Towards Stopwords Identification in Tamil Text Clustering

M.S. Faathima Fayaza[1]
Department of Information Technology
South Eastern University of Sri Lanka
Oluvil, Sri Lanka

F. Fathima Farhath[2]
Department of Computing
Informatics Institute of Technology
Colombo, Sri Lanka

***Abstract*—Now-a-days, digital documents have become the primary source of information. Therefore, natural language processing is widely utilized in information retrieval, topic modeling, document classification, and document clustering. Preprocessing plays a significant role in all of these applications. One of the critical steps in preprocessing is removing stopwords. Many languages have defined their list of stopwords. However, a publicly available stopwords list isn't available for the Tamil language since it is under-resourced. This study identified 93 general and some domain-specific stopwords for sports, entertainment, local and foreign news by analyzing more than 1.7 million Tamil documents with more than 21 million words. Also, this study shows that removing stopwords improves the accuracy of a Tamil document clustering system. It showed an improvement of 2.4%, 0.95% in the F-score for TF-IDF with one pass algorithm and FastText with the one-pass algorithm, respectively.**

Keywords*—Stopwords; Tamil; pre-processing; TF-IDF; clustering

I. Introduction

Usage of digital text information is growing exponentially in today's world, not only in English but also in other regional languages. Managing this data and extracting the relevant information has become a challenge. Henceforth, Natural Language Processing (NLP) has emerged as a new research field. Yet, the text is more challenging to manipulate and process than numerical data since it is unstructured, ambiguous, and difficult to manipulate. Information retrieval, document clustering, question and answering, document classification, sentiment analysis, and text summarization are some trivial applications of NLP. In every such application, the very first step is data preprocessing. Typically, data preprocessing includes tokenization, stemming, lemmatization, and stopwords removal [1]. Preprocessing eliminates the noise in the data. Further, preprocessing improves the performance of applications. More than 70% of the total text classification process comprises of preprocessing of text alone [2].

Stopwords are the frequently occurring words in a language containing very little or no meaning when used alone. They influence the syntax of a language rather than the semantics of a language [3]. "Are, is, be, a, the, an, of" are some examples of stopwords in English [3]. Therefore, removing stopwords shrinks the size of the text corpus by nearly 35-45% [4] by leaving only the semantically significant words. Also, this aids in improving the accuracy and efficiency of application as the emphasis is given to the semantic of the text. For instance, in document classification, the corpus size reduction reduced the time needed to train a model [5].

In the literature, there are several studies conducted for stopwords identification for many languages such as English, Hindi, Arabic, Chines [4], [6]. Yet, there is no publicly available list of stopwords for Tamil as less research has been done. This paper focuses on identifying stopwords for the Tamil language. Tamil is one of the highly agglutinative languages. The language is used in Sri Lanka, India, Canada, Malaysia, and many other parts of the world. Also, its use in the online platform has shown a hike in the recent past. Therefore, there is a notable need for NLP applications to make use of these available online resources. Therefore, developing and defining stopword lists will benefit the preprocessing step for Tamil language NLP applications. In this research, the authors identify 93 general stopwords for Tamil. Apart from that, the authors identify some domain-specific stopwords for the domain of sports, entertainment, international and local news. Further, by incorporating the preprocessing step of stopwords removal in Tamil text clustering, an improvement of 2.4 and 0.95 in F-score were noted for TF-IDF with one pass algorithm and FastText with one pass algorithm accordingly.

The remaining part of the paper is laid out as follows. Section II elaborates on the related work. Section III details the methodology of stopwords generation for the Tamil language. Section IV reports on the evaluation conducted. Section V analyzes the result and discusses the findings. Section VI conveys the conclusion of the research.

II. Related Work

Fox[4] created a stopword list for English using Brown Corpus of 1,014,000 words. Here the author manually added the words that appeared more than 300 times in the corpus in a list and finalized it by manually analyzing. The stopword list contained 421 words. This approach is domain-independent and is widely used in retrieval systems. Hao et al. [6] generated a stopword list using the weighted Chi-squared statistic technique for the Chinese language. Researchers observed that the suggested methodology effectively improved the F1 classification score by nearly 7%.

