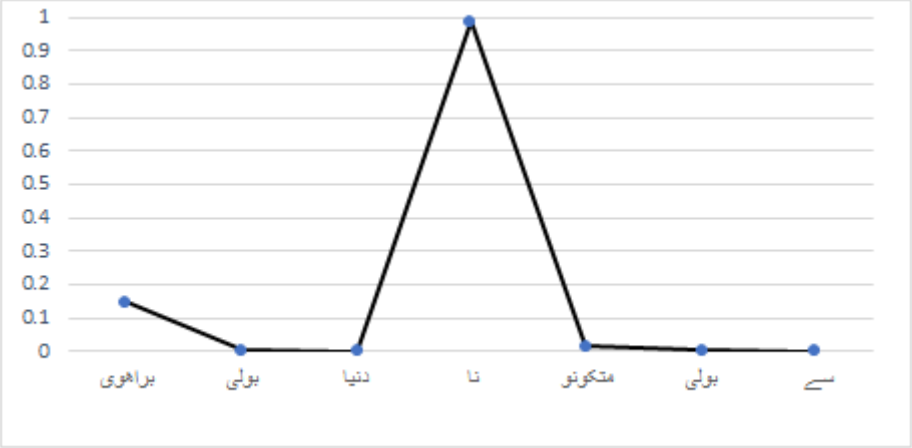

Figure 4: TF-IDF of the Brahui words and their association with the documents

2 DISCUSSION AND CHALLENGES

This study uses advanced computational techniques like TF-IDF and DTM analysis to develop a corpus of texts in the Brahui language and analyze its linguistic features. Over the past forty years, the field of language study has changed, moving from intuition-based methods to data-driven methodologies made possible by advancements in computing power and storage. The development of corpora, which serve as tools for language analysis, has caused this transformation. Improving study corpora's accuracy and reliability provides opportunities for advancement in languages like Brahui. Our study contributes to this expanding trend by building a corpus of Brahui language and using N gram analysis with TF IDF weighting. Using these methodologies, we have explored the statistical complexities in the Brahui language, such as analyzing the frequency distributions of lexicographic elements and their combinatorial arrangements. We were able to identify the most common terms as well as those that are more significant due to their particular context within the corpus by applying TF-IDF to analyze unigrams (single words), bigrams (two-word sequences), and trigrams (three-word sequences). Such quantitative data can prove to be extremely beneficial for a variety of natural language processing applications,

including machine translation that is specifically tailored for the Brahui language, sentiment analysis, and automatic text summarization. The developed corpus itself provides a basis for upcoming studies in computer language processing and Brahui linguistics. It is necessary to solve several constraints associated with the TF-IDF algorithm. The main drawback of TF-IDF is that it cannot distinguish between words even when their tenses are slightly different. For instance, it will treat "go" and "goes" as two distinct independent words; similarly, it will treat "play" and "playing," "mark" and "marking," "year" and "years," and so on. Because of this restriction, the TFIDF method occasionally produces unexpected results when it is applied [69]. Another drawback of TF-IDF is that it is only helpful up to the lexical level because it is unable to verify the semantics of the text in documents. Also, it is unable to verify word co-occurrences. there are numerous methods for improving accuracy and performance [70]. Similarly, the main limitation of TF-IDF is also faced during the implementation of TF-IDF on the Brahui Corpus in that the algorithm is unable to recognize the Brahui words even when their tense is slightly changed. like it will treat "بر" and "برنگ" as two independent words, and similarly, it will treat "هور" and "هورنگ" as distinct words. This restriction means that occasionally when the TF-IDF method is used, it produces some surprising results.

