Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

A Feature Based Approach on Automatic Stopword Detection

Tayfun Kucukyilmaz (kucukyilmaz@rsm.nl)
Erasmus University
Tayfun Akin
TED University

Research Article

Keywords: Stopword, Automatic Stopword Detection, Supervised Learning, Gradient Boosted Decision Trees, Machine Learning

Posted Date: August 29th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-1986294/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

A Feature Based Approach on Automatic Stopword Detection

Tayfun Kucukyilmaz [1*] and Tayfun Akin[2†]

[1*]Rotterdam School of Management, Erasmus University, Netherlands.
[2]Department of Computer Science, TED University, Turkey.

*Corresponding author(s). E-mail(s): kucukyilmaz@rsm.nl;
Contributing authors: tayfun.akin@gmail.com;
†These authors contributed equally to this work.

Abstract

In information retrieval, words that contribute little to the semantic information within a text are called stopwords. Identification of stopwords is considered an asset for many retrieval tasks as removal of such words from the text collection increase the regularization within the dataset and reduce the volume and computational complexity. Traditional techniques for identification of stopwords either involve manual construction of stopword lists by analyzing terms individually or employing sorting-based techniques that uses global term characteristics as a proxy. Due to the context-dependant and non-precise definition of what really classifies as a stopword, transfer and application of these traditional methodologies to different languages or domains leaves room for generalized and interpretable results. To address this concern, we propose a feature based supervised machine learning technique for automatic detection of stopwords. We have tested the validity of the proposed technique with extensive experiments and compared the results with a general English stopword list. Furthermore, we have evaluated the proposed technique with both formal written and social media text. The results reveal that the proposed technique leads to promising results and is also capable of addressing the changes in the dialect.

Keywords: Stopword, Automatic Stopword Detection, Supervised Learning, Gradient Boosted Decision Trees, Machine Learning

1 Introduction

In information retrieval, a commonly used representation for documents is the *Bag of Words* (BoW) model. In this model, a document is represented by the terms contained in the document, and term is positioning is usually disregarded. Elimination of words that contribute little to the semantic information within a text is a frequent practice in many retrieval tasks that use the BoW model. The motivation behind removal of such words is intuitive: removing the need of allocating computational power for less informative tokens will result in increased speed and efficiency for a variety of applications [1, 2]. Terms that are used to support the grammatical formation of a sentence in natural languages are usually referred to as "stopwords" and are considered as the best candidates for such removal.

Traditionally, there are two commonly used approaches for identifying stopwords in a language: employing hand-crafted compilation of term listings [3], or classifying terms via their global usage characteristics [4, 5]. Although the former method produce reliable results, such list compilations are very expensive and only available for frequently researched languages such as English, French, and German. Such listings are usually not available for low-resource languages. Additionally, such compilations are often static, and may become outdated due to the natural evolution of a language. The latter strategy involves characteristics of a specific language, and may not perform equally well with different languages. Although the literature survey reveals that using global term characteristics in different languages indicate promising results, the reasons of successful application is often not addressed. In this study, we propose a novel approach for automatic stopword detection. To this end, we propose a feature-based supervised classification technique. The refer to this proposed technique as the "Gradient Boosted Stopword Detection" (GBSD) as future reference in this paper. By using an extensive set of representative features, we model the problem of detecting stopwords as a classification task.

Although stopword detection is often considered in conjunction with a BoW model, using only global term characteristics for stopword identification has its own pitfalls. As BoW models often disregard relative term positioning in a text collection and such models fail to examine the availability of any intra-sentence information that can be used in the stopword detection problem. One of the contributions of this study is the additional use of sentence-level features in the proposed machine learning technique. As discussed earlier, terms that are characterized as stopwords are usually considered as grammatical facilitators within a sentence. Motivated by this fact, we hypothesize that the relative positioning of words may contain detectable patterns that can be used for stopword classification. To test this hypothesis, we propose using a set of novel features describing the associations of terms within a sentence.

The contributions of this study are as follows:

- We propose using a supervised machine learning model for automatic detection of stopwords.

- We propose using a set of novel, sentence-based features.
- By analyzing the proposed approach on both formal document collections and social media data, we examined the effects of evolution of a language on stopword detection.
- We examined some constructs commonly used in digital social media and evaluated their candidacy for being stopwords.

The rest of the paper is organized as follows: We provide a detailed literature survey of the related work in Section 2. In Section 3, we described the datasets and the extracted features that are used in this study. Section 4 presents our framework for solving the automatic stopword detection problem. Section 5 provides the results of a large number of experiments to evaluate the feasibility of the proposed approach. Finally, in section 6 we finalise the paper with a concluding discussion.