Raulji et al. [7] proposed a dictionary-based approach to remove the stopwords for the Sanskrit Language. They used a predefined word list, compared it with the targeted text, and removed the stopwords. The researchers stated that from

87,000 words corpus, 11,200 words were removed as stopwords. This reduced the corpus size by 13% and reduced the feature space and CPU cycle.

Saif et al. [8] generated stopwords for the twitter sentimental analysis using a semantic approach. The researchers concentrated on word semantics and contextual semantics of the words. For this study, six different datasets were used. It was noticed that using semantically found out stopwords improved the accuracy by 0.42% and F-score by 0.94% than using a classic stopword list. Further stopwords removal reduced the classification features by 48.34% and size by 1.17% compared to traditional approaches. Miretie and Khedkar[9] generated a stopword list for the Amharic language by applying aggregated term frequency, entropy and inverse document frequency. El-Khair [10] conducted a comparative study to determine the effect of stopwords removal for the Arabic language. The notion of this study is to combine statistical and linguistic approaches. This study used three stopword lists: general list, corpus-based list, and combined list. Also, they used inverse document frequency, probability weight, and statistical modeling approaches. It was concluded that the general stopword list performed better than the latter two.

Bouzoubaa et al. [11] standardized the Arabic stopword list. They [12] used a statistical approach to find stopwords in the Arabic language. They concluded that this approach increased the performance of the Artificial Neural Network (ANN) classifier than when the general stopword list was used.

Ghag and Shah [5] studied the consequences of stopwords elimination in sentiment analysis and reported that the traditional classifier accuracy improved from 50% to 58.6% when stopwords were removed. However, when applying for "Average Relative Term Frequency Sentiment Classifier," "Senti-Term Frequency Inverse Document Frequency," and "Relative Term Frequency Sentiment Classifier," the improvement was insignificant.

Gunasekara and Haddela [13] created a domain-specific stopword list for the Sinhala language. Reported improved precision, recall, F-score, and accuracy when removed the stopwords in Naïve Bayes and Maximum Entropy-based classifier to classify the Sinhala news articles.

Ladani and Desai [3] surveyed the available stopwords removal techniques for Indian and Non-Indian Languages. They classified stopwords into two main categories as general and domain-specific. They were reported that removing stopwords reduces the size and improves the accuracy of text classification.

Jha et al. [14] proposed a Deterministic Finite Automata (DFA) based stopwords elimination algorithm for the Hindi language. Used JSON objects to implement DFA. For this study total of 200 Hindi documents were used as input, including a movie review dataset gathered from the internet. Here the accuracy and efficiency improved for text preprocessing.

Jayaweera et al. [15] proposed a dynamic approach to find Sinhala stopwords. In this study, they argued the cutoff point is subjective to the dataset. This study used 90,000 documents.

Wijeratne and de Silva [16] collected the data from patent documents between 2010-2020 and created a corpus with 540,276 words of Sinhala text and listed the stopwords using term frequency. Sarica and Luo [17] identified the stopwords in technical language.

Rakholia et al. [18] proposed a rule-based approach to detect stopwords for the Gujarati language dynamically. They developed 11 static rules and used them to generate a stopword list at runtime. They attained 98.10% accuracy for generic stopwords detection and 94.08% for domain-specific stopwords detection.

Multiple approaches have been tried in different languages in the literature to identify the stopwords. Those are mainly: a manual, dictionary-based, rule-based, statistical approach, term frequency, weighted term frequency, inverse document frequency. Most of them are static approaches and data-dependent. Further, these approaches require an intense amount of resources to gain better accuracy. Tamil is a low-resourced and highly inflected language. Therefore, applying these techniques directly to the Tamil language will not be feasible. Hence this study presents a dynamic approach for Tamil stopwords identification.

III. Stopwords Identification for the Tamil Language

A. Data Collection and Preprocessing

The data collected by Fayaza and Ranatunga [19] is utilized in this study. Datasets contain more than 1.7 Tamil documents with more than 21 million words. The datasets contain two types of data. Namely.

