Journal of King Saud University – Computer and Information Sciences xxx (xxxx) xxx

Contents lists available at ScienceDirect

Journal of King Saud University – Computer and Information Sciences

journal homepage: www.sciencedirect.com

Performance evaluation of text-mining models with Hindi stopwords lists

Ruby Rani *, D.K. Lobiyal

School of Computer and Systems Sciences, Jawaharlal Nehru University, New Delhi, India

ARTICLE INFO

Article history:
Received 17 November 2019
Revised 23 February 2020
Accepted 4 March 2020
Available online xxxx

Keywords:
Stopwords
Hindi language
Text classification
Text clustering
Topic modeling

ABSTRACT

Nowadays, several news portals, government websites, and social media sites are generating a massive amount of digitalized Hindi textual information. Stopword removal is a significant factor in text mining tasks that helps the miner to enhance the performance of a system. This paper attempts to construct the corpus specific stopwords lists for Hindi text documents using statistical and knowledge-based methods. In order to prepare the stopwords list, the proposed method considers the ranking of the words given by different methods followed by normalization of the outcomes of these methods using the social choice theory based vote ranking method. Further, we propose an evaluation method to evaluate the prepared stopword lists and investigate their behavior using text mining models. We also compare our prepared stopword lists with the baselines and conclude that the technique which fetches the best features does not necessarily identify the candidate stop words. To the best of our knowledge, the proposed approach guarantees the removal of candidate stop words and has the least information dissipation.

1. Introduction

Recently, the exponential growth of digital Hindi text data generated by several websites, including social networking websites, government websites, and bloggers have ameliorated the user towards the internet platform (Gulati and Sawarkar, 2020). Nowadays, several Hindi-text based applications exist in the area of machine translation (Singh et al., 2017), information retrieval (Kumar et al., 2019), text summarization (Verma et al., 2019), named entity recognition (Thomas and Sangeetha, 2019), and other linguistic perspectives. There is a broad scope of classifying the extracted text of the Hindi document into pre-defined high-quality information, shortening the long piece of text into short informative data, and translate the text from one language to another languages. Generally, when dealing with the text, we encounter several words, including denotational reference or impact on the semantics of the documents, so-called stopwords.

Stopwords have no discriminating power and least predictive capability. It is of two types: generic and domain-specific stopwords. Generic stopwords are the grammatical words that help in sentence formation and have no independent significance (Petras et al., 2003). These are considered as the standard stopwords and mostly available in all categories of documents. Currently, many researchers and academicians who are working in natural language processing (NLP) and text mining confront such words very frequently (Choy, 2012). On the other side, the words with negligible discrimination power in the domain-specific documents are called as domain-specific stop words. These words vary from one document to another based on the domain of the document (Sinka and Corne, 2003). For instance, the word *'speech'* can be a stop word in politics, but it is a significant term in computer science. Like generic stop words, domain-specific stop words also depend on the sparsity of the document, vocabulary size, and the number of sub-domains present in one domain. There have been constructed some domain-specific stopword lists that include domains such as human resource, gene ontology, physics, bioinformatics, and history (Crow and DeSanto, 2004; Seki and Mostafa, 2005).

Irrespective of the different nature of generic and domain-specific stopwords, the stopwords removal has some merits and demerits according to their applications in different areas. The advantages of removing stopwords during text mining are to reduce the dimension of the text document (Ricardo and Modern,

* Corresponding author.
E-mail addresses: ruby73_scs@jnu.ac.in (R. Rani), dkl@jnu.ac.in (D.K. Lobiyal).
Peer review under responsibility of King Saud University.

Production and hosting by Elsevier

https://doi.org/10.1016/j.jksuci.2020.03.003
1319-1578/

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

1999), diminish the size of vocabulary up to 40%, reduce memory overhead, reduce noise and false positives and improve the performance of a system. For instance, suppose we search a query, e.g., "*what are the benefits of stopword removal in text mining*?" on any search engine with sample space of N web pages. If the search engine wishes to search web pages which contain the words "*what*", "*are*", "*the*", "*benefits*", "*of*", "*stopword*", "*removal*" "*in*" "*text*" "*mining*?" the search engine is going to search the web pages in the time of order $O(N^{10})$. These pages also contain the words "*what*", "*are*", "*the*", "*of*", "*in*" as they are so commonly used in the English language. However, if we search a query by alleviating these words, the search engine can extract the web pages which involve the words "*benefits*", "*stopword*", "*removal*", "*text*", "*mining*?" in time of order $O(N^5)$. These commonly used words, generally known as stopwords have negligible significance in text mining and so eliminate them from the text document will make the query simple.

Thus, the primary benefit of stopword removal is to improve the system performance in terms of memory reduction and saving processing time during indexing and searching respectively. Sometimes, reduction of stop words from user queries, generic documents, and labeled text documents removes the semantically important terms which in result hurts the IR models implementation. These meaningful terms include abbreviations and symbols (Song et al., 2005). Besides, in sentiment analysis where the term 'not', 'no' and 'never' have a significant role but due to linguistic correlation with many words, these are removed as stop words that don't guarantee execution enhancement. Accordingly, careful consideration should be taken by researchers for the removal of stop words (Riloff, 1995; Forman, 2003).

1.1. Motivation and contribution

From literature survey, as discussed in Section 2, we observe that several stopwords lists have been prepared for English, Russian, Arabic, and Chinese languages. It is found that nearly 341 million native speakers around the world speak and understand the Hindi language (WorldoMetero, 2019). Currently, many industries and academicians are working on Hindi text in various text mining areas such as sentiment analysis (Joshi et al., 2016; Akhtar et al., 2016), text clustering (Jain et al., 2016; Tayal et al., 2015), document summarization (Kumar et al., 2015; Kumar et al., 2015), text categorization (Harikrishna and Rao, 2015), and entity extraction (Thomas and Sangeetha, 2019; Singh et al., 2018; Rao et al., 2015). Several foreign languages construct stopword lists based on Zipf's law (Riloff, 1995; Forman, 2003; WorldoMetero, 2019; Joshi et al., 2016; Jain et al., 2016; Zipf, 1932). According to Zipf's law (Zipf, 1932), stopwords are the terms that carry high document frequency. Although, it has been reported that this observation does not fit well in some real-time applications, where the documents are not uniformly distributed over categories. Thus, the unequal distribution of documents across different categories, in the Hindi language restricts Zipf's law (Zipf, 1932) for constructing stopword lists. No such domain-specific stopword lists are available for the Hindi text documents. In (Jha et al., 2016; Pandey and Siddiqui, 2009; Rani and Lobiyal, 2018), some generic stopword lists are available but they are not suitable for domain-specific text mining tasks. In the last decade, Zou et al. (2006) proposed four approaches for preparing a generic stopwords for Chinese language. However, they did not evaluate their stopwords list. Thus, the unavailability of stopwords list for Hindi Language and Zou et al. (2006) approach motivate us for constructing stopwords list for Hindi Language.

In this paper, we borrowed the notion of Zou et al. (2006) approach in order to address the issue for Hindi Language. In addition, the proposed work has the following contributions.

- We prepare the real-time domain-specific dataset for the Hindi Language, where we collect real-world data of various domains from different web sources such as online websites, news portals, and blogs. We consider the Indian language (IL) Crawler and IL-Sanitizer as the software tools to crawl and sanitize the dataset respectively.
- We construct automatic domain-based stop word lists on sanitized labeled corpus using ranking measures, such as modified conventional statistical techniques and knowledge model-based techniques.
- We also stabilize the outcomes of different ranking measures using social choice theory based voting method in order to output the normalized-stopwords lists.
- Further, we propose a new Netting Ranked Performance Evaluation (NRPE) approach, which evaluates the domain-based as well as normalized-stopwords lists, and examines their behavior. The NRPE extracts the terms from the proposed domain-based and normalized stopword lists, followed by it evaluates the prepared stop words list based on the performance of the K-nearest neighbor (KNN) classifier, K-Means clustering, and Latent Dirichlet Allocation (LDA) topic modeling methods. It also validates the strength of the corpus-based as well as normalized-stopwords lists in order to identify whether the term is either feature term or stopword.
- Moreover, we use the proposed NRPE approach to evaluate the baselines methods and compare them with the prepared stopword lists. The performance analysis demonstrates that proposed automatic domain-based stop word lists cover almost all of the stop words, given in currently available online lists (Taranjeet, 2018; Ranks, 2018; GitHub, 0000).
- The experimental section illustrates that the removal of proposed stopword lists from the text documents has a significant impact on performance enhancement of text classification, text clustering, and topic modeling models.

The remaining paper is arranged as follows. In Section 2, we give the related work. Section 3 discusses the background. Section 4 explains the proposed automatic domain-based stop list construction. In Section 5, we propose a new performance evaluation method. Section 6 gives the experimental results and discussion. The conclusion and future work are given in Section 7.

2. Related work

This section briefly gives the literature survey of methods of constructing stopwords list for English, web-based, Non-English and Hindi languages.

Traditional stopwords lists. In 1957, H. P. Luhn (Luhn, 1957) discussed the notion of stopwords, where he investigates the text document based on the statistical probability. In approach (Luhn, 1957), the author addresses the level of subjects specificity and those words that convey the most meaning. In 1979, Van Rijsbergen (Van Rijsbergen, 1986) first proposed an approach for stopwords extraction for the English language, which is one of the most suitable stopwords list used in applications such as NLP and information retrieval. Later, Fox (Fox, 1992) prepared the most adoptable stopwords list for the English language, which discussed the design and implementation of a lexical analyzer and stopword lists for information retrieval. Fox (Fox, 1992) found that incorporating stopword lists with a lexical analyzer is an efficient way for constructing a stopword list. In this similar manner, Francis et al. (Francis and Kucera, 1982) prepared the 425 stopword lists for the English language. In (Makrehchi and Kamel, 2008), Makrehchi et al. discussed the extraction of domain-based stopword lists from a labeled corpus, in which the authors consider the idea of back-

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

ward filter-level performance and sparsity of trained data to extract the stopwords. Although, the stopword extraction, given in (Makrehchi and Kamel, 2008) is based on document frequency and their ranking is evaluated using a traditional method known as backward filter-level performance method. Recently, Makrehchi et al. (Makrehchi and Kamel, 2017) extend the work, given in (Makrehchi and Kamel, 2008). White et al. (White et al., 2007) built a corpus of 36,000 products in English language and constructed a domain-based stopwords lists after linguistic analysis. In (Crow and DeSanto, 2004; Seki and Mostafa, 2005) authors manually created stopwords lists for different domains like physics, bioinformatics, and human resource management.

Web-based stopwords lists. Sinka et al. (Sinka et al., 2003) generate the stopwords lists for web-based documents using the unsupervised methods, i.e., word entropy followed by they evaluate the lists by web clustering scheme. In (Sinka and Corne, 2003), Sinka et al. optimized the stopwords generated in (Sinka et al., 2003) using a stochastic search algorithm and k-means clustering. In (Kawahara and Kawano, 2001), Kawahara et al. suggested a Receiver Operating Characteristics (ROC) analysis-based association methods for generating stopwords lists.