2 Related Work

Traditionally, stopword identification is as a sorting problem: ordering the words in the vocabulary of a text collection with respect to some attribute reflecting term usage and marking highly ranked terms as stopwords. Using attributes such as term frequency (TF) and normalized term frequency (NTF) are among the most popular techniques employed for the stopword detection problem [6]. Integrating term importance to this formulation lead to use of document-based metrics such as inverse document frequency (IDF) [7]. There are also several studies that utilised a combination of multiple attributes, integrating both TF and IDF into a single metric [8, 9]. Other techniques include basic voting systems that aggregate the results of a multiple sorted lists [10]. While the attributes used in the literature show a lot of variation, most studies employ the single attribute, sorting-based strategy. One of the important drawbacks of using such techniques is that, the observations are limited to only a single aspect of the language behaviour.

Another important drawback of attribute based models is the selection criteria for stopwords. In studies that utilize these metrics, the resulting lists were created by marking a predetermined number words in the vocabulary as stopwords [11]. Such a strategy is usually not suitable for practical applications as the selection of the number of stopwords can have significant effects on the quality of the results, and requiring prior knowledge about the text collection to be accurate. Although applying machine learning techniques would also suffer from problems such as the skew in the data, such techniques would not require prior knowledge during setup, ultimately being more suitable for generalization and practical application.

Some studies also implement vector space and similarity-based strategies [12] where, similarity based comparisons are used to determine the classification of a term in a supervised setting. The implications of the aforementioned skewness problem also leads to unorthodox applications even for attribute- and sorting-based strategies, such as requiring manual editing of the text collections

prior to list generation [13]. This is highly undesirable for practical results and low-resource language studies, as it hinders the possibility of generalization and autonomy.

A survey into the past literature also reveals that there have been a few studies that apply machine learning for the stopword detection problem [14]. However, due to complications such as the skewness of the data and under-representation of stopwords, these applications only have marginal success compared to attribute and sorting-based techniques.

Although many studies in the literature agree that frequency-based stopword detection consistently produce high quality results, many studies examine the performance of stopword detection under the influence of various attributes. To the best of our knowledge, the highest number of attributes using in a single study is 6 where, the authors compare the performance of these attributes in a sorting-based single attribute stopword detection scenario [10].

To improve the representable capability of the proposed machine learning technique, we included 12 different attributes in this study. To this end, we also proposed sentence-based features where, the relative position of terms in relation to grammatical artifacts in a sentence is also considered as a candidate attribute for stopword detection problem. Table 1 presents a summary of the attributes deployed in the literature. In the table, the adopted attributes are presented with abbreviations. These abbreviations and a detailed description of these attributes will be presented in Section 3.2.

Table 1 A comparison of features used in related studies.

	TF	IDF	TF.IDF	MP	VP	E	IM	WP
Zou et al. (2005)	✓	✓	✓	✓	✓	✓		
Rani et al. (2011)	✓	✓	✓				✓	
Medhat et al. (2012)	✓							
Lo et al. (2013)	✓	✓						
Alajimi et al. (2013)	✓			✓	✓	✓		
Chekima et al. (2015)					✓	✓		
Kumova et al. (2017)	✓		✓					
GBSD	✓	✓		✓	✓	✓	✓	✓

3 Dataset and Features

We conduct extensive experiments on three realistic datasets to evaluate the performance of the proposed technique. In this section, we first introduce the text collections that we have used in this study. Then, we present the features that are extracted from these text collections in order to represent the terms in the proposed models.

3.1 Dataset

We have used three text datasets in this study. First, in order to examine the results of stopword detection problem on a more formal setting, we have collected the text from a classic book "Tom Sawyer" [15]. In this study, we also try to examine the impact of digital communications on the stopword detection problem. To this end, we have used two text collections representing text from various social media outlets. These datasets are referred to as "Twitter" and "Reddit" text collections in the study. The Twitter dataset was gathered from [16] and contains generic tweets with no contextual limitations. The Reddit dataset was gathered from [17], which includes posts from 38 different SubReddits, who's topics range from chatting and hobbies to science and professions.

Table 2 summarize statistical information about the datasets used in this study. In the Table, the *#Sentences* column represent the unit of information, i.e., a document, in the dataset. This definition of a document is required for the calculation of many of the features in this work, and will be used later, in Section 3.2. For all datasets, the unit of information is considered as a sentence. The table also summarize the total number of distinct terms and number of terms that are considered as stopwords in each dataset. In this work, we establish the ground truth for stopwords by using the stopword list in the NLTK library [18]. The respective column in the table identifies the availability of these words in the respective datasets.

We have also performed several preprocessing operations to each of the datasets. Sentences that are shorter than 2 words are eliminated. Non ASCII characters and trailing whitespaces are also eliminated while special markers such as hashtags, mentions, links and emojis are tokenized for regularization.

Table 2 Summary of the datasets and their composition.

Collection Tag	# Sentences	# Distinct Terms	# Stopwords
Twitter	2.434.494	102.885	146
Reddit	194.904	53.555	146
Tom Sawyer	4.681	7.789	139

3.2 Features

As the statistical representations of a term, we used several features. This study has adopted a set of features from the literature as well as proposing a set of new and innovative features. Table 1 summarize the attributes used in this study and their past adaptations. For each term in the text collection, the below mentioned features are extracted, and each term is represented with a vector of features. In the term vectors, each feature, in normalized form, constitutes to a separate dimension. Below, these features are described in more detail:

- **Frequency (F):** Frequency represents the number of occurrences of a term in a collection. It is one of the most commonly used features in the literature.