1) Online news data: The online news data was collected from nine news providers and covered the local, international, sport, business, and entertainment news. Each news consisted of title, body, URL, and date published. Every news domain is identified using a URL, and domain-specific groups were created as the first step. Using this data following datasets were created.

Dataset 1: International news.

Dataset 2: Sports news.

Dataset 3: Local news.

Dataset 4: Entertainment news.

The following procedures were carried out as preprocessing for all the datasets created above. Using <TITLE> and <BODY> tags, the news title and body were identified. Then the data was tokenized into distinct terms using the white space characters like space, tab, newline/carriage return, and punctuation marks. This was followed by the removal of non-Tamil characters and punctuation marks.

2) General data: It contains randomly collected data from multiple sources. The same preprocessing steps carried out on the previous data set were carried out.

B. Stopword Identification

Several approaches were implemented to identify the stopwords in different languages in the literature. However, most of these approaches are based on the term frequency of the text. In this study, term frequency (TF), inverse document frequency (IDF), and term-frequency-inverse-document-frequency (TF-IDF) are calculated for every term in the dataset. Authors select this approach since it is independent of the dataset size and domain, and it has been used in many other low-resource languages [15]. Further authors conducted multiple executions with different values to define the threshold. From that, the generated threshold value is selected by analyzing the stopword lists.

This automatic identification process consists of the following set of procedures:

- Calculate term frequency (TF) for each term in the document ($TF_{t,d}$).
- Calculate the document frequency (DF) for each term.
- Calculate the Inverse document frequency (IDF_t) : log10(N/dft) (N: total number of documents in the dataset).
- Calculate the TF*IDF for each term.
- Calculate the average TF*IDF for each term.
- List the TF*IDF in order.
- If TF*IDF is lower than threshold value term, added to stopword list.
- Identify intersect words in the lists and create a general stopword list.

It was identified that some stopwords are common among all the domain-specific stopword lists. Therefore, this study categorizes those words under general stopwords for the Tamil language.

IV. Evaluation

To date, there is no published work or publicly available stopword list for Tamil. This study created two types of stopwords lists (One general and four domain-specific). Three individuals manually evaluated these generated lists. Fleiss' Kappa statistic [20] was used to assess the agreement between the evaluators, which was 93.0.

Stopwords removal is one of the basic preprocessing steps in NLP. To evaluate the impact of stopwords removal on system performance of the NLP system, the generated lists were utilized for stopword removal in clustering for Tamil news, using ten datasets as of [19]. The incorporation was done over the same two approaches used in [19]. Those are:

1) TF-IDF with one pass algorithm [19].
2) FastText with one pass algorithm [19].

The system without removing stopwords was used as the baseline for this experiment. Then the stopwords removed datasets were tested. Generated results are compared against the manual clusters created in [19], and Pairwise F-scores [19] were calculated.

V. Result and Discussion

This study paved the way to create two types of stopwords lists, a general one and a domain-specific one. The general stopword list contains 93 words. Fig. 1 depicts the effectiveness of the clustering system with and without stopwords. Fig. 2 list downs the general stopwords. Domain-specific stopwords are listed in Fig. 3, Fig. 4, Fig. 5 and Fig. 6. Fig. 3 is the list of stopwords for the domain of International news. Fig. 4 is the same for the domain of Local news. Fig. 5 is for the Sports news, and Fig. 6 is for the Entertainment news.

Four experiments were conducted under this study. They are:

1) TF-IDF with one pass algorithm (TFIDF-OPA) clustering using the dataset with stopwords.
2) FastText with one pass algorithm (FT-OPA) clustering using the dataset with stopwords.
3) TF-IDF with one pass algorithm (TFIDF-OPA) clustering using the dataset without stopwords (stopwords removed).
4) FastText with one pass algorithm (FT-OPA) clustering with the datasets without stopwords (stopword removed.).

Table I describes the results obtained for the above 4 experiment setups. There is a significant improvement in the F-score when the dataset is used after removing stopwords with regard to that of the data set with the stopwords TFIDF-OPA increased by 2.4% and FT-OPA increased by 0.95%. This shows the impact of stopwords removal is higher in TF-IDF than that of FastText.

