Sādhanā (2022)47:17
https://doi.org/10.1007/s12046-021-01731-z

Effect of stopwords in Indian language IR

SIBA SANKAR SAHU* and SUKOMAL PAL*

Department of Computer Science and Engineering, Indian Institute of Technology (BHU), Varanasi, India
e-mail: sibasankarsahu.rs.cse17@itbhu.ac.in; spal.cse@itbhu.ac.in

MS received 1 November 2020; revised 6 July 2021; accepted 25 August 2021

Abstract. We explore and evaluate the effect of stopwords in retrieval performance of different Indian languages such as Marathi, Bengali, Gujarati and Sanskrit. The issue was investigated from three viewpoints. Is there any impact of non-corpus-based stopword removal on chosen Indian languages (if yes, to what extent)? Can we recommend, based on experiment, a number of stopwords for chosen Indian languages that are good enough from retrieval point of view? Is there any relationship of stopwords with average document length from retrieval perspective? It is observed that the stopword removal generally improves mean average precision (MAP) significantly compared with the case when it is not done. For each language, different lengths of the stopword list are explored and evaluated that lead to suggesting its optimal length. We also study the effect of stopwords on retrieval performance over document length. The effect of stopwords is generally found to be quite low in short documents compared with their long counterparts across the four Indian languages.

Keywords. Stopword; Bengali; Marathi; Gujarati; Sanskrit; evaluation.

1. Introduction

In natural languages, a large portion of words are used as only syntactic units to complete sentences but do not carry much information. An observation in Brown Corpus shows that 42% of the words belong to a set of 100 frequently used words in the corpus and they form only 0.1% in the lexicon. However, 5.7% of the words in the corpus belong to 58% in lexicon [1]. Also, it is observed that top-ranked documents are retrieved due to the presence of semantically informative words rather than frequently used words or stopwords. As the frequency of stopwords is very high in a document, they can affect system performance during document retrieval. In most document retrieval techniques a high-frequency word is given more weight compared with less-frequent words, so removal of stopwords gives more importance to other less-frequent words. Moreover, queries containing stopwords complicate the retrieval process. Hence it is a common practice for an information retrieval (IR) system to remove the stopwords for improving performance of IR systems. Doing so not only improves performance but also reduces index size by 30–50%.

Although the effect of stopword has been evident, there is no clear-cut guideline for developing stopword lists [2]. Hence different systems use different sizes of stopword list for different languages. For English language, SMART system suggested a stopword list of 571 words while Fox proposed a size of 421 words [2] and commercial system like DIALOG information service [3] used a smaller stopword list-only 9 words ('an', 'and,' 'with', 'by,' 'for,' 'of,' 'the,' 'to,' 'from'). Dolamic and Savoy [4] show that a short stopword list (9 words) gives similar performance as that of a longer stopword list (571 words). However different sizes of stopword list have been explored and evaluated in the English language but not in other languages, especially Indian languages.

Broad objective of this paper is to analyze and evaluate different sizes of stopword list for Indian languages (Marathi, Bengali, Gujarati and Sanskrit) and to suggest an optimal size of stopword list based on experiment over a wide range of queries. We also evaluate the effect of stopwords on retrieval performance in short and long documents. In particular, we try to explore the following research questions (RQs).

RQ1: At the gross level, is there any impact of non-corpus-based stopword removal on chosen Indian languages (if yes, to what extent)?
RQ2: Stopwords are taken from a web-source (non-corpus-based) that is independent of collection in hand. Is number of stopwords (or length of stopword list) a determining factor in retrieval performance? Can we experimentally find out recommended length of stopword list?
RQ3: Do stopwords have any relationship with average document length from the perspective of retrieval performance? In other words, how does retrieval performance change with number of stopwords and document length?

*For correspondence
Published online: 10 January 2022

To the best of our knowledge, no such study has been carried out in Indian languages so far. Dolamic and Savoy [5] proposed a corpus-based stopword list for Marathi and Bengali languages and evaluated its effectiveness from IR point. However those experiments use a smaller length of stopwords list, i.e. 114 for Bengali and 99 for Marathi. Our aim here is to study in more detail the effect of stopwords in Indian languages.

Rest of the paper is organized as follows. Section 2 describes the effect of stopword in different languages. Section 3 provides a brief description about different IR models applied in experiments. Different evaluation metrics used in experimentation are described in Section 4. Section 5 depicts the characteristics of test collections. We provide a brief description about experimentation in Section 6 . Section 7 describes the evaluation of different IR models to answer the RQs. Section 8 provides a brief discussion on our observations followed by conclusion.

2. Background and related work

Several studies on stopwords were carried out in European languages [2–4]. In recent years a rapid growth of e-content in low-resource languages (non-English) sets the premise for redoing similar exercise in these languages as well. We cannot make a conclusion from the evaluation of European languages for low-resource languages because each language has its morphological variation and other language-specific features. Therefore, an extensive study on stopword list for other low-resource languages is required.

In literature, it is reported that different stopword lists are generated on the basis of word statistics, corpus-specific, domain-specific, web-specific, etc. Different researchers also reported that each stopwords list enhances the performance of system in different perspectives. Rachel Tsz-Wai Lo *et al* [6] proposed a new approach called a term-based random sampling approach for automatically generating a stopword list from the given collection. They show that the proposed approach gives comparable performance to those of the baseline approaches, with a lower computational overhead. Manning and Schutze [7] and Konchady [8] extract the most common word from a document by different frequency-based measures such as term frequency, document frequency and inverse document frequency. The term frequency with inverse document frequency (tf–idf) combination is an implicit approach for creating a stop words list [8]. Ayral and Yavuz [9] proposed domain-specific stopwords to improve the classification of natural language content. Khalifa and Rayner [10] proposed an aggregation method for construction of Malay stopword list. The aggregate method is based upon word frequency inspired by Zipf's law, words distribution against documents using variance measure, and entropy value.

Mireti and Khedkar [11] proposed an automatic identification of stopwords for Amharic text by an aggregate method of word frequency, inverse document frequency and entropy value. Choy [12] proposed a combinatorial value to automatically generating a stop word list from Twitter data. They observed that the proposed approach outperforms other tf–idf and variants by a fair margin. Asubiaro [13] proposed an entropy-based algorithm to identify generic stopwords for the Yoruba language. They extracted two sets of stopwords list from the diacritized and undiacritized versions of the corpus. In both the versions, stopword removal reduces index size substantially. Kaur and Saini [14] proposed a stopword list for the Punjabi language. They proposed a stopword list in Gurmukhi script and transliterated into Shahmukhi and Roman script. Raulji and Saini [15] investigated a dictionary-based stopword removal in Sanskrit. Sinka and Corne [16] explored the usage of classical stopwords in web-specific. They proposed a new stopword list based on word-entropy over modern collections of documents. They show that the proposed stopword list gives better performance than classical stopword list. Lazarinis [17] proposed a stopword list for the Greek language. They observed that stopword removal improves performance of web retrieval.

Savoy [18] proposed a general stopword list for French corpora. They observed that stopword removal improves performance of retrieval. Dolamic and Savoy [4] split the stopword list into short and long parts and evaluate the effect of stopword on retrieval performance for different languages such as Hindi, English, Persian and French. Dolamic and Savoy [5] proposed a stopword list of 165 for Hindi, 114 for Bengali and 99 for Marathi. They show that the stopword removal improves retrieval performance higher in Hindi compared with Bengali and Marathi. Ghosh and Bhattacharya [19] investigate the effect of stopword removal in verbose queries. They show that stopword removal does not give noticeable difference in retrieval performance when compared to not done. In Chinese text retrieval Zou *et al* [20] show that stopword removal is an important pre-processing for Chinese word segmentation, which improves retrieval performance. Davarpanah *et al* [21] proposed a stopword list for Farsi language. They show that stopword removal improves the efficiency of Farsi IR system. Moreover, stopword plays an important role in Farsi text segmentation. Yaghoub-Zadeh-Fard *et al* [22] proposed a part of speech-tagging-based automatically building stop-word lists for Persian IR systems. They show that the proposed approach enhances average precision (AP), reduces index size and improves response time. El-Khair [23] evaluates the effect of stopwords in

Arabic IR. They investigated three types of stopword lists, i.e. general, corpus-based and combined. They show that the performance of a general stoplist is better than those of the other two lists.

Stopword removal not only improves performance in text retrieval but also it is used in different computational tasks like text classification, text categorization, text summarization, sentiment classification and machine translation. Al-Shargabi *et al* [24] show that stopword removal improves performance of Arabic text classification. They evaluate the performance of classifier in terms of precision, recall, the percentage split, *K*-fold cross-validation and time needed for classification. Jayashree *et al* [25] demonstrate that stopword removal improves classification performance for both Naïve Bayes upbeatable and naïve Bayes complement methods, but not for Naïve Bayesian. Zin *et al* [26] evaluate the effect of pre-processing in the classification of online movie reviews. They observe that pre-processing strategies give a significant impact on the classification process. Saif Hassan *et al* [27] show that using a pre-compiled stopword list negatively impacts the twitter sentiment classification, whereas the dynamic generation of stopword lists improve classification performance. Saif *et al* [28] find that semantically identified stopwords improve binary sentiment classification more than the pre-compiled stopword list.