- **Term Frequency (TF):** TF is a document-specific representation of frequency. Unlike other features, TF values are calculated separately for each document and serve as a fundamental component of many other features. Equation 1 provides a formal definition of TF where, i, j, and M_j represents the ith term, j^{th} document, and the total document size of the jth document, respectively.

$$TF_{i,j} = \frac{F_{i,j}}{M_j} \tag{1}$$

- **Inverse Document Frequency (IDF):** IDF is a measure that evaluates the consistency of a terms usage across multiple documents. Low values of IDF are associated with common use of a word while high values are associated with specialized use. In that regard, IDF is a widely adopted feature for the stopword detection problem. Equation 2 describes IDF where, N is the total number of documents and N_i is the total number of documents that the term i can be found in.

$$IDF_i = \frac{N}{N_i} \tag{2}$$

- **Mean Probability (MP):** MP represents the probability of a term occurring in a given document, and is the mean of the TF values across multiple documents. Equation 3 summarize MP where, $TF_{(i,j)}$ is the term frequency of the term i in document j, N is the total number of documents, and N_i is the total number of documents that the term i appears.

$$MP_i = \frac{\sum_{j=0}^{n_i} TF_{i,j}}{N} \tag{3}$$

- **Variance Probability (VP):** VP of a term represents the variation of a term's use over multiple documents. A higher VP value indicates an inconsistent usage of the term while a lower value indicates a consistent use pattern. Equation 4 presents the definition of VP where, $TF_{(i,j)}$ is the term frequency of term i in document j, and $\overline{TF_i}$ is the mean of the TF_i and N is the total number of documents in a collection.

$$VP_i = \frac{\sum_{j=0}^{n_i} (TF_{i,j} - \overline{TF_i})}{N} \tag{4}$$

- **Entrophy (E):** Entropy of a word is used to predict the informativeness of a term. E represents the amount of irregularity and unpredictability of a term's use. We expect the words with lower importance/informativeness to have a high entropy value. Equation 5 represent E where, $TF_{(i,j)}$ is the term frequency of term i in document j.

$$E_i = \sum_{j=0}^{N_i} TF_{i,j} log(\frac{1}{TF_{i,j}}) \tag{5}$$

- **Information Model (IM):** Info model is a measure for the ratio of a term's occurrence relative to the total number of terms. We would expect IM to take higher values for stopwords as they are generally expected to be used more often. Equation 6 represent IM where, F is the frequency of a term M is the total number of term in the vocabulary of a collection.

$$IM_i = \frac{F_i}{M} \tag{6}$$

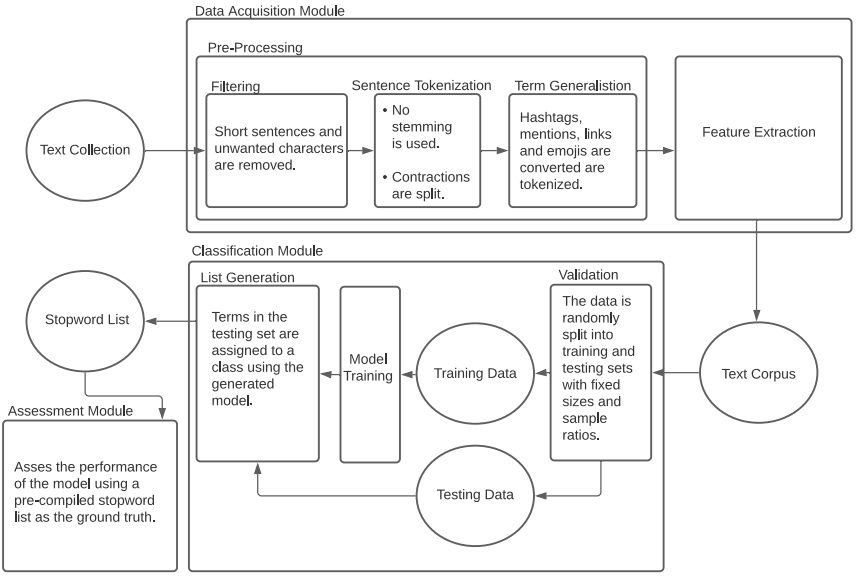
- **Word Positioning (WP):** The above features indicate that, in the past literature, frequent, common, and regular use of a term is associated with its eligibility for being a stopword. Literature also indicate that such features are employed effectively for the stopword detection problem. In this study we hypothesize that if frequent, common, and regular use of a term is indeed an indicator of a term being a stopword, sentence-level analysis of a term's use can be a used as proxy for a feature being a stopword or not. To this end, we propose using a set of novel features that we term as "Word Positioning" (WP) features. WP is a measure that signifies the distance of a term in relation to certain part-of-speech artifacts within a sentence. The part-of-speech artifacts that we use in this study are nouns, verbs, adjectives, pronouns, conjunctions, and interjections. We hypothesize that the relative positioning of a stopword with respect to such artifacts in a sentence should show less variance than terms that convey relatively more semantic information. Equation 7 summarize the calculation of WP for a specific part-of-speech artifact where, $P_{(i,k)}$ is the position of the term i in a document, P_k is the position of the part-of-speech artifacts of type k in a document, and N_i is the number of documents containing term i. Informally, WP is calculated by taking the sum of distances of a term to a specific part-of-speech artifact type, normalized by the total number of documents that contain that specific term.