Tamil is a grammar-rich and highly inflected language. Even though some words are in stopwords, inflections of them are not included in the list. For example, அணி (Team – ani) is a stopword in the sports domain. It has the following inflected term: அணிகளுக்கு (for teams – Aṇikaḷukku), அணிக்கு (To the team - Aṇikku), அணியின் (Of the team- Aṇiyiṉ), அணிக்கும் (for the Team- Aṇikkum). TF-IDF fails to identify all these forms as stopwords in the clustering process. But Fasttext was able to handle this in the clustering process. Because Fasttext include representing sentences with bag-of-words and bag-of-n–grams, as well as using subword information, and sharing information across classes through a hidden representation. Further, TF-IDF considers inflected terms as different words.

Also, another challenge is some terms are written in different styles by different news providers. For example: கிரிக்கட் (Cricket -Kirikkat), கிரிக்கெட் (Cricket -Kirikket). TF-IDF considers these as two different terms.

All the following news samples reports on different arrests instances. All these news get grouped into one cluster without the stopwords removed if the clustering is performed. However, when the stopwords were removed, the system could cluster them into three different clusters based on the reason for the arrest. The first, second, and sixth reference the same instance about an arrest associated with drugs. The fourth and fifth are about an arrest related to an accident; they are

clustered together. The third news is in another cluster, which is about an arrest related to smuggling. Since the word கைது (arrested – kaitu) was in the stopword list, the stopword removal process removes it from the dataset. Therefore, it was possible to distinguish these news articles, increasing the clustering accuracy.

- 12 இலட்சம் ரூபா பெறுமதியான போதை மாத்திரைகளுடன் மூவர் கைது (12 Ilaṭcam rūpā peṟumatiyāṉa pōtai māttiraikaḷuṭaṉ mūvar kaitu) –Three arrested with drugs worth 12 lakh rupees.
- 12 லட்சம் பெறுமதியான போதை மாத்திரைகளுடன் மூவர் கைது (12 Laṭcam peṟumatiyāṉa pōtai māttiraikaḷuṭaṉ mūvar kaitu) - Three arrested with Rs 12 lakh worth of drugs.
- ஒருதொகை தங்காபரணங்களுடன் சிங்கப்பூர் பிரஜை கைது (Oru tokai taṅkāparaṇaṅkaḷuṭaṉ ciṅkappūr pirajai kaitu) - Singaporean national arrested with gold jewelry.
- மூவர் கைது (Mūvar kaitu) - Three arrested.
- 3 பேர் கைது (3 Pēr kaitu) - 3 people arrested.
- ஹெரோயின் வியாபாரத்தில் ஈடுபட்ட நபர் கைது (Herōyiṉ viyāpārattil īṭupaṭṭa napar kaitu)- The person involved in the heroin trade was arrested.

Further, the following is a set of news related to the same day. The first and the third related to a discussion on bus fare reduction, while the second is regarding a meeting between the president and the TNA parliamentarians. When they are clustered without the stopwords removal, all three are placed into one cluster. However, when classified after the stopword removal, the first and the third are allocated to the same cluster while the second is set onto another separate cluster. Since the word இன்று (today – inru) that was in the stopword list was removed in the stopword removal process, the documents similarly between first and third increased while for the second one reduced..

- பஸ் கட்டணம் குறைக்கப்படுமா ? இன்று கலந்துரையாடல் (Pas kaṭṭaṇam kuṟaikkappaṭumā? Iṉṟu kalanturaiyāṭal) - Will bus fares be reduced? Discussion today.
- தமிழ் தேசிய கூட்டமைப்பின் பாராளுமன்ற உறுப்பினர்களுக்கும் ஜனாதிபதி மைத்திரிபால சிறிசேனவுக்கும் இடையில் இன்று சந்திப்பு ஒன்று (Tamiḻ tēciya kūṭṭamaippiṉ pārāḷumaṉṟa uṟuppiṉarkaḷukkum jaṉātipati maittiripāla ciṟicēṉavukkum iṭaiyil iṉṟu cantippu oṉṟu) - Today a meeting between TNA parliamentarians and President Maithripala Sirisena.
- பஸ் கட்டணத்தைக் குறைப்பது குறித்த இறுதித்தீர்மானம் தொடர்பிலான கலந்துரையாடல் (Pas kaṭṭaṇattaik kuṟaippatu kuṟitta iṟutittīrmāṉam toṭarpilāṉa kalanturaiyāṭal)- Discussion on the final decision on reducing bus fares.

