The object of the research is the method of natural language processing (NLP) with balanced parameters of the BestMatch25 (BM25) algorithm to recognize and classify fake news based on natural language processing (NLP). The unsatisfactory accuracy and speed of existing methods for detecting fake news in unstructured input data demanded the development of a new approach for their effective detection.

The study investigated the BM25 algorithm, methods for selecting parameters k1 and b, and their impact on the algorithm's effectiveness in detecting fake news. It was established that precise and detailed adjustment of these parameters is crucial in achieving optimal accuracy and data processing speed.

The results showed that the successful selection of BM25 parameters improves the model's accuracy by up to 14 % compared to standard term frequency – inverse document frequency (TF-IDF) calculations. These results were made possible by experimentally tuning different combinations of k1 and b parameters, in which the algorithm shows the best speed indicator or the most accurate estimate of the importance of a term in a document. Balanced values of k1 and b parameters were identified, leading to the algorithm's optimal speed and accuracy in assessing word importance considering the input data's peculiarities.

The balanced setting of the BM25 algorithm parameters explains the obtained results. They can be used for automated recognition and analysis of news and information on social media based on natural language processing. However, in practice, the effectiveness of the set of parameters depends on linguistic variations, content, and the theme within new input data sets

Keywords: BestMatch25, term frequency – inverse document frequency, natural language processing, fake news

UDC 004.91

DOI: 10.15587/1729-4061.2023.293513

RECOGNIZING FAKE NEWS BASED ON NATURAL LANGUAGE PROCESSING USING THE BM25 ALGORITHM WITH FINE-TUNED PARAMETERS

Liudmyla Mishchenko
Corresponding author
Postgraduate Student*
E-mail: liudamishchenko@gmail.com
Iryna Klymenko
Doctor of Technical Sciences, Professor*
*Department of Computer Engineering
National Technical University of Ukraine
"Igor Sikorsky Kyiv Polytechnic Institute"
Beresteiskyi (Peremohy) ave., 37, Kyiv,
Ukraine, 03056

Received date 09.10.2023
Accepted date 15.12.2023
Published date 28.12.2023

***How to Cite**: Mishchenko, L., Klymenko, I. (2023). Recognizing fake news based on natural language processing using the BM25 algorithm with fine-tuned parameters. Eastern-European Journal of Enterprise Technologies, 6 (2 (126)), 33–40. doi: https://doi.org/10.15587/1729-4061.2023.293513*

1. Introduction

The proliferation of fake news in the digital world poses a serious challenge to the verification and veracity of information. The constant spread of misleading or fabricated information through various online platforms requires the use of sophisticated methods to combat the spread of misinformation. Existing methods are usually unable to cope with the volume of fake news generated daily. They are not adapted to the topic, or the model is not trained on similar texts, which affects the effectiveness of the method. In response to this critical need, the proposed research examines a comprehensive method for detecting fake news by using natural language processing (NLP) and the BM25 algorithm.

Using the BM25 algorithm, which is known for its effectiveness in information retrieval tasks, simplifies the task of distinguishing real news articles from fabricated ones. And the combination of it and NLP helps achieve the goal of this work and devise a reliable and adaptive solution to the problem of spreading false information.

The current paper examines a step-by-step method of analyzing the text of news related to the ongoing war in Ukraine using natural language processing. The parameters of the ranking function are adjusted and adapted to the specifics of the input data. The subtleties of the proposed method for recognizing fake news are considered, namely the effect of changing the parameters of the BM25 algorithm on the speed and accuracy of processing.

Scientific research on this topic is important because existing methods of automatic verification of news lose significantly in time and accuracy of text processing compared to the speed of their creation and distribution. The results of such studies are needed in practice because they ensure the construction of a reliable and effective method for combating disinformation. They complement and strengthen the existing ways of checking information in the digital environment for reliability and truthfulness.

2. Literature review and problem statement

More and more researchers around the world are raising the issue of combating misleading information. For this, all possible variants of automatic verification of texts are involved, using machine learning algorithms and improving them in various possible ways, in particular mathematical solutions such as the Naive Bayes classifier or BM25. In particular, work [1] presents an overview of different approaches to checking news for truth and provides an understanding of the main characteristics of NLP and machine learning (ML). There is a separate stage of word classification using three approaches: Passive Aggressive, Naïve Bayes

classifier, and Support Vector Machine. It is shown that a simple classification is not completely effective for detecting fake news. The reason for this may be the peculiarities of the set of unstructured input data. An option to overcome these difficulties can be optimally selected classification parameters.

