THE *African Stopwords* PROJECT: CURATING STOPWORDS FOR AFRICAN LANGUAGES

Anonymous authors
Paper under double-blind review

ABSTRACT

Stopwords are fundamental in Natural Language Processing (NLP) techniques for information retrieval. One of the common tasks in preprocessing of text data is the removal of stopwords. Currently, while high-resource languages like English benefit from the availability of several stopwords, low-resource languages, such as those found in the African continent, have none that are standardized and available for use in NLP packages. Stopwords in the context of African languages are understudied and can reveal information about the crossover between languages. The *African Stopwords* project aims to study and curate stopwords for African languages. In this paper, we present our current progress on ten African languages as well as future plans for the project.

1 INTRODUCTION AND MOTIVATION

Stopwords are uninformative common words in a text. When analysing text data and building various NLP models, stopwords might not add much value to the meaning of the document (Singh, 2019) depending on the NLP task (like text classification). Words such as articles and some verbs are usually considered stopwords because they don't usually determine the context or the true meaning of a sentence – they are words that can be removed without any negative consequences to the final NLP model training. Key to note also is that the removal of stopwords could improve a model training time owing to the reduced data size: the model will improve efficiency due to the reduced number of tokens involved in the training process (Singh, 2019).

However, stopword removal is highly dependent on the language, domain and task (Vallantin, 2020). Therefore, the use of one 'standard' stopword list is problematic because it ignores the domain-knowledge specificity of stopwords (Lo et al., 2005) and because it is language-specific (Zou et al., 2006). Researchers (Sarica & Luo, 2020; Wilbur & Sirotkin, 1992; Gerlach et al., 2019) have shown that domain-specific and language-specific stopwords can make significant impact both in general tasks (like spam filtering, caption generation, language classification and auto-tag generation) and in domain-specific tasks (like NLP in the medical field or with the Chinese language(Zou et al., 2006)). This shows the need for packages or techniques that can be utilised to effectively remove stopwords and enable building of effective NLP models.

There are many available libraries for stopwords removal including Natural Language Toolkit(NLTK) (Bird & Loper, 2004), spaCy (Honnibal et al., 2020), Gensim (Řehůřek & Sojka, 2010), among others. Notwithstanding the ubiquity of stopwords packages in popular languages like English, Spanish, German, they do not support any African language (to the best of our knowledge). In this light, the main aim of our project is to curate stopwords for various African languages in order to positively contribute towards the advancement of natural language processing for African languages.

2 *African Stopwords*

The *African Stopwords* project aims to systematically gather stopwords for African languages – starting with 13 African languages.

2.1 CURRENT PROGRESS

Through the help of contributors, we have gathered some stopwords for the 13 focus African languages in Table 1. We use a Github repository to host our code, contribution guides, curated stopwords, and linguistic discussions where there is controversy over potential stopwords candidates for a language.

Lang.	**#**	**Source**		**Lang.**	**#**	**Source**
Afrikaans	51	(Tatman, 2017)		Yoruba	60	(Tatman, 2017)
Hausa	322	(Tatman, 2017; Abdulmumin & Galadanci, 2019)		isiZulu	29	(Tatman, 2017)
Nigerian Pidgin	34	(Muhammad et al., 2022)		kiSwahili	103	(Tatman, 2017; David, 2020)
Kirundi	59	(Niyongabo et al., 2020)		Igbo	-	-
Kinyarwanda	80	(Niyongabo et al., 2020)		Shona	-	-
Somali	30	(Tatman, 2017)		Amharic	-	-
Sesotho	31	(Tatman, 2017)				

Table 1: Focus African languages and current number of stopwords curated. '-' denotes languages designated for future work.

2.2 FUTURE WORK: LEVERAGING MONOLINGUAL DATA

One of the current directions of the project is in investigating the feasibility of harvesting online texts from multiple domains for curating better (language-specific and domain-relevant) African stopwords. There are numerous sources of monolingual data for African languages. Efforts for other languages have used monolingual knowledge sources such as Brown Corpus (Fox, 1989; Maverick, 1969), 20 newsgroup corpus (Gerlach et al., 2019), books corpus (MONTEMURRO & ZANETTE, 2010), etc. In that respect, we plan to identify stopwords from monolingual data (of multiple domains) in the following steps:

1. **Gather monolingual data:** The first part involves gathering a list of monolingual sources for the focus African languages. In order to ensure diverse, multi-domain stopwords, we will focus on getting data from many domains.

2. **Using statistical methods to automatically identify candidate stopwords:** Research on stopwords identification have employed various statistical metrics, such as term-frequency-inverse-document-frequency (TF-IDF) (Wilbur & Sirotkin, 1992; Lo et al., 2005), entropy (MONTEMURRO & ZANETTE, 2010), information gain (Makrehchi & Kamel, 2008) and Kullback-Leibler divergence (Lo et al., 2005). We plan to use these statistical methods too to automatically identify candidate stopwords.

3. **Human evaluation:** As a final step in ensuring the automatically curated stopwords are actually in line with the language, we will employ a number of human evaluators to review our stopwords. Only the stopwords that pass the evaluation will be published.

4. **Open-sourcing the stopwords:** Finally, we will either integrate the African stopwords into popular NLP processing toolkits (like NLTK) or create a separate Python package for it.

3 CONCLUSION

Although the NLP community focuses on research in improvements of model development and data collection, there is still limited work on the inclusion of low resource languages in natural language processing toolkits. In this project, we set out to overcome one of those challenges by setting a framework to curate and accumulate stopwords for African languages. We show that stopwords have a huge significance in NLP. Finally we share our current progress on ten African languages as well as our plan for including more African languages.