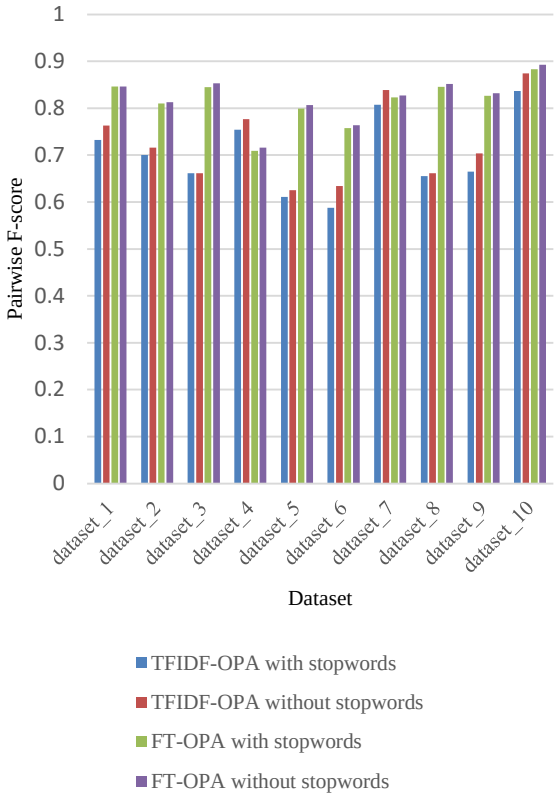

Fig. 1. Effectiveness of Clustering with and without Stopwords.

ஒரு	மற்றும்	அது	பெரும்	அனைத்து	வேண்டும்	அவர்	போல்
என்று	இல்லை	மூலம்	என்று	இன்று	மாற்றம்	போன	இங்கு
இந்த	என்று	தன்	உள்ளது	ஆண்டு	மிக	அல்ல	கூட
என	என்ற	என்ன	என்பது	ஆம்	தனது	எங்கும்	போல
இது	அந்த	என்	தான்	மீண்டும்	அவர்	இதோ	கைது
முதல்	அல்லது	ஏன்	கொண்டு	முதல்	பெற்றது	செய்து	இடையிலான
பல	வரும்	ஒரே	அதன்	என்ற	விசேட	சிறு	எதிரான / எதிராக
சில	மேலும்	யார்	நாள்	போது	ஒருவர்	மிகவும்	பலர்
இருந்து	ஆனால்	இருந்தது	பிரபல	ஊடக	இருந்து	தன	உள்ள
நாள்	வரை	பெற	தமது	கடந்த	எந்த	வந்த	காரணமாக/காரணம்
உடனான	சிறந்த	பெரும்	ஏற்பட்ட	விட	புதிய	ஆகிய	அதை
ஒன்றில்	கடந்த	தொடர்பில்	தொடர்பான	மீது			

Fig. 2. Stopwords for the Tamil Language.

தாக்குதல்	இடம்பெற்ற	தீவிரவாதிகள்	காவல்துறையினர்	பாரிய	நாட்டு
காணொளி	எச்சரிக்கை	தொடர்பான	கொல்லப்பட்டனர்.	வயது	செய்த
தலைவர்	குண்டு	நடத்தப்பட்ட	ஏற்பட்ட	பகுதி	தற்கொலை
பல்வேறு	வந்த	வகையில்	தெரிவித்துள்ளார்	கொலை	ஜனாதிபதி

Fig. 3. International News Stopwords for the Tamil Language.