$$WP_{i,k} = \frac{P_i - min(P_k)}{N_i} \tag{7}$$

4 Framework

In this section, we provide an overview of the framework we developed for solving the stopword detection problem. Figure 1 summarizes the classification procedure used in our framework. The framework consists of three modules that work in a pipelined fashion. These modules are: Data Acquisition Module, Classification Module, and Assessment Module.

Fig. 1 Framework of the experimental process



The Data Acquisition module summarizes the preprocessing operations that are performed on the text collections and feature extraction. For this study, each sentence in the collection is considered as a document. In the module, each text collection is first filtered, sentences that are shorter than 2 words are eliminated, contractions are split into multiple terms (e.g.: "don't" into "do" and "not"), and character use is standardized as detailed in Section 3. In addition, platform-specific artifacts such as hashtags, links, mentions, and emojis are converted are tokenized by using the tokenization feature of the NLTK library [18]. We would also like to note that no stemming is applied in the Data Acquisition Module.

The classification module uses the output of the Data Acquisition Module as its input. In this module, the data collection is divided into a training and testing set. This division is performed by randomly splitting the dataset into two partitions. While selecting the testing set, instead of using a percent of data, we placed some compositional restrictions on the number of selected

terms. The motivation behind this decision is that, the existence of imbalanced classes is a crucial problem in text classification [19]. If the number of representative instances selected from each class are not roughly equal in the training set, the classifiers may be biased, favoring more populated classes. To counter the imbalanced data in the training set, each stopword instance was oversampled, and inserted 10.000 times to the training data. When selecting the testing set composition, the stopword to non-stopword ratio is selected as 50/20.000, 50/20.000 and 25/2500 for the Twitter, Reddit and Tom Sawyer corpora, respectively. The remaining terms in the text collection are kept as candidates of the training set.

As our classifier model, we used a decision tree-based machine learning technique, Gradient Boosted Decision Trees (GBDT) [20]. The GBDT model provides a substantial improvement over the regular decision tree algorithm. In GBDT, each additional tree is created to as a revision over the previous trees. After the generation of each tree, the algorithm uses the outputs of the previous trees for training the next one in an iterative manner, leading to the final forest containing simple trees that cover the data efficiently and provide great coverage of the expected cases. We used the XGBoost implementation of GBDT [21], as it is a highly optimised version of the base algorithm. Table 3 summarizes the hyper-parameters used in this study.

The assessment module is responsible for evaluating the performance of the models. This is done by comparing the outcome of the models against the stopword list in the NLTK library [18] which is used as the ground truth. In this study, we compare the performance of the models using a wide range of metrics. For each experiment precision, recall, accuracy, F-score, and NDCG values are used as the performance metrics. These metrics are calculated as follows:

- **Precision:** Precision is the fraction of relevant instances among the retrieved instances.
- **Recall:** Recall is the fraction of relevant instances that were retrieved.
- **Accuracy:** Accuracy is the fraction of correct classifications over all classification attempts. We would also like to note that, due to heavy skew in the data collections, accuracy metric may often be biased towards decisions for the non-stopword label.
- **F-Score:** F-Score is a combined measure that evaluates the precision and recall together as a single numeric measure. For this purpose, harmonic mean of precision and recall is often used.
- **Normalised Discounted Cumulative Gain (NDCG):** NDCG is a measure that is commonly used in information retrieval and recommendation systems research in order to sort and order the utility of a retrieval decision [22, 23]. In this study, NDCG is adopted as measure that is used to score the confidence of decisions given by a model during classification (i.e., assigning stopword labels). Equation 8 presents NDCG where, REL_p is size of the stopword list, i is the index of the term i, and rel_i is the relevance of

Table 3 Hyper-parameters used in GBDT models.

Parameter Name	Value
colsample_bynode	0.8
learning_rate	0.3
max_depth	2
num_paralell_tree	5
objective	binary:logistic
subsample	0.8
tree_method	gpu_hist
num_boost_round	12

the term i. In our metric, relevancy is selected as 1 if a term is a stopword, and 0 otherwise.

$$NDCG = \sum_{i=1}^{REL_p} \frac{rel_i}{log(i+1)} \tag{8}$$

5 Experimental Results

We have used several baseline methods in order to compare the performance of the proposed algorithm. In these experiments, we used 3 techniques as the baseline: IDF sorted [7], Frequency sorted [10] and Entropy sorted [13]. In our experiments, the baseline methods are configured to sort collection vocabularies according to the reported metric and picking the first N terms as stopwords. Notice that, in order for baseline methods to work, an N needs to be established, limiting their automated use potential even during the setup stage. We have selected N as 20 for all the baseline methodologies for cleaner presentation and to keep the tests as consistent as possible against the proposed technique GBSD.