4. CONCLUSION AND FUTURE DIRECTION

There are more linguistic features available for the Brahui text corpus. This study has demonstrated that the corpus contains a rich variety of intricate terms, providing a comprehensive foundation for advanced linguistic analysis. The text corpus's characteristics are subjected to study. While initial analyses focused on a single plain text corpus, future research should consider a broader range of texts to capture more linguistic nuances. On a single plain text corpus, analysis was conducted. The extensiveness of the Brahui texts corpus comes from the abundance of complicated terms. The N-gram model was essential in preparing the corpus for Document Term Matrix (DTM) formation and filtering significant patterns within the corpus. The N-gram model prepares the Corpus for Document Term Matrix creation. The Brahui texts are categorized and recognized by the DTM, and their frequency in various texts is displayed using the N-gram model. The TF-IDF matrix provided superior results. This method highlighted key terms and their importance across different documents, enhancing the analysis of the corpus. The Word-to-document correspondence in a text corpus is important, as demonstrated by the TF-IDF. The significance of the Brahui text corpus is useful for information retrieval which is confirmed by applying DTM and TF-IDF. This initial research provided insightful information but more research is required for the advancement of the Brahui language such as Advanced methods including Word2Vec, sentiment analysis, cluster analysis, word similarity analysis, and topic modeling are included in future work. The development and analysis of the Brahui text corpus are the main objectives of this initial research. These additional methods will offer a deeper understanding of the Brahui language. These additional analyses will significantly enrich NLP and computational linguistics studies, enhancing the understanding and application of the Brahui text corpus. Future research in these areas will build on the foundational work presented here, contributing to the broader field of computational linguistics and offering valuable insights for NLP applications. Future NLP and computational linguistics studies will benefit from the research's contribution to the Brahui text corpora.

5. REFERENCE

[1] B. Rooman, Anwar, Prof., "Urdu ki Lisani wa saqafati rawabit, in Pakistan min Urdu," *Islamabad, National Language Authority, 2006, p.213.*, vol. (Vol. II), 2006.

[2] L. Pagani, V. Colonna, C. Tyler-Smith, and Q. Ayub, "An ethnolinguistic and genetic perspective on the origins of the Dravidian-speaking Brahui in Pakistan," *Man India*, vol. 97, no. 1, pp. 267–277, 2017.

[3] A. A. Sanjrani, "Multilingual OCR systems for the regional languages in Balochistan," *Indian J Sci Technol*, vol. 13, no. 21, pp. 2157–2168, Jun. 2020, doi: 10.17485/IJST/v13i21.2.

[4] M. A. Dootio and A. I. Wagan, "Development of Sindhi text corpus," *Journal of King Saud University - Computer and Information Sciences*, vol. 33, no. 4, pp. 468–475, 2021, doi: 10.1016/j.jksuci.2019.02.002.

[5] M. Turchi, T. De Bie, and N. Cristianini, "Learning performance of a machine translation system," no. June, pp. 35–43, 2008, doi: 10.3115/1626394.1626399.

[6] D. Padhya and J. Sheth, "A Review of Machine Translation Systems for Indian Languages and Their Approaches," *Advances in Intelligent Systems and Computing*, vol. 841, pp. 103–110, 2019, doi: 10.1007/978-981-13-2285-3_13.

[7] J. Mirzakhalov *et al.*, "A Large-Scale Study of Machine Translation in the Turkic Languages," *EMNLP 2021 - 2021 Conference on Empirical Methods in Natural Language Processing, Proceedings*, pp. 5876–5890, 2021, doi: 10.18653/v1/2021.emnlp-main.475.

[8] B. Ji, Z. Zhang, X. Duan, M. Zhang, B. Chen, and W. Luo, "Cross-lingual pre-training based transfer for zero-shot neural machine translation," *AAAI 2020 - 34th AAAI Conference on Artificial Intelligence*, pp. 115–122, 2020, doi: 10.1609/aaai.v34i01.5341.

[9] J. J. Zhang and C. Q. Zong, "Neural machine translation: Challenges, progress and future," *Sci China Technol Sci*, vol. 63, no. 10, pp. 2028–2050, 2020, doi: 10.1007/s11431-020-1632-x.

[10] M. Ali and A. Imdad, "Sentiment Summerization and Analysis of Sindhi Text," *International Journal of Advanced Computer Science and Applications*, vol. 8, no. 10, pp. 296–300, 2017, doi: 10.14569/ijacsa.2017.081038.

[11] N. Ahmed, M. A. Khouro, A. Khan, M. Dawood, M. A. Dootio, and N. U. Jan, "Student textual feedback sentiment analysis using machine learning techniques to improve the quality of education," *Pakistan Journal of Engineering, Technology & Science*, vol. 11, no. 2, pp. 32–40, Dec. 2023, doi: 10.22555/pjets.v11i2.1039.

[12] M. U. Rahman, "Linguistics and Literature Review (LLR)," vol. 6510, no. 1, 2015.

[13] S. Hussain, "Resources for Urdu Language Processing.," *Ijcnlp*, no. 2003, pp. 99–100, 2008.