Medhat *et al* [29] proposed a stopword list from online social network corpora in Egyptian dialect. They show that Egyptian dialect stopwords gives better performance than the Modern Standard Arabic stopwords in the sentiment analysis. Silva C and Ribeiro B [30] evaluate the effect of stopword removal in text categorization. They evaluate the performance of the classifier in terms of precision, recall, *f*-measure and accuracy. Xia *et al* [31] investigate the effect of stopword in English text categorization. Azmi and Al-Thanyyan [32] show that stopword removal improves the performance of Arabic text summarization. Schofield *et al* [33] investigate the effect of stopwords in topic models. They observe that the stopword removal improves model quality. Different cross-language IR systems such as Japanese–English [34], Bengali–Hindi [35] and Turkish–English [36] show that stopword removal improves performance. Chong *et al* [37] show the effect of stopword in Statistical Machine Translation (SMT).

In recent years many modern language applications such as Natural Language Toolkit (NLTK[1]), CLTK[2] and Scikit-learn[3] provide different stopword lists for different languages. The NLTK provides stopword lists for 21 languages. The CLTK provides stopword lists for different historical languages. The Scikit-learn, by default, supports an English stopword list. Researchers apply machine-learning algorithms through Scikit-learn.

From this analysis, we conclude that the different stopword removal methods improve performance in text retrieval, text classification, text categorization, sentiment classification and machine translation. However this observation comes from the experiments done in European languages and a few Asian languages. How far they hold good in south Asian languages, especially in Bengali, Marathi, Gujarati and Sanskrit, has not been investigated yet. Hence we study the effect of stopword list in Indian languages on document retrieval. We are also interested in recommending a length of stopwords list for different south Asian languages, for example a minimum length of stopword list for efficient document retrieval. Moreover, we investigate the effect of stopwords on average document length. We believe that these observations can be applied to other morphologically rich languages.

Our work is in line with the earlier work of Dolamic and Savoy [5], Dolamic and Savoy [4] and Singhal *et al* [38]. Our effects of stopwords on retrieval experiments are in line with those of Dolamic and Savoy [5]. The recommended length of stopword list experiment is motivated by Dolamic and Savoy [4]. The implemented methodology is quite different, but the objective of works remains the same. The effects of stopwords on document length experiments are motivated by Singhal *et al* [38].

3. IR framework

We use an open-source search engine called Terrier[4] IR platform for indexing and retrieval of the document collection. The main aim of indexing is to structure, organize and store statistical information about the collection and support efficient search. Stopwords are removed during indexing of the collection. The user expresses his information need in terms of a query; retrieval model matches the query term with document term in the collection and retrieves a set of documents from the collection. For a particular query q, the relevance score of a retrieved document (d) is given by

$$score(d,q) = \sum_{t \epsilon q} score(t \epsilon d) \tag{1}$$

where score ($t\epsilon d$) represents weight of a term calculated by a particular retrieval model. We apply different stopword lists to a set of different retrieval models, to understand effect of stopwords on retrieval performance.

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

[2] http://cltk.org/

[3] https://scikit-learn.org/

[4] http://terrier.org/

3.1 *tf–idf model*

In this model, the relevance score of a document for a given query is calculated based on *term frequency* and *inverse document frequency*. The *term frequency* indicates the number of times a term is present in a given document and *inverse document frequency* indicates the number of documents that contains the given term; tf–idf weighting model within Terrier uses Robertson's *tf* and Sparck Jones *idf* [39]:

$$w(t,d) = Robertson\_tf \cdot idf \tag{2}$$

where

$$Robertson\_tf = \frac{tf_d}{k_1((1-b)+b\frac{dl}{avdl})+tf_d}$$
$$idf = \log(N/d_f+1)$$

tf_d : term frequency of term t in document d
dl : document length in number of terms
$avdl$: average document length
d_f : document frequency of term t
N : total number of documents in the collection
n : number of documents containing at least one term t
k_1 : term-frequency parameter, constant
b : document length normalization parameter, constant

3.2 *BM25 model*

We consider a representative probabilistic model as BM25. BM stands for 'best matching'. For a given query term t, its score in document d is given by Equation 3:

$$w(t,d) = tf_d\left(\frac{\log(\frac{N-n+0.5}{n+0.5})}{k_1((1-b)+b\frac{dl}{avdl}+tf_d)}\right) \tag{3}$$

We also consider other probabilistic models like In_expB2, In_expC2 and InL2. These models come from the Divergence From Randomness (DFR) family [40]. The DFR models are based on the following idea: the more the divergence of the within-document term frequency from its frequency within the collection, the more information carried by the word t in the document d [41].

3.3 *In_expB2 model*

In inverse expected document frequency model for randomness, the relevance score of a document is given by the ratio of two Bernoulli's processes for first normalization, and normalization 2 for term-frequency normalization shown in Equation 4:

$$w(t,d) = \frac{F+1}{n_t(tfn+1)}\left(tfn \cdot \log_2\frac{N+1}{n_e+0.5}\right) \tag{4}$$

Notations used by In_expB2 retrieval models are as follows:

F : frequency of term t in the collection
N : total number of documents in the collection
n_t : document frequency of the term t
n_e : $N(1-(1-\frac{n_t}{N})^F)$
tfn : normalized term frequency. It is given by the normalization 2:

$$tfn = tf \cdot \log_2(1+c\frac{avg\_l}{l}) \tag{5}$$

c : free parameter
avg_l : average document length in the collection

3.4 *In_expC2 model*

In this model, the relevance score of a document is calculated by Equation 6:

$$w(t,d) = \frac{F+1}{n_t(tfn_e+1)}\left(tfn_e \cdot \log_2\frac{N+1}{n_e+0.5}\right) \tag{6}$$

tfn_e denotes the normalized term frequency. It is given by a modified version of the normalization 2:

$$tfn_e = tf \cdot \log_e(1+c\frac{avg\_l}{l}) \tag{7}$$

3.5 *InL2 model*

In this model, the relevance score of a document is given by the ratio of two Laplace processes for first normalization, and normalization 2 for term-frequency normalization as shown in Equation 8:

$$w(t,d) = \frac{1}{tfn+1}\left(tfn \cdot \log_2\frac{N+1}{n_t+0.5}\right) \tag{8}$$

3.6 *Hiemstra_language model*

Finally, we explore a non-parametric probabilistic model known as language model proposed by Djoerd Hiemstra [42]. The probability estimation depends upon the term frequency in the document d_i or in the entire corpus. In this

model, a smoothing parameter λ uses the default value 0.15. Similarity between a query and a document is represented by generation probability as given in Equation 9:

$$P(d_i|q) = P(d_i)\prod_{t_j \epsilon q}[\lambda P(t_j|d_i) + (1-\lambda)P(t_j|c)] \quad (9)$$

where

$$P(t_j|d_i) = \frac{tf_{ij}}{l_i} \quad (10)$$

$$P(t_j|c) = \frac{df_j}{l_c}(l_c = \sum_{ij} tf_{ij}) \quad (11)$$

λ is the smoothing factor and set value = 0.15;
l_c is length of corpus in terms of number of words.

4. Evaluation measures

We evaluate the performance of different retrieval models by following evaluation measures.

4.1 *Precision*

It is the fraction of relevant documents retrieved among the retrieved documents:

$$\text{precision} = \frac{\text{number of relevant documents retrieved}}{\text{number of retrieved documents}} \quad (12)$$

4.2 *Recall*

It is the fraction of relevant documents retrieved from a set of relevant documents:

$$\text{recall} = \frac{\text{number of relevant documents retrieved}}{\text{number of relevant documents in the collection}} \quad (13)$$

4.3 *Precision @ k*

In web-scale IR, recall is not considered as meaningful metric because each query contains thousands of relevant documents and very few users try to read all of them. Precision at k documents ($P@k$) is still a meaningful metric because it tries to read the top 10 or 20 documents (e.g., $P@10$ defines number of relevant document retrieved in top 10 documents). The advantage of this metric is that the total number of relevant documents in the collection is not required a priori whereas disadvantage is that it is the least stable evaluation metric.

$$\text{Precision@}k = \frac{\text{number of relevant documents retrieved in top } k \text{ documents}}{k} \quad (14)$$

4.4 *R-prec*

It is the precision when Rel number of documents are retrieved:

$$\text{R-prec} = \frac{\text{number of relevant retrieved } (r)}{\text{Rel}} \quad (15)$$

Rel : number of relevant documents in the collection.