This approach is described in [2]. This method uses the Naive Bayes Document Classifier together with TF-IDF, presenting a novel text classification scheme that learns from datasets and accurately classifies unstructured text into True and False groups. However, the issues of text specificity or linguistic features remain unresolved.

The option to overcome these limitations is eliminated in work [3]. The FDCD-TF-IDF algorithm is presented, an improved feature weighting technique that combines word frequency distribution and category distribution information. This approach improves the representation of the importance of feature elements by taking into account the relationships between elements and categories, as well as the characteristics of the category. However, it does not address the performance issue of large input data, which potentially increases the computational load during real-time news text analysis.

An option to overcome such difficulties can be a solution that does not take into account the distribution of words in the text. This is the approach used in work [4]. It emphasizes the problem of inadequate extraction of eigenvalues in non-unified text classification, referring to the traditional TF-IDF word frequency ranking model. The authors offer an improved solution – the extended TF-IDF algorithm, which eliminates the irregular distribution of the text. Experimental results demonstrate that using this enhancement outperforms the original algorithm in terms of accuracy and recall for classification tasks. This innovation aims to address the shortcomings of traditional TF-IDF, presenting the potential for significant improvements in real-world text classification scenarios.

Another approach to solving the problem of summarizing and losing the weight of important words in the text is proposed by researchers in the method [5] of weighting mTF-IDF-Assoc terms. This approach integrates a modified association concept. It considers document length (DL) to normalize term frequency by dividing it by the length of the document vector, subsequently including IDF and Assoc when calculating word weights. This approach improves the accuracy of multiclass classification. However, the obtained results are justified only on short news texts from the Twitter network, the problem of processing full news texts remains unsolved.

An alternative effective method of detecting fake news based on NLP together with the use of a naive Bayes classifier is research [6]. A hybrid algorithm is proposed, which deploys a simple Bayesian classifier on each final node of the constructed decision tree. The results of the research will demonstrate the excellent classification efficiency. However, high accuracy rates remained an unresolved issue.

Solving this issue, as well as the limitations of the TF-IDF method in terms of understanding the essence of the sentence, was considered in study [7], focusing exclusively on word segmentation. The authors modified the method to capture language features for classification purposes. In addition, work [8] analyzed in detail the influence of various term frequency factors on seven controlled term weighting schemes. The results showed that modifying the term frequency factor improved the performance of almost all weighting schemes. Such a modification refers to the frequency of collections of words, and not to a single term, which makes relevant studies inexpedient.

Another approach to solving the problem is described in the work on the optimization of the BM25 algorithm [9]. The main contribution of that paper is a new method for information content field weighting (ICFW). It applies weights to the structure without optimization and overcomes the problems faced by some existing SDR models, including the issue of term frequency saturation in documents. The authors emphasize that they managed to achieve not only a balanced value of the criteria of the BM25 algorithm but also a significant increase in performance for non-optimized search. However, the result depends on the structure of the documents, which indicates the impossibility of applying it to the analysis of news on the Internet.

An option to overcome difficulties with unstructured data is the selection of optimal limits of the BM25 algorithm. In study [10], a number of experiments were conducted to accurately select the parameters of the BM25 algorithm to achieve the fastest possible result on large language models (LLM) in the field of natural language processing (NLP). The work states that large volumes of input data are not a guarantee of their fast processing, a well-chosen pair of coefficients has a much greater impact. However, the selection of BM25 algorithm parameters depends solely on the input data set and the needs of the result.

All this allows us to state that it is appropriate to conduct a study aiming to search for optimal parameters of the BM25 ranking algorithm for recognizing fake news based on natural language processing.

3. The aim and objectives of the study

The purpose of our study is to improve the method of recognizing fake news based on natural language processing due to the optimal selection of parameters of the BM25 word ranking algorithm. This will make it possible to increase the accuracy of the assessment of the importance of words in the text and their classification without losing the speed of news analysis.

