A corpus-based approach to identifying stopwords in multiple languages

Bronwyn Bowles-King[1], Tanja Gaustad[2], Muzi Matfunjwa[1], Nomsa Skosana[1], Benito Trollip[1], Michelle White[1], Menno van Zaanen[1]

[1]South African Centre for Digital Language Resources, Potchefstroom, South Africa
[2]Centre for Text Technology, North-West University, Potchefstroom, South Africa

Correspondence: menno.vanzaanen@nwu.ac.za

Abstract

Stopwords are generally considered common, high-frequency words that carry little semantic information, and they are often filtered out during preprocessing, for example, to save computational power, improve model performance, or support focused analysis. However, not all languages have a stopword list readily available. This study identifies statistical word features in general-purpose corpora to evaluate existing Afrikaans, English, isiZulu, and Sesotho stopword lists. We consider whether stopword features differ for these languages compared to features typically used to identify English stopwords, such as word length, word frequency, document frequency, and combinations of these features. Although stopwords are ubiquitous in the corpora, they are a tiny minority group in terms of the number of unique words. This skewed distribution raises challenges in statistical modelling as it biases the classifier towards the larger group of content words. We therefore applied elastic net logistic regression with feature ablation for 25 features. The most effective class differentiation was achieved using a reduced trivariate model including word length, log term frequency, and log document frequency, with AUCs between 98.20 and 99.99 for the four languages. Out-of-fold evaluations for this core model demonstrated low precision (0.87–6.14) and F1-scores (1.73–11.53), but high recall (91.36–100.00). Our findings suggest that these three statistical measures for stopword identification behave consistently across the target languages, regardless of orthography or corpus size, and they display the potential to filter candidate stopwords for subsequent evaluation and curation.

1 Introduction

A large amount of Digital Humanities (DH) research uses text as its primary object of investigation, be it for quantitative or qualitative analyses. In order to apply digital methods, these texts need to be available in digitised form and ideally prepared for digital methodologies. Other disciplines, such as corpus linguistics, computational linguistics, and natural language processing (NLP), also have a long-standing tradition of working with digital text. One of the techniques employed in those fields of research is the application of stopword lists. In many corpus linguistics and NLP tasks, stopwords are removed from texts to mitigate the skewing effect of high-frequency words. As they are high-frequency words that carry little semantic or unique information, they do not contribute much to the content and are often filtered out during preprocessing. As such, stopword lists are a prerequisite for different corpus investigations or downstream NLP tasks such as Information Retrieval. Stopword lists, however, are language-dependent. To create such a list for a language that does not yet have one may sound conceptually like a simple task, but automatically identifying stopwords is far from a solved problem.

In an effort to construct a reusable method for automatically extracting a stopword list from texts, we will be using readily accessible corpora to calculate specific word-based features for a few languages. English, by way of the Brown corpus, will serve as the reference language to identify relevant stopword features. By testing and evaluating different combinations of the chosen features, we hope to gain an understanding of which word-based statistics most accurately identify stopwords. Following this, the same features will be applied to Afrikaans, isiZulu, and Sesotho corpora (in the form of Wikipedia data) to test the reusability of the identified stopword features for those languages even if they have corpora of different sizes and varying orthography. The ultimate aim is to determine whether successful measures are language-independent or differ per language or language family, and if automatic stopword identification is possible for the studied languages.

This work is licensed under CC BY SA 4.0. To view a copy of this license, visit https://creativecommons.org/licenses/by-sa/4.0/deed.en
The copyright remains with the authors.

2 Background

2.1 Previous research on stopwords

Generally speaking, a stopword is a word that is omitted when working with corpus concordances or frequency lists. As such, stopwords are common words that are needed to form syntactically correct sentences, but do not contribute much to the content of the sentences. For example, words like "the" and "is" seem to provide less information to the meaning of a sentence compared to content words like "computer" or "bicycle". Generally, they contribute less to the intrinsic meaning or other distinguishing characteristics of a text (Turki et al., 2024). Even though *stopword* is a term sometimes used in passing as if it has a universally agreed-upon definition, that is not the case. The meaning of the term *stopword* can range from being synonymous with a closed grammatical class word (Gandotra and Arora, 2018; Kucukyilmaz and Akin, 2024) to words with little to no lexical content (Lo et al., 2005; Sarica and Luo, 2021), or simply short and frequent words (Fox, 1989; de Keyser, 2012).

Stopword definitions vary by research goals and corpus properties and no single stopword list has blanket applicability. Acknowledging this context-dependency, our approach does not seek an overall ground-truth stopword list for the languages under study. Instead, we evaluate how useful statistical features are in helping to develop independently curated stopword lists.

