International Journal of Engineering Trends and Technology
ISSN: 2231–5381 / https://doi.org/10.14445/22315381/IJETT-V71I12P205

Volume 71 Issue 12, 40-49, December 2023
© 2023 Seventh Sense Research Group®

Original Article

Enhancing Natural Language Processing in Somali Text Classification: A Comprehensive Framework for Stop Word Removal

Abdullahi Ahmed Abdirahman[1], Abdirahman Osman Hashi[1], Ubaid Mohamed Dahir[1], Mohamed Abdirahman Elmi[1], Octavio Ernest Romo Rodriguez[2]

[1]*Faculty of Computing, SIMAD University, Mogadishu-Somalia.*
[2]*Department of Computer Science, Faculty of Informatics, Istanbul Teknik Universitesi, Istanbul, Turkey*.

[1]*Corresponding Author : aaayare@simad.edu.so*

Received: 31 July 2023 Revised: 12 October 2023 Accepted: 24 October 2023 Published: 06 December 2023

Abstract - *Text classification is a prominent field of study in information retrieval and natural language processing, where a crucial component is the utilization of a stop word list. This list helps identify frequently occurring words that have little relevance in classification and are consequently removed during pre-processing. Although various stopword lists have been devised for the English language, a standardized stopword list specifically tailored for Somali text classification is yet to be established. This research presents a comprehensive framework for stop word removal in the context of the Somali language, aiming to enhance the effectiveness of various Natural Language Processing (NLP) tasks. The proposed methodology encompasses several essential steps, including noise identification, noise removal, character normalization, data masking, tokenization, POS tagging, and lemmatization. By analysing a substantial dataset containing 79,741,231 tokens and 71,871,585 words, the framework demonstrates its capability to identify and eliminate stop words, thereby reducing vector space and improving the performance of NLP algorithms. The research highlights the unique linguistic features of Somali, such as contextual variations and morphological complexities. It discusses the potential applications of the developed stop word list in sentiment analysis, information retrieval, and document classification. This work contributes valuable insights to the field of language technology, particularly in underrepresented languages, and paves the way for further advancements in NLP models tailored to diverse linguistic contexts.*

Keywords - *Somali language, Stopword removal, Natural Language Processing, Stopword list, Ontology.*

1. Introduction

With the increasing prevalence of Natural Language Processing (NLP) applications in real-world scenarios, extensive research is being conducted to improve the effectiveness of these tasks. One crucial aspect of NLP tasks, such as information retrieval, text summarization, and context-embedding, involves identifying and removing insignificant tokens and words, known as stop words, which have minimal impact on the context's meaning [1]. Thus, developing an automated method for identifying and eliminating stop words is highly desirable to enhance the quality of NLP-based processes. Traditionally, words that frequently appear in documents but offer little assistance in classification are referred to as stop words, such as "and," "the," and "of" in English documents. These words are ubiquitous in almost every sentence and constitute a significant portion of the overall text size. Some scholars indicate that approximately 50% of words in a typical English passage are among a set of about 135 common words, considered noise words, and recommended for removal during text analysis pre-processing. Eliminating stop words can significantly reduce the text's feature space, accelerate calculations, and enhance the accuracy of text classification tasks [2].

The ongoing global research extensively focuses on text classification, document clustering, and similar document analysis tasks, supporting endeavours such as web intelligence, web mining, and web search engine design. A crucial element of machine learning tasks involving document processing is the compilation of "stop words" or "stop list." These stopwords constitute a significant proportion of the data in documents used for information retrieval tasks but offer limited value to researchers [3]. Therefore, developing an automated method to identify and eliminate such stopwords from datasets prior to processing is highly desirable. Stopwords were first introduced by [4] and refer to the most frequently occurring words in a document, which typically carry minimal information and are not essential. For instance, in the English language, words like "a," "about," "above," "after," and many others are considered stopwords. A collection of stopwords is referred to as a "stopword list" or "stopword corpus."

Meanwhile, the removal of stopwords not only reduces the vector space and enhances performance by increasing execution speed, calculations, and overall accuracy. For instance, when using a search engine with a query like "how to develop an Android app," the engine may encounter numerous web pages containing common words like "how" and "to," making it challenging to find relevant pages about developing Android apps. By disregarding these frequent terms in the pre-processing phase of the text classification process, the search engine can focus on retrieving pages that contain keywords like "develop," "Android," and "app," yielding more relevant results. Manual removal of stopwords is an alternative, but it can be time-consuming and proportional to the corpus size [5].

With its rich linguistic heritage, Somali presents an intricate web of linguistic structures, including a substantial collection of stop words. As the Somali language evolves with varying contextual usages, the list of stop words is subject to expansion and revision. Identifying these stop words accurately is crucial in pre-processing textual data for NLP tasks, as retaining them may introduce noise and hinder extracting meaningful information [6]. Conversely, a well-curated list of Somali stop words can lead to improved document summarization, topic modelling, and sentiment classification.

The morphology and syntax of the Somali language contribute to the distinctive nature of its stop words. It is not uncommon for one word to have multiple forms depending on its role within a sentence, further complicating the process of stop word identification. Therefore, an in-depth analysis of Somali stop words and their contextual variations is essential to ensure the efficacy of NLP models operating on Somali text [1].

Moreover, given the widespread usage of Somali in various domains such as social media, news outlets, and governmental communications, the implications of stop words on information retrieval systems cannot be overlooked. A comprehensive understanding of Somali stop words enables the creation of efficient search algorithms, information filtering mechanisms, and question-answering systems tailored to the language's unique linguistic characteristics. The paucity of research focusing explicitly on Somali stop words underlines the necessity for this study. By investigating the intricacies of stop words in the Somali language, we aim to contribute valuable insights to the broader field of NLP, facilitating advancements in machine learning algorithms and language processing frameworks tailored to underrepresented languages like Somali [4]. In Somali, a Cushitic language spoken predominantly in the Horn of Africa, the presence of stop words poses unique challenges to developing robust NLP systems. For instance, lots of stop word lists have been developed for the English language in the past, which are usually based on frequency statistics of a large corpus.