To achieve the goal, the following tasks are set:

– to perform an analysis of the relevance of the BM25 algorithm to the assessment of the importance of words in the text in the method of recognizing fake news;

– to devise a method for determining the balanced parameters k_1 and b of the BM25 algorithm, which allows improving the method of recognizing fake news to increase the efficiency of processing texts with dynamically changing length, as well as unbalanced data and topics.

4. The study materials and methods

4. 1. The object and hypothesis of the study

Given the speed at which news spreads, standard natural language processing analysis does not perform well. It needs improvement and different methods of speeding up and increasing accuracy. Therefore, the object of the proposed

research is the method of natural language processing (NLP) for automatic recognition of fake news. The main hypothesis of the research is the possibility of optimizing the method of news analysis due to the application of the BM25 algorithm at the stage of evaluating the importance of words in the text.

4. 2. Modified method of detecting fake news

A well-known method for identifying fake news using natural language processing (NLP) is described in [7]. Processing of input data according to the described method is based on generally accepted methods of data collection and formatting for their further analysis and modeling. The stages of data processing according to the given method of identifying fake news are given below:

1. Data collection: The first step is to select diverse, reliable sources for data collection. The input data sets are taken from a database of the European Union, which contains reliable news sources, the texts of news and posts of users on social networks, as well as websites that commonly publish false information. This provides a variety of input data, allows the model to work with different content, includes links to verified sources and facts. This approach to the formation of the array of data helps determine real or fake news, compare the results of the analysis, and make an assessment of the correctness of the algorithm.

2. Labeling: Accurate labeling or labeling is essential for model training and evaluation. Each article in the data set is automatically labeled as "real" or "fake" based on the assessment of fact-checkers from a database of verified news.

3. Balancing: Ensuring a balanced data set with nearly equal numbers of genuine and fake articles is necessary to prevent bias in the model and optimize performance.

4. Data pre-processing: starts with text cleaning. Elements such as HTML tags, URLs, and special characters are removed to focus the NLP model on the semantic content of the text. This is followed by converting all to lowercase, tokenizing, extracting stop words, stemming, lemmatization, processing numeric data, removing null values from the database, working with unbalanced classes, and encoding labels.

5. After completing the data collection and pre-processing, we receive a cleaned and structured data set, ready for extracting the features of the text and modeling.

6. Feature Extraction: Pre-processed text data is converted into numerical values that are used for machine learning. The TF-IDF algorithm transforms text into a high-dimensional vector representation, preserving semantic information.

7. Linguistic and contextual feature engineering: at this stage, additional linguistic features are identified (tags of parts of speech, syntactic dependencies, named entities) responsible for detailed linguistic characteristics. Contextual features (sentiment analysis scores) – emotional coloring of the text.

8. Classification and evaluation of the importance of words in the text: standard approaches use classification algorithms such as Naïve Bayes, the method of support vectors. TF-IDF is the most widespread among them.

9. Model training: based on the vector representation of real and fake news tokens, the selected models go through the training stage. Hyperparameter tuning is also done to optimize performance.

10. Cross-validation and evaluation of results: Evaluating the model's performance using metrics such as accuracy, speed, and recall measure its effectiveness in separating real from fake news.

11. Robustness testing: models are robustly tested on news sets not previously trained. It is important that the datasets differ in content and linguistic features.

Our paper proposes to improve the method of recognizing fake news at the stage of classification and assessment of the importance of words in the text. It is assumed that the use of the BM25 algorithm will allow balancing the parameters of the ranking function with the aim of processing arrays of text with a dynamically changing length with increased accuracy without affecting the speed of the algorithm.

4. 3. Analysis of the effectiveness of TF-IDF and BM25 algorithms for evaluating the importance of words in the text

The TF-IDF and BM25 algorithms are widely used in various natural language processing tasks, such as text analysis, information retrieval, keyword extraction, and others. The use of one of the algorithms plays an extremely important role in the modified method of recognizing fake news to solve the following problems:

1) identification of the importance of the term;
2) selection of key terms;
3) reduction of auxiliary words in the text;
4) preservation of contextually important words;
5) improvement of classification accuracy.