As an example of the varied ways in which this term has been conceptualised by researchers, Wilbur and Sirotkin (1992, p. 45) define a stopword (in the context of the field of information retrieval) from a practical, probabilistic, and text-based perspective as having "the same likelihood of occurring in those documents not relevant to a query as in those documents relevant to the query". In the context of spoken language, Saal and Fourie (2021, p. 1120) refer to stopwords in a more qualitative sense as fillers that help speakers formulate their thoughts and as "*verbleek*" [bleached].[1]

Considering how often stopwords are employed, it is reasonable to conclude that they have certain characteristics as well as practical implications. Researchers, such as Kaur and Buttar (2018) or Lo et al. (2005), regularly define them by what they lack, i.e., as words that do not indicate the subject matter of a text on their own. This is true for general-purpose stopwords at least. Perusing lists from various languages, such as the Porter (2001) "Snowball" lists or the ISO stopwords dataset (Diaz, 2016), shows minimal indication of subject matter. As they tend to carry low semantic weight and are frequent across most documents, regardless of purpose or genre, stopwords affect information retrieval processes and must be handled efficiently (Kaur and Buttar, 2018).

Makrehchi and Kamel (2008) observe that there are two types of stopwords: general and domain-specific. General stopwords are language-specific and applied across contexts or genres, while domain-specific stopwords are relevant to a specific field (see also Ladani and Desai (2020); Sarica and Luo (2021)). Stopwords can be identified manually or curated based on direct examination and researcher discretion. More specifically, they can be determined and classified linguistically, according to grammatical, lexical, and semantic classes, as suitable to the research data (Ladani and Desai, 2020), and this has been effectively combined with automatic part-of-speech tagging (Matrane et al., 2025; Van Huyssteen, 2021). In practice, different methods are often combined, and the final step involves evaluation and approval by a native speaker or linguistic expert, an approach supported by researchers working across diverse contexts (Agbeyangi, 2026; Dubit and Lamar, 2024; Emezue et al., 2023; Muhammad et al., 2023).

Ferilli (2021) notes that, for a given corpus, term frequency (TF), the number of times a word occurs in a document, and document frequency (DF), the number of documents in which a word appears, are two basic statistical methods that can yield acceptable results for stopword identification, even for small corpora and under-resourced languages. Statistical methods based on Zipf's Law may also incorporate normalised term frequency (NTF), inverse document frequency (IDF), and term frequency-inverse document frequency (TF-IDF) (Ladani and Desai, 2020). TF-IDF has been reported as an effective metric across languages because it identifies ubiquitous words that are likely good stopword candidates (Achsan et al., 2023; Bashir et al., 2024; Thu and Supnithi, 2023). However, some researchers have investigated alternative approaches that rely on word frequency, including Van Zaanen and Witzlack-Makarevich (2025), who

[1] Saal and Fourie (2021, p. 1120) define stopwords as: "woorde wat semanties byna geheel en al verbleek het en deur sprekers gebruik word om hulself tyd te gee om hul woorde agtermekaar te kry" [words that are semantically almost totally bleached and are used by speakers to give themself time to get their words together].

studied stopwords for Khoekhoe, a low-resource language spoken in Southern Africa (see also Dubit and Lamar (2024)).

There are several more sophisticated computational methods for determining stopwords, as well as for testing their features and determining their efficacy, including the use of word network graph features (Ali et al., 2024), probability statistics such as entropy (Lo et al., 2005), or the Poisson distribution (Baradad and Mugabushaka, 2015). While these computationally-intensive approaches have seen success in well-resourced languages, their transferability to typologically diverse, lower-resource languages remains an open challenge. In particular, frequency thresholds, document distribution, and orthographic consistency can vary widely between languages. Evaluating how statistical stopword features perform across diverse Southern African language morphologies is important in the development of new language technologies.

2.2 Stopwords in South African and other low-resource languages

Research focused on stopwords for South African languages is scarce, not necessarily focused on official languages (Van Zaanen and Witzlack-Makarevich, 2025), and, to a certain extent, outdated (Taljard and De Schryver, 2002). In many cases, stopwords are mentioned in passing and removed from a corpus before it is studied or analysed (Oriola and Kotzé, 2020; Saal and Fourie, 2021).

Emezue et al. (2023) describe an African Stopwords Project aimed at investigating, developing, and curating stopword lists for African languages. This initiative is meant to address the dearth of standardised stopword resources for African languages with low resources and proposed a systematic approach involving the collection of diverse monolingual data from multiple domains, the automatic identification of candidate stopwords using statistical techniques such as TF-IDF, entropy, information gain, and Kullback–Leibler divergence, followed by human evaluation to validate the linguistic appropriateness of the identified candidates. The project further envisages integrating the resulting resources into existing NLP toolkits such as NLTK[2] or developing a dedicated Python package. This work is significant for African languages because it demonstrates an open and collaborative approach to stopword development, combining computational frequency-based methods with human linguistic validation, and community contributions. Since its inception, this project curated stopword lists for Afrikaans, Hausa, isiZulu, Kinyarwanda, Kirundi, Kiswahili, Nigerian Pidgin English, Sesotho, Somali, and Yorùbá, while the proposed resources for Amharic, Igbo, and Shona were unavailable at the time of writing. The project adopts an open-source approach through its African Stopwords GitHub repository[3], which provides language-specific stopword lists, extraction scripts, contribution guidelines, and a discussion platform where researchers can debate potentially controversial stopword candidates, as well as expand, review, and improve stopword lists for under-resourced African languages. The project, unfortunately, does not currently seem to be active.