Table 4 presents the results of the experiments. In the table, the columns represent the criterion used for comparisons and results of the techniques on three different data collections are organized as rows. As the results in the table indicate, in terms of recall, GBSD performs significantly better than baseline techniques while providing competitive precision values. Informally, the results show that GBSD is capable of correctly classifying the terms (as indicated with precision) while correctly classifying more than half of the stopwords correctly (as indicated with recall). Here, we would also like to note that, due to the skewness between stopwords and non-stopwords in the dataset, recall provides a better comparative value than precision or overall accuracy.

The Table also indicate that the proposed technique also performed better in terms of F-Score and Normalized Document cumulative Gain function, a measure widely used for assessing the ranking quality. Informally, these metrics

Table 4 Comparative results of the experiments.

Technique	Dataset	Precision	Recall	Accuracy	F-Score	NDCG
GBSD	Twitter	1.00	0.56	1.00	0.72	1.00
	Reddit	0.98	0.61	1.00	0.76	0.99
	Tom Sawyer	0.96	0.37	0.99	0.54	0.97
IDF Sorted	Twitter	0.90	0.19	1.00	0.31	0.82
	Reddit	0.95	0.20	1.00	0.33	0.96
	Tom Sawyer	1.00	0.17	0.98	0.30	1.00
Freq Sorted	Twitter	0.85	0.18	1.00	0.29	0.72
	Reddit	0.95	0.20	1.00	0.33	0.86
	Tom Sawyer	0.95	0.17	0.98	0.28	0.96
Entropy Sorted	Twitter	0.90	0.19	1.00	0.31	0.82
	Reddit	1.00	0.21	1.00	0.34	1.00
	Tom Sawyer	0.95	0.17	0.98	0.28	0.96

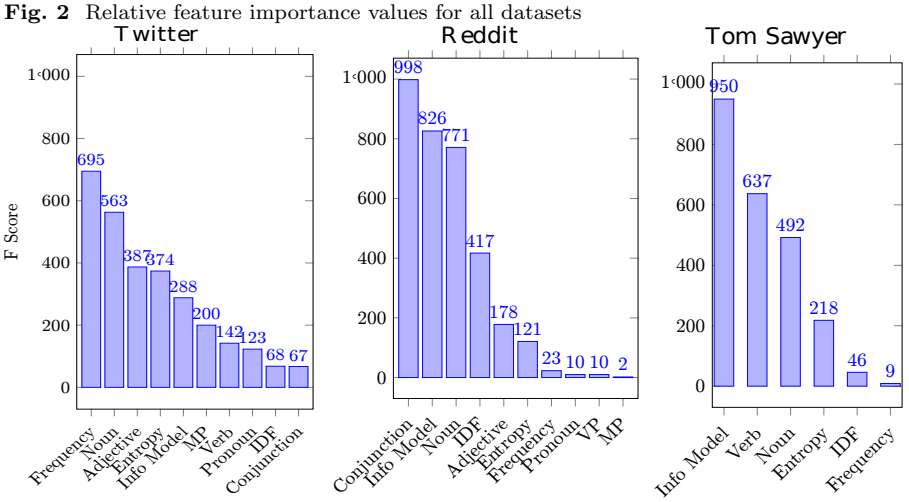
depict how credible the proposed rankings are when a term is labeled as stopword. The results reveal that GBSD again performed better compared to the baseline methods, revealing that the selection rankings also scored relatively better than the baseline methods.

As an added analysis of the proposed machine learning models, we also examine the contribution of representative features in the model. Figure 2 presents the relative feature importance of the features for the models trained using different datasets. The figure reveals that the proposed "Word Positioning" attributes are consistently ranked within top 3 most important features in the models trained using all 3 data collections. In the figure, these attributes are termed with a tag indicating the name of the part-of-speech artifact (e.g., "Noun" or "Adjective"). We can also observe from the results that the Twitter dataset generated weaker feature scores compared to the other two data collections. This can be attributed to the size of the Twitter dataset: The volume of the Twitter dataset is much larger than the other datasets; hence it is possible that the Twitter dataset contains a much larger variation and diversity in both the content and context within the text contained in the dataset. The figure also presents that the previously used metrics are also selected as other important features, validating the literature.

In order to compare our findings in a more concrete fashion, we compared the Top 20 terms chosen as stopwords for the data collection. Tables 5, 6, and 7 present the top 20 candidate stopwords for the Twitter, Reddit, and Tom Sawyer datasets, respectively. In the Tables, the ranking of meta-terms such as links, mentions, and Hastags are also reported explicitly except for the Tom Sawyer dataset where, such meta-terms are unavailable due to the formal writing style of the aforementioned book (and their unavailability in the collection).