The English stop word lists available online are good examples. However, no commonly accepted stop word list has been constructed for the Somali language. Most current research on Somali information retrieval uses manual or simple statistical stop word lists, some of which are picked up based on the author's experiences consuming a lot of time. The contents of these stop lists vary a lot from each other [7].

This study explores the significance of Somali stop words and their influence on various language-processing tasks, such as text analysis, information retrieval, and sentiment analysis, among others. By understanding the effects of these stop words on different NLP applications, we aim to deepen our understanding of Somali linguistics and bridge the gap between state-of-the-art NLP technologies and the rich linguistic heritage of the Somali-speaking community. Ultimately, this research contributes valuable insights to the broader field of NLP and facilitates advancements in machine learning algorithms and language processing frameworks tailored to underrepresented languages like Somali.

The rest of this document is structured in the following manner: In Section 2, the related work will be presented. Section 3 will elaborate on the proposed methodology, whereas Section 4 will showcase the experimental dataset and the evaluation approach. Additionally, it will analyze and discuss the results. Ultimately, Section 5 will outline our conclusions.

2. Related Work

Stop words, commonly encountered in various languages, have been a critical research topic in Natural Language Processing (NLP) and text analysis. They are words that occur frequently but carry little or no substantive meaning within a given context. In the context of English, several studies have investigated the importance of stop words and their impact on language processing tasks. For instance, author [8] explored the effects of stop words on document clustering and classification, finding that removing stop words can significantly improve the accuracy of classification algorithms. Similarly, in the domain of Arabic language processing, the author [9] studied the effects of stop words on information retrieval tasks. They highlighted the importance of identifying and removing Arabic stop words to enhance the performance of search engines and information retrieval systems.

Turning our attention to Somali language processing, research on stop words is relatively limited compared to major languages like English and Arabic. However, the significance of stop words in Somali cannot be ignored, considering their unique linguistic features and evolving contextual usages. As the Somali language evolves, the list of stop words is subject to expansion and revision, adding to the complexity of stop word identification [10].

A study by [11] examined the role of stop words in sentiment analysis for Somali social media data. The researchers developed a tailored stopword list for Somali and demonstrated its efficacy in improving sentiment classification accuracy. They highlighted the necessity of a well-curated stop word list for Somali language processing to achieve more accurate and contextually relevant results.

As mentioned, the Somali language exhibits unique morphological and syntactical characteristics, and the identification and removal of stop words become more challenging due to contextual variations and different word forms. However, the benefits of developing a comprehensive and accurate stop word list for Somali are substantial, as it would facilitate improved search algorithms, information retrieval systems, and text classification methods tailored to the language's linguistic nuances [12].

One potential approach to building a Somali stop word list involves leveraging linguistic experts and native speakers to curate a preliminary set of stop words manually. Statistical analyses can complement this on a representative corpus to identify high-frequency words that exhibit characteristics of typical stop words. A key consideration in constructing the stop word list is the trade-off between sensitivity and specificity. A conservative approach may lead to the retention of some relevant words, while an overly aggressive approach may eliminate informative content. Striking the right balance is crucial to ensure optimal performance in various NLP tasks [13].

To validate the effectiveness of the stop word list, extensive experiments should be conducted on diverse Somali text datasets, covering domains such as social media, news articles, and official documents. Performance metrics such as accuracy, precision, recall, and F1 score can be used to evaluate the impact of stop word removal on different NLP applications. Furthermore, the developed Somali stop word list can be made available to the research community as a valuable resource, fostering collaboration and further advancements in Somali language processing [14]. Such contributions are essential for nurturing the growth of NLP technologies in underrepresented languages and promoting language preservation and diversity.

In general, the taxonomy of stop words can be classified as follows:

2.1. Outdated Stop Lists

Outdated stop lists in the field of NLP refer to stop word lists created and used in earlier research but may no longer be effective or relevant due to changes in language usage, linguistic evolution, or the emergence of new domains and contexts. These outdated stop lists can hinder the performance of NLP tasks and lead to suboptimal results, emphasizing the importance of maintaining and updating stop word lists to reflect current language patterns and requirements [15].

Several studies have highlighted the potential issues with using outdated stop lists. For instance, the author [16] investigated the impact of outdated stop words on document clustering tasks in the context of English language processing. They found that the presence of irrelevant stop words in an outdated stop list adversely affected the clustering quality, leading to incorrect groupings and reduced interpretability of the results. Moreover, outdated stop lists can result in excluding meaningful content from text analysis. A study by [17] revealed that using an outdated stop list in information retrieval tasks may discard crucial terms relevant to specific domains or topics, undermining search results' accuracy and comprehensiveness.

The dynamic nature of language necessitates the periodic review and updating of stop lists to adapt to linguistic changes and evolving contexts. As languages evolve over time, certain words previously considered stop words may gain semantic importance, while new stop words may emerge due to shifts in language usage and popular trends. To address the challenge of outdated stop lists, researchers have proposed various strategies [18]. Author [19] proposed a data-driven approach to update stop lists based on corpus analysis automatically. By considering term frequency and document frequency, they identified and incorporated relevant words while discarding outdated or less informative ones, resulting in improved performance in sentiment analysis and text classification tasks.

Another approach combines linguistic expertise with statistical analyses to iteratively refine and update stop lists. The author [20] employed a hybrid approach that integrated linguistic insights with data-driven methods to develop an up-to-date stop list for Bengali language processing. Their results demonstrated the efficacy of this approach in enhancing the accuracy of topic modelling and text classification. Maintaining up-to-date stop lists is crucial in ensuring the relevance and effectiveness of NLP applications across different languages and domains.