Van Zaanen and Witzlack-Makarevich (2025) investigated how to compile a stopword list for Khoekhoe using word-frequency patterns across different text genres. Two corpora were used: a Khoekhoe corpus comprising six genres (conversation, fiction, learned, miscellaneous, press, and religion) and the Brown Corpus (Kučera and Francis, 1967) as an English reference corpus. For each genre, the researchers ranked words by frequency and selected the top N words. The overlap between genre-specific lists was then measured using the Jaccard coefficient, rather than the more traditional TF-IDF. The findings showed that both the corpus variable (English versus Khoekhoe) and the genre pair (the two genres being compared) had statistically significant effects on the overlap of frequent words, with corpus accounting for a larger proportion of the variance than genre pair. While English displayed a relatively high degree of overlap in frequent words across genres, Khoekhoe showed considerably lower overlap. Van Zaanen and Witzlack-Makarevich (2025) suggest that this difference may be related to the typological characteristics of Khoekhoe, making the identification of stopwords through frequency-based methods more challenging. It was noted that although word frequency can contribute to the development of stopword lists, relying solely on frequency is unlikely to be sufficient for Khoekhoe, and additional techniques may be required to develop an effective stopword list for this low-resource language.

[2] https://www.nltk.org/

[3] https://github.com/masakhane-io/masakhanePreprocessor/tree/main/african-stopwords

2.3 Linguistic background

As discussed in Section 2.1, there are numerous definitions of stopwords as well as approaches to identifying them. South African languages offer an interesting additional dimension of investigation. They represent various language families with different linguistic characteristics (Germanic, Sotho, Nguni, and others), but these languages also use different orthographic writing styles (disjunctive versus conjunctive), both of which might influence how to best identify stopwords. Furthermore, South African languages are generally considered low-resource, meaning the amount of available data is limited. When using statistical, data-driven methods, limited resources can pose an additional challenge.

Next to English, the current investigation includes three of the twelve official languages of South Africa:

- Afrikaans, which is a Germanic (fusional) language similar to English, but with very productive compounding;
- isiZulu, a Nguni language with high agglutination and a conjunctive writing style, which means that words tend to be longer, less frequent, and more complex morphologically;
- Sesotho, a Sotho language with disjunctive orthography, resulting in many shorter, highly ambiguous words.

By applying the same statistical measures and combinations of measures to these different languages, we hope to identify successful strategies for the identification of stopwords in all these linguistically diverse languages.

3 Methodology

The aim of this research is to help develop stopword lists for languages that do not yet have such lists. Here, we investigate whether features extracted from a corpus can help with identifying (at minimum, a preliminary list of) stopwords. To evaluate our approach, we use existing stopword lists, and to better understand the impact of orthographic as well as linguistic properties on the usefulness of the features, we test our approach on four different languages, namely English (for initial evaluations), Afrikaans, isiZulu, and Sesotho (see Section 2.3 for more detail). We begin with an existing general reference corpus and stopword list for each language and extract statistical word features (word length, TF, NTF, DF, IDF, the log-transformed values of these metrics, and various forms of TF-IDF). Based on these features, we train statistical models to identify stopwords automatically. We then evaluate the models against existing stopword lists to determine which features contribute most to the identification of stopwords.

3.1 Data

3.1.1 Corpora

For English, the Brown Corpus (the Standard Corpus of Present-Day Edited American English) was used to compute the features. It was the first computer-readable general text corpus, compiled by Kučera and Francis (1967). It contains 933,698 words (after filtering punctuation and words containing numbers) of edited US English prose printed in 1961 and covers fifteen genres, e.g., press reportage, editorials, religion, skills and hobbies, and science fiction. The corpus data is divided into 500 documents with an average of 1867.4 (standard deviation (SD): 57.4) filtered words. Even though it is an older corpus, its structure lends itself to testing different TF-IDF variants that require document-level data linked to word statistics (as described in Section 3.2).