Non-English Stopwords lists. Other than the English language, stopwords lists for other languages have been issued, including Russian (Petras et al., 2003), Arabic (El-Khair, 2006), and Farsi (Taghva et al., 2004). Zou et al. (Zou et al., 2006) constructed two lists of stopwords for the Chinese, language based on statistical and knowledge-based models. Another stopwords list for the Chinese language is developed by Hao et al. (Hao and Hao, 2008), this list is based on the weighted Chi-squared approach.

Hindi Language stopwords lists. Recently, papers (Harikrishna and Rao, 2015); (White et al., 2007) discussed the social influence language that is the outcomes of social networking tools such as Facebook, Twitter, Chatting apps, and emails. Some stopwords lists for the Hindi language has also been noted in (Jha et al., 2016; Pandey and Siddiqui, 2009; Singh and Siddiqui, 2012); (Sharma and Namita Mittal, 2019). Recently, Rani et al. (Rani and Lobiyal, 2018; Rani and Lobiyal, 2018) discussed generic stopwords lists and domain-specific stopwords lists respectively. In this paper, we extend (Rani and Lobiyal, 2018) to construct domain-specific stopwords list for Hindi corpora by employing the methodology given by Zou et al. (Zou et al., 2006) and evaluate their performance using text mining and machine learning models.

3. Background

In this section, we discuss the concise outline of software tools used for Hindi corpora creation, online Hindi web sources for text collection, a methodology for the preparation of Hindi corpora and characteristics of created corpora.

3.1. Software tools

Here, we give a brief overview of the Indian Language crawler (IL-crawler) and sanitizer (IL-sanitizer) tools, and functionality of these tools in automatic crawling and cleaning of the huge size of the corpus from different web sources.

Indian Language Crawler: The Computational linguistics R&D group implements the Indian language crawler tool at Jawaharlal Nehru University, India. It is a variant of web crawler that browses and extracts the useful information for Indian languages such as Bengali, Bhojpuri, Hindi, Odia, Urdu, and Maithili from the different websites. During crawling, it avoids useless data like advertisements and promotional content. Based on the structure of the webpage, it uses a depth-first search or breadth-first search approach to crawl the data (Choudhary and Jha, 2011).

Indian Language Sanitizer: The Indian Language sanitizer, a cleaner tool separates the undesired and redundant or harsh kinds of stuff from the corpora. For example, it removes the header of top web-pages like education, sports, news, date, and author name from web pages. Further, it creates systematic files containing valuable information that can be used to serve various areas, such as, text mining, information retrieval, and NLP tasks (Choudhary and Jha, 2011).

3.2. Borda's vote count approach

Borda's count method (Myerson, 2013) determines the result of the discussion of an election by providing each candidate some points or preferences. In this one winner election scheme, each voter casts his vote based on his different preferences for candidates. For example, the voter casts his vote 'n' to his most favorite candidate, $'n-1'$ to the second most favorite candidate and so on. Eventually, the sum of the preferences given to the candidate by all voters is the final rank of him, and he is declared as the final winner with the first rank.

3.3. Supporting evaluation techniques

Here, we discuss the supporting evaluation methods to validate the prepared stopwords lists. For validation, we consider the following three text mining models as external influencers: KNN text classifier, K-Means text clustering, and LDA-topic modeling. The rationale behind the selection of these models is that the accuracy of these influencers is affected by the presence or absence of stopwords.

- *KNN-Classifier*: It is a text classification approach, simple in implementation and requires less training (Guo et al., 2006). The K-fold cross-validation method has been adopted to measure the accuracy of KNN (Kevin, 2019).
- *K-Means Clustering*: It is a document clustering analysis method and very sensitive towards noise such as irrelevant terms (stop words). We consider the Elbow curve to find the optimal number of clusters and, precision, recall, and F-Score as the accuracy measures (Wikipedia, 2019).
- *LDA-Topic Modeling*: This is a statistical approach in text mining to find hidden thematic patterns from the document which highly depend on the nature of the data. Generally, LDA finds the qualitative topics from the data free from stop words (Blei, 2012). Further, the quality of the thematic patterns (topics) is determined by log-perplexity. Log-perplexity is a measure to capture the 'uncertainty' of the model in predicting useful information based on the assigned probabilities. The lower value of perplexity implies less uncertainty of the model in prediction (Benjamin, 2018).

4. Construction of stopword lists

In this section, we prepare the stopwords lists for the Hindi language by tweaking the ranking measures, such as conventional Statistical and Knowledge-based methods. Each ranking measure produces a different result. In order to achieve the final stopword list, we normalize the overall result prepared from all lists using R.B. Myerson's social choice method (Myerson, 2013).

4.1. Notation and abbreviation

Table 1 describes the notations and abbreviation used in this paper.

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

Table 1
Representation and Illustration.

Representation	Illustration
T	Total count of distinct terms
Y	Total count of files in the corpora
$\boldsymbol{w_i}$	The i^{th} term in the file, $1 \leq i \leq T$
$\boldsymbol{D_j}$	The j^{th} file in the corpora, $1 \leq j \leq Y$
$\|\|D\|\|$	Total count of terms in file D
$\boldsymbol{MLT}(\boldsymbol{w_i})$	Mean Probability of i^{th} term
$\boldsymbol{TF_{ij}}$	Term-Frequency of the i^{th} term in the j^{th} file
$\boldsymbol{VLT}(\boldsymbol{w_i})$	Variance Probability of i^{th} term
$\boldsymbol{MVR}(\boldsymbol{w_i})$	Mean-Variance Ratio of i^{th} term
$\boldsymbol{MAD}(\boldsymbol{w_i})$	Mean Absolute Deviation probability of i^{th} term
$\boldsymbol{MDR}(\boldsymbol{w_i})$	Mean-Deviation Ratio of i^{th} term
$\boldsymbol{TE}(\boldsymbol{w_i})$	Term entropy of an i^{th} word

4.2. Proposed ranking measures

4.2.1. Statistical models

Here, we constructed the stopwords lists by tweaking numerous existing statistical techniques. Statistically, several researchers argue that the significance of a term can be defined by their frequency in the document, that we call the *Term-Frequency (TF).* The paper (Shannon, 1948) extracts the noisy words based on their frequency in the documents. The paper (Kantor and Lee, 1986) states that the word with high frequency is correlated to stopword but this rule doesn't apply in all cases. Inspired from (Shannon, 1948) and (Kantor and Lee, 1986), we prepared a stopwords list for Hindi language by tweaking the five traditional statistical technique as follows: Mean of Log-TF (MLT), Variance of Log-TF (Var), Mean-variance ratio (MVR), Mean Absolute Deviation (MAD) and Mean Absolute Deviation Ratio (MDR).

- *Mean of Log-TF (MLT):* A variant of statistical mean, MLT computes value of each word w_i is the ratio of aggregate normalized TF of word w_i in all files to the total number of documents 'Y' in the dataset, as given in Eq. (1).

$$MLT(w_i) = \frac{|\sum_{j=1}^{j=Y} Log_e(TF_{ij})|}{Y} \tag{1}$$

- *Variance of Log-TF (*Var*):* Var measures the displacement of word w_i in document D_j from MLT value and provides the stable distribution of the word, as defined in Eq. (2).

$$Var(w_i) = \frac{\sum_{j=1}^{j=Y}(MLT(w_i) - TF_{i,j})^2}{Y} \tag{2}$$

- *Mean-variance ratio (MVR):* MVR computes the ratio of MLT and Var values using Eq. (3)

$$MVR(w_i) = \frac{MLT(w_i)}{Var(w_i)} \tag{3}$$

- *Mean Absolute Deviation (MAD):* MAD is the mean of the absolute deviation of data value from its central value. It is a slight variant of original mean absolute deviation method. It is calculated by Eq. (4).

$$MAD(w_i) = \frac{\sum_{j=1}^{j=Y}|MLT(w_i) - TF_{ij}|}{Y} \tag{4}$$

- *Mean Absolute Deviation Ratio (MDR):* In MDR, ratio between MLT and MAD is calculated using Eq. (5).

$$MDR(w_i) = \frac{MLT(w_i)}{MAD(w_i)} \tag{5}$$

4.2.2. Knowledge-based entropy model

In the Hindi language, there are some common characters, like 'क', 'म', 'ह', 'प', and 'र' while some are rarely used, 'ङ', 'ञ', and 'ण'. The random nature of the upcoming Hindi character to be employed in an encoded string makes the task of text mining more difficult. Claude E. Shannon (Shannon, 1948) states that entropy is an expression of the randomness of the system that plays an important role in information theory. We assume the distribution of Hindi words in our prepared dataset as a Shannon channel, in which only the highest information (feature terms) can be passed through the channel by filtering noise (stop words) in the channel. According to this theory in text mining applications such as IR, NLP and linguistics applications, more the number of stop words available in the document lesser efficient is the text mining systems and vice versa (Kantor and Lee, 1986). Inspired from the randomness of Hindi corpus, we evaluate the corpus based on the entropy of a word (term), simply we call it as "Term-Entropy" (TE). Now, we define the TE of a word w_i in terms of term frequency$TF_{i,j}$, given in Eq. (6).

$$TE(w_i) = \sum_{j=1}^{j=Y} TF_{i,j} \times \log(\frac{1}{TF_{i,j}}) \tag{6}$$

4.3. Election voting method for ranking

It is observed from the previous stopwords lists constructed by statistical and knowledge-based methods that each technique outcomes the resultant words in different ranking order. It is not a good idea to consider the stop words list prepared by one technique as a standard list. Here, we normalize all lists to obtain the final unbiased stopwords list. Borda count method (Myerson, 2013; Myerson, 1996) is explained in Section 3.2 is used to achieve normalized list from all lists and give the final stopwords list. As we can see in Table 2, top-10 terms are ranked in ascending order by different techniques. We assume each method as a voter and words as the candidates while rank assigned to the word is the rank assigned by the certain statistical and knowledge-based ranking measure to the word. In the end, the final rank is computed using Borda's count method in which, a word is ranked based on the sum of weights given to it by each technique and the final stop words list is prepared by putting a threshold constraint. For example: if the word 'और' (and) is ranked '1′ by MLT while ranked '2′ by TE in constructed stop list. Hence, the final rank given to the word 'और' (and) is 3.

5. Proposed performance evaluation method

It is well-known that the performance of the text model is potentially defined by the presence or absence of stopwords. The validity of these stopwords lists is computed by measuring the performance of information mining systems.

Makrehchi and Kamel (2017), in their paper, measure their prepared stopwords lists performance using text classification method. Although, text classifier (Rocchio classifier) was not giving better results as it is sensitive to noise and costly as much as possible. In this paper, we use three performance measure models, i.e., K-nearest neighbor (KNN) classifier, K-Means clustering, and Latent Dirichlet Allocation (LDA) topic modeling to evaluate the prepared stopwords lists.