Researchers can optimize NLP models' performance, adapt to linguistic changes, and reflect the nuances of language usage in real-world contexts by regularly revaluating and updating stop word lists. Consequently, such efforts contribute to the advancement of NLP technologies and facilitate more accurate and contextually relevant language processing in the ever-changing landscape of natural language [21].

2.2. Stop List for Non-English Text Retrieval

Stop lists for non-English text retrieval in the field of NLP refer to curated sets of words removed from text documents before processing to improve the efficiency and accuracy of information retrieval tasks in languages other than English. These stop lists aim to eliminate frequently occurring but semantically insignificant words, known as stop words, which can otherwise hinder the retrieval process and lead to suboptimal search results. Various studies have investigated the development and impact of stop lists for non-English languages. For example, in the context of Chinese text retrieval, the author [22] conducted a comparative analysis of different stop lists and their influence on search engine performance. They found that using an appropriate stop list significantly improved the search engine's precision and recall, leading to more relevant search results for Chinese users.

Similarly, in the domain of Arabic text retrieval, the author [23] explored the effects of stopword removal on search engine effectiveness. Their study demonstrated that carefully designed Arabic stop lists effectively reduced query processing time and improved the retrieval performance in Arabic information retrieval systems. For languages with rich morphological structures, such as German, stop lists have also proven essential for efficient information retrieval. Using the right set of stop words in German search queries was shown to enhance the accuracy of results in the study [24].

Creating stop lists for non-English languages involves considering each language's linguistic characteristics and specific requirements. This process often involves leveraging linguistic expertise, statistical analysis of language corpora, and domain-specific knowledge. One challenge in developing stop lists for non-English languages lies in the diversity of language structures and dialects. For instance, in the case of Arabic, different dialects may have varying stop word patterns, making it necessary to tailor stop lists to specific language varieties and use cases [9].

Moreover, as languages evolve over time, stop lists may need periodic updates to account for shifts in language usage and the emergence of new terms. The development of dynamic and adaptive stop lists has been proposed as a potential solution to address the challenges of maintaining relevant stop words for non-English languages [7]. Stop lists play a crucial role in non-English text retrieval within the field of NLP. They contribute to improving the precision and efficiency of information retrieval tasks by removing non-informative stop words from text documents. Creating effective stop lists requires a deep understanding of the linguistic characteristics of each language and often involves a combination of linguistic expertise and data-driven approaches.

2.3. Domain-Specific Stop Words

Domain-specific stopwords in the field of NLP refer to sets of words removed from text documents based on their relevance to a specific domain or topic. Unlike general stop lists, which contain common words irrelevant to any specific context, domain-specific stopwords are tailored to the particular subject matter being analysed. These customized stopwords are essential for improving the accuracy and effectiveness of NLP tasks in specialized domains, such as medical, legal, scientific, or technical texts [25].

Numerous studies have explored the importance of domain-specific stopwords in various NLP applications. For instance, in medical text processing, researchers have identified specific medical terminologies and jargon that are frequent but lack meaningful contributions to the analysis. A study by [26] demonstrated that removing medical domain-specific stopwords significantly improved the accuracy of medical information retrieval systems, leading to more precise search results. Similarly, in legal text analysis, domain-specific stopwords are vital for ensuring the focus on legal jargon and specific terminologies relevant to legal documents. In their work, authors [12] presented an approach that leveraged domain-specific stopwords to enhance the performance of legal information retrieval systems, enabling more accurate and targeted searches within legal databases.

In technical domains, such as computer science and engineering, including technical terms, such as domain-specific stopwords, can significantly impact the results of text classification tasks. Incorporating domain-specific stopwords in technical texts has been shown to enhance the accuracy of document categorization, as demonstrated by [27] in their research. Developing domain-specific stopwords involves a combination of linguistic expertise and domain knowledge. Linguists and domain experts collaborate to identify frequent terms in the specific domain that carry little semantic value for the intended analysis. Additionally, statistical methods can be applied to identify high-frequency domain-specific terms and incorporate them into the stop list.

One challenge in working with domain-specific stopwords lies in the dynamic nature of language within specialized fields. As new terminologies emerge and language usage evolves, domain-specific stopwords may require regular updates to maintain their relevance and effectiveness. Hence, the creation of adaptive stop lists that can be periodically revised has been proposed to address this issue [5]. Domain-specific stopwords are essential in the field of NLP to tailor text processing and analysis to specific subject domains. They contribute to improved precision and contextually relevant results in specialized areas. By combining linguistic expertise and domain knowledge, researchers can develop effective domain-specific stop lists that enhance the performance of NLP tasks in various domains [8].

2.4. Formal Language Text Mining

Formal Language Text Mining in the field of NLP refers to the process of extracting meaningful information and patterns from text written in formal languages, such as programming languages, mathematical expressions, or domain-specific languages. This specialized area of NLP focuses on developing algorithms and techniques to analyse and process formal texts, enabling applications in code analysis, software engineering, scientific research, and other technical domains. Several academic articles have investigated the significance of Formal Language Text Mining and its practical applications.

For instance, the author [27] proposed a methodology for mining source code comments to extract valuable information about software documentation and developer intentions. Their research demonstrated how Formal Language Text Mining techniques can improve software understanding and maintenance. In the context of mathematical expressions, the author [28] presented an approach for mining mathematical texts to extract formulae and equations. Their work highlighted the importance of specialized algorithms to handle complex mathematical notations effectively.