The most common algorithm is TF-IDF (Term Frequency-Inverse Document Frequency). The result of his work is a numerical value that is used in text analysis to evaluate the importance of a word in a document compared to a collection of documents. TF-IDF helps highlight words that are characteristic or unique to the document, while reducing the importance of auxiliary words ("the", "and", ...). In this way, TF-IDF detects keywords and reduces the influence of generic terms that have no linguistic meaning in a particular document or context.

The BM25 (Best Match 25) algorithm is an improved version of TF-IDF, used to solve identical problems. However, BM25 has a different mechanism for evaluating the importance of words and takes into account the weight parameters of terms k_1 and b.

During the implementation of the fake news recognition method, the TF-IDF algorithm is used, but in our study, it is proposed to apply BM25. To justify the choice, a comparative theoretical analysis of the standard characteristics of the two algorithms was conducted in the context of their use to solve the problem of classification and assessment of the importance of words in the text. The results of the analysis are given in Table 1.

Considering the above comparative characteristics, the TF-IDF algorithm is fundamental and widely used to estimate the importance of words in text. However, the BM25 algorithm eliminates such problems of TF-IDF as term frequency normalization, dynamic document length, processing of common and rare terms. Based on these characteristics, the BM25 performs better in word processing tasks, namely with documents that vary significantly in length or contain repeated terms.

Table 1

Comparison of characteristics of TF-IDF and BM25 algorithms

Comparison sign	TF-IDF	BM25
Basic concept	Evaluates the importance of a term in a document relative to a set of documents. Based on the calculation of the frequency of the term in the document (TF) and the inverse frequency of the term in the entire array of documents (IDF)	An extended version of TF-IDF, focused on solving the term saturation problem. BM25 takes document length into account and modifies the term frequency component of TF-IDF to prevent over-importance of repeated terms
Determination of the periodicity of the term in the text	Directly uses the frequency term in a document without considering the length of one or the average length of all documents in the text set	Includes term frequency normalization based on document length, making it less sensitive to length variations
Calculation of the inverse frequency of documents	Uses a logarithmic transformation to assign a weighting factor to reduce the importance score to terms that occur frequently in the text. However, it does not always handle very common or rare terms effectively	Uses a more complex IDF calculation that does not depend solely on the logarithm of the ratio of the total number of documents to those containing the term. BM25 can handle both common and rare terms
Scalability	Does not count because it requires TF and IDF values to be calculated for each term in the document without first calculating them	Takes into account Calculations include more complex components to account for the length of documents. Although this may increase the time spent on calculations
Effectiveness of calculating the importance of words in the text	Works well in many scenarios, but may not handle certain issues like term frequency normalization and variations in document length	Known for its reliability in processing documents of different lengths and changes in the frequency of deadlines
Use and application	Widely used in search engines and applications for finding or analyzing text due to its simplicity and efficiency	Increasingly popular in word processing tasks due to its ability to more efficiently cope with the complexity of long documents and varying term frequency

4. 4. Adjusting the parameters of the BM25 ranking algorithm

The well-known word ranking algorithm BM25 has defined execution stages [11] involving the necessary parameters, depending on the input data set:

1. Term frequency adjustment (TF): BM25 takes into account the frequency of the term but adjusts its fullness in the document. However, the algorithm introduces term frequency normalization to mitigate the effects of those that occur too many times in the input data. Normalization is achieved through parameter k_1 in formula (1):

$$TF_i' = \frac{TF_i \times (k_1 + 1)}{TF_i + k_1 \left(1 - b + b \times \frac{DL}{AvgDL}\right)}, \quad (1)$$

where TF_i is the normalized term frequency, TF_i' is the adjusted term frequency, k_1 regulates the saturation of the term frequency normalization, b affects the ratio of the document length to the term frequency normalization, DL is the document length, $AvgDL$ is the average document length.

2. Inverse document frequency (IDF): similar to TF-IDF, BM25 considers the importance of a term in the entire document corpus. However, the IDF in BM25 has been modified to provide more accurate results. This reduces the influence of very common and very rare terms and prevents bias in the results. The IDF component for term i is given by formula (2):

$$IDF_i = \log\left(\frac{N - n_i + 0.5}{n_i + 0.5}\right), \quad (2)$$

where N is the total number of documents, n_i is the number of documents with term i.