5.1. Netting ranking performance evaluation

Suppose, we have two lists, say $S_X = \{w_{X1}, w_{X2}, \cdots, w_{Xn}\}$ and $T_X = \{w_{Xn}, w_{X(n-1)}, \cdots, w_{X1}\}$ denoted as the ranked stop words lists in ascending order and descending order for techniques X

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

Table 2
Top-10 Ranked Words by ranking measures and Borda's Count methods.

Statistical methods					Knowledge-based method	Borda count method
MLT	Var	MVR	MAD	MDR	TE	Final Rank
इस (is)	प्रियंका(priyanka)	के (ke)	समाचार(samaachaar)	के (ke)	के (ke)	के (ke)
भी (bhee)	बल्की (balki)	और (aur)	हुई (huee)	की (kee)	मे (mein)	है (hain)
पर (par)	क्षेत्र (kshetr)	मे (mein)	मे (mein)	होती(hoteen)	है (hai)	उसे (use)
तो (to)	सवाल (savaal)	है (hai)	फरि (phir)	हौव्वा(hauvva)	की (ki)	कर (kar)
नहीं(nahin)	जल (jal)	को (ko)	जो (jo)	को (ko)	को (ko)	कुछ (kuchh)
हैं (hain)	वाली (vaalee)	पर (par)	इन (in)	की (ki)	से (se)	किया (kiya)
यह (yah)	ज्यादा (jyaada)	की (ki)	जब (jab)	ने (ne)	की(kee)	होता(hota)
लिए (lie)	साफ (saaph)	का (ka)	रहे (rahe)	है (hai)	का (ka)	बात (baat)
और (aur)	जबकी (jabaki)	तथा(tatha)	उसे (use)	पर (par)	पर (par)	दिया (diya)
हो (ho)	उसका (usaka)	भी (bhee)	देने (dene)	का (ka)	यह (yah)	करना (karana)

respectively. Let w_{Xi} denotes the i^{th} ranked-word w in technique X, where X represents the different techniques, i.e., MLT, Var, MVR, MAD, MDR, TE and Borda's count. We define w_{X1} as a word with poor knowledge (stopword) and w_{Xn} as a word with rich knowledge (feature) for technique X. We select the words based on their ranking by applying the best-m rule, in which first m words are omitted and rest are retained. We consider *a* as the band size that approaches the value of m, known as the threshold value.

Algorithm 1. Leading Net Ranked Performance Evaluation (L-NRPE)

***Input**: On given n words (number of words in the dataset)*
***Output**: Give Combined Net Ranked Performance Evaluation in the Leading manner for different performance measure models.*
1. *Omit first-m words from S_X, wherem $\leq n$.*
2. *Input "$n - m$" remaining words from the S_X as the features.*
3. ***Repeat for each performance measure models***
4. *i = a*
5. *$CLB_X = 0$*
6. ***While** (i<=m)*
7. *$CLB_X = \frac{CLB_X + Prf(S_X(n) - S_X(a))}{m/a}$*
8. *i = i + d*
9. ***End while***
10. ***End foreach***

Based on choosing M-rich knowledge and M-poor knowledge words, two different ways of performance evaluation methods are setup. The first method is known as the Leading-NRPE that evaluates the performance based on selecting M-rich knowledge words from the vocabulary S_X and another is known as the Trailing-NRPE works on M-poor knowledge words from the vocabulary T_X. The final performance of three text mining influencers with different levels of the band (defined over proposed stopwords lists) is known as the netting ranking performance evaluation (NRPE). There are two objectives that NRPE achieved. The proposed NRPE extracts the features terms for different text mining models followed by evaluates and validates the prepared stop words lists based on the performance of classifier, clustering and topic modeling. It also validates the strength of the domain-based as well as normalized-stopwords lists in order to identify whether the term is either feature term or stopword.

In the simulation, we assume the threshold value m = 1000 words for both lists S_X and T_X and the band size a = 100 words. Suppose d = 100 is an incrementing factor, that is defined as the difference between two band sizes, i.e., $d = a_2 - a_1$. *Algorithm 1* computes the Combined Leading Band (CLB_X) NRPE for each ranking measures, defined in Section 4.2. In *Algorithm 1*, the method omits m words from the set of the dataset and remaining 'n-m' words is considered as the features words. The method computes the performance of a list of feature words using *Prf*() function, which evaluates the accuracy of a particular influencer model on the given band size. For example, precision is considered as the accuracy measure of K-means clustering. Finally, method computes the Combined Leading Band (CLB_X) NRPE for each ranking measures. Similarly, *Algorithm 2* computes the Combined Trailing Band (CTB_X) NRPE for each ranking measures. In *Algorithm 2*, the method computes the performance of list of feature words using *Prf*() in trailing manner, which also evaluates the accuracy of the used text mining influencer method. In the end, method computes the Combined Trailing Band (CTB_X) NRPE for each ranking measures. Table 6 illustrates the performance areas covered under CLB and CTB over different ranking technique curves versus different external text mining influencers as described in Section 3.3.

Algorithm 2. Trailing Net Ranked Performance Evaluation (T-NRPE)

***Input**: On given n words (number of words in the dataset)*
***Output**: Give Combined Net Ranked Performance Evaluation in Trailing manner for different performance measure models.*
1. *Omit first-m words from T_X, wherem $\leq n$.*
2. *Input "$n - m$" remaining words from the T_X as the features.*
3. ***Repeat for each performance measure models***
4. *i = a*
5. *$CTB_X = 0$*
6. ***While** (i<=m)*
7. *$CTB_X = \frac{CTB_X + Prf(T_X(n) - T_X(a))}{m/a}$*
8. *i = i + d*
9. ***End while***
10. ***End foreach***

6. Experiments results and discussions

6.1. Benchmark datasets

To the best of our knowledge, no sufficient authentic domain-specific Hindi textual data is available on the market. Therefore, we first prepare the domain-specific corpora by collecting real-world data from different online portals. Some of the portals in document collection includes the "*NaiDunia*", "*Vigyandunia*", "*Center for Advanced Study of India*", and "*Kisanhelp*". The collected data is aggregated from more than 4200 articles of different 11,000 online web pages. The collected dataset covers four different domains: politics (PL), agriculture (AG), economy (EC), and entertainment (ENT).

We simulate the experiment in Java 1.2 run on Intel(R) Core (TM) i3 3110 M CPU @ 2.4 GHz, 64-bit window 10 operating sys-

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

Table 3
Description of the prepared dataset for Hindi text.

Parameters	Description
Total Number of Text Files	Around 18,400 documents
Input Corpora Size	Around 231 MB
Average number of words in a document	580 words
Largest file size	20 KB
Smallest file size	6 KB
Domains Traversed	AG, EC, ENT, PL
Number of files of agriculture domain	Number of Crawled documents –4832, Number of Cleaned documents –4329
Number of files of the economy domain	Number of Crawled documents –5136, Number of Cleaned documents –3524
Number of files of entertainment domain	Number of Crawled documents –3523, Number of Cleaned documents –2895
Number of files of politics domain	Number of Crawled documents –3825, Number of Cleaned documents –3021
Number of terms in whole corpora	63,54,785
Number of Distinct terms	21,17,208

Table 4.1
Top-10 Borda's Count ranked words of each domain in leading fashion.

Agriculture (AG)	Economy (EC)	Entertainment (ENT)	Politics (PL)
के, (ke), (of)	के, (ke), (of)	के, (ke), (of)	के, (ke), (of)
की, (ki), (of)	और, (aur), (and)	है, (hai), (is)	भी (bhee), (also)
करना (karana)	में, (mein), (in)	की, (ki), (of)	है, (hai), (is)
दुधाची (Dudhācī) (Milk)	की, (ki), (of)	में, (mein), (in)	ऐसी (aisee), (such)
हाइड्रोमैट्रिक (haidromaitrik) (Hydromatic)	सवीता (Savita) (Savita)	परन्तू, (parantoo), (but)	में, (mein), (in)
रिपोर्ट: (riporth) (report)	योजनावार, (yojanaavaar), (Plan Period)	आर्यभट, (aaryabhat), (Aryabhatta)	का, (ka), (of)
बरखा (barakha) (monsoon)	वाले (vaalen) (Walon)	मन्दार, (Mandāra), (Malar)	से, (se), (from)
जबीन(jabeen) (jabeen)	शिक्षावल्ली, (shikshaavallee), (Education School)	एलेक्जेंडर (elekjendar) (Alexander)सुधाकर, (Sudhakar), (Sudhakar)	कि, (ki), (that)
वचिार (vichaar) (idea)	सरोवरे, (Sarōvarē), (lakes)	एड्मशिन, (edmishan), (Admission)	की, (ki), (of)
वर्ष: (Varṣaḥ) (Year)	हास्यस्पद, (Hā syaspada), (funny)	अग्रानीत, (agraaneet), (Afaritan)	को, (ko), (to)

tem, 8 GB RAM. First, we retrieve the data from numerous online web sources using the IL crawler on different time duration, for example, data collected from "NaiDunia Editorial's" web page's URL is from 1 September 2015 to 31 March 2019. Once the data is extracted, it is cleaned by the IL sanitizer tool and arranged into systematic files. To discard the compatibility difficulties, we encode the files into UTF-8 format. Table 3 summarizes the statistics of the dataset used in this experiment.

6.2. Evaluation of proposed lists

In this section, we compute the performance of our prepared stopwords lists on text classification (KNN), text clustering (K-Means) and topic modeling (LDA-topic modeling).

First, we construct the domain-specific stopwords list using the aforementioned ranking measures for Hindi text on large corpora (231 MB) for four domains including AG, EC, ENT, and PL. Second, we normalize the prepared lists using the Borda count method to give the final list as demonstrated in Section 4. The experiments have been performed in two phases. Third, the performance of the stopwords lists prepared using proposed ranking measures is evaluated. Four, comparison of the proposed stopwords lists against the stopwords lists prepared by baseline methods is done. In both kinds of experiments, text mining models including KNN text classifier, K-Means text clustering, and LDA topic modeling have been employed as external influencers for performance evaluation of prepared stopwords lists. The baselines considered for comparison are No-filter (no stopwords removal), standard stopwords list (Ranks, 2018), TF-IDF (Rajaraman, 2011), Generic stop list (Rani and Lobiyal, 2018), and High-Low word count (terms with highest and lowest count).

Now, we compute the performance of domain-wise (AG, EC, ENT, PL) stopwords lists built using Borda's count along with the statistical and knowledge-based methods in both leading and training orders using *Algorithm 1* and *Algorithm 2* respectively. Domain-specific (prepared on given corpus) stopwords lists generated by Borda's count method is given in Table 4.1(leading fashion) and 4.1(trailing fashion) respectively. The impact of prepared stopwords lists using external influencers (text mining models) is shown in Fig. 1–3 (Table 4.2).