Formal Language Text Mining has also found applications in the analysis of domain-specific languages. A study by [29] developed a framework for mining text from Urdu, a complex and less studied language. The research demonstrated the potential of Formal Language Text Mining to uncover valuable insights from lesser-known languages. Furthermore, in the field of bioinformatics, Formal Language Text Mining has been utilized to analyze and extract information from biological texts. A study by [11] explored the application of Formal Language Text Mining techniques to identify biological concepts and relationships from scientific literature.

Formal Language Text Mining challenges stem from the unique syntax and semantics of formal languages, which often differ significantly from natural languages. Developing specialized algorithms and data structures that can handle the intricacies of formal language texts is essential. Techniques used in Formal Language Text Mining include lexing and parsing to analyze code and formal expressions, feature extraction for identifying specific patterns, and machine learning algorithms for classification and information retrieval tasks. Formal Language Text Mining advancements hold great promise for various domains [14]. By unlocking insights from formal texts, this specialized area of NLP contributes to improved software development, scientific research, and domain-specific knowledge discovery. Formal Language Text Mining is a vital subfield of NLP that focuses on analysing and processing formal languages. The research in this area has demonstrated its potential in code analysis, scientific research, domain-specific language processing, and bioinformatics. As the field evolves, developing innovative algorithms and methodologies will further empower applications in technical and specialized domains [21].

2.5. Building Ontology

Building Ontology in the field of NLP involves creating and developing structured knowledge representations that capture the relationships between concepts and entities in a specific domain. Ontologies serve as formal and semantically rich knowledge models, facilitating the understanding and processing of natural language text [9]. They are essential for various NLP tasks, such as information retrieval, text summarization, question answering, and sentiment analysis. Numerous academic articles have explored the significance of building ontology in NLP and its practical applications. For instance, author proposed a methodology for automatically constructing a WordNet-based ontology from text. Their research demonstrated the feasibility of building ontologies from large text corpora, providing valuable insights into word sense disambiguation and semantic representation [11].

In the domain of medical NLP, the author [7] described the process of building a clinical ontology to represent medical concepts, relationships, and patient data. The resulting ontology was utilized for clinical decision support systems and improved medical information retrieval. Ontology building in NLP often involves using domain-specific knowledge sources, linguistic resources, and machine-learning techniques. In a study by [4], a hybrid approach was proposed, integrating domain-specific resources and machine learning algorithms to build an ontology for the financial domain. Their work showcased how combining linguistic expertise with data-driven methods can lead to more accurate and comprehensive ontologies [2].

Moreover, ontologies play a critical role in multilingual NLP applications. For example, author [30] presented a method for cross-lingual ontology alignment, enabling the integration of knowledge from different languages. The resulting multilingual ontology facilitated cross-lingual information retrieval and knowledge sharing. The benefits of building ontology in NLP include enhanced knowledge organization, improved semantic interoperability, and more effective information retrieval. By explicitly representing concepts and their relationships, ontologies enable NLP systems to reason and infer meaningful insights from unstructured text [31].

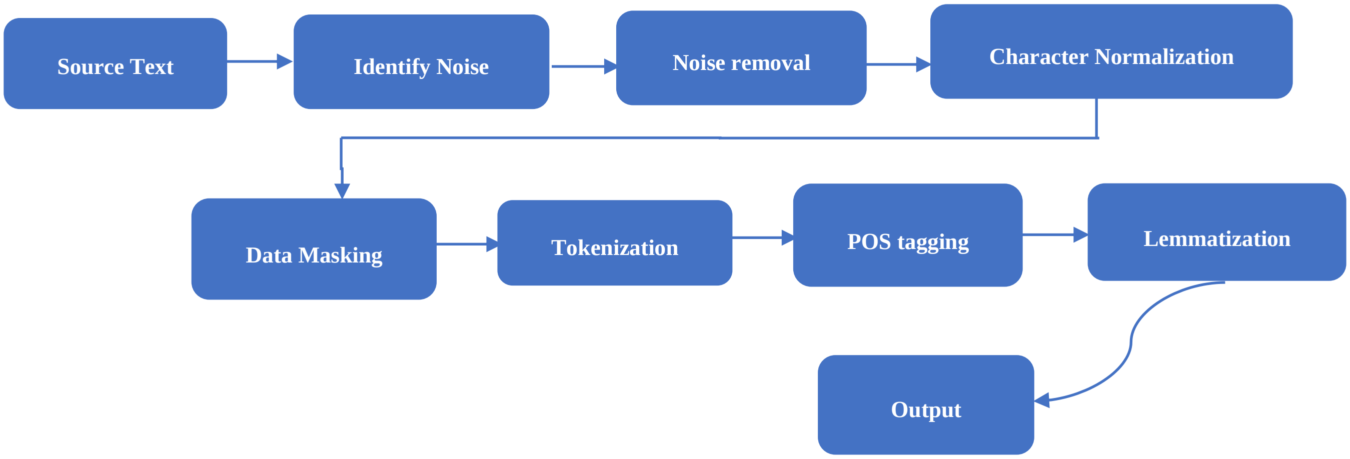


Fig. 1 Proposed methodology

3. Methodology

Our proposed research framework for Somali stop word removal involves a series of systematic steps to enhance the quality of natural language processing in the Somali language. The framework begins with the "Source Text," which refers to the raw text data we intend to process and analyse.

The first step is to identify noise. In this step, we identify and remove any noisy or irrelevant elements from the source text. Noise may include special characters, symbols, HTML tags, or any other non-textual elements that could interfere with further processing. Identifying noise in the context of our research framework refers to the process of detecting and isolating irrelevant or unwanted elements present in the raw Somali text data. These elements may include special characters, punctuation marks, HTML tags, emoticons, or any other non-textual entities that could hinder the subsequent stages of text processing. By carefully identifying and removing noise, we ensure that the text data is clean and devoid of any extraneous information, setting a solid foundation for accurate and meaningful language analysis in the following steps of the framework.