4.5 *AP*

Average precision (AP) is the average of the precision value obtained for the set of top 'k' retrieved documents (R_k) determined after each relevant document is retrieved. Here, the total number of relevant documents for a given query is 'n':

$$\text{average precision (AP)} = \frac{1}{n}\sum_{k=1}^{n} \text{precision}(R_k) \quad (16)$$

4.6 *MAP*

In recent years, the most standard evaluation measure among TREC[5] community is mean average precision (MAP). It is widely used in last 25 years because of good discrimination and stability [43]. The MAP value does not have a direct interpretation for the end-user. It is computed as the mean of the precision scores obtained after each relevant document is retrieved, using zero as the precision for relevant documents that are not retrieved. We computed the MAP values by TREC_EVAL software, based on a maximum of 1,000 retrieved documents. Mean as a performance measure signifies that we give equal importance to all queries. Comparison between two IR models should not be based on a single query; it should be based on a set of queries to give a meaningful conclusion.

$$\text{mean average precision (MAP)} = \frac{1}{|Q|}\sum_{t=1}^{|Q|} \text{AP}(t) \quad (17)$$

$|Q|$: set of queries

[5] https://trec.nist.gov/

Table 1. Statistics of test collection.

	Marathi	Bengali	Gujarati	Sanskrit
Size (MB)	514.8	2600	1700	11
# of documents	1,00,137	4,43,697	2,42,115	7,057
Number of indexing terms per document				
Mean	279.35	335.21	432.49	56.04
Median	235	285	304	44
Standard deviation	367.93	299.15	397.61	90.34
Maximum	5308	18318	8972	2788
Minimum	18	8	10	14
Total number of tokens	854145	1324941	2045445	109027
Number of topics	39	50	46	50
Number of relevant docs	621	2455	580	427
Mean reldoc	15.92	49.1	12.60	8.54
Length of stopword list	216	398	210	522

5. Test collections

In this work, we used the collection of Marathi, Bengali and Gujarati languages built during FIRE[6] evaluation campaign. We also built a small test collection in Sanskrit language and experimented with it. These corpora consist of news articles extracted from different resources. The Marathi articles are extracted from *'Maharashtra Times'* and *'Sakal'* (articles span the period April 2004 through September 2007), Bengali articles from *'CRI'* and *'Anandabazar Patrika'* (a newspaper edited by ABP Ltd.) and Gujarati articles from archives of the daily newspaper *'Gujarat Samachar'* from 2001 to 2010. Moreover, we extract the Sanskrit news data from *'All India Radio News'* and *'Samprativartah'* news from the period 2015 to 2019. In these collections, both topics and documents use UTF-8 encoding system.

Table 1 shows the statistics of four text corpora. Bengali corpus is the largest in size (MB) with a good number of documents, and Sanskrit is the smallest containing very small number of documents. Gujarati corpus has a greater mean document length (on the basis of mean number of indexed terms per document) whereas Sanskrit has the smallest mean document length.

Marathi and Gujarati collections have 39 and 46 topics, respectively. However, Bengali and Sanskrit collections comprise 50 topics each. Based on the TREC model, each topic comprises three logical sections: a brief title (under the <TITLE> tag), containing two to four words, followed by description tag (<DESC>tag) containing one-sentence user's information need, and narrative tag (<NARR> tag) describing relevance assessment criteria. The example of query representation of Bengali, Marathi, Gujarati and Sanskrit is depicted in Figure 1, whereas document representation of Marathi language is shown in Figure 2. In our experiments we consider all the sections of a query, i.e. title, description and narrative.

[6]http://fire.irsi.res.in/fire/static/data

<Top lang='Mar'>
<NUM>79</NUM>
<TITLE>चीन आणि माऊंट एव्हरेस्ट दरम्यान रस्ता बांधणे</TITLE>
<DESC> चीनपासून माऊंट एव्हरेस्टपर्यंत रस्ता बांधण्याची योजना </DESC>
<NARR> संबंधित कागदपत्रात चीनपासून माऊंट एव्हरेस्टपर्यंत रस्ता बांधण्याच्या योजनेचे चित्रण करायला हवे. भारतीय आणि चीनी अधिकाऱ्यांच्या ह्या मुद्याबाबच्या चर्चादेखील संबंधित आहेत. </NARR>
</TOP>

<TOP>
<NUM>79</NUM>
<TITLE> Building a road between China and Mount Everest </TITLE>
<DESC> Road from China to Mount Everest
</DESC>
<NARR> Relevant documents should outline plans to build a road from China to Mount Everest. Discussion between Indian and Chinese officials on the issue are also relevant. </NARR>
</TOP>

<TOP>
<TOP lang='Guj'>
<NUM>176</NUM>
<TITLE>વાય.એસ.આર. રેડ્ડી ની મૌત</TITLE>
<DESC>આંધ્ર પ્રદેશ ના મુખ્ય મંત્રી વાય.એસ.આર. રેડ્ડી ની મૌત </DESC>
<NARR>સંબંધિત દસ્તાવેજો એક હેલિકોપ્ટર અકસ્માત માં આંધ્ર પ્રદેશ મુખ્ય પ્રધાન વાય.એસ.આર. રેડ્ડી ની મૃત્યુ વિશે જાણકારી સમાવતા હોવા જોઈએ.</NARR>
</TOP>
<TOP>
<TOP lang='bn'>
<NUM>176</NUM>
<TITLE>ওয়াই এস আর রেড্ডির মৃত্যু</TITLE>
<DESC>অন্ধ্র প্রদেশের মুখ্যমন্ত্রী ওয়াই এস আর রেড্ডির মৃত্যু</DESC>
<NARR>অন্ধ্র প্রদেশের মুখ্যমন্ত্রী ওয়াই এস আর রেড্ডির মৃত্যু হেলিকপ্টার দুর্ঘটনায় হয়েছে , প্রাসঙ্গিক নথিতে এই সংক্রান্ত তথ্য প্রয়োজনীয় </NARR>
</TOP>

<TOP>
<TOP lang='en'>
<NUM>176</NUM>
<TITLE>YSR Reddy death</TITLE>
<DESC>Death of Andhra Pradesh Chief Minister YSR Reddy</DESC>
<NARR>Relevant documents should contain information about Andhra Pradesh Chief Minister YSR Reddy's death in a helicopter crash.</NARR>
</TOP>

Figure 1. Shows example of topic description for Marathi, Gujarati, Bengali, Sanskrit languages along with their English translation.

<TOP lang='Sans'>
<NUM>2</NUM>
<TITLE> दक्षिण–अफ्रीकायाः दशम–ब्रिक्स–सम्मेलनम् </TITLE>
<DESC> दक्षिण–अफ्रीकायाः जोहान्सबर्ग्गे दशम–ब्रिक्स–सम्मेलनं भविष्यति । </DESC>
<NARR> दक्षिण–अफ्रीकायाः जोहान्सबर्ग्गे पञ्चानां ब्रिक्स–राष्ट्रप्रमुखाणाम् अध्यक्षतायाम् आयोजितस्य दशमब्रिक्स–सम्मेलनस्य सम्बन्धिनः विषयाः अत्र भवेयुः । भारतस्य सुदृढ–पारस्परिक–सम्बन्धार्थम् एतत् सम्मेलनम् अति–महत्वपूर्णं विद्यते । अन्यत् किमपि राष्ट्रियम् अन्ताराष्ट्रियं वा सम्मेलनम् अत्र प्रासङ्गिकं नास्ति । </NARR>
</TOP>
<TOP lang='Sans'>
<NUM>2</NUM>
<TITLE> 10th Brics summit at South Africa </TITLE>
<DESC> 10th Brics summit will be held in Johannesburg of South Africa </DESC>
<NARR> Relevant documents should outline 10th Brics summit held in Johannesburg of South Africa. Discussion about other international summits are irrelevant. </NARR>
</TOP>

Figure 1. continued

<doc>
<docno>Solapur61B5F4CF38.htm.txt</docno>
<text> जगदंबा सूत गिरणीतून पाच लाखाच्या मालाची चोरी माढा, ता. १ – येथील जगदंबा सूत गिरणीच्या गोदामातून सुमारे पाच लाखाचा माल चोरीस गेल्याचा तक्रारी अर्ज मालेगाव येथील एस.एम. एन्टरप्राईजचे मालक महेंद्रकुमार शंकरलाल मोदी यांनी माढा पोलिसांना दिला असून पोलिस निरीक्षक सुरेश गुरव याबाबत चौकशी करीत असून अद्याप गुन्हा नोंद केला नसल्याचे सांगितले. .....
</text>
</doc>
<doc>
<docno>Solapur61B5F4CF38.htm.txt</docno>
<text> Theft of goods worth Rs 5 lakh from Jagdamba Yarn Mill Madha, Tal. 1 – Complaint of theft of goods worth around Rs 5 lakh from the warehouse of Jagdamba Yarn Mill here. Enterprise owner Mahendra Kumar Shankarlal Modi has handed over the case to Madha police and police inspector Suresh Gurav is investigating the matter and no case has been registered yet. </text>
</doc>

Figure 2. An example of Marathi document and its translation

We extract the stopword lists from GitHub[7][8][9] for different languages. The extracted stopword lists contain duplicate words that are removed to consider unique stopwords only.