[14] N. Ahmed, "A review of existing Machine Translation Approaches, their Challenges and Evaluation Metrics," *Pakistan Journal of Engineering, Technology & Science*, vol. 11, no. 1, pp. 29–44, Dec. 2023, doi: 10.22555/pjets.v11i1.1002.

[15] M. A. Dootio and A. I. Wagan, "Unicode-8 based linguistics data set of annotated Sindhi text," *Data Brief*, vol. 19, pp. 1504–1514, 2018, doi: 10.1016/j.dib.2018.05.062.

[16] M. A. Dootio and A. I. Wagan, "Syntactic parsing and supervised analysis of Sindhi text," *Journal of King Saud University - Computer and Information Sciences*, vol. 31, no. 1, pp. 105–112, 2019, doi: 10.1016/j.jksuci.2017.10.004.

[17] R. Tsarfaty, D. Seddah, S. Kübler, and J. Nivre, "Parsing morphologically rich languages: Introduction to the special issue," *Computational Linguistics*, vol. 39, no. 1, pp. 15–22, 2013, doi: 10.1162/COLI_a_00133.

[18] N. Ahmed, M. Gul Bizanjo, A. Khan, S. Gull, and S. khaliq, "Analysis of Textual Feedback of Students for Course Evaluation in Universities Through Machine Learning Algorithms," 2024.

[19] Korean Society of Speech Sciences, International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques. Oriental Chapter., and Institute of Electrical and Electronics Engineers, *20th 2017 Conference of The Oriental Chater of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA) : 1-3rd of November 2017, SNU Hoam Faculty House, Seoul, Korea*.

[20] Korean Society of Speech Sciences, International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques. Oriental Chapter., and Institute of Electrical and Electronics Engineers, *20th 2017 Conference of The Oriental Chater of International Committee for Coordination and Standardization of Speech Databases and Assessment Techniques (O-COCOSDA) : 1-3rd of November 2017, SNU Hoam Faculty House, Seoul, Korea*.

[21] T. Arkhangelskiy, "Corpora of social media in minority Uralic languages."

[22] R. Agerri, X. Gómez Guinovart, G. Rigau, M. Anxo, and S. Portela, "Developing New Linguistic Resources and Tools for the Galician Language." [Online]. Available: http://clarin.eu

[23] A. Ghafoor *et al.*, "The Impact of Translating Resource-Rich Datasets to Low-Resource Languages through Multi-Lingual Text Processing," *IEEE Access*, vol. 9, pp. 124478–124490, 2021, doi: 10.1109/ACCESS.2021.3110285.

[24] I. Abdulmumin *et al.*, "Hausa Visual Genome: A Dataset for Multi-Modal English to Hausa Machine Translation," 2022. [Online]. Available: https://github.com/abumafrim/

[25] I. Haq *et al.*, "The Pashto Corpus and Machine Learning Model for Automatic POS Tagging," 2023, doi: 10.21203/rs.3.rs-2712906/v1.

[26] I. Antonellis, E. Gallopoulos, I. Antonellis, and E. Gallopoulos, "Exploring term-document matrices from matrix models in text mining Exploring term-document matrices from matrix models in text mining *," 2006. [Online]. Available: www.ceid.upatras.gr

[27] R. K. Kevat and Dr. S. D. Degadwala, "Developing Gujarati Article Summarization Utilizing Improved Page-Rank System," *International Journal of Scientific Research in Computer Science, Engineering and Information Technology*, vol. 10, no. 2, pp. 293–299, Mar. 2024, doi: 10.32628/cseit2410222.

[28] IEEE Staff, *2013 IEEE International Advance Computing Conference*. IEEE, 2013.

[29] E. S. Tellez, S. Miranda-Jiménez, M. Graff, D. Moctezuma, O. S. Siordia, and E. A. Villaseñor, "A case study of Spanish text transformations for twitter sentiment analysis," *Expert Syst Appl*, vol. 81, pp. 457–471, Sep. 2017, doi: 10.1016/j.eswa.2017.03.071.