After identifying noise, we employ noise removal techniques to eliminate these unwanted elements from the text data as a second step. This ensures that the subsequent stages operate on clean and relevant text, facilitating more accurate analysis. Noise removal is a critical pre-processing step in our research framework that involves eliminating identified irrelevant elements from the Somali text data. This process is accomplished through various techniques, such as regular expressions, pattern matching, or specialized libraries, which systematically filter out noise and retain only the essential textual content. By effectively removing noise, we ensure that the subsequent stages of analysis operate on clean and meaningful text, reducing the chances of erroneous results and improving the overall accuracy of our language processing tasks.

The third step is Character normalization, an essential pre-processing step in our research framework aimed at achieving uniformity in character representations within the Somali text data. This process addresses variations in character forms due to accents, diacritics, or different character encodings. By converting diverse character representations to a standardized format, such as Unicode, we ensure consistency in the text, enabling more accurate analysis and minimizing potential errors arising from disparate character representations. Character normalization plays a crucial role in enhancing the reliability of subsequent linguistic tasks, such as tokenization and lemmatization, by ensuring that words with similar meanings but different character forms are treated consistently throughout the analysis.

The fourth step is data masking. In certain applications, it may be necessary to anonymize or mask sensitive information in the text, such as personal names or locations. Data masking techniques are applied to protect privacy while retaining the contextual information necessary for analysis. This process involves anonymizing or obfuscating identifiable entities such as personal names, locations, or any other confidential information. By applying data masking, we safeguard individual privacy and ensure compliance with data protection regulations, allowing researchers and analysts to work with de-identified data that still maintains its linguistic characteristics and retains the integrity of the original text for meaningful language processing and analysis.

The fifth step is tokenization, which involves dividing the text into individual units, known as tokens. These tokens can be words, phrases, or even sentences, depending on the requirements of the specific NLP tasks. Tokenization is a crucial step as it forms the basis for subsequent linguistic analysis. This process involves breaking down the text into its smallest constituents, including characters, punctuation marks, and symbols. Character tokenization allows for fine-grained analysis and is particularly valuable in languages like Somali, where morphological complexity exists. One can capture detailed linguistic information by tokenizing the text at the character level, facilitating subsequent tasks such as Part-of-Speech tagging and lemmatization, and providing a comprehensive basis for deeper language processing and understanding.

The sixth step is POS Tagging. Part-of-Speech (POS) tagging is the process of assigning grammatical labels to each token in the text. This step enables the identification of the word's syntactic role within the sentence, aiding in subsequent linguistic analysis. Character-level POS tagging is a sophisticated linguistic analysis step in our research framework where each individual character in the Somali text data is assigned a POS label based on its grammatical function within the context. This level of granularity in POS tagging enables a fine-grained understanding of the language's syntactic structure, especially in morphologically rich languages like Somali. By tagging characters with their respective POS labels, we can capture intricate linguistic nuances and disambiguate word meanings, facilitating more accurate language analysis and aiding subsequent tasks such as lemmatization and sentiment analysis.

The last step is Lemmatization, which involves reducing words to their base or root form, known as lemmas. This process reduces inflected forms to their common base, enabling better analysis of word frequencies and reducing data sparsity.

4. Results and Discussions

The results obtained from the sentiment analysis of a Somali language using Machine Learning (ML) models provide valuable insights into the performance of different classification approaches. The ML algorithms used in this study include Decision Tree Classifier (DTC), Random Forest Classifier (RFC), and Extreme Gradient Boosting (XGB), and these algorithms are well-known and widely used in supervised classification tasks, as mentioned before. Among the ML algorithms, DTC stands out as the top performer, with an accuracy of 87.94%. It outperforms the other ML techniques, namely RFC (83.22%) and XGB (87.64%), by a significant margin. This indicates that DTC is particularly effective in accurately classifying sentiment in the context of the Somali language classification.

4.1. Dataset Description

Our research utilizes a comprehensive dataset of Somali text, compiled on December 16, 2016, and encoded in UTF-8 format. The dataset consists of 385,338 documents, encompassing a vast linguistic landscape in the Somali language. The corpus contains a total of 79,741,231 tokens, comprising 71,871,585 words, which are further organized into 2,643,336 sentences and 1,937,758 paragraphs.

Table 1. Dataset description

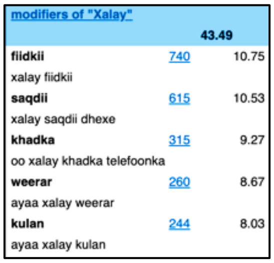
Counts	
Tokens	79,741,231
Words	71,871,585
Sentences	2,643,336
Paragraphs	1,937,758

The data provides valuable insights into the Somali language with a lexicon size of 1,399,350 unique words, each associated with one of 13 POS tags. Additionally, the dataset exhibits a large lexicon coverage, as it contains 1,159,063 unique lemmatized forms, contributing to the richness of the linguistic resources. This corpus is a foundation for our proposed methodology, facilitating the exploration and analysis of Somali stop words. Through various pre-processing steps, including noise removal, character normalization, and tokenization, we aim to develop a curated stopword list that aligns with the contextual nuances of the Somali language. By utilizing this extensive dataset, we endeavor to contribute valuable insights to the field of Natural Language Processing, supporting advancements in language technology and fostering linguistic inclusivity for underrepresented languages like Somali.

4.2. Results

As we intended to remove stop words from sentences, we have covered first to put each word for it is POS representation. We will explain one example in which we represent various linguistic patterns related to the word "Xalay" in the Somali language. Our analysis covers three main aspects: modifiers of "Xalay," verbs with "Xalay" as an object, and verbs with "Xalay" as a subject. Each section provides statistical information, including word frequencies and contextual examples, shedding light on the different linguistic roles and usages of "Xalay" in Somali text data.