6. Experimental setup

We provide a brief description about experimental setup to see the effect of stopwords in the Indian languages Marathi, Bengali, Gujarati and Sanskrit. Stopwords are the most frequently used words like articles, pronouns, conjunctions, prepositions, prefixes, adverbs and adjectives. The stopword lists are used during indexing the collections using Terrier. We also use a set of retrieval models supported in Terrier retrieval system.

Experiments are conducted to see the effect of stopwords from three view points as described earlier.

1. *[RQ1] At the gross level, is there any impact of stopword removal on chosen Indian languages?*
   Here, we see the effect of stopwords on retrieval performance in different Indian languages by not removing the stopwords (not using any stopword list during indexing) and then removing them.
2. *[RQ2] Stopwords are taken from a web-source that is independent of collection in hand. Is number of stopwords (or length of stopwords list) a determining factor in retrieval performance? Can we experimentally find out recommended length of stopwords list?*
   To find answers we use different length of stopwords list for each language considered, at $10\%, 20\%, 30\%, \ldots, 90\%$, and see the variations in overall retrieval performance. Based on performance, we also recommend length of stopwords list for each language.
3. *[RQ3] Do stopwords have any relationship with average document length from the perspective of retrieval performance? In other words how does retrieval performance change with number of stopwords and average document length?*
   Here we divide each corpus into two parts, i.e. short and long documents, in such a way that each part contains almost equal number of documents. In each subpart, we evaluate the effect of stopwords on retrieval performance for different languages.

All experiments are conducted in a personal laptop system with core i3 processor and 8 GB RAM.

7. Evaluation

To address the RQs discussed in Section 1, we experiment and evaluate in different Indian languages in the following way.

[7] https://github.com/gujarati-ir/Gujarati-Stop-Words

[8] https://gist.github.com/Akhilesh28/sanskrit stopwords

[9] https://github.com/stopwords-iso/stopwords-bn.txthttps://github.com/stopwords-iso/stopwords-bn.txt

Table 2. MAP, R-prec and *P*@10 without and with stopword removal in Marathi language (39 TDN queries).

	Without stopword removal			With stopword removal		
Retrieval model	MAP	R-prec	*P*@10	MAP	R-prec	*P*@10
BM25	0.3232	0.32	*0.3769*	*0.3258**	0.3159	*0.3744*
tf_idf	*0.324*	*0.321*	0.3718	0.3233*	*0.3195*	0.3728
In_expC2	0.2603	0.2559	0.3205	0.2638	0.2561	0.3282
In_expB2	0.2803	0.278	0.3462	0.2816	0.2738	0.3487
InL2	0.2802	0.2761	0.3487	0.2824	0.2718	0.3513
Hiem_LM	0.2571	0.2544	0.3256	0.2578*	0.2492	0.3231
Mean	0.2875	0.2842	0.3483	0.2891	0.281	0.3497
% Change				+.56%	−1.1%	+.42%

Italic character defines the best performing retrieval model among different retrieval model evaluation

Table 3. MAP, R-prec and *P*@10 without and with stopword removal in Bengali language (50 TDN queries).

	Without stopword removal			With stopword removal		
Retrieval model	MAP	R-prec	*P*@10	MAP	R-prec	*P*@10
BM25	0.2668	*0.3047*	*0.452*	*0.271**	*0.3112*	0.448
tf_idf	*0.267*	0.302	0.448	0.2706*	0.3093	*0.45*
In_expC2	0.2188	0.2638	0.398	0.2206	0.2614	0.396
In_expB2	0.2385	0.2831	0.412	0.2388	0.2832	0.414
InL2	0.2434	0.2865	0.43	0.239	0.2873	0.432
Hiem_LM	0.2057	0.2518	0.39	0.2109*	0.2558	0.392
Mean	0.24	0.282	0.4217	0.2418	0.2847	0.422
% Change				+.74%	+.96%	+0.07%

Italic character defines the best performing retrieval model among different retrieval model evaluation

Table 4. MAP, R-prec and *P*@10 without and with stopword removal in Gujarati language (46 TDN queries).

	Without stopword removal			With stopword removal		
Retrieval model	MAP	R-prec	*P*@10	MAP	R-prec	*P*@10
BM25	0.302	*0.3134*	*0.3*	*0.3132**	*0.3252*	*0.3087*
tf_idf	0.2957	0.3059	0.2935	0.3107*	0.3225	0.3
In_expC2	0.2892	0.3022	0.2957	0.2958	0.3059	0.2978
In_expB2	*0.3068*	0.3091	0.2957	0.3086	0.3164	0.3022
InL2	0.3013	0.3001	0.287	0.2999	0.297	0.3065
Hiem_LM	0.2079	0.2318	0.237	0.2201*	0.2425	0.2391
Mean	0.2838	0.2938	0.2848	0.2914	0.3016	0.2924
% Change				+2.66%	+2.67%	+2.65%

Italic character defines the best performing retrieval model among different retrieval model evaluation

7.1 *Effect of stopword on retrieval*

In the first set of experiments, we see the effect of non-corpus-based stopwords on retrieval performance in different Indian languages. In Marathi, the MAP, R-prec and *P*@10 evaluated without stopword removal and with stopword removal are shown in table 2. It is observed that in most of the retrieval models, the MAP and *P*@10 scores increase after stopword removal. However, the R-prec scores decrease after stopword removal. We conduct similar experiments for the other three languages: Bengali, Gujarati and Sanskrit languages, as shown in tables 3, 4 and 5, respectively. In all the tables, the best performance by a given retrieval model is shown in italics.

For most of the retrieval models the MAP, R-prec and *P*@10 scores increase after stopword removal. We also observe that in Marathi and Bengali languages the *P*@10 values are quite high, which signify that more number of relevant documents are retrieved at early ranks. In Gujarati, the MAP, R-prec and *P*@10 values are quite similar. In Sanskrit, the *P*@10 values are quite low. There are two reasons for this: a) the number of relevant documents in Sanskrit is less compared with other languages as the Sanskrit dataset is the smallest in size and b) the relevant

Table 5. MAP, R-prec and *P*@10 without and with stopword removal in Sanskrit language (50 TDN queries).

	Without stopword removal			With stopword removal		
Retrieval model	MAP	R-prec	*P*@10	MAP	R-prec	*P*@10
BM25	0.405	0.3807	0.218	0.4209*	0.4016	0.224
tf_idf	0.4023	0.376	0.212	0.421*	0.3931	0.224
In_expC2	0.4077	*0.3988*	0.222	0.4205*	*0.4087*	0.23
In_expB2	*0.4091*	0.3892	*0.226*	*0.4232**	0.4037	*0.234*
InL2	0.3913	0.3697	0.212	0.403*	0.3756	0.214
Hiem_LM	0.3581	0.3505	0.174	0.3877*	0.3772	0.196
Mean	0.3956	0.3775	0.2107	0.4127	0.3933	0.2203
% Change				+4.33%	+4.19%	+4.58%

Italic character defines the best performing retrieval model among different retrieval model evaluation

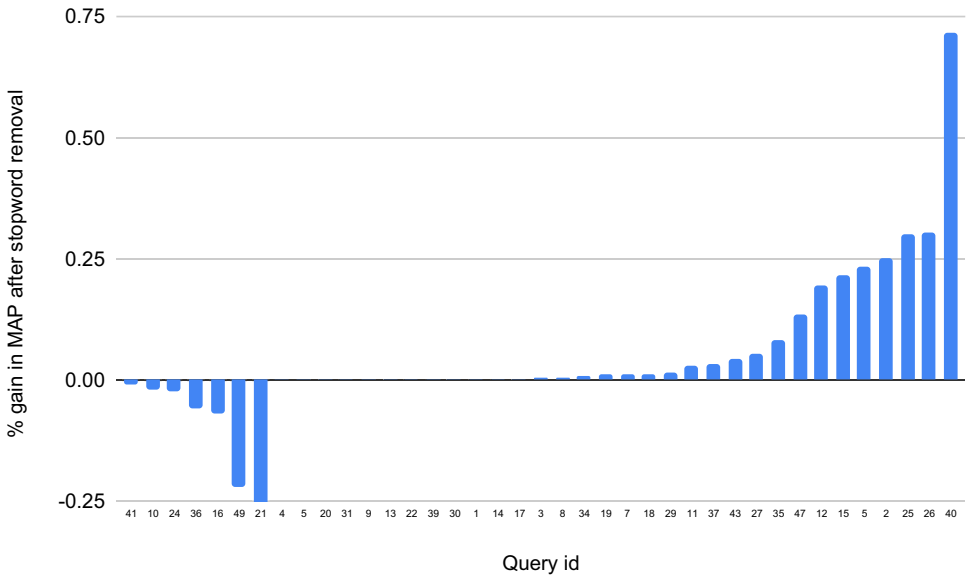

Figure 3. A query-by-query evaluation in Marathi language by BM25 model.

documents are retrieved at later ranks. On closer observation, we see that the MAP seems to have the least variation among the metrics as MAP is a stabler metric in comparison with the other two.