6.2.1. Text classifier

This section evaluates the performance of prepared stopword lists using proposed NRPE approach by employing external influencer i.e., KNN classifier. Fig. 1.1 illustrates the correlation between the Leading-NRPE and Netting bands for the proposed ranking measures, studied in this paper. Fig. 1.2 illustrates the correlation between the Trailing-NRPE and Netting bands for the proposed ranking measures. Each graph has approximately 350 experiments, including 10 numbers of bands, 5 k-fold cross-validations and 7 number of ranking measures. During experiments, the KNN classifier is used due to its simple implementation and susceptible behavior towards noisy words. From Fig. 1.1, we predict that band size inversely impacts the accuracy of the KNN over MVR method, that is, as the band size increases, the graph shows the consistent fall in accuracy of the classifier. Thus, we can say that the top-ranked terms in the MVR method are stopwords. Unlike MVR, MLT and TE methods show the random behavior and thus are not considered as reliable as MVR. Also, in terms of area coverage, MVR covers the largest CLB area while MLT covers the least area. So, MVR can be considered as the good method for stopwords extraction. From Fig. 1.2, we observe that the MLT method has consistent lowest accuracy as compared to other methods. Thus, MLT is not good for stop word extraction, in reverse, it can be viewed as the good feature selection approach on the same dataset. Conversely, MVR is good in fetching candidate stop words. Due to the unstable nature of MDR and MAD methods, they could not be considered for extracting candidate stopwords. The larger area covered under combined trailing band (CTB) in T-NRPE implies the high capability of ranking measures in stop word extraction, as summarized in Table 5.

Now, we compare the NRPE of the proposed ranking measures that are MLT, var, MVR, MAD, MVR, TE and final rank with the considered baseline methods such as with the baselines no filter, standard list, TF-IDF, generic list, and high-low in a leading as well as trailing manner. Fig. 1.3 illustrates the comparison of Leading-NRPE of proposed ranking measures, with the baseline methods. The figure shows that MVR and Borda's techniques are more effective in candidate stop words selection. The traditional TF-IDF technique shows satisfactory results but we cannot get good classification results by considering traditional lists or the generic stop words list on the given dataset. Moreover, the removal of stop words based on its extremely high and low frequency do not lead to significant improvement in classifier accuracy. Fig. 1.4 demon-

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

Table 4.2
Top-10 Borda Count ranked Words of each domain in trailing fashion.

Agriculture	Economy	Entertainment	Politics
जैवरसायनकिी, (jaivarasaayanikee), (Biochemistry)	अक्षुण्ण, (akshunn), (intact)	क्रोधकि, (krodhik), (Angry)	एयरबेस, (eyarabes), (air Base)
मूंगडाइन्चा, (moongadaincha), (Moongdine)	योजनाकार, (Yōjanākāra), (Planner)	कपलि, (Kapila), (Kapila)	नागरकि, (naagarik), (Citizen)
हड्डीरहति, (haddeerahit), (Boneless)	बढ़ायी, (badhaayee), (enhanced)	काशीनाथ, (kaasheenaath), (Kashinath)	प्रधानमंत्री, (pradhaanamantree), (Primeminister)
कार्बेन्डाजमि, (kaarbendaajim), (Carbendazim)	रक्षामंत्री, (rakshaamantree), (defence minister)	इंसाफ, (insaaph), (Justice)	अर्थतंत्र, (arthatantr), (Economy)
मूल्यवान्, (Mūlyavān), (Value)	फोक्सवैगन, (phoksavaigan), (Volkswagen)	जसबीर,(jasabeer), (Jasbir)	आतंकवादी, (aatankavaadee), (Terrorist)
तलिाड़ी, (tilaadee), (Tahadi)	रोज़गारोन्मुख, (rozagaaronmukh), (job oriented)	बालमणी, (Bālamaṇi), (Balamani)	वोटबैंक, (votabaink), (votebank)
गरिावत, (Girāvata), (falling)	अचीवमेंट, (acheevament), (achievement)	फुटबाल, (futabaal), (Football)	इंडयिा, (indiya), (India)
भूमि, (bhumi), (land)	शक्षिाक्षेत्र, (shikshaakshetr), (study field)	आयुर्वेज़्ञानकि, (aayurvaigyaanik), (Anaerologist)	चर्चा, (charcha), (Discussion)
कृषि, (krshi), (Agriculture)	कॉंटेस्ट, (Kŏn̤tēsṭa), (Contest)	पंजाबीपीडयिा, (panjaabeepeediya), (PanjabiPedia)	हंदि, (hind), (Hind)
परविहन, (parivahan), (Transportation)		धर्मजीवन, (dharmajeevan), (Religion Life)	सपा-कांग्रेस, (sapa-kaangres), (SP-Congress)

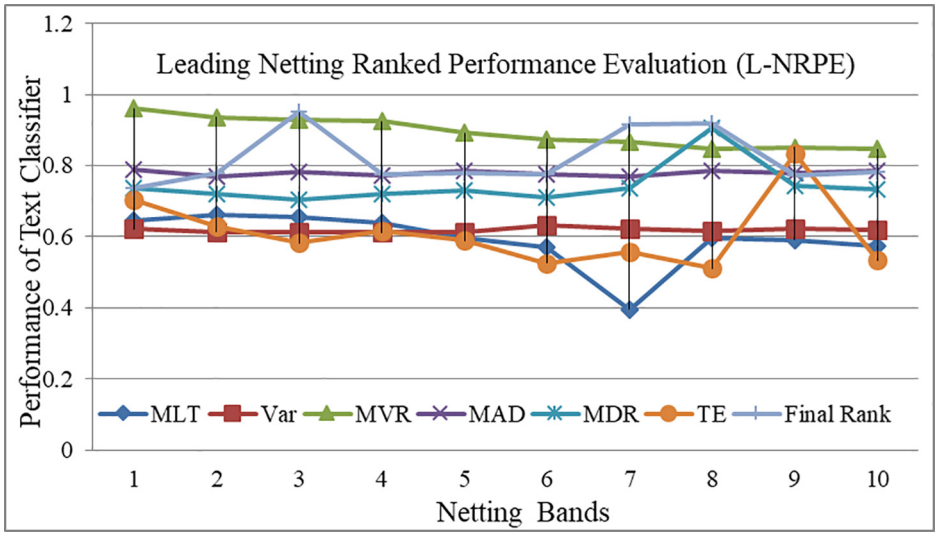

Fig 1.1. Leading Ranked Bands performance of the ranking measures on whole corpora using **KNN.**

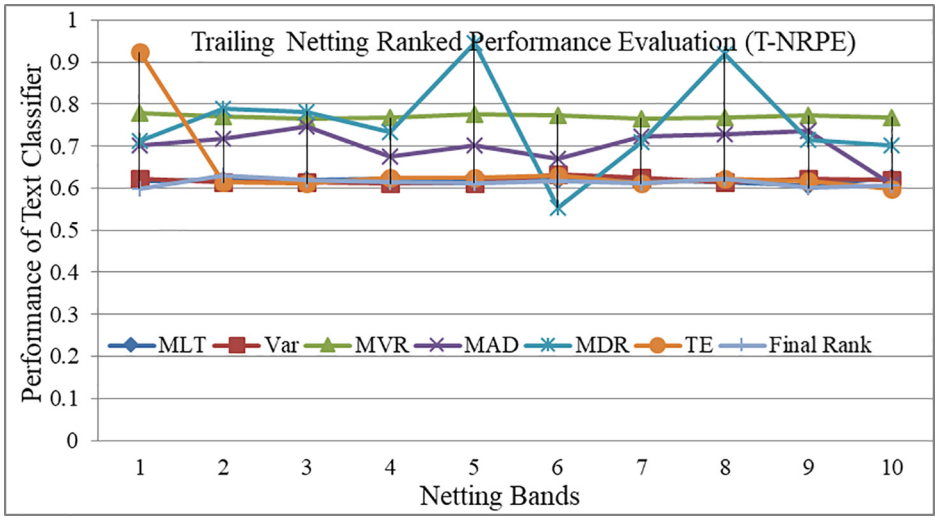

Fig 1.2. Trailing Ranked Bands performance of the ranking measures on whole corpora using **KNN.**

strates the comparison of the Trailing-NRPE of proposed ranking measures, with the baseline methods, in which we found that TF-IDF shows competitive results with the MVR and MDR in stop words extraction. Therefore, it is observed from the results and discussions that ranked words in ascending order by different ranking measures lead to candidate stopword removal.

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

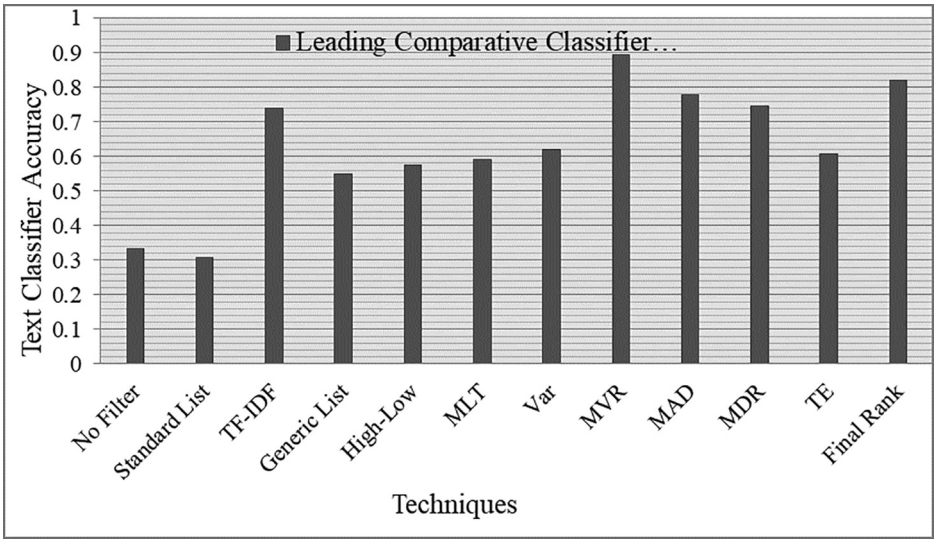

Fig 1.3. Leading comparative performance evaluation of proposed ranking methods against the baselines using **KNN classifier**.

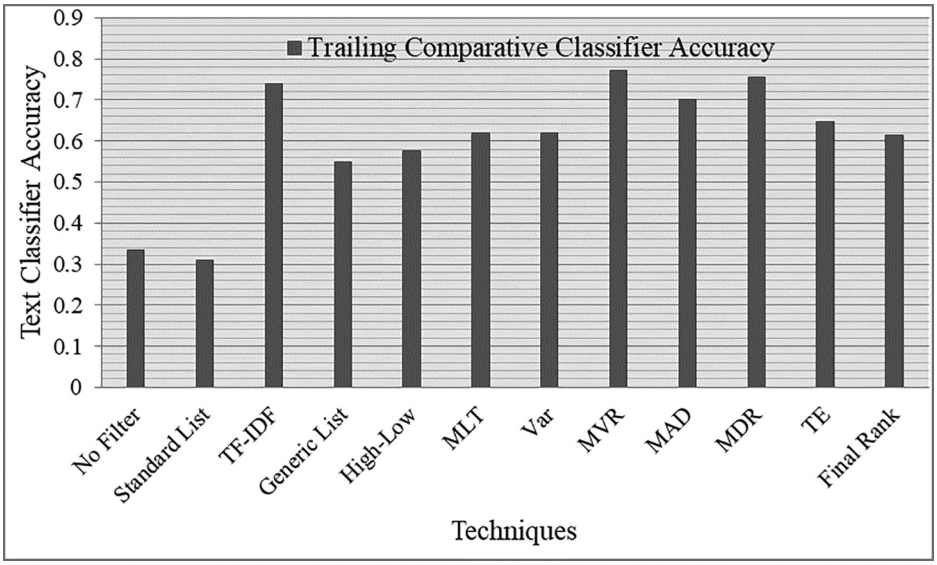

Fig 1.4. Trailing comparative performance evaluation of proposed ranking methods against the baselines using the **KNN classifier.**

6.2.2. Text clustering

Here, we discuss the performance of prepared stop lists on the second utilized external influencer i.e., K-Means text clustering using the NRPE method. We consider the precision, recall, and F-score as the accuracy measures for K-Means text clustering approach.