3. Term frequency normalization: BM25 normalizes term frequency based on document length. This is done to balance the impact of long documents that have a higher frequency of terms compared to shorter ones. This normalization factor helps reduce the bias towards longer documents. In this step, the TF and IDF components are combined in formula (3) to calculate the weight of term i in the document:

$$Weight_i = TF_i' \times IDF_i. \quad (3)$$

4. Adjustment factors k_1 and b: two parameters k_1 and b are introduced into the algorithm to adjust the results. The parameter k_1 controls the change in the rate of saturation of the frequency of the term in the text. The b parameter controls the effect of document length normalization.

5. Evaluation of the value of the term in the document: the evaluation of the relevance of the document to a certain iteration of the algorithm is calculated according to formula (4) by summing the value of each term present in the document:

$$Score_{doc} = \sum_i Weight_i. \quad (4)$$

In all iterations, BM25 generates a score for each document based on the relevance of terms to it. Documents with higher scores are considered more relevant to the iteration request.

6. Ranking: at the end, papers are ranked based on their scores. The higher the score, the more important the document is considered for the conditions of a particular iteration.

At the third stage of the BM25 algorithm, normalization factors are introduced for the TF component, improving the accuracy of term frequency estimation in the ranking process. In the fourth step, the parameters k_1 and b are adjusted for specific characteristics of the data set, which makes it more complex compared to traditional TF-IDF models and more effective for the task of evaluating text for fakeness or truthfulness.

At the same time, the methods of selecting parameters can be completely different and do not affect the final result. Thus, the researchers considered various methods of selection of coefficients, for example, a probabilistic interpretation of TF BM25 normalization and its parameter k_1 based on a logarithmic model for the probability of meeting

a document in the collection with a given level of TF [12]. In study [13], a framework was proposed for searching articles by checking the context of the original tweet, which may contain misinformation. The scientific value and relevance of this work lies in the experimental selection of parameters of the BM25 algorithm for effective analysis of tweets. This technique is used to validate both texts and images and achieves high results on real datasets.

The optimal selection of coefficients k_1 and b of the BM25 function depends on the input data and cannot be taken as a universal value for decisions with different criteria [14]. Therefore, the proposed technique takes into account unstructured, unbalanced news texts of different lengths. The machining process should have optimal indicators between accuracy and processing speed.

4. 5. Technique for selection of algorithm parameters

For the effective operation of the BM25 algorithm with the task of distinguishing real news from fake news and ensuring a balance between accuracy and speed of data set processing, it is necessary to adjust the parameters and evaluate their effectiveness. This process is based on the following input criteria:

– texts of different lengths;
– an unbalanced set of input data;
– topics related to the Russian-Ukrainian war;
– text processing should be done as quickly as possible.

Based on the defined criteria, a technique for selecting and validating the parameters of the ranking algorithm is proposed. This approach involves the following stages:

1. Handling text length variations: k_1 parameter must be adjusted to account for changes in text length. Higher values of k_1 (within a reasonable range) can be useful because they affect the normalization of the frequency of the terms. However, to account for different lengths of texts in the data set, it is necessary to experiment with the values of k_1.

2. Management of unbalanced data: parameter b affects the normalization of document length. Therefore, decreasing the value of b can help balance the effect of document length on the IDF component.

3. Balancing between accuracy and speed: accuracy requirements should not have a negative impact on computing speed. To increase the speed without significant damage to the accuracy, it is necessary to adjust k_1 and b moderately, and not to set extreme values.

4. Selection of a validation strategy: cross-validation methods are chosen to fine-tune the parameters, given the size-proportional input data sets.

5. Iterative experiment: Several experiments were performed, systematically adjusting k_1 and b in small steps. This helps understand their impact on the accuracy and efficiency of calculations.

6. Optimization and verification: a configuration of parameters (k_1, b) is chosen that maximizes accuracy while meeting speed requirements. This choice is tested on a new test data set to ensure its universality.

7. Final model selection: The selection results in a combination of parameters that provides the best balance between accuracy and speed on the test set and takes into account high performance requirements.