In this work, the statistically significant differences are detected by a one-sided *t*-test (significance level $\alpha = 5\%$). We use without stopword removal as a baseline (tables 2, 3, 4 and 5 and the statistically significant differences are denoted by the symbol '$*$'. In Marathi and Bengali, stopword removal improves MAP and *P*@10 scores in different retrieval models but DFR-based models do not produce statistically significant results. In Gujarati, stopword removal improves performance equally in different retrieval models but DFR-based models do not produce statistically significant results. In Sanskrit, stopword removal improve performance in different retrieval models and they produce statistically significant results.

To get more insights, we also perform a query-by-query analysis. Here, we consider the BM25 retrieval model for Marathi and Bengali and In_expB2 model for Gujarati and Sanskrit as they are found to be best performing models for the respective languages. On closer observation, in Marathi, stopword removal improves performance for 29 topics and reduces performance for 7 topics. The performance of each query is shown in figure 3. For example, in Topic 12, सियाचिन सभोवतीच्या सैन्याच्या स्थानाविषयी मनमोहन सिंह आणि परवेझ मुशर्रफ मोनिका बेदी आणि बनावटी पारपत्र खटला (Manmohan Singh, Pervez Musharraf discuss troop position around Siachen), stopword removal improves performance compared to when it is not done by 19.56%. A similar observation is found in Topic 25 ह्यांच्यामधील चर्चा (Monika Bedi and fake passport law suit) (improvement of 30.4%). Likewise, in Bengali, Gujarati and Sanskrit, stopword removal improves performance in 41, 36 and 36 topics respectively, and reduces performance in 9, 10 and 6 topics respectively. The percentage changes in performance due to stopword removal at per-query-level are shown in Figs. 4, 5 and 6.

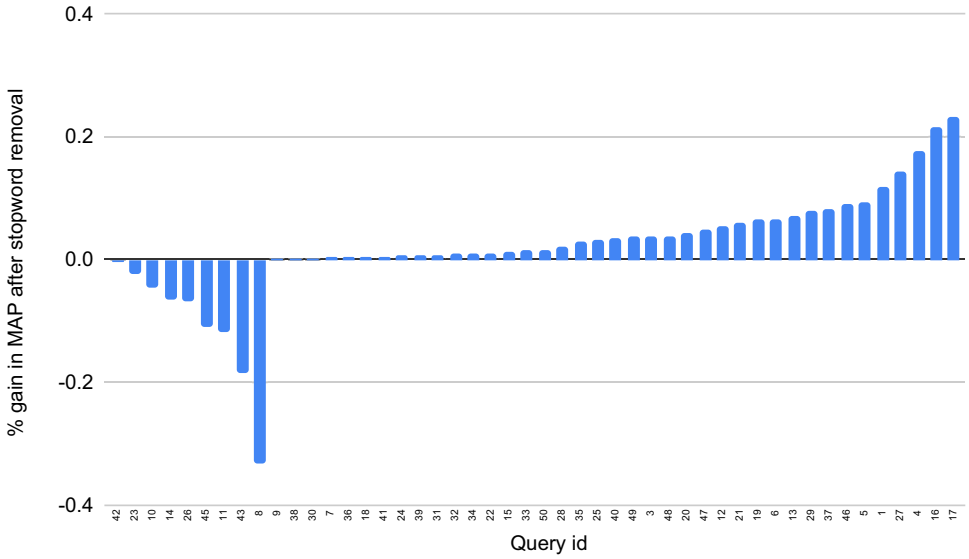

Figure 4. A query-by-query evaluation in Bengali language by BM25 model.

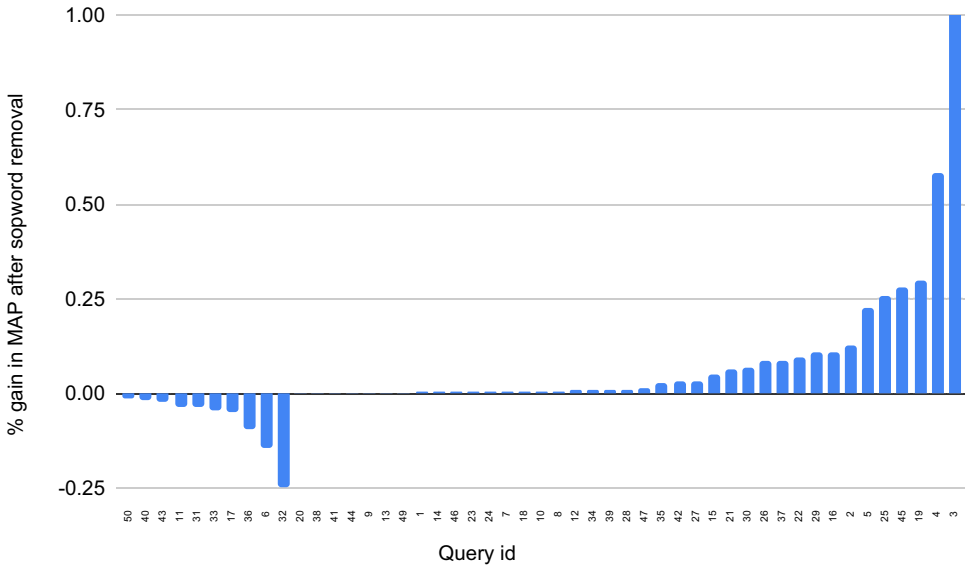

Figure 5. A query-by-query evaluation in Gujarati language by In_expB2 model.

In all the languages, performance gains outweigh the losses both in terms of number of topics and amount of gains.

7.2 *Length of stopword list*

The second set of experiments are conducted to determine the length of stopwords list for different Indian languages. We randomly sample stopword list for different Indian languages at $10\%, 20\%, \ldots, 90\%$, up to full length of stopword list, and then compute MAP. For Marathi, MAP values at different lengths of stopwords list are shown in table 6. We actually take 10 random samples of stopword list (with replacement) at each %-point of stopword length and the MAP distribution of Marathi (for BM25 model) is summarized in a box plot shown in Figure 7. In a box plot, the box signifies the range of values from first quartile to third quartile and the whiskers go from each quartile to the minimum or maximum. The horizontal lines inside the boxes denote mean-value of the samples. From table 6 and Figure 7, we see that the MAP values change non-uniformly at different lengths of stopwords list. On closer examination we find that a small stopword length, i.e. 10% of the total length of stopwords list (number of stopwords =

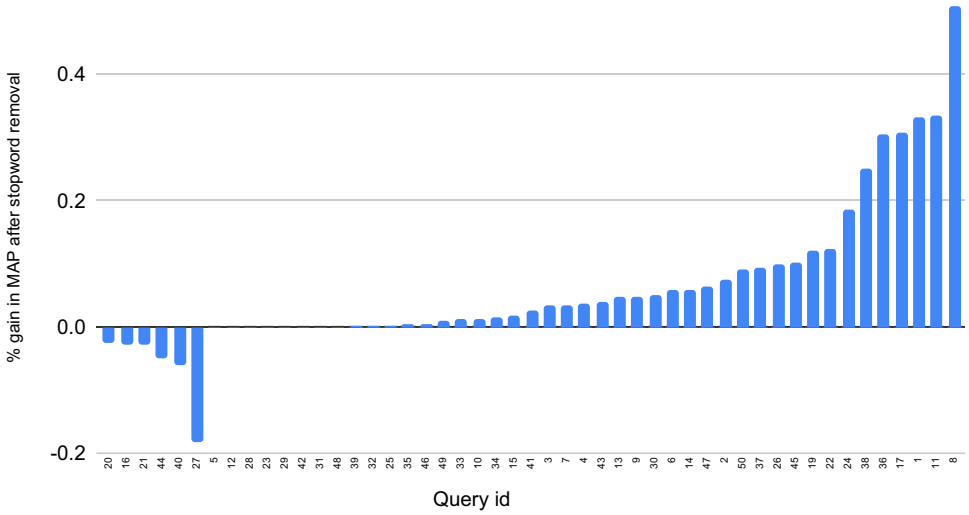

Figure 6. A query-by-query evaluation in Sanskrit language by In_expB2 model.

Table 6. MAP scores for different sizes of stopword length in Marathi language (39 TDN queries).

Retrieval model	None	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
BM25	0.3232	*0.3279*	0.3258	*0.3276*	*0.3269*	*0.3265*	*0.3258*	*0.3247*	0.3241	0.324	*0.3258*
tf_idf	*0.324*	0.3268	*0.3268*	0.3254	0.3256	0.3252	0.3251	0.3239	*0.3245*	*0.3246*	0.3233
In_expC2	0.2603	0.2639	0.2639	0.2635	0.2637	0.2638	0.2644	0.2632	0.2615	0.2614	0.2638
In_expB2	0.2803	0.2824	0.2811	0.2819	0.2822	0.2812	0.2809	0.2811	0.28	0.2804	0.2816
InL2	0.2802	0.2823	0.2804	0.2824	0.2824	0.2814	0.2808	0.2812	0.2815	0.2814	0.2824
Hiem_LM	0.2571	0.257	0.2606	0.2584	0.2573	0.257	0.2557	0.2542	0.2543	0.255	0.2578

Italic character defines the best performing retrieval model among different retrieval model evaluation

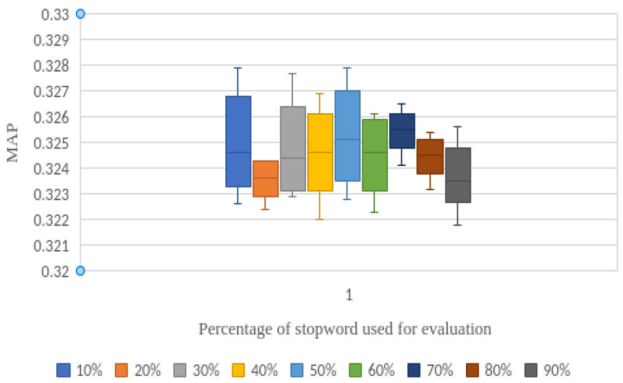

Figure 7. Box plot for Marathi language by BM25 model.