அதிகரிப்பு	அமைச்சு	தெரிவித்துள்ளது	நடவடிக்கை	விசாரணை	கட்சி	திணைக்களம்	செயலாளர்
விலை	வைத்து	இடம்பெற்ற	முன்னாள்	திட்டம்	குழு	எதிர்வரும்	மோசடி
ஒன்றை	ஆர்ப்பாட்டம்	பிரதேசத்தில்	போராட்டம்	தேசிய	அதிகாரிகள்	சேர்ந்த	செய்யப்பட்டுள்ளனர்

Fig. 4. Local News Stopwords for Tamil Language.

TABLE I. STATISTICAL ANALYSIS OF OBTAINED PAIRWISE F-SCORE WITH STOPWORDS AND WITHOUT STOPWORDS

Document Representation Techniques with Clustering Algorithms	**TFIDF-OPA with stopwords**	**FT-OPA with stopwords**	**TFIDF-OPA without stopwords**	**FT-OPA without stopwords**
Mean (Average)	70.1	81.5	**72.5**	**82.0**
Median	68.2	82.5	**70.9**	**83.2**
Minimum	58.8	70.9	**62.5**	**71.6**
Maximum	83.7	88.3	**87.4**	**89.2**
Standard Deviation	7.7	4.7	**5.7**	**4.6**

In the following scenario from entertainment deomain, the first news about Deepika Ranveer Wedding Date Announcement, the second talks about the second single released in the movie “Kattrin moli,” third is about “Sarkar,” movie story released. All these sentences are clustered into one group before removing the stopword வெளியாகியுள்ளது (Released - veḷiyākiyuḷḷatu). After removing the stopword, all three news clustered into three different clusters.

- தீபிகா ரன்வீர் திருமண திகதி அறிவிப்பு ! (tīpikā raṉvīr tirumaṇa tikati aṟivippu!) - Deepika Ranveer Wedding Date Announcement!
- காற்றின் மொழி திரைப்படத்தின் 2_வது சிங்கிள் இன்று (Kāṟṟiṉ moḻi tiraippaṭattiṉ 2_vatu ciṅkiḷ iṉṟu) - second single of the “karrin moli” movie today
- சர்கார் படத்தின் கதை வெளியானது (carkār paṭattiṉ katai veḷiyāṉatu) – Sarkar film story released

அணி	வீரர்	போட்டி	ஆட்டம்	உலக
இறுதி	வெற்றி	தொடர்	சர்வதேச	

Fig. 5. Sports News Stopwords for the Tamil Language.

இயக்குனர்	நடிகை	வெளியாகியுள்ளது
நடிகர்	ரசிகர்கள்	நடிப்பில்
தற்போது	திடீர்	படங்களில்

Fig. 6. Entertainment News Stopwords for the Tamil Language.

VI. CONCLUSION

This paper presents an approach to list out the stopwords in Tamil, which is a low-resource language. So far, there is no predefined published stopword list for Tamil. The widely used technique for stopwords identification is based on term frequency. In this study, TF*IDF with threshold value is used to identify the stopwords for Tamil. The research resulted in the generation of stopword lists for general domain and domain-specific ones for local, international, sport, and entertainment domains. To evaluate its impact on Tamil NLP, it was used in document clustering using TF-IDF with one pass algorithm and FastText with the one-pass algorithm. The results revealed that the removal of stopwords at the preprocessing stage improved F-score, mean, median, and standard deviation in both the approaches.

REFERENCES

[1] M. Anandarajan, C. Hill, and T. Nolan, Cluster Analysis: Modeling Groups in Text. 2019.

[2] S. C. Satapathy, Advances in Intelligent Systems and Computing 1177 Intelligent Data Engineering and Analytics, vol. 2, no. Ficta. 2020.

[3] D. J. Ladani and N. P. Desai, "Stopword Identification and Removal Techniques on TC and IR applications: A Survey," 2020 6th Int. Conf. Adv. Comput. Commun. Syst. ICACCS 2020, pp. 466–472, 2020, doi: 10.1109/ICACCS48705.2020.9074166.