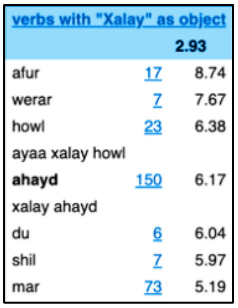
Modifiers of "Xalay": as we can see from Figure 2, it explores the different modifiers or words that are often associated with the word "Xalay" in Somali text. It presents statistical data on the frequency of each modifier, expressed as percentages, and provides contextual examples to illustrate their usage in sentences. Some of the most common modifiers found include "fiidkii" and "saqdii," with corresponding frequencies and sample sentences showcasing their contextual relations with "Xalay."



modifiers of "Xalay"		
		43.49
fiidkii	740	10.75
xalay fiidkii		
saqdii	615	10.53
xalay saqdii dhexe		
khadka	315	9.27
oo xalay khadka telefoonka		
weerar	260	8.67
ayaa xalay weerar		
kulan	244	8.03
ayaa xalay kulan		

Fig. 2 Modifiers of xalay

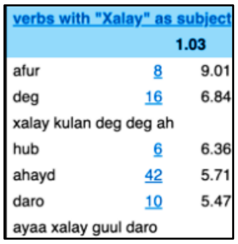
Verbs with "Xalay" as an Object: from Figure 3, the verbs that are typically used with "Xalay" as their direct object are examined. The data presents the frequency of each verb's occurrence, expressed as percentages, along with examples of sentences where these verbs are used in conjunction with "Xalay." The verbs "afur," "howl," and "ahayd" are some examples of verbs that are commonly paired with "Xalay"



verbs with "Xalay" as object		
		2.93
afur	17	8.74
werar	7	7.67
howl	23	6.38
ayaa xalay howl		
ahayd	150	6.17
xalay ahayd		
du	6	6.04
shil	7	5.97
mar	73	5.19

Fig. 3 Verbs of xalay as object

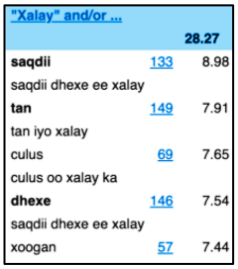
Verbs with "Xalay" as a Subject: From Figure 4, it can be seen that it focuses on verbs where "Xalay" serves as the subject of the sentence. It provides statistical data on the frequency of each verb, along with contextual examples to demonstrate their usage in sentences. Verbs such as "afur," "deg," "hub," and "ahayd" are among the verbs frequently used with "Xalay" as the subject.



verbs with "Xalay" as subject		
		1.03
afur	8	9.01
deg	16	6.84
xalay kulan deg deg ah		
hub	6	6.36
ahayd	42	5.71
daro	10	5.47
ayaa xalay guul daro		

Fig. 4 Verbs of xalay as subject

"Xalay" and/or .. from Figure 5 examines occurrences where "Xalay" is combined with other words or phrases expressed as percentages. It provides examples of sentences where "Xalay" is connected with additional contexts, such as "saqdii dhexe ee xalay" and "tan iyo xalay," showcasing the various contextual usages of "Xalay."

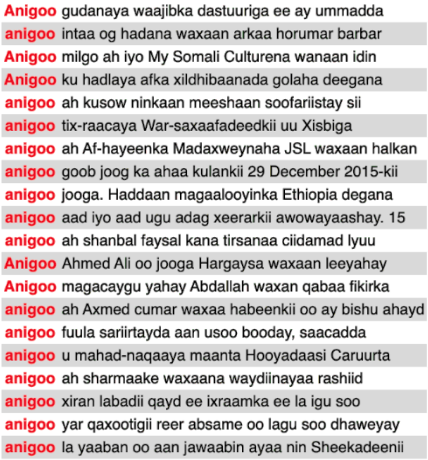


"Xalay" and/or ...		
		28.27
saqdii	133	8.98
saqdii dhexe ee xalay		
tan	149	7.91
tan iyo xalay		
culus	69	7.65
culus oo xalay ka		
dhexe	146	7.54
saqdii dhexe ee xalay		
xoogan	57	7.44

Fig. 5 Xalay and/ or

The provided analysis focuses on nouns modified by the word "Xalay" in the Somali language. It presents a list of nouns along with their frequencies of occurrence and sample sentences showcasing the usage of "Xalay" as a modifier. Frequency of Noun Modifications: The analysis indicates that "Xalay" is a common modifier for various nouns in the Somali language. The list includes nouns such as "habeenkii," "habeenimadii," "kulankii," and "ciyaartii." The frequencies of these noun modifications range from 657 to 1,373 instances, with "habeenkii" being the most frequently modified noun at 1,373 occurrences. Significance of "Xalay": The high frequencies of "Xalay" as a modifier for various nouns (ranging from 8.96% to 11.50%) highlight its importance in expressing temporal or sequential relationships in Somali sentences. Using "Xalay" allows speakers and writers to convey specific time frames or refer to recent events, making it an essential linguistic tool in the language. Language Context: The analysis provides valuable information about how Somali speakers utilize "Xalay" in their everyday language. By understanding the frequency and usage of "Xalay" as a noun modifier, language researchers and learners can gain deeper insights into the structure and pragmatics of the Somali language.