For the three South African languages, we extracted articles from the 2025 Wikipedia archive.[4] Wikipedia is a collaborative encyclopedia covering a wide range of concepts and domains, and it is accessible to both humans and computers. This makes it a good general-purpose corpus (Mehdi et al., 2017). Each article is interpreted as an individual document. The articles were extracted from the XML version, converted to plain UTF-8 text documents, and punctuation and numerals were removed. Table 1 provides an overview of the four corpora, including their source, language, number of documents included, total word count, and the mean and SD for the number of words per document.

3.1.2 Stopword lists for evaluation

To evaluate the effectiveness of different features and feature combinations for identifying (potential) stopwords, we use existing stopword lists as Gold standard data. For English, we opted to use the standard NLTK stopword list with 127 terms (Bird

[4] https://dumps.wikimedia.org/

Table 1: Overview of corpora used, total number of documents, words per corpus, and mean and SD of the number of words per document (after removing punctuation and words containing numbers) for each of the languages.

Language	Source	Documents	Words	Mean	SD
English	Brown Corpus	500	933,698	1867.4	57.4
Afrikaans	Wikipedia 2025	108,313	25,039,146	231.2	625.7
isiZulu	Wikipedia 2025	12,037	539,699	44.8	144.7
Sesotho	Wikipedia 2025	1,630	282,781	173.5	610.7

Table 2: Overview of the number of words on the stopword lists used for evaluation in our experiments per language and source.

Language	Source	# Stopwords
English	NLTK	127
Afrikaans	Tatman (2017)	51
	Van Huyssteen (2021)	253
isiZulu	Tatman (2017)	29
Sesotho	Tatman (2017)	31

et al., 2009), a widely used resource in the NLP community.

As discussed in Section 2.2, not many stopword lists are available for South African languages. One collection of stopword lists for South African languages is that collected by Tatman (2017) using the African Storybook Project[5] as a corpus to extract frequency lists and identify stopwords (also cited in Emezue et al. (2023)). Unfortunately, no further detail is given on the criteria used to add a word to the stopword list. African Storybook contains children's picture storybooks aimed at improving early literacy. This might limit the lexical diversity and richness of the source corpus. The resulting lists contain a small number of stopwords: 51 for Afrikaans, 31 for Sesotho, and 29 for isiZulu (see Table 2).

For Afrikaans, Van Huyssteen (2021) presents an alternative source with four different types of stopword lists based on the Taalkommissie corpus (Taalkommissie van die Suid-Afrikaanse Akademie vir Wetenskap en Kuns, 2011), a stratified corpus of 60 million words featuring a wide variety of genres and domains. The selection criteria used to identify stopwords include, among others, word length, word frequency, part-of-speech information, and Zipf value, as well as qualitative curation. As an additional evaluation list, we use the comprehensive list of 253 stopwords of Van Huyssteen (2021), which includes some non-Afrikaans terms.

[5] https://www.africanstorybook.org

3.2 Experimental setup and features

The experiments described here consist of two steps. First, for each word in the corpus, we compute 25 features, namely nine main features as listed in Table 3 and sixteen pair-wise products of the term frequency (tf_n, tf_l, ntf_n, and ntf_l) and document frequency (df, log_df, idf, and log_idf) features as interaction measures. In addition to the features, we include information on whether the word can be found in the stopword list under consideration as a target variable. Second, we create logistic regression models with these features. Varying the features used and employing elastic net logistic regression (Zou and Hastie, 2005) allows us to test for the contributions (or effectiveness) of the different features to identify stopwords. The evaluation metric initially used is the Area Under the Receiver Operating Characteristic Curve (AUC), which gives an indication of a classification model's ability to distinguish between classes, in this case, separating stopwords from non-stopwords (Fawcett, 2006). However, after finding the most suitable features, we provide the more detailed precision, recall, and F1-scores of the best-performing models.

After identifying the most effective model, we then also compare the results of the models with the same features on the different languages. This provides insight into the stability of the features between the languages.

3.3 Feature selection on English data

The feature selection process consists of two phases. First, we analyze all models on the English Brown corpus and the corresponding NLTK stopword list. Based on the results of these analyses, we apply the settings of the best models (based on the English data) to the three African language corpora and stopword lists.

For the English models, we fit elastic net logistic

Table 3: Overview of the nine main-effect features.

Feature	Description
length	Number of characters in a word
tf_n	Word frequency in the entire corpus as an absolute number
tf_l	Log of tf_n
ntf_n	Word frequency normalised over the total number of words in the corpus
ntf_l	Log of ntf_n
df	Number of documents with the word divided by the total number of documents
log_df	Log of df
idf	Total number of documents divided by the number of documents with the word
log_idf	Log of idf

regression models (with the α mixing parameter set to 0.5) to predict stopword status based on the list of the nine main-effect features expanded with sixteen interaction features (the multiplication of the four word frequency and the four document frequency features) (Table 3). Many of these features are collinear, since they are transformations or rescalings of the same underlying frequencies, and the interaction terms are products of them. Elastic net regression is suitable as the ridge component shrinks the coefficients of correlated features together and stabilises their estimates. The lasso component can set redundant coefficients to zero. As the elastic net penalty is sensitive to the scale of the features, all features are standardised internally by the glmnet package and return coefficients on the original scale.