[4] C. Fox, "A Stop List for General Text," ACM SIGIR Forum, vol. 24, no. 1–2, pp. 19–21, 1989, doi: 10.1145/378881.378888.

[5] K. V. Ghag and K. Shah, "Comparative analysis of effect of stopwords removal on sentiment classification," IEEE Int. Conf. Comput. Commun. Control. IC4 2015, pp. 2–7, 2016, doi: 10.1109/IC4.2015.7375527.

[6] L. Hao and L. Hao, "Automatic identification of stop words in chinese text classification," Proc. - Int. Conf. Comput. Sci. Softw. Eng. CSSE 2008, vol. 1, pp. 718–722, 2008, doi: 10.1109/CSSE.2008.829.

[7] J. K. and J. R., "Stop-Word Removal Algorithm and its Implementation for Sanskrit Language," Int. J. Comput. Appl., vol. 150, no. 2, pp. 15–17, 2016, doi: 10.5120/ijca2016911462.

[8] H. Saif, M. Fernandez, and H. Alani, "Automatic stopword generation using contextual semantics for sentiment analysis of Twitter," CEUR Workshop Proc., vol. 1272, pp. 281–284, 2014.

[9] S. Girmaw and V. Khedkar, “Automatic Generation of Stopwords in the Amharic Text,” Int. J. Comput. Appl., vol. 180, no. 10, pp. 19–22, 2018, doi: 10.5120/ijca2018916161.

[10] I. A. El-Khair, "Effects of stop words elimination for arabic information retrieval: A comparative study," arXiv, no. December, 2017.

[11] K. Bouzoubaa, H. Baidouri, T. Loukili, and T. El Yazidi, “Arabic stop words: Towards a generalisation and standardisation,” Knowl. Manag. Innov. Adv. Econ. Anal. Solut. - Proc. 13th Int. Bus. Inf. Manag. Assoc. Conf. IBIMA 2009, vol. 3, no. November 2009, pp. 1844–1848, 2009.

[12] A. Alajmi and E. mostafa Saad, "Toward an ARABIC Stop-Words List Generation Toward an ARABIC Stop-Words List Generation," vol. 46, no. January 2012, pp. 8–13, 2018.

[13] S. V. S. Gunasekara and P. S. Haddela, "Context aware stopwords for Sinhala Text classification," 2018 Natl. Inf. Technol. Conf. NITC 2018, pp. 2–4, 2018, doi: 10.1109/NITC.2018.8550073.

[14] V. Jha, N. Manjunath, P. D. Shenoy, and K. R. Venugopal, "HSRA: Hindi stopword removal algorithm," Int. Conf. Microelectron. Comput. Commun. MicroCom 2016, 2016, doi: 10.1109/MicroCom.2016.7522593.

[15] A. A. V. A. Jayaweera, Y. N. Senanayake, and P. S. Haddela, "Dynamic Stopword Removal for Sinhala Language," 2019 Natl. Inf. Technol. Conf. NITC 2019, pp. 8–10, 2019, doi: 10.1109/NITC48475.2019.9114476.

[16] Y. Wijeratne and N. de Silva, "Sinhala Language Corpora and Stopwords from a Decade of Sri Lankan Facebook," arXiv, 2020, doi: 10.2139/ssrn.3650976.

[17] S. Sarica and J. Luo, "Stopwords in Technical Language Processing," arXiv, no. June, 2020.

[18] R. M. Rakholia, and J. R. Saini, "A Rule-Based Approach to Identify Stop Words for Gujarati Language," In Proceedings of the 5th International Conference on Frontiers in Intelligent Computing: Theory and Applications, 2017, pp. 797-806, Springer, Singapore.

[19] M. S. Faathima Fayaza and S. Ranathunga, "Tamil News Clustering Using Word Embeddings," MERCon 2020 - 6th Int. Multidiscip. Moratuwa Eng. Res. Conf. Proc., pp. 277–282, 2020, doi: 10.1109/MERCon50084.2020.9185282.

[20] J. L. Fleiss, "Measuring nominal scale agreement among many raters," Psychological Bulletin, vol. 76, no. 5, pp. 378–382, 1971.