21), gives best or near-best MAP across different retrieval models. This shows that retrieval performance is not significantly affected if a smaller length of the stopwords list is used during retrieval.

We conduct similar experiments for the other three languages as well. For Bengali, Gujarati and Sanskrit languages, the evaluation of different lengths of stopword lists is shown in tables 7, 8 and 9 and the MAP distribution of Bengali, Gujarati and Sanskrit languages are summarized in box plots shown in Figures 8 (with BM25 model), 9 (In_expB2 model) and 8 (In_expB2), respectively. Linguistically, there are two important reasons for varying performance during stopword removal of different Indian languages. Primarily, the usage of stopwords varies from one language to another. Hence their removal will have different magnitudes of effect in retrieval performance. Secondly we experimented and evaluated different non-corpus-based stopword lists taken from the web, in Indian languages. The non-corpus-based stopword list is extracted from web that is independent of collection in-hand and comprises a wide range of vocabulary. Hence its removal affects the retrieval performance disproportionately across languages. A corpus-based stopword list has been seen to give better retrieval performance in different European languages. Hence, we can hypothesize that a similar study in Indian language will show similar performance improvement in Indian languages as well. However, this is yet to be experimented.

Table 7. MAP scores for different sizes of stopword length in Bengali language (50 TDN queries).

Retrieval model	None	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
BM25	0.2668	*0.2668*	*0.2668*	*0.2664*	*0.2667*	*0.2669**	*0.2668*	*0.2688*	*0.2701*	*0.2709*	*0.271*
tf_idf	*0.267*	0.266	0.2662	0.266	0.2662	0.2669	0.2662	0.2682	0.2697	0.2701	0.2706
In_expC2	0.2188	0.2164	0.217	0.2168	0.2186	0.218	0.2183	0.2195	0.2206	0.2205	0.2206
In_expB2	0.2385	0.2375	0.2378	0.237	0.2375	0.2363	0.2358	0.2359	0.2372	0.238	0.2388
InL2	0.2434	0.2424	0.2424	0.2426	0.2406	0.2393	0.2388	0.2387	0.239	0.2412	0.239
Hiem_LM	0.2057	0.2052	0.2061	0.2062	0.2062	0.2084*	0.2068	0.2092	0.2091	0.2102	0.2109

Italic character defines the best performing retrieval model among different retrieval model evaluation

Table 8. MAP scores of different sizes of stopword length in Gujarati language (46 TDN queries).

Retrieval model	None	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
BM25	0.302	0.303	0.3049	0.3015	0.3008	0.3051*	0.3042	0.3048	0.3111	0.3104	*0.3132*
tf_idf	0.2957	0.2978	0.2955	0.2995	0.2952	0.3032	*0.306*	*0.3064*	*0.3139*	0.3092	0.3107
In_expC2	0.2892	0.2892	0.2897	0.2895	0.2894	0.2891	0.2909	0.2907	0.2953	0.2959	0.2958
In_expB2	*0.3068*	*0.3077*	*0.3093*	*0.3046*	*0.3056*	*0.3062*	0.3008	0.303	0.3093	*0.3108*	0.3086
InL2	0.3013	0.3047	0.3051	0.3021	0.3029	0.3	0.3008	0.3009	0.3037	0.2988	0.2999
Hiem_LM	0.2079	0.2121	0.2118	0.2094	0.2104	0.2123*	0.2105	0.2116	0.2138	0.2185	0.2201

Italic character defines the best performing retrieval model among different retrieval model evaluation

Table 9. MAP scores of different sizes of stopword length in Sanskrit language (50 TDN queries).

Retrieval model	None	10%	20%	30%	40%	50%	60%	70%	80%	90%	100%
BM25	0.405	0.4061	0.4051	0.406	0.408	0.4152	0.4153	0.4161	0.4148	0.4171	0.4209
tf_idf	0.4023	0.4028	0.4021	0.4018	0.4072	0.4161	0.416	0.4153	0.4137	0.419	0.421
In_expC2	0.4077	0.4084	*0.408*	0.4049	0.4112	0.4193	0.419	0.4192	0.4181	0.4183	0.4205
In_expB2	*0.4091*	*0.4085*	0.4077	*0.4084*	*0.4125*	*0.4201*	*0.4192*	*0.4193*	*0.42*	*0.422*	*0.4232*
InL2	0.3913	0.391	0.3908	0.3914	0.3951	0.4017	0.4008	0.3995	0.4005	0.4009	0.403
Hiem_LM	0.3581	0.3599	0.3592	0.359	0.368	0.3816*	0.3812	0.382	0.3775	0.3807	0.3877

Italic character defines the best performing retrieval model among different retrieval model evaluation

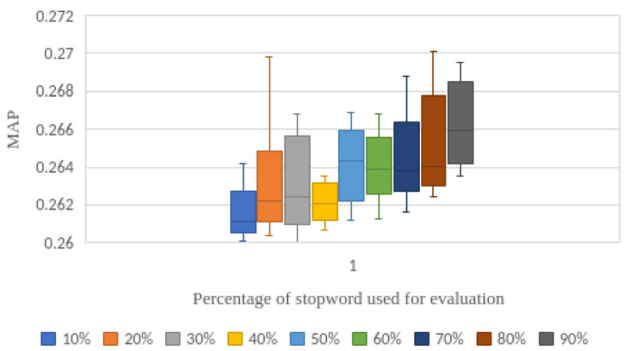

Figure 8. Box plot for Bengali language by BM25 model.

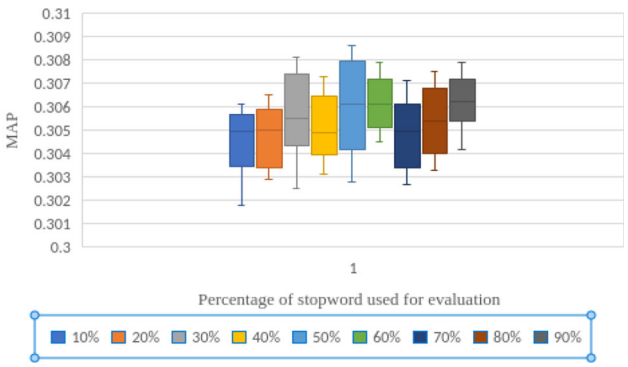

Figure 9. Box plot for Gujarati language by In_expB2 model.

We also look at statistical significance with respect to full length of stopwords list (tables 6, 7, 8 and 9) and the statistically significant differences are denoted by the symbol '$*$'. It is observed that 50% of the total length of stopwords list, i.e. length 249 of stopwords list for Bengali, length 113 of stopwords list for Gujarati and length 261 of stopwords list for Sanskrit, gives no significant difference against the total length of stopwords list. For Marathi it is even as small as 10%, i.e. a set of mere 21 stopwords. A similar observation can be drawn by examining figures 8, 9

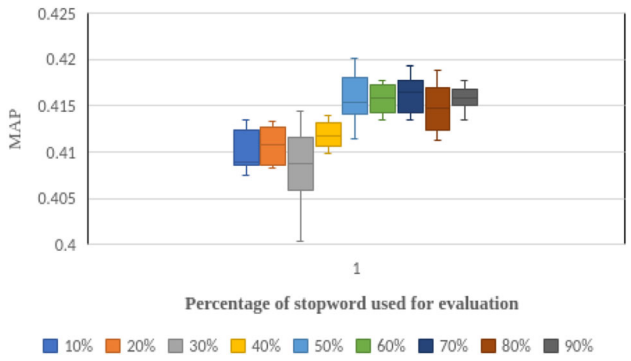

Figure 10. Box plot for Sanskrit language by In_expB2 model.

Table 10. Suggested length of stopwords list for different languages.

Language	Suggested length of stopwords list
Marathi	21
Bengali	249
Gujarati	113
Sanskrit	261

and 10 as well. The set of experiments also reveals an important feature of the languages concerned. Marathi needs actually a small set of stopwords; rather, use of long list of stopwords affects retrieval performance. For Gujarati, about 100 stopwords are fine. However, for Bengali and Sanskrit, quite long (about 250 or more) stoplists are necessary for good retrieval performance.