As stopwords only constitute a small portion of the words in the corpus (0.32%), stratified five-fold cross-validation is used to preserve the class ratio within each fold, and inverse-frequency class weighting is applied during model fitting to counteract this imbalance (Kuhn and Johnson, 2013). Feature sets were compared by their mean cross-validated AUC across ten seeds, each seed producing a different random stratified fold assignment; for each model we used the maximum of the mean cross-validated AUC over the λ path.

A primary model containing all nine main-effect features is compared against the full model, which also includes the sixteen interaction terms. The simpler primary model outperforms the full model slightly on cross-validated AUC (99.22 and 99.21, respectively). Inspection of the fitted coefficients shows that both models converge on the same underlying features, particularly the word length feature, a frequency measure, and a document-frequency measure. (The weights of the models are provided in Table 8 in Appendix A.) The full model distributes this signal across interaction features rather than revealing additional structure (which is likely due to the fact that the interaction features do not provide any inherently new information with respect to the main effect features). In view of this, subsequent analyses use only the primary model containing the nine main-effect features.

Focusing on the nine main-effect features, six combinations of these features are compared, including raw, normalised, and log-transformed versions of the frequency and document-frequency features (see Table 4). Predictive performance is highly consistent across combinations, with the AUC ranging narrowly between 99.04 and 99.22. Models including normalised or log-transformed features consistently match or somewhat exceed their raw counterparts, therefore prompting a further search for the smallest feature set which retains comparable performance.

To find the smallest feature set that still performs well, five further reduced feature combinations are compared: three models with two features each and two models with three features each. As we can see in Table 5, a three-feature model including word length, log term frequency, and log document frequency, referred to as the core model, performed equivalently to the nine-feature primary model (AUC 99.20 versus 99.22) and to a variant substituting inverse document frequency for a log version (model core_idf 99.21). Removing length (model tf_log_df_log) reduced performance a bit more (AUC 99.10), suggesting that length should be preserved as a feature.

Table 4: Six combinations of the main effect predictors with mean AUC.

Model	Features	AUC
primary	length + tf_n + tf_l + ntf_n + ntf_l + df + log_df + idf + log_idf	99.22
raw_normalised	length + tf_n + tf_l + ntf_n + ntf_l + df + log_df	99.19
normalised	length + ntf_n + ntf_l + df + log_df	99.19
raw_log	length + tf_n + tf_l + df + log_df	99.19
log	length + tf_l + log_df	99.20
basic	length + tf_n + df	99.04

Table 5: Reduced feature combinations with mean AUC.

Model	Features	AUC
core	length + tf_l + log_df	99.20
core_idf	length + tf_l + idf	99.21
length_tf_log	length + tf_l	99.18
length_df_log	length + log_df	99.17
tf_log_df_log	tf_l + log_df	99.10

Table 6: Mean cross-validated AUC for the core versus primary models, by language.

Language	Core	Primary
English	99.20	99.22
Afrikaans (Tatman)	99.99	99.99
Afrikaans (van Huyssteen)	98.62	98.59
isiZulu	99.65	99.64
Sesotho	99.64	99.78

4 Results, evaluation and discussion

After the feature selection on the English data, we now report our results. As described above, both the primary model with all nine main features as well as the core model with the three most predictive features perform well with AUC 99.22 and 99.20 for English and thus succeed at identifying stopwords. For Afrikaans, isiZulu, and Sesotho, the primary (nine features) and core (three features) models were run for AUC comparison. The core model performed equivalently to the primary model across all languages, with minor differences in the mean AUC (see Table 6).

The AUC result of 99.99 for the Tatman (2017) Afrikaans stopword list obtained should be interpreted with caution. Inspection against corpus frequency rank showed that 40 out of the 50 stopwords fell within the 150 most frequent word types, with the remaining stopwords distributed over a long tail (up to rank 873). The stopword list is therefore not defined by a strict frequency cutoff, but due to it being heavily weighted toward the most frequent words in the corpus, the frequency-derived features are highly predictive of list membership. This near-perfect performance should not be interpreted as evidence of a more coherent stopword criterion; rather, it reflects the compositional properties of the list itself.