[30] R. Kaur and V. Bhardwaj, "Gurmukhi Text Emotion Classification System using TF-IDF and N-gram Feature Set Reduced using APSO," *International Journal on Emerging Technologies*, vol. 10, no. 3, pp. 352–362, 2019, [Online]. Available: www.researchtrend.net

[31] P. Kherwa and P. Bansal, "Topic Modeling: A Comprehensive Review," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 24, pp. 1–16, 2020, doi: 10.4108/eai.13-7-2018.159623.

[32] L. Flowerdew, "Applying corpus linguistics to pedagogy," *International Journal of Corpus Linguistics*, vol. 14, no. 3, pp. 393–417, Sep. 2009, doi: 10.1075/ijcl.14.3.05flo.

[33] Ö. F. Kaya, "Using corpora for language teaching and assessment in L2 writing: A narrative review," *Focus on ELT Journal*, pp. 46–62, 2022, doi: 10.14744/felt.2022.4.3.4.

[34] N. Vyatkina and A. Boulton, "Corpora in language teaching and learning Corpora in language learning and teaching," *Language Learning & Technology*, vol. 21, no. 3, pp. 1–8, 2017, doi: 10.1017/S0261444817000167ï.

[35] Q. Li, S. Li, S. Zhang, J. Hu, and J. Hu, "A review of text corpus-based tourism big data mining," Aug. 01, 2019, *MDPI AG*. doi: 10.3390/app9163300.

[36] A. Kilgarriff *et al.*, "The sketch engine: Ten years on," *Lexicography*, vol. 1, no. 1, pp. 7–36, 2014, doi: 10.1007/s40607-014-0009-9.

[37] S. S. Agrawal, Abhimanue, S. Bansal, and M. Mahajan, "Statistical analysis of multilingual text corpus and development of language models," *Proceedings of the 9th International Conference on Language Resources and Evaluation, LREC 2014*, vol. 1, pp. 2436–2440, 2014.

[38] F. Baseer, A. Habib, and J. Ashraf, "Romanized Urdu Corpus development (RUCD) model: Edit-distance based most frequent unique unigram extraction approach using real-time interactive dataset," *2016 6th International Conference on Innovative Computing Technology, INTECH 2016*, pp. 513–518, 2017, doi: 10.1109/INTECH.2016.7845117.

[39] K. Chapman, J. Bernhard, and D. Klakow, "CoLi at UdS at SemEval-2020 Task 12: Offensive Tweet Detection with Ensembling," Online, 2020. [Online]. Available: https://github.com/jb-1811/NLP_with_NN_ws1920_OffensEval2020

[40] Z. Zainol, M. T. H. Jaymes, and P. N. E. Nohuddin, "VisualUrText: A Text Analytics Tool for Unstructured Textual Data," in *Journal of Physics: Conference Series*, Institute of Physics Publishing, Jun. 2018. doi: 10.1088/1742-6596/1018/1/012011.

[41] P. Sheridan and M. Onsjö, "The hypergeometric test performs comparably to TF-IDF on standard text analysis tasks," Feb. 2020, [Online]. Available: http://arxiv.org/abs/2002.11844

[42] M. Alkaff, A. Rizky Baskara, and Y. Hendro Wicaksono, "Sentiment Analysis of Indonesian Movie Trailer on YouTube Using Delta TF-IDF and SVM," in *2020 5th International Conference on Informatics and Computing, ICIC 2020*, Institute of Electrical and Electronics Engineers Inc., Nov. 2020. doi: 10.1109/ICIC50835.2020.9288579.

[43] S. Qaiser, U. Utara, M. Sintok, M. Kedah, A. Ramsha, and T. Analytics, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents Text Mining," 2018.

[44] N. S. Dash and B. B. Chaudhuri, "The Process of Designing a Multidisciplinary Monolingual Sample Corpus."

[45] S. Avasthi, R. Chauhan, and D. P. Acharjya, "Extracting information and inferences from a large text corpus," *International Journal of Information Technology (Singapore)*, vol. 15, no. 1, pp. 435–445, Jan. 2023, doi: 10.1007/s41870-022-01123-4.

[46] B. F. Klimova, "Using Corpus Linguistics in the Development of Writing," *Procedia Soc Behav Sci*, vol. 141, pp. 124–128, Aug. 2014, doi: 10.1016/j.sbspro.2014.05.023.

[47] N. Sekhar and B. B. Chaudhuri, "WHY DO WE NEED TO DEVELOP CORPORA IN INDIAN LANGUAGES ?"