Precision. First, we evaluate the performance of proposed stopword lists extracted from different ranking measures using precision as the accuracy measure. Generally, precision p is the parts of achieved relevant information among the selected information. In our proposed work, the precision measures the strength of the word, which qualifies for stopword among the selected words that are identified by different ranking measures. Figs. 2.1 and 2.2 show the precision p of K-Means text clustering on different ranking measures over different bands in leading and trailing NRPE way respectively. In the L-NRPE approach as shown in Fig. 2.1, MDR fetches comparable stopwords while Var misunderstands the feature terms as stop words. In the T-NRPE approach exhibited in Fig. 2.2, MDR shows good results and MVR shows poor results for extracting stopwords against other methods. Each graph has approximately 70 experiments, including 10 numbers of bands, and 7 number of ranking measures.

Recall. The second accuracy measure for K-Means text clustering is the recall. The recall r defines the part of information retrieved as relevant over the total amount of needed relevant information. In our proposed work, the recall r evaluates the number of words those qualify for the stopwords among the total selected stopwords obtained from different ranking measures. We compute the r value for each ranking measure. Figs. 2.3 and 2.4 show the recall score of K-Means text clustering for different netting bands on the dataset in leading and trailing NRPE approaches respectively. From Figs. 2.3 and 2.4, we observe that the K-Means performance of the current netting band is affected by both the previous and current bands.

F-score. F-score is a measure to test the accuracy (*fscore*) which is defined as the harmonic mean of precision and recall, where$0 \leq fscore \leq 1$. Figs. 2.5 and 2.6 illustrate the *fscore* of

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

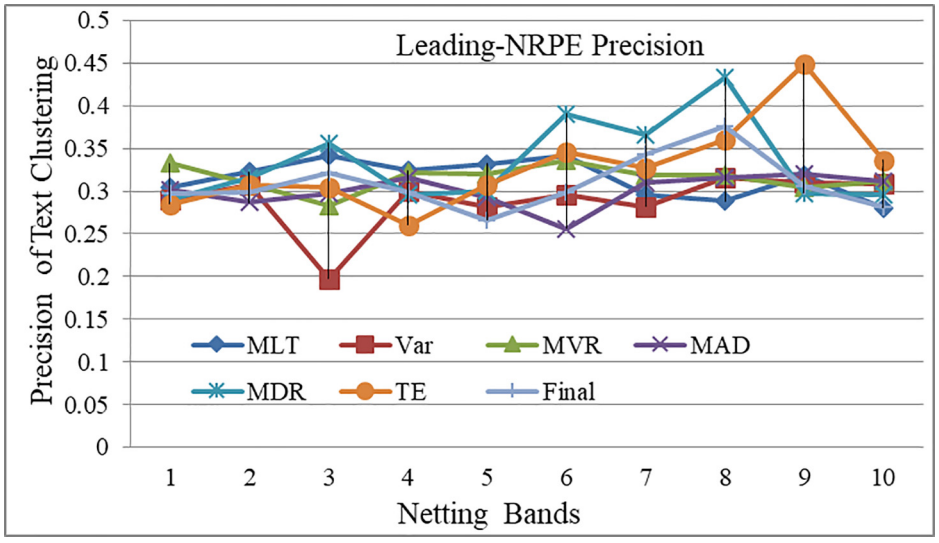

Fig 2.1. Leading-NRPE **precision** of K-Means text clustering for different netting bands on the whole dataset.

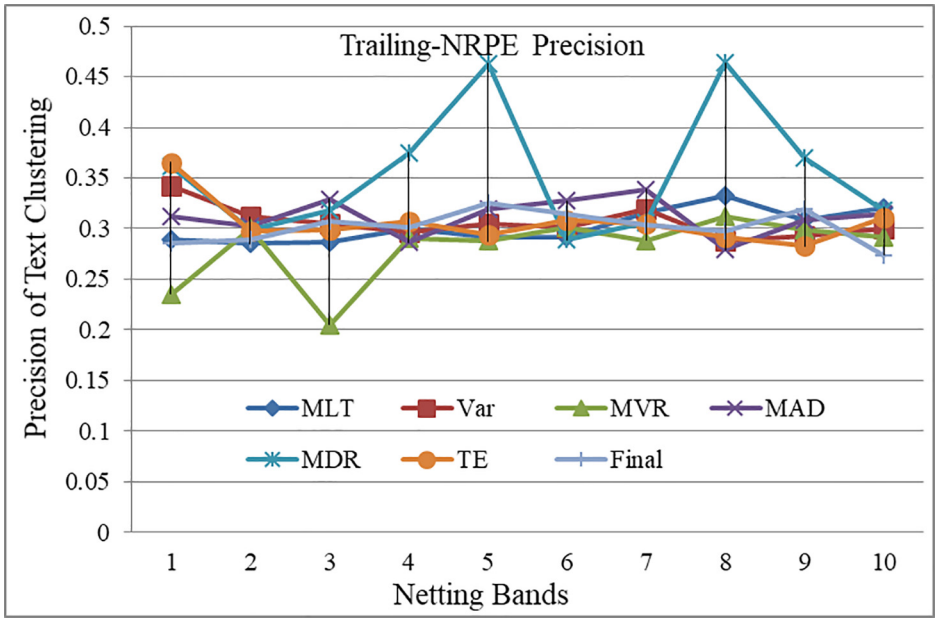

Fig 2.2. Trailing-NRPE **precision** of K-Means text clustering for different netting bands on the whole dataset.

K-Means text clustering for different netting bands on the dataset in leading and trailing NRPE models respectively.

Based on their F-score values, we compare our proposed ranking measures with five traditional baseline models. Figs. 2.7 and 2.8 show the comparison of proposed ranking measures against the baselines in leading and trailing order respectively. The traditional TF-IDF proves to be the most satisfying technique for stop words list building for the text clustering approach. The removal of stop words based on High-Low frequency from the vocabulary set also improves the text clustering performance on the remaining set of features.

6.2.3. LDA-Topic modeling

The LDA topic model eliminates Top-m stopwords of band size 'a' defined by different ranking measures and extracts topics, that is, topic '0', topic '1' and so on. In LDA topic modeling, we consider the log-perplexity score as its accuracy measure. Lower value of log-perplexity score shows the good quality topics extraction. Fig. 3.1 illustrates the most coherent topics for Top-300 stopwords, extracted by TE method in the leading phase. For instance, after the removal of Top-300 stopwords, Fig. 3.1 shows that the topic '0' contains the name of states, topic '1' covers the politician's name and topic '3' is related to entertainment. Similarly, Fig. 3.2 shows the most coherent topics for Top-300 stopwords, extracted by the MDR method in the trailing phase. Fig. 3.2 demonstrates that the topic '2' refers to the science and technology domain after the removal of top-300 terms. Each graph has approximately 420 experiments, including 10 numbers of bands, 6 different set of topics and 7 number of ranking measures.

Based on extracted topics, we evaluate the performance of proposed ranking measures using the L-NRPE and T-NRPE methods. Figs. 3.3 and 3.4 show the NRPE of proposed ranking measures on LDA topic modeling in leading and trailing manner, respectively. In Fig. 3.3, we analyze that removal of stopwords extracted by TE,

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

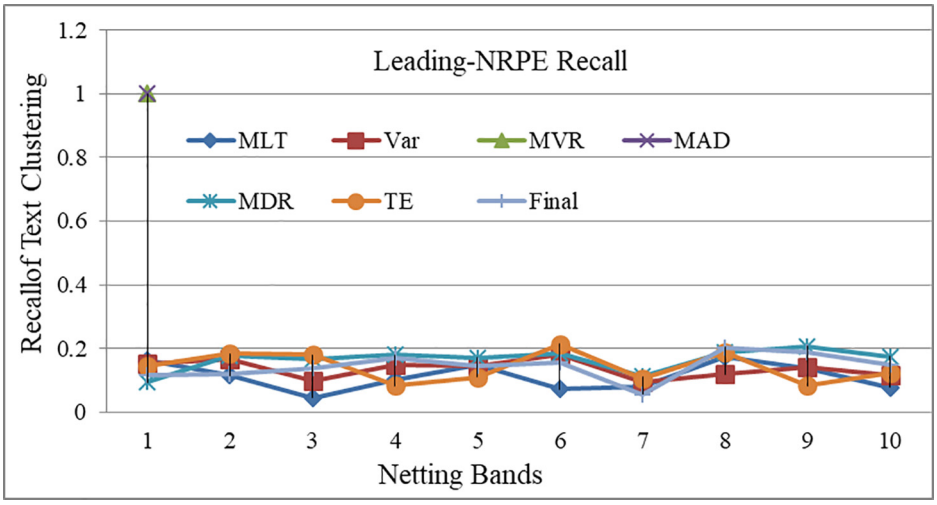

Fig 2.3. Leading-NRPE **recall** of K-Means text clustering for different netting bands on the whole dataset.

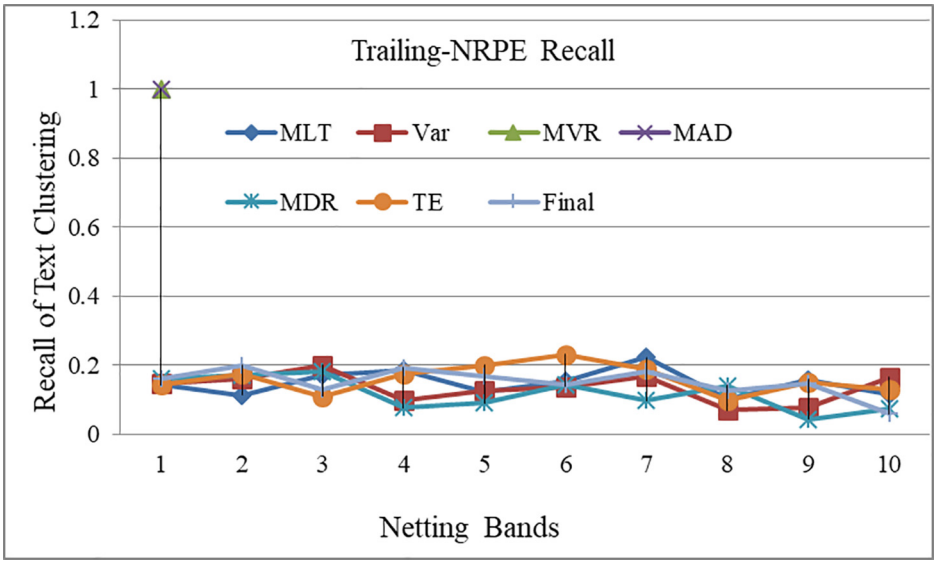

Fig 2.4. Trailing-NRPE **recall** of K-Means text clustering for different netting bands on the whole dataset.