With the low prevalence of stopwords in the data, the AUC metric alone is not an optimal predictor of performance. Therefore, nested out-of-fold evaluations of the three-features core model are run. For each of ten random seeds, words were divided into five stratified outer folds. Within each outer fold, λ was selected using five-fold inner cross-validation on the training data only, and predictions were generated for the held-out fold. These predictions were pooled to obtain one out-of-fold probability per word per seed. ROC-AUC, precision, recall and F1-score were then calculated. Precision, recall and F1-score were calculated using a probability cutoff of 0.5: words with predicted probabilities ≥ 0.5 were classified as stopwords, and those with probabilities < 0.5 as non-stopwords. Inverse-frequency class weighting was used during model fitting to address class imbalance. The random seeds determined the fold assignments, and reported values are means across the ten seeds. Table 7 presents the results for each language and stopword list.

The mean for the English core model[6] across the ten seeds was 6.14 for precision, 93.46 for recall, and 11.53 for the F1-score (see Table 7). This indicates that most words flagged by the model as stopwords were false positives (low precision), but at the same time, the model succeeds in identifying most true stopwords (high recall). As with the Brown corpus for English, recall remained con-

[6]Note that the results in Table 7 are calculated on unseen, held-out data, compared against the results in the earlier tables that show the maximum of the mean cross-validated error for each λ value, which leads to optimistic results.

Table 7: Out-of-fold performance of the core model, mean across ten seeds, by language.

Language	AUC	Precision	Recall	F1-score
English	98.67	6.14	93.46	11.53
Afrikaans (Tatman)	99.99	4.03	100.00	7.75
Afrikaans (van Huyssteen)	98.20	0.95	91.36	1.87
isiZulu	99.50	0.87	95.17	1.73
Sesotho	99.38	5.09	93.55	9.66

sistently high across all three African languages (91.36–100), while precision was low throughout and notably lower than English for isiZulu and the Afrikaans Van Huyssteen (2021) stopwords (<1). Considering the very small number of stopwords for Sesotho (n = 31) and isiZulu (n = 29), these precision estimates should be treated cautiously. The pattern of high recall and low precision, however, is consistent with the low prevalence of stopwords for the different languages, therefore suggesting that the models might be better suited to ranking candidate stopwords for manual review than to unsupervised binary classification (Kumova Metin and Karaoğlan, 2017).

Overall, the statistical models for stopword features tested in this study behave consistently for the four languages studied. This has several implications. First, it confirms the influence of several general stopword features widely reported in the existing literature, including word length, word frequency, and document frequency (de Keyser, 2012; Fox, 1989; Kucukyilmaz and Akin, 2024; Lo et al., 2005). These relatively basic features already help finding potential stopwords, but the features are not particularly strong in that many non-stopwords are marked as stopwords.

Second, the distinguishing features are consistent across the different language corpora tested. This indicates that even the corpus of the smallest size, the Sesotho corpus (±280,000 words), is sufficient to yield reliable results for identifying stopwords. A reason for this may be that most stopwords are found in the head of the Zipfian curve. A small corpus already provides a reliable head, making the identification of stopwords in that part of the curve possible.

Third, the most important features are also consistent across languages. This suggests, surprisingly, that the orthography of the language is not an important factor in the approach. Whereas orthography can be expected to have a major impact on the Zipfian curve of the language (e.g., disjunctive languages will likely have short, non-content words in the head of the Zipfian curve, while conjunctive languages will most likely already have content words in the head of the curve), for stopwords this does not seem to have a major impact. However, it may also be that in this case the document frequency feature has a larger effect as stopwords are typically found across documents.

5 Conclusion

In this article, we presented an investigation into finding word statistics in a corpus that allow us to automatically identify stopwords for English, Afrikaans, isiZulu, and Sesotho. In total, 25 features (nine single features and sixteen combined features) were tested with logistic regression models to identify the best predictive features for stopwords. The study established that stopwords represent a minority group within natural corpora in terms of the number of unique words. Standard evaluation and fitting routines on such a distribution will be skewed towards the majority class. The research found that this can be managed with an elastic net approach combined with inverse-frequency class weighting preferential towards stopwords, and stratified cross-validation across multiple rounds.

Our investigation showed that (combinations of) the selected nine single features are useful predictors, and a model containing only three features, namely word length, a word frequency measure, and a document-frequency measure, yield good identification results. Furthermore, both for English and the three South African languages, Afrikaans, isiZulu, and Sesotho, when using the same features the models produce similar results. This is regardless of the amount of data available (corpus size) as well as the type of orthography. For all four languages, recall (identification of all stopwords) is high, but precision (choosing cor-

rect stopwords) is rather low. This means that the models can be used as a filter to identify potential stopwords, but for the final selection, human validation is needed.

Future work can extend this research by investigating other statistical measures, such as Kullback-Leibler divergence, entropy, or network graphs, to possibly improve the precision when applying statistical models to corpora for automatic stopword identification.

References

Harry Tursulistyono Yani Achsan, Heru Suhartanto, Wahyu Catur Wibowo, Deshinta A. Dewi, and Khairul Ismed. 2023. Automatic extraction of Indonesian stopwords. *International Journal of Advanced Computer Science and Applications*, 14(2):166–171.