[48] A. Introduction, "Kennedy, G., 2014. An Introduction to Corpus Lin," p. 2014, 2014.

[49] N. Sekhar Dash, "LANGUAGE CORPORA: PRESENT INDIAN NEED."

[50] T. Mcenery and A. Wilson, "Teaching and language corpora (TALC)," 1997.

[51] J. R. Finkel and C. D. Manning, "Nested Named Entity Recognition," 2009. [Online]. Available: http://www.cs.rhul.ac.uk/home/alexc/RHUL/Downloads.html

[52] R. Tsarfaty, D. Seddah, and J. Nivre, "Parsing Morphologically Rich Languages: Introduction to the Special Issue," 2013. [Online]. Available: http://direct.mit.edu/coli/article-pdf/39/1/15/1798976/coli_a_00133.pdf

[53] S. Ma and C. Zhang, "LNAI 8801 - Automatic Collection of the Parallel Corpus with Little Prior Knowledge," 2014. [Online]. Available: http://www.swd.gov.hk/vs/english/police.html

[54] M. Jordan, D. M. Blei, A. Y. Ng, and J. B. Edu, "Latent Dirichlet Allocation Latent Dirichlet Allocation Michael I. Jordan," 2003. [Online]. Available: https://www.researchgate.net/publication/221620547

[55] "54".

[56] D. Kučera and M. R. Mehl, "Beyond English: Considering Language and Culture in Psychological Text Analysis," *Front Psychol*, vol. 13, Mar. 2022, doi: 10.3389/fpsyg.2022.819543.

[57] D. Kučera and M. R. Mehl, "Beyond English: Considering Language and Culture in Psychological Text Analysis," *Front Psychol*, vol. 13, Mar. 2022, doi: 10.3389/fpsyg.2022.819543.

[58] K. Spärck Jones, "A statistical interpretation of term specificity and its application in retrieval," *Journal of Documentation*, vol. 60, no. 5, pp. 493–502, Oct. 2004, doi: 10.1108/00220410410560573.

[59] I. Alsmadi *et al.*, "Adversarial Machine Learning in Text Processing: A Literature Survey," *IEEE Access*, vol. 10, pp. 17043–17077, 2022, doi: 10.1109/ACCESS.2022.3146405.

[60] S. E. Robertson and K. Sparck Jones, "Relevance Weighting of Search Terms."

[61] I. T. Joliffe and B. Morgan, "Principal component analysis and exploratory factor analysis," *Stat Methods Med Res*, vol. 1, no. 1, pp. 69–95, 1992, doi: 10.1177/096228029200100105.

[62] C. Bosco, V. Patti, and A. Bolioli, "Developing corpora for sentiment analysis: The case of irony and senti-TUT," *IEEE Intell Syst*, vol. 28, no. 2, pp. 55–63, 2013, doi: 10.1109/MIS.2013.28.

[63] "Sentiment Analysis and Topic Modeling on News Headlines by Vijay Yadav Thesis Supervisor Prof . Dr . Subarna Shakya A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science in Computer System and Knowledge Engineer," 2022.

[64] T. Widiyaningtyas, I. Ari, E. Zaeni, and R. Al Farisi, "Sentiment Analysis Of Hotel Review Using N-Gram And Naive Bayes Methods."

[65] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int J Comput Appl*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.

[66] B. Trstenjak, S. Mikac, and D. Donko, "KNN with TF-IDF based framework for text categorization," in *Procedia Engineering*, Elsevier Ltd, 2014, pp. 1356–1364. doi: 10.1016/j.proeng.2014.03.129.

[67] Institute of Electrical and Electronics Engineers. DMI College of Engineering Student Branch, Institute of Electrical and Electronics Engineers. Madras Section., and Institute of Electrical and Electronics Engineers, *International Conference on Electrical, Electronics, and Optimization Techniques (ICEEOT) - 2016 : 3rd-5th, March 2016*.

[68] J. Gautam and E. Kumar, "An Integrated and Improved Approach to Terms Weighting in Text Classification," 2013. [Online]. Available: www.IJCSI.org

[69] J. Ramos, "Using TF-IDF to Determine Word Relevance in Document Queries."

[70] S. Qaiser and R. Ali, "Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents," *Int J Comput Appl*, vol. 181, no. 1, pp. 25–29, Jul. 2018, doi: 10.5120/ijca2018917395.