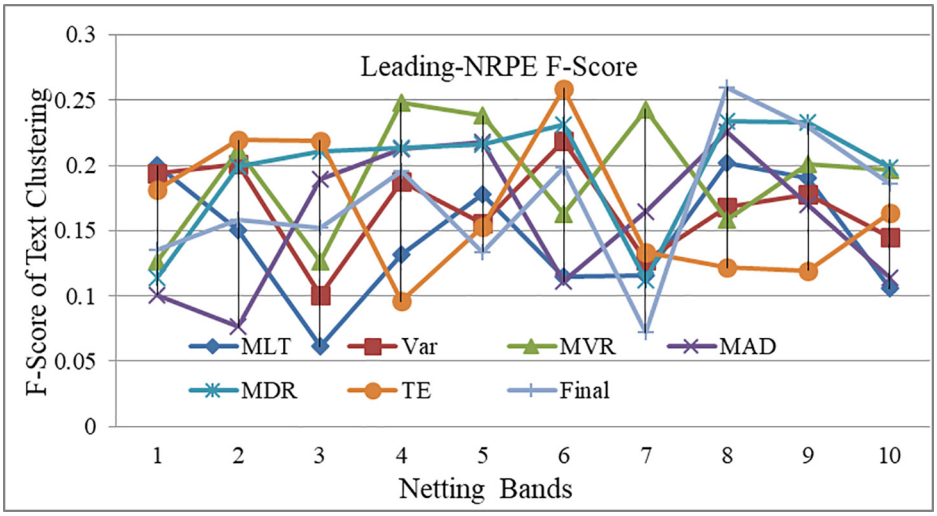

Fig 2.5. Leading-NRPE **F-score** of K-Means text clustering for different netting bands on the whole dataset.

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

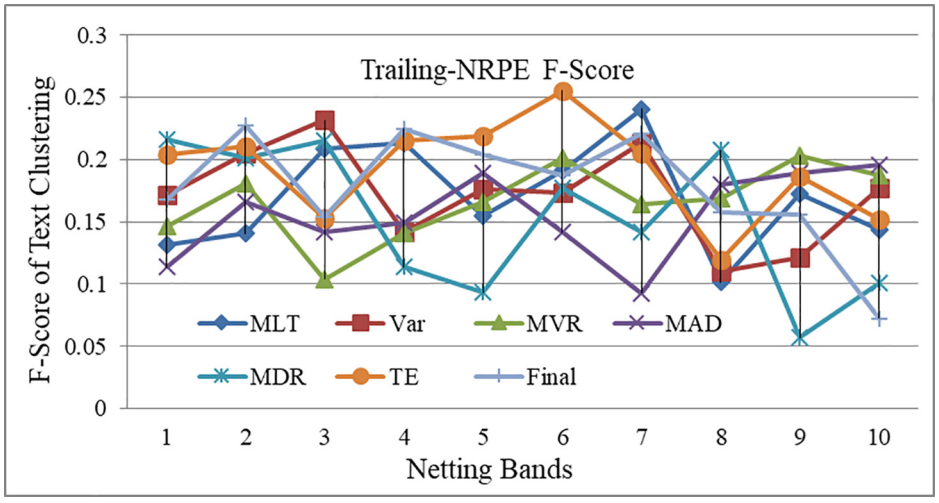

Fig 2.6. Trailing-NRPE **F-score** of K-Means text clustering for different netting bands on the whole dataset.

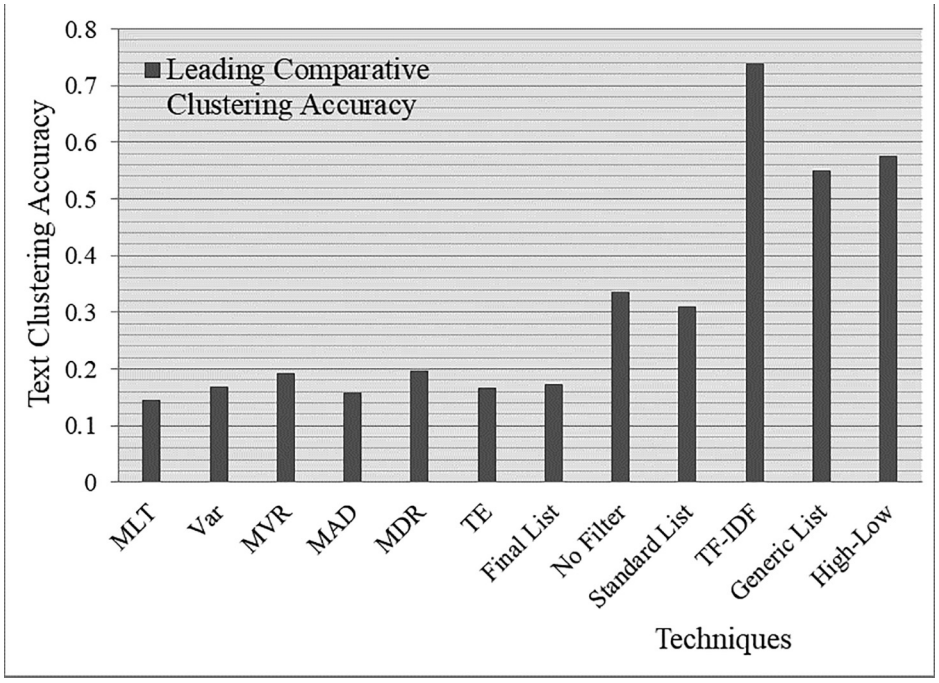

Fig 2.7. Leading comparative performance evaluation of proposed methods against the baselines using K-Means text clustering on F-Score.

MVR and MLT methods fetch more qualitative topics as compared to other methods. While the removal of stop words extracted by the Var method gives poor quality topics. It is also demonstrated that the ranking measures, such as Var, MDR, MAD and Borda's final lists have constant behavior and high log-perplexity value. We can conclude from the above observations that the words filtered in different netting bands are not the stop words and their inclusion in the vocabulary set might be helpful in improving log-perplexity value. In Fig. 3.4, Var again shows the high value of log-perplexity which makes it inefficient for selecting good topics. MDR outperforms amongst all of the ranking measures and it can be considered as the most suitable technique to build the stop words list for topic modeling.

Now, we compare the proposed ranking measures with the baseline approaches in terms of the log-perplexity score. Fig. 3.5 shows the comparison of the proposed ranking measures with the baseline approach in terms of L-NRPE. In the leading phase as shown in Fig. 3.5, we notice that the TE method outperforms amongst all the proposed ranking methods and baselines. Fig. 3.6 shows the comparison of the proposed ranking measures with the baselines in terms of T-NRPE. In trailing the NRPE approach shown in Fig. 3.6, baselines, including TF-IDF and High ranked terms give more qualitative topics than the proposed ranking measures. In proposed approaches, only MDR shows low log-perplexity value resulting in fair quality of topics. From Figs. 3.5 and 3.6, we suggest that the terms ranked by ranking measures in the leading approach qualify the criteria to become a stop word but not in trailing approach.

Table 5 gives the combined band performance area covered under the curves of different ranking measures and Borda's count

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

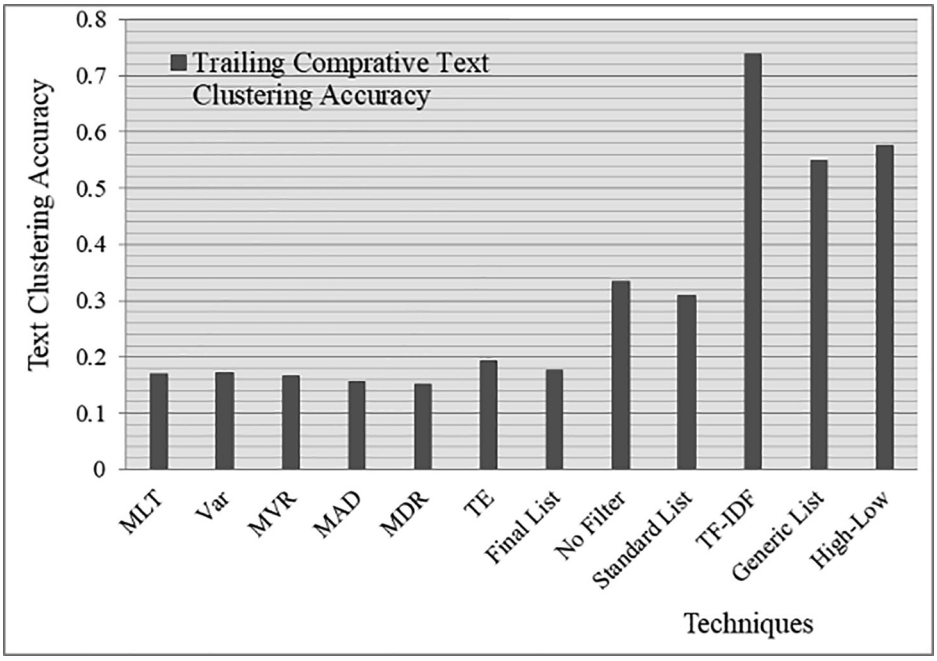

Fig 2.8. Trailing comparative performance evaluation of proposed methods against the baselines using K-Means text clustering on F-Score.

```
100-[(0, '0.001*"दिल्ली"  0.001*"नाम" + 0.001*"लोग"'),
(1, '0.001*"दिल्ली" + 0.001*"चुनाव" + 0.001*"लोग"'),
(2, '0.002*"ज्ञान" + 0.001*"साहित्य" + 0.001*"भाषा"')]
200-[(0, '0.001*"कला" + 0.001*"शिक्षा" + 0.001*"पूर्व"'),
(1, '0.001*"कला" + 0.001*"शिक्षा" + 0.001*"जीव "'),
(2, '0.001*"महाराष्ट्र" + 0.001*"स्थायी_प्रविष्टि" + 0.001*"शिक्षा"')]
```

Fig 3.1. Topics extracted using LDA model in the leading phase using Term Entropy (TE) knowledge-based ranking method.

```
100-[(0, '0.001*"भाषा" + 0.001*"दिल्ली" + 0.001*"राष्ट्रीय"'),
    (1, '0.004*"पर्यटन" + + 0.003*"पर्यटन_स्थल" + 0.003*"धर्म"'),
    (2, '0.005*"पाकिस्तान"  + 0.002*"मोदी" + 0.002*"प्रधानमंत्री"')]
200-[(0, '0.001*"साहित्य" + 0.001*"स्थल" + 0.001*"पर्यटन"'),
    (1, '0.003*"सुप्रीम_कोर्ट" + 0.002*"कांग्रेस" + 0.002*"चुनाव"'),
```

Fig 3.2. Topic Models extracted using LDA model in the trailing phase using MDR statistical measure.

for exploited influencers such as KNN text classifier, K-Means text clustering and LDA topic modeling in leading and trailing fashion. Here, CLB tells the combined performance of each technique in leading fashion while the CTB denotes the combined trailing band performance which is the sum of the performances of the ranking measures in all 10 bands.