Abayomi O Agbeyangi. 2026. Text detoxification in isiXhosa and Yorùbá: A cross-lingual machine learning approach for low-resource African languages. *arXiv preprint arXiv:2601.05624*.

Felermino DMA Ali, Gabriel de Jesus, Henrique Lopes Cardoso, Sérgio Nunes, and Rui Sousa-Silva. 2024. Network-based approach for stopwords detection. In *Proceedings of the 16th International Conference on Computational Processing of Portuguese*, volume 2, pages 55–63.

Vicenç Parisi Baradad and Alexis-Michel Mugabushaka. 2015. Corpus specific stop words to improve the textual analysis in scientometrics. In *Proceedings of ISSI*, pages 999–1005.

Abubakar Salisu Bashir, Abdulkadir Abubakar Bichi, and Alhassan Adamu. 2024. Automatic construction of generic Hausa language stop words list using term frequency-inverse document frequency. *Journal of Electrical Systems and Information Technology*, 11(58).

Steven Bird, Ewan Klein, and Edward Loper. 2009. *Natural language processing with Python: analyzing text with the natural language toolkit*. O'Reilly Media, Inc.

Piet de Keyser. 2012. Techniques applied in automatic indexing of text material. In *Indexing*, Chandos Information Professional Series, chapter 3, pages 65–78. Chandos Publishing.

Gene Diaz. 2016. Stopwords ISO project.

Rachel E. Dubit and Annie K. Lamar. 2024. Era- and genre-specific stop word lists for low-resource computational research: A classical Latin exemplum. *Journal of Open Humanities Data*, 10.

Chris Emezue, Hellina Nigatu, Cynthia Thinwa, Helper Zhou, Shamsuddeen Muhammad, Lerato Louis, Idris Abdulmumin, Samuel Oyerinde, Benjamin Ajibade, Olanrewaju Samuel, Oviawe Joshua, Emeka Onwuegbuzia, Handel Emezue, Ifeoluwatayo A. Ige, Atnafu Lambebo Tonja, Chiamaka Chukwuneke, Bonaventure F. P. Dossou, Naome A. Etori, Mbonu Chinedu Emmanuel, and 3 others. 2023. The African stopwords project: curating stopwords for African languages. *arXiv preprint arXiv:2304.12155*.

Tom Fawcett. 2006. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874.

Stefano Ferilli. 2021. Automatic multilingual stopwords identification from very small corpora. *Electronics*, 10(17):2169.

Christopher Fox. 1989. A stop list for general text. *SIGIR Forum*, 24(1–2):19–21.

Sonam Gandotra and Bhavna Arora. 2018. Functional words removal techniques: A review. In *Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)*, pages 546–550. IEEE.

Jashanjot Kaur and P Kaur Buttar. 2018. A systematic review on stopword removal algorithms. *International Journal on Future Revolution in Computer Science & Communication Engineering*, 4(4):207–210.

Tayfun Kucukyilmaz and Tayfun Akin. 2024. A feature-based approach on automatic stopword detection. In *Intelligent Systems and Applications (IntelliSys)*, pages 51–67. Springer Nature Switzerland.

Max Kuhn and Kjell Johnson. 2013. Remedies for severe class imbalance. In Max Kuhn and Kjell Johnson, editors, *Applied Predictive Modeling*, pages 419–443. Springer New York.

Senem Kumova Metin and Bahar Karaoğlan. 2017. Stop word detection as a binary classification problem. *Anadolu University Journal of Science and Technology A - Applied Sciences and Engineering*, 18(2):346–359.

Henry Kučera and W. Nelson Francis. 1967. *Computational analysis of present-day American English*. Providence, Brown University Press, Providence, USA.

Dhara J. Ladani and Nikita P. Desai. 2020. Stopword identification and removal techniques on TC and IR applications: A survey. In *6th International Conference on Advanced Computing and Communication Systems (ICACCS)*, pages 466–472. IEEE.

Rachel Tsz-Wai Lo, Ben He, and Iadh Ounis. 2005. Automatically building a stopword list for an information retrieval system. *Journal on Digital Information Management*, 5:17–24. Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR).

Masoud Makrehchi and Mohamed S Kamel. 2008. Automatic extraction of domain-specific stopwords from labeled documents. In *European Conference on Information Retrieval*, pages 222–233. Springer.

Yassir Matrane, Faouzia Benabbou, Zineb Ellaky, and Chaimae Zaoui. 2025. An automatic stop words removal in Maghrebi Arabic dialect text classification using part of speech tagging. In *Innovations in Smart Cities Applications*, volume 8, pages 187–196. Springer Nature Switzerland.