REFERENCES

Idris Abdulmumin and Bashir Shehu Galadanci. hauwe: Hausa words embedding for natural language processing. In *2019 2nd International Conference of the IEEE Nigeria Computer Chapter (NigeriaComputConf)*, pp. 1–6, 2019. doi: 10.1109/NigeriaComputConf45974.2019.8949674.

Steven Bird and Edward Loper. NLTK: The natural language toolkit. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, pp. 214–217, Barcelona, Spain, July 2004. Association for Computational Linguistics. URL https://aclanthology.org/P04-3031.

Davis David. Swahili : News classification dataset, 2020. URL https://zenodo.org/record/5514203.

Christopher Fox. A stop list for general text. *SIGIR Forum*, 24(1–2):19–21, sep 1989. ISSN 0163-5840. doi: 10.1145/378881.378888. URL https://doi.org/10.1145/378881.378888.

Martin Gerlach, Hanyu Shi, and Luís A. Nunes Amaral. A universal information theoretic approach to the identification of stopwords. *Nature Machine Intelligence*, 1(12):606–612, Dec 2019. ISSN 2522-5839. doi: 10.1038/s42256-019-0112-6. URL https://doi.org/10.1038/s42256-019-0112-6.

Matthew Honnibal, Ines Montani, Sofie Van Landeghem, and Adriane Boyd. spaCy: Industrial-strength Natural Language Processing in Python. 2020. doi: 10.5281/zenodo.1212303.

Rachel Tsz-Wai Lo, Ben He, and Iadh Ounis. Automatically building a stopword list for an information retrieval system. *J. Digit. Inf. Manag.*, 3:3–8, 2005.

Masoud Makrehchi and Mohamed S. Kamel. Automatic extraction of domain-specific stopwords from labeled documents. In Craig Macdonald, Iadh Ounis, Vassilis Plachouras, Ian Ruthven, and Ryen W. White (eds.), *Advances in Information Retrieval*, pp. 222–233, Berlin, Heidelberg, 2008. Springer Berlin Heidelberg. ISBN 978-3-540-78646-7.

George V. Maverick. Computational analysis of present-day american english. henry kučera , w. nelson francis. *International Journal of American Linguistics*, 35(1):71–75, January 1969. doi: 10.1086/465045. URL https://doi.org/10.1086/465045.

MARCELO A. MONTEMURRO and DAMIÁN H. ZANETTE. Towards the quantification of the semantic information encoded in written language. *Advances in Complex Systems*, 13(02): 135–153, Apr 2010. ISSN 1793-6802. doi: 10.1142/s0219525910002530. URL http://dx.doi.org/10.1142/S0219525910002530.

Shamsuddeen Hassan Muhammad, David Ifeoluwa Adelani, Sebastian Ruder, Ibrahim Said Ahmad, Idris Abdulmumin, Bello Shehu Bello, Monojit Choudhury, Chris Chinenye Emezue, Saheed Abdullahi Salahudeen, Aremu Anuoluwapo, Alípio Jeorge, and Pavel Brazdil. Naijasenti: A nigerian twitter sentiment corpus for multilingual sentiment analysis. *CoRR*, abs/2201.08277, 2022. URL https://arxiv.org/abs/2201.08277.

Rubungo Andre Niyongabo, Qu Hong, Julia Kreutzer, and Li Huang. KINNEWS and KIRNEWS: Benchmarking cross-lingual text classification for Kinyarwanda and Kirundi. In *Proceedings of the 28th International Conference on Computational Linguistics*, pp. 5507–5521, Barcelona, Spain (Online), December 2020. International Committee on Computational Linguistics. doi: 10.18653/v1/2020.coling-main.480. URL https://aclanthology.org/2020.coling-main.480.

Serhad Sarica and Jianxi Luo. Stopwords in technical language processing. *CoRR*, abs/2006.02633, 2020. URL https://arxiv.org/abs/2006.02633.

Shubham Singh. How To Remove Stopwords In Python | Stemming and Lemmatization. *Analytics Vidhya*, August 2019. URL https://www.analyticsvidhya.com/blog/2019/08/how-to-remove-stopwords-text-normalization-nltk-spacy-gensim-python/.

Rachael Tatman. Stopword lists for african languages, Jul 2017. URL https://www.kaggle.com/rtatman/stopword-lists-for-african-languages.

Lima Vallantin. Why is removing stop words not always a good idea. *Medium*, June 2020. URL https://medium.com/@limavallantin/why-is-removing-stop-words-not-always-a-good-idea-c8d35bd77214.

W. John Wilbur and Karl Sirotkin. The automatic identification of stop words. *Journal of Information Science*, 18(1):45–55, 1992. doi: 10.1177/016555159201800106. URL https://doi.org/10.1177/016555159201800106.

Feng Zou, Fu Lee Wang, Xiaotie Deng, Song Han, and Lu Sheng Wang. Automatic construction of chinese stop word list. In *Proceedings of the 5th WSEAS International Conference on Applied Computer Science*, ACOS'06, pp. 1009–1014, Stevens Point, Wisconsin, USA, 2006. World Scientific and Engineering Academy and Society (WSEAS). ISBN 9608457432.

Radim Řehůřek and Petr Sojka. Software Framework for Topic Modelling with Large Corpora. Proceedings of LREC 2010 workshop New Challenges for NLP Frameworks, pp. 45–50, Valetta, MT, 5 2010. University of Malta. URL http://is.muni.cz/publication/884893/en.