As Table 5 indicates, the top stopwords generated by each method is very competitive, all containing only stopwords, almost exclusively. However, we

Fig. 2 Relative feature importance values for all datasets



would like to note that, meta-terms are scored relatively high in the results of the baseline methods. Although, it is possible to argue that meta-terms would be expected to act similar to stopwords, one may also argue that such terms actually contain valuable information. Although these terms are also ranked very highly using GBSD, they are not selected as candidate stopwords, validating possible benefits of the proposed technique.

Table 6 reveal similar results for the Reddit dataset. All techniques select Top 20 candidate terms that are almost entirely composed of stopwords. However, compared to the results of the Twitter dataset, meta-terms are relatively ranked lower and mostly are not classified as stopwords. It is possible to attribute this to the data collection procedure employed while collecing the Reddit dataset: As the messages in the Reddit dataset specialize on certain topics (depending on the Subreddit/post) it is possible that the vocabulary use in this dataset is more restricted. Another important aspect of the Reddit dataset that should be considered is the relative freedom of its users compared to our other datasets: In Twitter, users have a very strict character limit per post, and for the Tom Sawyer dataset, the story-telling aspect of writing limit the usage of the language and vocabulary. As Reddit has no such limitations, the users' need for meta-terms may be less of a necessity.

The more formal and story-focused nature of the Tom Sawyer dataset reveal that the previous methods may be prone to producing flawed results. As indicated by the results in Table 7, a story-critical term "Huck" is selected as a stopword by the frequency and entrophy sorted stopword identification techniques. The results indicate that due to the frequent usage of the name of one of the characters in the book resulted in wrong classification. As GBSD also facilitates word positioning to part-of-speech artifacts, the generated models successfully classifies the aforementioned term without fault.

Table 5 Top 20 candidate stopwords for the Twitter dataset.

Twitter Dataset Top Stopwords								
#		GBSD		IDF		Frequency		Entropy
1	✓	Be	✗	MENTION	✗	MENTION	✗	MENTION
2	✓	Up	✓	Be	✗	LINK	✓	Be
3	✓	Too	✓	Up	✓	Be	✓	Up
4	✓	Are	✓	Too	✓	Up	✓	Too
5	✓	What	✓	Are	✓	Too	✓	Are
6	✓	We	✓	What	✓	Are	✓	What
7	✓	From	✓	Had	✓	What	✓	Had
8	✓	Had	✓	From	✗	HASHTAG	✓	From
9	✓	Has	✓	We	✓	We	✓	We
10	✓	Your	✗	LINK	✓	From	✗	LINK
11	✓	More	✓	Has	✓	Had	✓	Has
12	✓	Off	✓	When	✓	Has	✓	Your
13	✓	Some	✓	Your	✓	Again	✓	Off
14	✓	Very	✓	Off	✓	Your	✓	Again
15	✓	Only	✓	Some	✓	More	✓	More
16	✓	When	✓	More	✓	Off	✓	When
17	✓	She	✓	If	✓	Some	✓	Some
18	✓	An	✓	Again	✓	Very	✓	Only
19	✓	If	✓	Only	✓	Only	✓	An
20	✓	Then	✓	An	✓	When	✓	If
Generalised Terms								
MENTION	✗	54	✓	0	✓	0	✓	0
LINK	✗	55	✓	9	✓	1	✓	9
HASHTAG	✗	58	✗	23	✓	7	✗	25

Looking at these tables, one can also observe the similarities between the top 20 words generated by the different methods. Especially within the first 5 words, the results consistently include the same terms. This finding also hints to the universality of some term behaviour.

As a final experiment, we tested whether the trained models are transferable or not. In this experiment, we trained the machine learning models using one of the three datasets and test its stopword detection performance on the other datasets. Table 8 presents the result of this experiment.

In the table, the columns indicate the selected training and testing sets, and rows indicate comparison metrics. The Table 8 reveals that, our method's performance on utilising models trained from different datasets generally remain competitive, but heavily impacted by the model and testing sets. The results highlight the importance of the dialect form adopted in both sets. The Table also reveals some interesting results on online- vs literature-based datasets: The models trained on online resources perform successfully on online resources while performing poorly on the formal, literature-based dataset. Similarly, model trained on the formal, literature-based dataset perform relatively poorly on online datasets. This result is especially important for this study, as it

Table 6 Top 20 candidate stopwords for the Reddit dataset.