Mohamad Mehdi, Chitu Okoli, Mostafa Mesgari, Finn Årup Nielsen, and Arto Lanamäki. 2017. Excavating the mother lode of human-generated text: A systematic review of research that uses the Wikipedia corpus. *Information Processing and Management*, 53(2):505–529.

Shamsuddeen Hassan Muhammad, Idris Abdulmumin, Abinew Ali Ayele, Nedjma Ousidhoum, David Ifeoluwa Adelani, Seid Muhie Yimam, Ibrahim Sa'id Ahmad, Meriem Beloucif, Saif Mohammad, Sebastian Ruder, Oumaima Hourrane, Pavel Brazdil, Alípio Jorge, Felermino Dário Mário António Ali, Davis David, Salomey Osei, Bello Shehu Bello, Falalu Ibrahim, Tajuddeen Gwadabe, and 8 others. 2023. Afrisenti: A Twitter sentiment analysis benchmark for African languages. In *Proceedings of the 2023 conference on Empirical Methods in Natural Language Processing*, pages 13968–13981. Association for Computational Linguistics.

Oluwafemi Oriola and Eduan Kotzé. 2020. Evaluating machine learning techniques for detecting offensive and hate speech in South African tweets. *IEEE Access*, 8:21496–21509.

Martin F Porter. 2001. Snowball: A language for stemming algorithms.

E. Saal and A. Fourie. 2021. "dit en dat" en "al daai goed": 'n korpuslinguistiese ondersoek na pragmatiese merkers in gesproke Afrikaans. *Tydskrif vir Geesteswetenskappe*, 61(4-1):1112–1128.

Serhad Sarica and Jianxi Luo. 2021. Stopwords in technical language processing. *PLOS ONE*, 16(8):e0254937.

Taalkommissie van die Suid-Afrikaanse Akademie vir Wetenskap en Kuns. 2011. Taalkommissiekorpus 1.1. Technical report, Centre for Text Technology (CText), North-West University, Potchefstroom, South Africa.

Elsabé Taljard and Gilles-Maurice De Schryver. 2002. Semi-automatic term extraction for the African languages, with special reference to Northern Sotho. *Lexikos*, 12(1):44–74.

Rachael Tatman. 2017. Stopword lists for African languages.

Ye Kyaw Thu and Thepchai Supnithi. 2023. Embedding meets frequency: Novel approaches to stopword identification in Burmese. In *18th International Joint Symposium on Artificial Intelligence and Natural Language Processing (iSAI-NLP)*, pages 1–6. IEEE.

Houcemeddine Turki, Naome A Etori, Mohamed Ali Hadj Taieb, Abdul-Hakeem Omotayo, Chris Chinenye Emezue, Mohamed Ben Aouicha, Ayodele Awokoya, Falalu Ibrahim Lawan, and Doreen Nixdorf. 2024. Text categorization can enhance domain-agnostic stopword extraction. In *International Conference on Language Processing and Knowledge Management*, pages 122–132. Springer.

Gerhard Van Huyssteen. 2021. Afrikaans stoplist for corpus research 1.0.

Menno Van Zaanen and Alena Witzlack-Makarevich. 2025. Stop words in Khoekhoe. *Journal of the Digital Humanities Association of Southern Africa (DHASA)*, 6(2).

W. John Wilbur and Karl Sirotkin. 1992. The automatic identification of stop words. *Journal of Information Science*, 18(1):45–55.

Hui Zou and Trevor Hastie. 2005. Regularization and variable selection via the elastic net. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 67(2):301–320.

A Appendix

Table 8: Elastic net logistic regression coefficients: primary and full models.

Feature	Type	Primary	Full
	intercept	-1.681	0.426
length	main-effect	-0.326	-0.316
tf_n	main-effect	0	0
tf_l	main-effect	0.261	0
ntf_n	main-effect	0	0
ntf_l	main-effect	234,784.4	0
df	main-effect	7.090	5.698
log_df	main-effect	-0.164	0.278
idf	main-effect	-0.0177	-0.0128
log_idf	main-effect	0.0990	-0.263
tf_n*df	interaction	–	0
tf_n*log_df	interaction	–	0.0000681
tf_n*idf	interaction	–	0
tf_n*log_idf	interaction	–	-0.0000815
tf_l*df	interaction	–	0.277
tf_l*log_df	interaction	–	-0.0397
tf_l*idf	interaction	–	0
tf_l*log_idf	interaction	–	0.0237
ntf_n*df	interaction	–	0
ntf_n*log_df	interaction	–	64.329
ntf_n*idf	interaction	–	0
ntf_n*log_idf	interaction	–	-72.027
ntf_l*df	interaction	–	263,496.6
ntf_l*log_df	interaction	–	-70,083.24
ntf_l*idf	interaction	–	0
ntf_l*log_idf	interaction	–	41,176.50