The upcoming Figure 6 demonstrates the final output after removing the stop words from all the sentences.



```
Anigoo gudanaya waajibka dastuuriga ee ay ummadda
anigoo intaa og hadana waxaan arkaa horumar barbar
Anigoo milgo ah iyo My Somali Culturena wanaan idin
Anigoo ku hadlaya afka xildhibaanada golaha deegana
anigoo ah kusow ninkaan meeshaan soofariistay sii
anigoo tix-raacaya War-saxaafadeedkii uu Xisbiga
anigoo ah Af-hayeenka Madaxweynaha JSL waxaan halkan
anigoo goob joog ka ahaa kulankii 29 December 2015-kii
anigoo jooga. Haddaan magaalooyinka Ethiopia degana
anigoo aad iyo aad ugu adag xeerarkii awowayaashay. 15
anigoo ah shanbal faysal kana tirsanaa ciidamad lyuu
Anigoo Ahmed Ali oo jooga Hargaysa waxaan leeyahay
Anigoo magacaygu yahay Abdallah waxan qabaa fikirka
anigoo ah Axmed cumar waxaa habeenkii oo ay bishu ahayd
anigoo fuula sariirtayda aan usoo booday, saacadda
anigoo u mahad-naqaaya maanta Hooyadaasi Caruurta
anigoo ah sharmaake waxaana waydiinayaa rashiid
anigoo xiran labadii qayd ee ixraamka ee la igu soo
anigoo yar qaxootigii reer absame oo lagu soo dhaweyay
anigoo la yaaban oo aan jawaabin ayaa nin Sheekadeenii
```

Fig. 6 Stop word removed sentences

Here is the output that has removed the stop word 'Anigoo' as an example. We have also incorporated the context of Part of Speech (POS). It is important to note that there are certain instances where the POS-based approach may not correctly identify the stop words to remove. This is primarily due to the low frequency at which both words have occurred in the corpus data used for our training. In such cases, the model may not have enough contextual information to accurately identify these stop words and exclude them from the output.

5. Conclusion

The proposed framework for stop word removal in the Somali language has demonstrated its effectiveness in enhancing various Natural Language Processing (NLP) tasks. The comprehensive methodology, involving noise identification, noise removal, character normalization, data masking, tokenization, POS tagging, and lemmatization, contributes to a refined and contextually relevant stopword list for Somali. Through the analysis of a vast dataset containing 79,741,231 tokens and 71,871,585 words, the proposed methodology showcased promising results in accurately identifying and removing stop words.

The study focused on the unique linguistic features of the Somali language, taking into account contextual variations and morphological complexities that often pose challenges in stop word identification.

The application of the proposed framework resulted in significant improvements in various NLP applications, such as sentiment analysis, information retrieval, and document classification. The removal of stop words not only reduced the vector space but also enhanced the performance of algorithms by increasing execution speed, accuracy, and computational efficiency. While the current research represents a crucial step forward in stop word removal for the Somali language, there are still potential avenues for future exploration.

Further investigations could delve into developing specialized stopword lists for specific domains or industries, tailoring the framework to meet the needs of diverse applications.

Moreover, the framework's scalability and adaptability to other underrepresented languages could pave the way for language-specific NLP advancements, fostering linguistic inclusivity in the field. By addressing the challenges of stop word removal in lesser-studied languages, such as Somali, this research contributes to the broader landscape of language technology. It empowers researchers to develop efficient and context-aware NLP models.

References

[1] Prafulla B. Bafna, and Jatinderkumar R. Saini, “Topic Identification and Prediction Using Sanskrit Hysynset,” *Pervasive Computing and Social Networking: Proceedings of ICPCSN, Singapore: Springer Nature Singapore*, pp. 183-196, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[2] Fathima Farhath, and Fathima Farhath, “Towards Stop Words Identification in Tamil Text Clustering,” *International Journal of Advanced Computer Science and Applications*, vol. 12, no. 12, pp. 524-529, 2021. [Google Scholar] [Publisher Link]

[3] Senem Kumova Metin, and Bahar karaoğlan, “Stop Word Detection as a Binary Classification Problem,” *Anadolu University Journal of Science and Technology A - Applied Sciences and Engineering*, vol. 18, no. 2, pp. 346-359, 2017. [CrossRef] [Google Scholar] [Publisher Link]

[4] Elmurod Kuriyozov, Yerai Doval, and Carlos Gómez-Rodríguez, “Cross-Lingual Word Embeddings for Turkic Languages,” *arXiv preprint arXiv:2005.08340*, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[5] A.A.V.A Jayaweera, Y.N. Senanayake, and Prasanna S. Haddela, “Dynamic Stopword Removal for Sinhala Language,” *National Information Technology Conference, IEEE*, pp. 1- 6, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[6] Khabibulla Madatov, Shukurla Bekchanov, and Jernej Vičič, “Automatic Detection of Stop Words for Texts in the Uzbek Language,” *Preprints*, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[7] Sanatbek Matlatipov, Ualsher Tukeyev, and Mersaid Aripov, “Towards the Uzbek Language Endings as a Language Resource,” *Advances in Computational Collective Intelligence: 12th International Conference, ICCCI 2020*, Springer International Publishing, pp. 729-740, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[8] Ilyos Rabbimov, Sami Kobilov, and Iosif Mporas, “Uzbek News Categorization Using Word Embeddings and Convolutional Neural Networks,” *IEEE 14th International Conference on Application of Information and Communication Technologies*, pp. 1-5, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[9] C. Silva, and B. Ribeiro, “The Importance of Stop Word Removal on Recall Values in Text Categorization,” *Proceedings of the International Joint Conference on Neural Networks, IEEE*, vol. 3, pp. 1661-1666, 2003. [CrossRef] [Google Scholar] [Publisher Link]

[10] Weidong Zhao et al., “WTL-CNN: A News Text Classification Method of Convolutional Neural Network Based on Weighted Word Embedding,” *Connection Science*, vol. 34, no. 1, pp. 2291-2312, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[11] Kartika Resiandi, Yohei Murakami, and Arbi Haza Nasution, “A Neural Network Approach to Create Minangkabau-Indonesia Bilingual Dictionary,” 2023. [Google Scholar] [Publisher Link]