Table 6 illustrates the rank of different ranking measures and Borda's count methods both in leading and trailing fashion based on the performance of KNN text classifier, K-Means text clustering and LDA Topic Modeling.

6.2.4. Comparison with Zou et al. Approach

In (Zou et al., 2006), Zou et al. have constructed a generic stopwords for Chinese language by employing four methods such as Mean, Variance, Mean-Variance Ratio and entropy. Further, they prepared a normalized list using vote ranking method. We observe that Zou et al. approach is to generate generic stopwords for Chinese language only. Another problem with their approach is the lack of stopwords list evaluation. In this paper, we have extended Zou et al. (2006) approach to construct the domain-specific stopwords lists for Hindi language by aiding two more methods such as MAD and MDR. In addition, we have proposed NRPE approach that evaluates the resultant stopwords lists of the proposed models in terms of the performances of the text mining models such as text classifier, text clustering and topic modeling.

7. Conclusion and future work

In this paper, we have presented a method for constructing the automatic domain-based Hindi stopword lists. First, we prepared the real-time domain-specific dataset for the Hindi Language followed by constructed the automatic domain-based stop word lists using different ranking measures. We use the social choice theory based voting method on prepared stopword lists to obtain the normalized list. In order to validate the potential of proposed lists, we proposed a new net ranking performance evaluation (NRPE) method. We also evaluate the baselines using proposed NRPE approach and compare them with the prepared stopword lists. The comparison shows that proposed lists perform better than

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

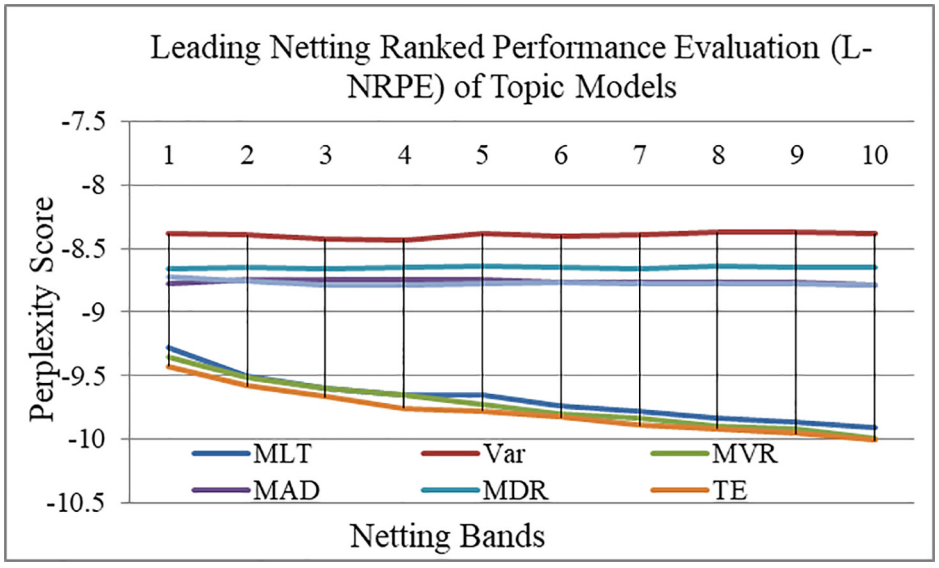

Fig 3.3. CLB-NRPE of proposed ranking methods using LDA on the log-Perplexity score.

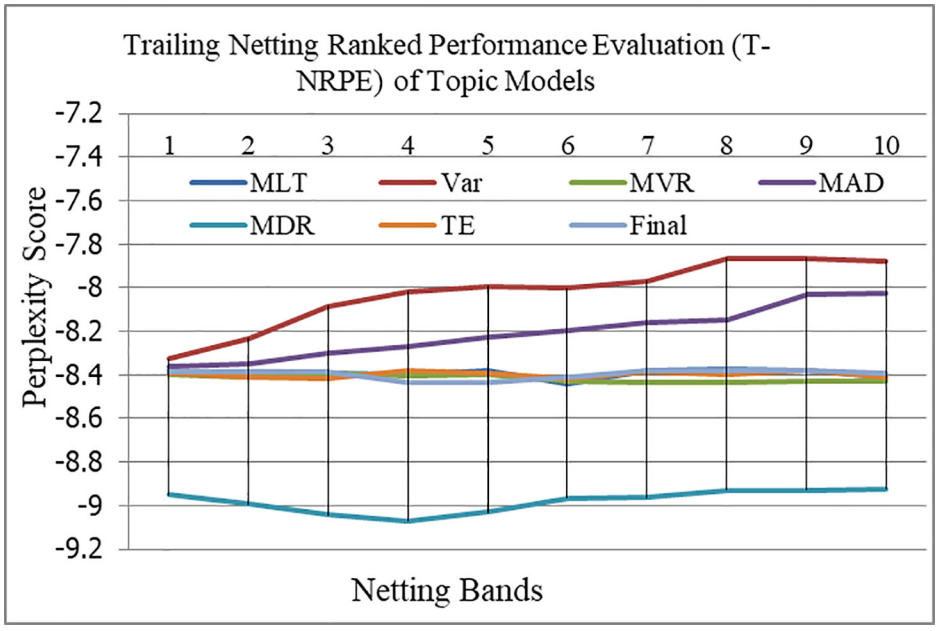

Fig 3.4. CTB-NRPE of proposed ranking methods on LDA using the log-Perplexity score.

the baselines except in the topic modeling. Our proposed stop word lists cover almost all of the stop words from existing Hindi stop word lists. In the future, authors would like to extend the work for constructing automatic text summarizer for the Hindi language.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Funding

The research work proposed here is partially supported by UPE II and DST-Purse grants received from JNU.

Acknowledgment

The research work proposed here is partially supported by UPE II and DST-Purse grants received from JNU.

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

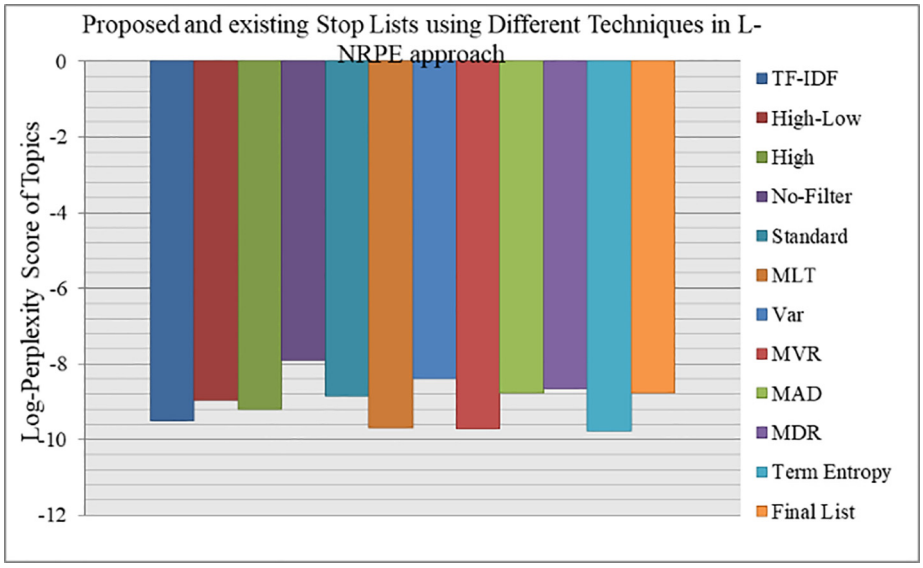

Fig 3.5. Leading comparative performance evaluation of proposed and baselines using LDA based on the log-perplexity value.

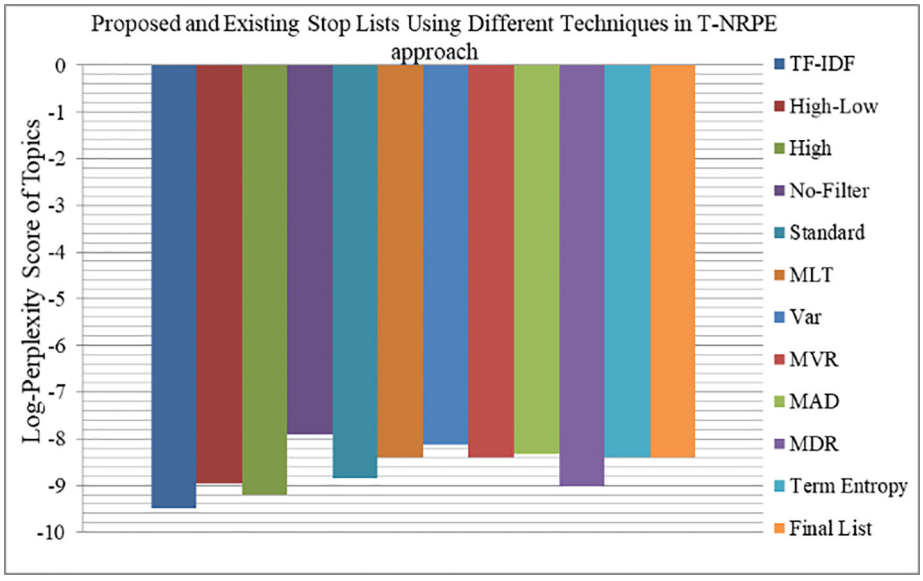

Fig 3.6. Trailing comparative performance evaluation of proposed and baselines using LDA based on the log-perplexity value.

Table 5
The combined band performance area covered under the curves of different ranking measures and Borda count methods for KNN text classifier, K-Means text clustering and LDA topic modeling in leading and trailing fashion.

Proposed Methods	KNN Text Classifier		K-Means Text Clustering		LDA Topic Modeling	
	CLB	CTB	CLB	CTB	CLB	CTB
MLT	5.92	6.18	1.45	1.7	−96.82	−180.76
Var	6.19	6.19	1.68	1.72	−83.95	−164.18
MVR	8.93	7.70	1.91	1.66	−97.30	−181.46
MAD	7.79	7.01	1.58	1.56	−87.63	−169.69
MDR	7.44	7.56	1.96	1.52	−86.52	−176.31
TE	6.08	6.48	1.66	1.92	−97.80	−181.79
Borda's Count	8.19	6.14	1.72	1.77	−87.73	−171.69

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

Table 6
The rank of different ranking measures (Leading, Trailing) and Borda count methods based on the performance of KNN text classifier, K-Means text clustering and Topic Modeling using LDA.