Based on our experiments, suggested lengths of the stopwords list for different languages are shown in table 10.

7.3 *Effect of stopwords on document length*

The third set of experiments are conducted to see the effect of stopwords on retrieval performance over document length. In a long document, diverse terms occur and each term occurs with high frequency. Any retrieval model prefers a long document over a short one because probability of a query term occurring in a long document is higher. To overcome retrieval bias towards longer documents, modern IR systems use document length normalization [38]. For a heuristic study we divide each corpus into two parts, viz. short and long documents, in such a way that each part contains an equal number of documents. We consider short documents of smaller file size, over larger file size. In each sub-part, we evaluate the effect of stopword on retrieval performance as shown in tables 11, 12, 13 and 14.

We use full length of stopwords list as a baseline (column of tables 11, 12, 13 and 14) and the statistically significant differences are denoted by the symbol $*$. In all the languages, most of the retrieval models give an comparable retrieval performance in both short and long documents. However, language models give poor performance in all the languages. It is also observed that in most of the retrieval models the suggested length of stopwords list gives no significant difference against total length of stopwords list in both short and long documents.

8. Discussion

In the first set of experiments (shown in tables 2, 3, 4 and 5), we observe that the non-corpus-based stopword removal improves performance in document retrieval and also give performance comparable to those of other corpus-based stopword removals [5]. On a closer observation, we also find that the effect of stopwords varies from one language to another. In Marathi and Bengali, the effect of stopword removal on retrieval performance is quite low (less than 1%). However in Gujarati and Sanskrit, it is higher (more than 2%). It is also observed that in Marathi and Bengali, the BM25 and tf–idf models give better MAP than DFR-based retrieval models. In Gujarati and Sanskrit, the performances of different retrieval models are quite similar. Among the different retrieval models, the performance of the language model is quite poor for all languages. In summary, we find that the effect of non-corpus-based stopword improves performance in document retrieval but the performance is quite low compared with other corpus-based stopword removals in European languages. A corpus-based stopword list may improve performance in South-Asian languages like other European languages, but is yet to be ascertained.

In the second set of experiments (shown in tables 6, 7, 8 and 9), we see that in most of the retrieval models there are no significant differences between the suggested length of stopwords list and full length of stopwords list. In Marathi, the suggested length of the stopwords list gives near-best MAP over higher length of stopwords list. In Bengali, Gujarati and Sanskrit also, the difference in MAP scores between the suggested length of stopwords list and full length of stopwords list are quite low. In summary, if one wants to minimize the computational effort, one can use a smaller length of stopwords list as suggested, instead of large ones. This observation is quite similar to the findings in European languages by Dolamic and Savoy [4].

In the third set of experiments (shown in tables 11, 12, 13 and 14), we observe that the effect of stopwords is quite low in short documents compared with long ones. This can be explained by the fact that a long document has many stopwords and each such stopword possibly with higher frequency compared with a short one. In Sanskrit, the effect

Table 11. MAP scores of effect of stopword in short and long documents in Marathi collection.

	Short document length			Long document length		
Retrieval model	None	Suggested stopword length	Full stopword length	None	Suggested stopword length	Full stopword length
BM25	*0.1314*	*0.1321*	*0.1321*	*0.2276*	*0.2274*	*0.2296*
tf_idf	0.1298	0.1278	0.1293	0.2249	0.2255	0.2257
In_expC2	0.1109	0.1107	0.1121	0.2224	0.2159	0.2248
In_expB2	0.1158	0.1154	0.1178	0.2254	0.2263	0.2278
InL2	0.1201	0.1197	0.1229*	0.2232	0.2256	0.225*
Hiem_LM	0.1094	0.1095	0.1119*	0.1913	0.1913	0.199*
Mean	0.1196	0.1192	0.121	0.2191	0.2187	0.222
% Change			+1.21%			+1.3%

Italic character defines the best performing retrieval model among different retrieval model evaluation

Table 12. MAP scores of effect of stopword in short and long documents in Bengali collection.

	Short document length			Long document length		
Retrieval model	None	Suggested stopword length	Full stopword length	None	Suggested stopword length	Full stopword length
BM25	*0.0843*	*0.0841*	*0.0843*	*0.199*	*0.2008*	0.2033*
tf_idf	0.0834	0.0832	0.0841	0.1978	0.1991	*0.2034**
In_expC2	0.0686	0.0681	0.0691	0.1739	0.1731	0.1753
In_expB2	0.075	0.0737	0.0745	0.1848	0.1843	0.1871*
InL2	0.0758	0.076	0.0764	0.1892	0.1875	0.1896
Hiem_LM	0.0683	0.0694	0.0706	0.1516	0.1529	0.1583*
Mean	0.0759	0.0758	0.0765	0.1827	0.183	0.1862
% Change			+.79%			+2.07%

Italic character defines the best performing retrieval model among different retrieval model evaluation

Table 13. MAP scores of effect of stopword in short and long documents in Gujarati collection.

	Short document length			Long document length		
Retrieval model	None	Suggested stopword length	Full stopword length	None	Suggested stopword length	Full stopword length
BM25	*0.1538*	*0.1518*	*0.1537*	0.1837	0.1865	0.1921*
tf_idf	0.148	0.1494	0.1521*	0.179	0.1851	*0.1923**
In_expC2	0.1366	0.1353	0.1369	0.166	0.1668	0.1736
In_expB2	0.1462	0.1455	0.1463	0.1769	0.1782	0.1849*
InL2	0.1424	0.1429	0.1434	*0.1845*	*0.1866*	0.1909
Hiem_LM	0.1186	0.125	0.1277	0.1202	0.1216	0.1285
Mean	0.1409	0.1417	0.1433	0.1683	0.17	0.1767
% Change			+1.71%			+5%

Italic character defines the best performing retrieval model among different retrieval model evaluation

of stopword removal is quite similar in both short and long documents —this can be a property of the languages concerned. However, we would like to mention that our Sanskrit collection contained small-length documents only with little variation in length. In Bengali and Gujarati, the effect of stopword removal in short documents is less than 2% whereas in long ones it is quite high. In Marathi and Bengali the differences in the MAP scores between short documents and long ones are quite high, whereas in Gujarati and Sanskrit the differences in the MAP scores are comparatively low. The main reason for this is that both Bengali and Marathi collections contain more number of

Table 14. MAP scores of effect of stopword in short and long documents in Sanskrit collection.

	Short document length			Long document length		
Retrieval model	None	Suggested stopword length	Full stopword length	None	Suggested stopword length	Full stopword length
BM25	0.1971	0.2135	0.2102*	*0.2355*	*0.2534*	*0.2536*
tf_idf	*0.2002*	*0.2144*	*0.2122**	0.2320	0.2510	0.2514
In_expC2	0.1939	0.2032	0.2035	0.2337	0.2423	0.2421*
In_expB2	0.1942	0.2047	0.205	0.2350	0.2440	0.2485
InL2	0.1959	0.2012	0.2018	0.2250	0.2322	0.2371
Hiem_LM	0.1589	0.1747	0.1738*	0.2194	0.2337	0.234*
Mean	0.19	0.2017	0.2017	0.23	0.2417	0.2433
% Change			+5.83%			+5.79%

Italic character defines the best performing retrieval model among different retrieval model evaluation

long documents as relevant compared with the short ones. However, in Gujarati and Sanskrit, the collection consists of nearly equal number of relevant documents in both long documents and short ones. We also observe that the different retrieval models prefer long documents at early ranks compared with short ones, which demonstrates that the evaluated retrieval models still preserve bias towards long documents.

9. Conclusion and future work

Stopword removal is an effective pre-processing step in IR. In the afore-mentioned experiments, we observed that the removal of stopwords improved MAP significantly compared with without stopword removal in general. In Marathi and Bengali, the BM25 and tf–idf models give better MAP than the DFR-based retrieval models. In Gujarati and Sanskrit, most of the retrieval models give similar performances. However, among them, language models are not that promising for the Indian languages considered. In Marathi, a small stopword length (21) gives the best or near-best MAP against longer stopword list (216). Similarly, for other three languages also (Bengali, Gujarati and Sanskrit), no significant differences are found over longer stopwords list. We also observe that the effect of stopword removal on retrieval performance is quite low in short documents compared with long ones. In Marathi and Bengali the differences in MAP scores between short and long documents are found to be quite high, whereas in Gujarati and Sanskrit they are comparatively low. In all the languages, most of the retrieval models give similar MAP scores in both short and long documents. In Sanskrit, the size of collection is very small compared with other three languages. Hence difference in length between short and long documents is very small. Experiments with larger collections for Sanskrit are therefore needed. Also, the effect of stopwords taken from the collections in-hand on retrieval performance is yet to be explored for collections of larger size of the languages considered and other languages as well.