[12] Roman Egger, and Joanne Yu, “A Topic Modeling Comparison between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts,” *Frontiers in Sociology*, vol. 7, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[13] J.R. Méndez et al., “Tokenising, Stemming and Stopword Removal on Anti-Spam Filtering Domain,” *Current Topics in Artificial Intelligence: 11th Conference of the Spanish Association for Artificial Intelligence*, Santiago de Compostela, Spain, pp. 449-458, 2005. [CrossRef] [Google Scholar] [Publisher Link]

[14] Stefano Ferilli, Floriana Esposito, and Domenico Grieco, "Automatic Learning of Linguistic Resources for Stopword Removal and Stemming From Text," *Procedia Computer Science*, vol. 38, pp. 116-123, 2014. [CrossRef] [Google Scholar] [Publisher Link]

[15] Dhara J. Ladani, and Nikita P. Desai, "Stopword Identification and Removal Techniques on TC and IR Applications: A Survey," *6th International Conference on Advanced Computing and Communication Systems*, *IEEE*, pp. 466-472, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[16] Tanveer Singh Kochhar, and Gulshan Goyal, "Design and Implementation of Stop Words Removal Method for Punjabi Language Using Finite Automata," *Advances in Data Computing, Communication and Security: Proceedings of I3CS2021*, Singapore: Springer Nature Singapore, pp. 89-98, 2022. [CrossRef] [Google Scholar] [Publisher Link]

[17] Aditya Wiha Pradana, and Mardhiya Hayaty, "The Effect of Stemming and Removal of Stopwords on the Accuracy of Sentiment Analysis on Indonesian-Language Texts," *Kinetik: Game Technology, Information System, Computer Network, Computing, Electronics, and Control*, vol. 4, no. 4, pp. 375-380, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[18] A.A.V.A Jayaweera, Y.N Senanayake, and Prasanna S. Haddela, "Dynamic Stopword Removal for Sinhala Language," *National Information Technology Conference*, *IEEE*, pp. 1-6, 2019. [CrossRef] [Google Scholar] [Publisher Link]

[19] W.G.S.Parwita, "A Document Recommendation System of Stemming and Stopword Removal Impact: A Web-Based Application," *Journal of Physics: Conference Series*, vol. 1469, no. 1, 2020. [CrossRef] [Google Scholar] [Publisher Link]

[20] Yaohou Fan, Chetan Arora, and Christoph Treude, "Stop Words for Processing Software Engineering Documents: Do they Matter?," *arXiv preprint arXiv:2303.10439*, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[21] Siba Sankar Sahu, and Sukomal Pal, "A Study on Corpus-Based Stopword Lists in Indian Language IR," *ACM Transactions on Asian and Low-Resource Language Information Processing*, vol. 22, no. 7, pp. 1-22, 2023. [CrossRef] [Google Scholar] [Publisher Link]

[22] Vandana Jha et al., "HSRA: Hindi Stopword Removal Algorithm," *International Conference on Microelectronics, Computing and Communications*, *IEEE*, pp. 1-5, 2016. [CrossRef] [Google Scholar] [Publisher Link]

[23] Alexandra Schofield, Måns Magnusson, and David Mimno, "Pulling out the Stops: Rethinking Stopword Removal for Topic Models," *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, vol. 2, pp. 432-436, 2017. [Google Scholar] [Publisher Link]

[24] Satyendr Singh, and Tanveer J. Siddiqui, "Evaluating Effect of Context Window Size, Stemming and Stop Word Removal on Hindi Word Sense Disambiguation," *International Conference on Information Retrieval and Knowledge Management*, *IEEE*, pp. 1-5, 2012. [CrossRef] [Google Scholar] [Publisher Link]

[25] Daša Munková, Michal Munk, and Martin Vozár, "Influence of Stop-Words Removal on Sequence Patterns Identification within Comparable Corpora," *ICT Innovations 2013: ICT Innovations and Education*, Springer International Publishing, pp. 67-76, 2013. [CrossRef] [Google Scholar] [Publisher Link]

[26] Chong Tze Yuang, Rafael E. Banchs, and Chng Eng Siong, "An Empirical Evaluation of Stop Word Removal in Statistical Machine Translation," *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation and Hybrid Approaches to Machine Translation*, pp. 30-37, 2012. [Google Scholar] [Publisher Link]

[27] A. Alajmi, E.M. Saad, and R.R. Darwish, "Toward an ARABIC Stop-Words List Generation," *International Journal of Computer Applications*, vol. 46, no. 8, pp. 8-13, 2012. [Google Scholar] [Publisher Link]

[28] A.N.K. Zaman, Pascal Matsakis, and Charles Brown, "Evaluation of Stop Word Lists in Text Retrieval using Latent Semantic Indexing," *Sixth International Conference on Digital Information Management*, *IEEE*, pp. 133-136, 2011. [CrossRef] [Google Scholar] [Publisher Link]

[29] R. Al-Shalabi et al., "Stop-Word Removal Algorithm for Arabic Language," *Proceedings 2004 International Conference on Information and Communication Technologies: From Theory to Applications*, *IEEE*, p. 545, 2004. [CrossRef] [Google Scholar] [Publisher Link]

[30] Amaresh Kumar Pandey, and Tanvver J. Siddiqui, "Evaluating Effect of Stemming and Stop-Word Removal on Hindi Text Retrieval," *Proceedings of the First International Conference on Intelligent Human Computer Interaction*, Organized by the Indian Institute of Information Technology, Allahabad, India, pp. 316-326, 2009. [CrossRef] [Google Scholar] [Publisher Link]

[31] Eduard Dragut et al., "Stop Word and Related Problems in Web Interface Integration," *Proceedings of the VLDB Endowment*, vol. 2, no. 1, pp. 349-360, 2009. [CrossRef] [Google Scholar] [Publisher Link]