Ranking Methods	KNN Text Classifier Accuracy	K-Means Text Clustering			LDA Topic Modeling Log-Perplexity
		Precision	Recall	F-Score	
MLT	(7,6)	(4,5)	(7,4)	(7,3)	(3,5)
Var	(5,5)	(7,4)	(5,5)	(4,4)	(7,7)
MVR	(1,1)	(3,7)	(2,2)	(2,5)	(2,3)
MAD	(3,3)	(6,2)	(6,6)	(6,6)	(5,6)
MDR	(4,2)	(1,1)	(1,7)	(1,7)	(6,1)
TE	(6,4)	(2,3)	(4,1)	(5,1)	(1,2)
Borda's Count	(2,7)	(5,6)	(3,3)	(3,2)	(4,4)

References

Gulati, A.N., Sawarkar, S.D., 2020. Comparative Analysis of Hindi Text Summarization for Multiple Documents by Padding of Ancillary Features. In: Performance Management of Integrated Systems and its Applications in Software Engineering. Springer, pp. 217–225.

Singh, S., Panjwani, R., Kunchukuttan, A., Bhattacharyya, P., . Comparing recurrent and convolutional architectures for english-hindi neural machine translation. In: Proceedings of the 4th Workshop on Asian Translation (WAT2017), pp. 167–170.

Kumar, V., Verma, A., Mittal, N., Gromov, S.V., 2019. Anatomy of preprocessing of big data for monolingual corpora paraphrase extraction: source language sentence selection. In: Emerging Technologies in Data Mining and Information Security. Springer, pp. 495–505.

Verma, P., Pal, S., Om, H., 2019. A comparative analysis on hindi and english extractive text summarization. ACM Trans. Asian Low-Resource Lang. Inf. Process. 18 (3), 30.

Thomas, A., Sangeetha, S., 2019. An innovative hybrid approach for extracting named entities from unstructured text data. Comput. Intell. 35 (4), 799–826.

Petras, V., Perelman, N., Gey, F.C., 2003. UC Berkeley at CLEF 2003-Russian Language Experiments and Domain-Specific Cross-Language Retrieval. CLEF (Working Notes).

Choy M. "Effective Listings of Function Stop words for Twitter," arXiv Prepr. arXiv1205.6396, 2012.

Sinka, M.P., Corne, D., 2003. Evolving Better Stoplists for Document Clustering and Web Intelligence. HIS, 1015–1023.

Crow, D., DeSanto, J. "A hybrid approach to concept extraction and recognition-based matching in the domain of human resources," in: Tools with Artificial Intelligence, 2004. ICTAI 2004. 16th IEEE International Conference on, 2004, pp. 535–541.

Seki, K., Mostafa, J., 2005. An application of text categorization methods to gene ontology annotation. In: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 138–145.

Ricardo, B.-Y. Modern information retrieval. Pearson Education India, 1999.

Song, S.K., Jin, Y., Myaeng, S.H. "Abbreviation disambiguation using semantic abstraction of symbols and numeric terms," in: Natural Language Processing and Knowledge Engineering, 2005. IEEE NLP-KE'05. Proceedings of 2005 IEEE International Conference on, 2005, pp. 14–19.

Riloff, E., 1995. Little words can make a big difference for text classification. In: Proceedings of the 18th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 130–136.

Forman, G., 2003. An extensive empirical study of feature selection metrics for text classification. J. Mach. Learn. Res. 3 (Mar), 1289–1305.

WorldoMetero, "World Population Clock: 7.7 Billion People (2019) - Worldometers," 2019. [Online]. Available: www.worldometers.info. [Accessed: 31-Mar-2019].

Joshi, A., Prabhu, A., Shrivastava, M., Varma, V., 2016. Towards sub-word level compositions for sentiment analysis of hindi-english code mixed text. In: Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers, pp. 2482–2491.

Akhtar, M.S., Ekbal, A., Bhattacharyya, P., 2016. Aspect based Sentiment Analysis in Hindi: Resource Creation and Evaluation. LREC.

Jain, A., Tayal, D.K., Yadav, S., 2016. Retrieving web search results using Max-Max soft clustering for Hindi query. Int. J. Syst. Assur. Eng. Manag. 7 (1), 70–81.

Tayal, D.K., Ahuja, L., Chhabra, S., 2015. Word Sense Disambiguation in Hindi Language Using Hyperspace Analogue to Language and Fuzzy C-Means Clustering. In: Proceedings of the 12th International Conference on Natural Language Processing, pp. 49–58.

Kumar, K.V., Yadav, D., Sharma, A., 2015. Graph Based Technique for Hindi Text Summarization. In: Information Systems Design and Intelligent Applications. Springer, pp. 301–310.

Kumar, K.V., Yadav, D., 2015. An improvised extractive approach to hindi text summarization. In: Information Systems Design and Intelligent Applications. Springer, pp. 291–300.

Harikrishna, D.M., Rao, K.S., 2015. Classification of children stories in hindi using keywords and POS density. In: Computer, Communication and Control (IC4), 2015 International Conference on, pp. 1–5.

Singh, V., Vijay, D., Akhtar, S.S., Shrivastava, M., 2018. Named entity recognition for hindi-english code-mixed social media text. In: Proceedings of the Seventh Named Entities Workshop, pp. 27–35.

Rao, P.R.K., Malarkodi, C.S., Ram, R.V.S., Devi, S.L., 2015. ESM-IL: Entity Extraction from Social Media Text for Indian Languages@ FIRE 2015-An Overview. FIRE Workshops, 74–80.

Jha, V., Manjunath, N., Shenoy, P.D., Venugopal, K.R., 2016. Hsra: Hindi stopword removal algorithm. In: Microelectronics, Computing and Communications (MicroCom), 2016 International Conference on, pp. 1–5.

L. Hao and L. Hao, "Automatic identification of stop words in chinese text classification," in Computer Science and Software Engineering, 2008 International Conference on, 2008, vol. 1, pp. 718–722.

Zipf, K., 1932. Selective Studies and the Principle of Relative Frequency in Language. MIT Press Cambridge, MA.

Pandey, A.K., Siddiqui, T.J., 2009. "Evaluating effect of stemming and stop-word removal on hindi text retrieval", in. In: Proceedings of the First International Conference on Intelligent Human Computer Interaction, pp. 316–326.

R. Rani and D. K. Lobiyal, "Automatic Construction of Generic Stop Words List for Hindi Text," in Procedia Computer Science Elsevier Journal, 2018, pp. 1–7.

Zou, F., Wang, F.L., Deng, X., Han, S., Wang, L.S., 2006. Automatic construction of chinese stop word list. In: Proceedings of the 5th WSEAS international conference on Applied computer science, pp. 1010–1015.

Taranjeet, "Hindi stopwords." [Online]. Available: https://github.com/Alir3z4/stop-words/blob/master/hindi.txt. [Accessed: 20-Dec-2018].

Ranks, "Ranks: Hindi Stopwords." [Online]. Available: https://www.ranks.nl/stopwords/hindi. [Accessed: 20-Dec-2018].

GitHub, "Hindi stopword list." .

Luhn, H.P., 1957. A statistical approach to mechanized encoding and searching of literary information. IBM J. Res. Dev. 1 (4), 309–317.

Van Rijsbergen, C.J., 1986. A non-classical logic for information retrieval. Comput. J. 29 (6), 481–485.

Fox, C.J. "Lexical Analysis and Stoplists." 1992.

Francis, W. Kucera, H. "Frequency analysis of English usage," 1982.

Makrehchi, M., Kamel, M.S. "Automatic extraction of domain-specific stopwords from labeled documents," in: European Conference on Information Retrieval, 2008, pp. 222–233.

Makrehchi, M., Kamel, M.S., 2017. Extracting domain-specific stopwords for text classifiers. Intell. Data Anal. 21 (1), 39–62.

White, B.J., Fortier, J., Clapper, D., Grabolosa, P., 2007. The impact of domain-specific stop-word lists on ecommerce website search performance. J. Strateg. E-Commerce 5 (1/2), 83.

Sinka, M.P., Corne, D.W. "Towards modernised and web-specific stoplists for web document analysis," in Web Intelligence, 2003. WI 2003. Proceedings. IEEE/WIC International Conference on, 2003, pp. 396–402.

Kawahara, M., Kawano, H., 2001. Mining association algorithm with threshold based on ROC analysis. In: Proceedings of the 34th Annual Hawaii International Conference on System Sciences, p. 8.

El-Khair, I.A., 2006. Effects of stop words elimination for Arabic information retrieval: a comparative study. Int. J. Comput. Inf. Sci. 4 (3), 119–133.

Taghva, K., Coombs, J., Pareda, R., Nartker, T. "Language model-based retrieval for Farsi documents," in: Information Technology: Coding and Computing, 2004. Proceedings. ITCC 2004. International Conference on, 2004, vol. 2, pp. 13–17.

Singh, S., Siddiqui, T.J., 2012. "Evaluating effect of context window size, stemming and stop word removal on Hindi word sense disambiguation. Information Retrieval & Knowledge Management (CAMP), 2012 International Conference on 2012, 1–5.

Sharma, V., Mittal, Namita, 2019. Refined stop-words and morphological variants solutions applied to Hindi-English cross-lingual information retrieval. J. Intell. Fuzzy Syst. 36 (3), 2219–2227.

Rani, R., Lobiyal, D.K., 2018. Social Choice Theory Based Domain Specific Hindi Stop Words List Construction and Its Application in Text Mining. In: International Conference on Intelligent Human Computer Interaction, pp. 123–135.

Choudhary, N., Jha, G.N., 2011. Creating multilingual parallel corpora in indian languages. Language and Technology Conference, 527–537.

Myerson, R.B., 2013. Fundamentals of social choice theory. Quart. J. Polit. Sci. 8 (3), 305–337.

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003

Guo, G., Wang, H., Bell, D., Bi, Y., Greer, K., 2006. Using kNN model for automatic text categorization. Soft Comput. 10 (5), 423–430.

Kevin, "Knn Classifier." [Online]. Available: https://kevinzakka.github.io/2016/07/13/k-nearest-neighbor/. [Accessed: 03-Jan-2019].

Wikipedia, "K-Means Clustering." [Online]. Available: https://en.wikipedia.org/wiki/K-means_clustering. [Accessed: 02-Jan-2019].

Blei, D.M., 2012. Probabilistic topic models. Commun. ACM 55 (4), 77–84.

Benjamin Soltoff, "Perplexity." [Online]. Available: https://cfss.uchicago.edu/fall2016/text02.html#perplexity. [Accessed: 31-Dec-2018].

Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Tech. J. 27 (3), 379–423.

Kantor, P.B., Lee, J.J., 1986. The maximum entropy principle in information retrieval. In: Proceedings of the 9th annual international ACM SIGIR conference on Research and development in information retrieval, pp. 269–274.

Myerson, R.B. Fundamentals of social choice theory. NorthWestern University, Center for Mathematical Studies in Economics and Management Sciences, 1996.

Rajaraman, A.U. "JD (2011)." Data Mining," Min. Massive Datasets, pp. 1–17.

Please cite this article as: R. Rani and D. K. Lobiyal, Performance evaluation of text-mining models with Hindi stopwords lists, Journal of King Saud University – Computer and Information Sciences, https://doi.org/10.1016/j.jksuci.2020.03.003