Reddit Dataset Top Stopwords								
#		GBSD		IDF		Frequency		Entropy
1	✓	You	✓	You	✗	LINK	✓	You
2	✓	Are	✓	Are	✓	You	✓	Are
3	✓	At	✓	At	✓	Are	✓	At
4	✓	Or	✓	Or	✓	At	✓	Or
5	✓	They	✓	If	✓	Or	✓	If
6	✓	If	✓	As	✓	They	✓	They
7	✓	As	✓	They	✓	If	✓	As
8	✓	From	✓	From	✓	As	✓	From
9	✓	He	✓	An	✓	Your	✓	Your
10	✓	There	✗	LINK	✓	From	✓	An
11	✓	No	✓	Your	✓	An	✓	He
12	✓	Any	✓	Been	✓	He	✓	There
13	✓	Been	✓	When	✓	There	✓	Been
14	✓	More	✓	There	✓	No	✓	When
15	✓	Now	✓	Because	✓	Any	✓	No
16	✓	Some	✓	He	✓	Been	✓	More
17	✓	When	✓	More	✓	More	✓	Some
18	✓	Has	✓	Some	✓	Now	✓	Has
19	✓	Did	✓	Has	✓	Some	✓	Now
20	✓	Because	✓	Now	✓	When	✓	Because
Generalised Terms								
MENTION	✗	52	✗	74	✗	74	✗	74
LINK	✗	110	✓	9	✓	0	✗	21
HASHTAG	✗	115	✗	52	✗	34	✗	50

validates the necessity to build automated, robust models for the stopword identification, especially for different communication platforms.

6 Conclusion and Future Work

In this paper, we propose an automatic stopword detection technique using gradient boosted decision trees. By using an extensive set of features to represent each term, we successfully predict a term's behaviour and attain better classification results across multiple domains. We also compared our results against an extensive set of baseline techniques, comparing the results across multiple metrics. Using Normalized Discounted Cumulative Gain (NDCG) as a comparative metric for evaluating the relative ranking of selected terms is also a novel contribution of this study.

For natural language processing and information retrieval, having a reliable list of stopwords that are suitable for the domain is extremely important as removal of such terms reduce the volume and computational cost of data-centered applications. The proposed machine learning technique provides an

Table 7 Top 20 candidate stopwords for the Tom Sawyer dataset.

Tom Sawyer Dataset Top Stopwords								
#		GBSD		IDF		Frequency		Entropy
1	✓	It	✓	To	✓	It	✓	To
2	✓	To	✓	It	✓	To	✓	It
3	✓	He	✓	Of	✓	He	✓	Of
4	✓	Of	✓	Was	✓	Was	✓	Was
5	✓	In	✓	He	✓	Of	✓	He
6	✓	His	✓	In	✓	Do	✓	In
7	✓	But	✓	His	✓	In	✓	His
8	✓	Will	✓	But	✓	His	✓	But
9	✓	Now	✓	Had	✓	But	✓	Do
10	✓	Him	✓	Him	✓	Will	✓	Had
11	✓	All	✓	Do	✗	Huck	✓	Him
12	✓	There	✓	On	✓	Now	✓	Will
13	✓	On	✓	At	✓	Him	✗	Huck
14	✓	This	✓	All	✓	All	✓	Now
15	✓	At	✓	There	✓	Had	✓	On
16	✓	Me	✓	This	✓	There	✓	All
17	✓	Not	✓	Now	✓	On	✓	There
18	✓	We	✓	Be	✓	This	✓	At
19	✓	By	✓	Not	✓	Is	✓	This
20	✓	If	✓	Will	✓	At	✓	Be

Table 8 Comparative results of using the trained models on different datasets.

Model	Twitter		Reddit		Tom Sawyer	
Test Set	Reddit	Tom S.	Twitter	Tom S.	Twitter	Reddit
Precision	1.00	1.00	1.00	0.71	0.85	0.97
Recall	0.61	0.04	0.47	0.39	0.18	0.33
Accuracy	1.00	0.98	1.00	0.98	1.00	1.00
F-Score	0.76	0.08	0.64	0.51	0.29	0.50
NDCG	1.00	1.00	1.00	0.79	0.90	0.98

easy to use, automated strategy for compiling such lists for low resource languages and different communication media. The results of our experiments also prove that the results of the proposed technique generate competitive, and high quality results compared to the state-of-the-art, attribute-based techniques.

Statements and Declarations

Competing Interests

The authors did not receive support from any organization for the submitted work and all authors certify that they have no affiliations with or involvement

in any organization or entity with any financial interest or non-financial interest in the subject matter or materials discussed in this manuscript.

Author Contribution Statement

All authors contributed to the study conception and design. **Conceptualization:** Tayfun Kucukyilmaz; **Data Curation:** Tayfun Akin; **Methodology:** Tayfun Kucukyilmaz, Tayfun Akin; **Formal analysis and investigation:** Tayfun Akin; **Writing - original draft preparation:** Tayfun Akin; **Writing - review and editing:** Tayfun Kucukyilmaz; **Supervision:** Tayfun Kucukyilmaz.

Data Availability Statement

All datasets used in this study are acquired from publicly available data repositories and both the names of the datasets and their sources are cited in the References section.