Appendix

Table 15. Example of few non corpus-based stopword list used in experimentation

Marathi	Bengali	Gujarati	Sanskrit
या	অতএব	અથવા	कु
असो	অথচ	અને	एवम्
इये	অথবা	અમને	कस्मिन्
कमी	অনেকে	અમારું	कया
तेथ	অন্তত	અમે	बभूव
त्या	অনুযায়ী	અહીં	अस्मात्
होता	অনেক	આ	कच्चित्
अनेक	অন্য	આગળ	भवतः
हे	অবধি	આથી	सर्वम्
झाली	অবশ্য	આનું	तत्
ही	অর্থাত	આને	कस्मात्
माहिती	আগে	આપણને	युष्माकम्
येणें	আপনি	આપણું	एने
ते	আবার	ઉપર	अति
ऐसें	আমরা	આવે	अस्य
म्हणौनि	আর	એનાં	तदीयात्
आता	আমাকে	એવાં	तासु
तैसा	আমাদের	એવું	अमुना
का	উপর	ઓછું	सर्वेभ्यः
होते	আরও	અંગે	एनान्
करण्यात	ইহা	અંદર	तुभ्यम्

Acknowledgements

This research work was supported by IIT (BHU), Varanasi, India.

References

[1] Timothy B, Ian HW, John GC (1989) Modeling for text compression. ACM Computing Surveys (CSUR), 21(4):557–591

[2] Christopher F 1989 A stop list for general text. In *ACM SIGIR Forum*, vol. 24, pp. 19–21. ACM

[3] Stephen PH 1986 *Online information retrieval: Concepts, principles, and techniques*. Academic Press Professional, Inc.

[4] Ljiljana Dolamic and Jacques Savoy. When stopword lists make the difference. *Journal of the American Society for Information Science and Technology*, 61(1):200–203, 2010.

[5] Ljiljana Dolamic and Jacques Savoy. Comparative study of indexing and search strategies for the hindi, marathi, and bengali languages. *ACM Transactions on Asian Language Information Processing (TALIP)*, 9(3):11, 2010.

[6] Rachel T-WL, Ben H, and Iadh O 2005 Automatically building a stopword list for an information retrieval system. In *Journal on Digital Information Management: Special Issue on the 5th Dutch-Belgian Information Retrieval Workshop (DIR)*, vol. 5, pp. 17–24

[7] Christopher D 1999 Manning and Hinrich Schütze. *Foundations of statistical natural language processing*. MIT press

[8] Manu K 2006 *Text Mining Application Programming*. Charles River Media, Inc., USA, 1st edition

[9] Hakan A, and Sirma Y 2011 An automated domain specific stop word generation method for natural language text classification. In *2011 International Symposium on Innovations in Intelligent Systems and Applications*, pp. 500–503. IEEE

[10] Khalifa C, and Rayner A 2016 An automatic construction of malay stop words based on aggregation method. In *International Conference on Soft Computing in Data Science*, pp. 180–189. Springer

[11] Sileshi Girmaw Miretie and Khedkar. Automatic generation of stopwords in the amharic text. *International Journal of Computer Applications*, 180(10):19–22, 2018.

[12] Murphy C 2012 Effective listings of function stop words for twitter. *arXiv preprint* arXiv:1205.6396.

[13] Toluwase VA 2013 Entropy-based generic stopwords list for yoruba texts. *International Journal of Computer and Information Technology*, vol. 2, no. 5

[14] Jasleen K, and Jatinder kumar RS 2016 Punjabi stop words: A gurmukhi, shahmukhi and roman scripted chronicle. In *Proceedings of the ACM Symposium on Women in Research 2016*, pp. 32–37

[15] Jaideepsinh KR and Jatinder kumar RS (2016) Stop-word removal algorithm and its implementation for Sanskrit language. *International Journal of Computer Applications*, 150(2): 15–17

[16] Mark PS and David C 2003 Evolving better stoplists for document clustering and web intelligence. In *HIS*, pp 1015–1023

[17] Fotis Lazarinis. Engineering and utilizing a stopword list in greek web retrieval. *Journal of the American Society for Information Science and Technology*, 58(11):1645–1652, 2007.

[18] Jacques Savoy. A stemming procedure and stopword list for general french corpora. *Journal of the American Society for Information Science*, 50(10):944–952, 1999.

[19] Kripabandhu G and Arnab B 2017 Stopword removal: Why bother? a case study on verbose queries. In *Proceedings of the 10th Annual ACM India Compute Conference*, pp. 99–102

[20] Feng Z, Fu LW, Xiaotie D, and Song H 2006 Evaluation of stop word lists in Chinese language. In *LREC*, pp. 2497–2500

[21] Mohammad RD, Sanji M, and Aramideh M (2009) Farsi lexical analysis and stop word list. *Library Hi Tech*, 27(3):435–449, .

[22] Mohammad-Ali Y-Z-F, Behrouz M-B, Saeed R, and Saeed S. Pswg: An automatic stop-word list generator for persian information retrieval systems based on similarity function & pos information. In *2015 2nd International Conference on Knowledge-Based Engineering and Innovation (KBEI)*, pp. 111–117. IEEE

[23] Ibrahim AE-K 2017 Effects of stop words elimination for arabic information retrieval: a comparative study. *arXiv preprint* arXiv:1702.01925

[24] Bassam A-S, Fekry O, and Waseem ALR (2011) An experimental study for the effect of stop words elimination for arabic text classification algorithms. *International Journal of Information Technology and Web Engineering (IJITWE)* 6(2):68–75

[25] R Jayashree, K Srikanta Murthy, and BS Anami 2014 Effect of stop word removal on the performance of naïve bayesian methods for text classification in the kannada language. *International Journal of Artificial Intelligence and Soft Computing*, 4(2-3):264–282, .

[26] Harnani MZ, Norwati M, Masrah A AM, and Nurfadhlina MS 2017 The effects of pre-processing strategies in sentiment analysis of online movie reviews. In *AIP conference proceedings*, vol. 1891, p. 020089. AIP Publishing LLC

[27] Hassan S, Miriam F, Yulan H, and Harith A 2014 On stopwords, filtering and data sparsity for sentiment analysis of twitter. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pp. 810–817

[28] Hassan S, Miriam F, and Harith A 2014 Automatic stopword generation using contextual semantics for sentiment analysis of twitter. In *CEUR Workshop Proceedings*, vol. 1272

[29] Walaa Medhat, Ahmed Yousef, and Hoda Korashy. Egyptian dialect stopword list generation from social network data. *The Egyptian Journal of Language Engineering*, 2(1):43–55, 2015.

[30] Catarina S and Bernardete R 2003 The importance of stop word removal on recall values in text categorization. In *Proceedings of the International Joint Conference on Neural Networks, 2003.*, vol. 3, pp. 1661–1666. IEEE

[31] Feng X, Tian J, and Liu Z 2009 A text categorization method based on local document frequency. In *2009 Sixth International Conference on Fuzzy Systems and Knowledge Discovery*, vol. 7, pp. 468–471. IEEE

[32] Aqil A and Suha A-T 2009 Ikhtasir—a user selected compression ratio arabic text summarization system. In *2009 International Conference on Natural Language Processing and Knowledge Engineering*, pp. 1–7. IEEE

[33] Alexandra S, Måns M, and David M 2017 Pulling out the stops: Rethinking stopword removal for topic models. In *Proceedings of the 15th Conference of the European*

Chapter of the Association for Computational Linguistics: Vol. 2, Short Papers, pp. 432–436

[34] Yaoyong Li and John Shawe-Taylor. Using kcca for japanese–english cross-language information retrieval and document classification. *Journal of intelligent information systems*, 27(2):117–133, 2006.

[35] Debasis M, Mayank G, Sandipan D, Pratyush B, and Sudeshna S 2007 Bengali and hindi to english clir evaluation. In *Workshop of the Cross-Language Evaluation Forum for European Languages*, pp. 95–102. Springer

[36] Erbug C, Baturman S, and Burak G 2009 Turkish—english cross language information retrieval using lsi. In *2009 24th International Symposium on Computer and Information Sciences*, pp. 634–638. IEEE

[37] Chong TY, Rafael EB, and Chng ES 2012 An empirical evaluation of stop word removal in statistical machine translation. In *Proceedings of the Joint Workshop on Exploiting Synergies between Information Retrieval and Machine Translation (ESIRMT) and Hybrid Approaches to Machine Translation (HyTra)*, pp. 30–37. Association for Computational Linguistics

[38] Amit S, Chris B, and Mandar M 1996 Pivoted document length normalization. In *Proceedings of the 19th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '96, pp. 21–29, New York, NY, USA. Association for Computing Machinery

[39] Karen Sparck Jones. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21, 1972.

[40] Gianni Amati and Cornelis Joost Van Rijsbergen. Probabilistic models of information retrieval based on measuring the divergence from randomness. *ACM Transactions on Information Systems (TOIS)*, 20(4):357–389, 2002.

[41] Stephen ER, van Rijsbergen CJ, and Porter MF 1980 Probabilistic models of indexing and searching. In *SIGIR*, vol. 80, pp 35–56

[42] Djoerd H 2001 *Using language models for information retrieval*. Univ. Twente

[43] Chris B and Ellen MV 2017 Evaluating evaluation measure stability. In *ACM SIGIR Forum*, vol. 51, pp 235–242. ACM