References

[1] Silva, C., Ribeiro, B: The importance of stop word removal on recall values in text categorization. In: Proceedings of the International Joint Conference on Neural Networks, vol. 3, pp. 1661–1666 (2003). https://doi.org/10.1109/IJCNN.2003.1223656

[2] Gerlach, M., Shi, H., Amaral, L.A.N.: A universal information theoretic approach to the identification of stopwords. In: Proceedings of the International Joint Conference on Neural Networks, vol. 1, pp. 606–612 (2019). https://doi.org/10.1038/s42256-019-0112-6

[3] Desai, N.: Stopword identification and removal techniques on TC and IR applications: A survey. In: 6th International Conference on Advanced Computing and Communication Systems (ICACCS), pp. 466–472 (2020). https://doi.org/10.1109/ICACCS48705.2020.9074166

[4] Zipf, G.K.: Selected Studies of the Principle of Relative Frequency in Language. Harvard University Press, Cambridge, MA London, England (2013). https://doi.org/10.4159/harvard.9780674434929

[5] Baeza-Yates, R.A., Ribeiro-Neto, B.A.: Modern Information Retrieval. ACM Press, Addison-Wesley, United States (1999). isbn: 0-201-39829-X

[6] Hassan, Y.A., Medhat, W., Mohamed, H.K.: Egyptian dialect stopword list generation from social network data. The Egyptian Journal of Language Engineering **2** (2015). https://doi.org/10.21608/ejle.2015.60258

[7] Lo, R.T., He, B., Ounis, I.: Automatically building a stopword list for an information retrieval system. Journal Digital Information Management **3**, 3–8 (2005)

[8] Zou, F., Wang, F. L., Deng, X., Han, S., Wang, L. S.: Automatic construction of chinese stop word list. In: ACOS'06. Proceedings of the 5th WSEAS International Conference on Applied Computer Science, Hangzhou, China, pp. 1009–1014 (2006)

[9] Kumova M.S., Karaoğlan, B.: Stop word detection as a binary classification problem. Anadolu University Journal of Science and Technology A - Applied Sciences and Engineering **18**, 1–1 (2017). https://doi.org/10.18038/aubtda.322136

[10] Rani, R., Lobiyal, D.K.: Automatic construction of generic stop words list for hindi text. Procedia Computer Science **132**, 362–370 (2018). https://doi.org/10.1016/j.procs.2018.05.196

[11] Chekima, K., Alfred, R.: An automatic construction of malay stop words based on aggregation method. In: International Conference on Soft Computing in Data Science, vol. 652, pp. 180–189 (2016). https://doi.org/10.1007/978-981-10-2777-2-16

[12] Alshanik, F., Apon, A., Herzog, A., Safro, I., Sybrandy, J: Accelerating text mining using domain-specific stop word lists. In: IEEE International Conference on Big Data (Big Data), pp. 2639–2648 (2020). https://doi.org/10.1109/BigData50022.2020.9378226

[13] Alajmi, A., Saad, E., Darwish, R.R.: Toward an arabic stop-words list generation. International Journal of Computer Applications **46**, 8–13 (2012)

[14] Wibawa, A., Fithri, H., Zaeni, I., Nafalski, A.: Generating Javanese stopwords list using k-means clustering algorithm. In: Knowledge Engineering and Data Science, vol. 3, p. 106 (2020). https://doi.org/10.17977/um018v3i22020p106-111

[15] Joseph, E.: Project Gutenberg (21.04.2021). https://www.gutenberg.org/ebooks/74

[16] Go, A., Bhayani., Huang, L.: Twitter sentiment classification using distant supervision. CS224N Project Report, Stanford **1**, 2 (2009)

[17] Rathi, P.: Kaggle Online Data Repository: Reddit Data-Huge (20.03.2021). https://www.kaggle.com/prakharrathi25/reddit-data-huge

[18] Loper, E., Bird, S.: NLTK: the natural language toolkit. In: Computing

Research Repository, vol. vol: cs.CL/0205028 (2002). https://doi.org/10.3115/1118108.1118117

[19] Kubat, M., Matwin, S.: Addressing the curse of imbalanced training sets: One-sided selection. In: Proceedings of the Fourteenth International Conference on Machine Learning, pp. 179–186 (2000)

[20] Ye, J., Chow, J.H., Chen, J., Chen, J.: Stochastic gradient boosted distributed decision trees. In: Proceedings of the 18th ACM Conference on Information and Knowledge Management (CIKM'09), Hong Kong, China, pp. 2061–2064 (2009). https://doi.org/10.1145/1645953.1646301

[21] Chen, T., Guestrin, C.: XGBoost: A scalable tree boosting system. In: The 22nd ACM SIGKDD International Conference, pp. 785–794 (2016). https://doi.org/10.1145/2939672.2939785

[22] Järvelin, K., Kekäläinen, J.: IR evaluation methods for retrieving highly relevant documents. In: Proceedings of the 23rd Annual International ACM SIGIR Conference onResearch and Development in Information Retrieval, pp. 41–48 (2000). https://doi.org/10.1145/345508.345545

[23] Järvelin, K., Kekäläinen, J.: Cumulated gain-based evaluation of IR techniques. ACM Transactions on Information Systems, 422–446 (2002). https://doi.org/10.1145/582415.582418