This approach is necessary to take into account all the specified input data criteria and optimally selected parameters. Verification of the results is performed experimentally.

5. Results of investigating the recognition of fake news with natural language processing and balanced parameters of BM25

5. 1. Analyzing the relevance of the BM25 algorithm to the assessment of the importance of words or terms in the text

To justify the choice of using the word ranking algorithm, an analysis of the relevance of the BM25 algorithm to the assessment of the importance of words or terms in the text in comparison with TF-IDF was carried out. The results of this analysis are given in Table 2.

Our analysis theoretically substantiates the reliability of the assumption that the use of the BM25 algorithm affects the accuracy of the assessment of the importance of words in the text and their classification without losing the speed of news analysis. And it is the basis of further balancing of BM25 parameters based on the characteristics of the input data set.

Table 2

Comparison of TF-IDF and BM25 algorithms for estimating the importance of words in a text

Indicator	TF-IDF	BM25
Assessment of the importance of the term	Evaluates the importance of a term in a document and in the entire data set. Terms that occur frequently in a particular document, but less frequently in the entire corpus, are considered more important	Improves TF-IDF by accounting for document length and addressing term saturation, providing a more detailed assessment of term importance
Key terms	Identifies terms that are statistically significant in a document by comparing their frequency with the overall frequency in the entire set of documents. Unusual terms in the array, but frequent in the document, indicate specific relevance or uniqueness	Solves the problem of processing common and rare terms, allowing to distinguish key terms more efficiently and to take into account the length of a specific document and the characteristics of the array of all documents
Reducing noise and emphasizing contextual relevance	Helps filter out common words (such as «and», «that», «or», etc.) that appear frequently but do not have a significant meaning. This reduction of noise helps to focus on important terms	Similarly, it reduces noise by considering different document lengths. At the same time, it provides a more context-relevant evaluation of terms
Effective representation of functions	Converts text into a numeric representation (vectors) while preserving semantic information. Creates the basis of machine learning models for text analysis and classification	Improves text rendering by removing limitations in TF-IDF. Provides a more balanced and accurate representation of features for modeling
Increasing the accuracy of classification	Creates more informative and differentiated representations of texts. This helps build classification models that better distinguish between real and fake news based on an assessment of the importance of terms in a document	

5. 2. Results of investigating the technique of balancing the parameters k_1 and b of the BM25 algorithm for the implementation of the improved method

To obtain a balanced value of the parameters, a number of experimental studies were conducted for different data sets. Input samples were formed from news from social networks with a difference in the creation date of one day. Table 3 gives the results of selecting the configuration of parameters taking into account the priority of various criteria.

Table 3

Configuration results of parameters k_1 and b, according to different criteria

Criterion	k_1	b
Speed priority	from 1.0 to 1.5	from 0.3 to 0.5
Accuracy priority	from 2.5 to 3.5	from 0.6 to 0.8
Compromised values	from 1.5 to 2.5	from 0.5 to 0.8

In the first case, the parameters are selected according to the priority of calculation speed, while maintaining acceptable accuracy. Smaller values of b minimize the effect of document length on the IDF and therefore speed up the computation. In the second, the parameters emphasize accuracy and memorization, giving preference to calculation speed. A higher value of k_1 enhances the effect of term frequency normalization, while a slightly higher b maintains a certain balance in document length normalization. In the third, the values strike a balance between term frequency normalization (k_1) and document length normalization (b), offering a good compromise between accuracy and recall while maintaining a reasonable computational speed.

For a deeper understanding of the influence of coefficient variation, a chart of the BM25 evaluation function is plotted depending on the frequency of terms and taking into account the IDF (inverse frequency of documents) according to formula (1).

In Fig. 1, *a*, a chart of the function with coefficients k_1=1.0, b=0.3 is plotted; in Fig. 1, *b*, a chart of the BM25 evaluation function is plotted with parameters k_1=2.5, b=0.8.

At the same time, with the first combination, it was possible to achieve a full-fledged jump of the curve for the terms faster, and with the second – a noticeable effect of increasing the iterations for the accuracy of the ranking assessment.

Fig. 2 shows the evaluation function of BM25 with the value of the parameters k_1=1.7, b=0.9.

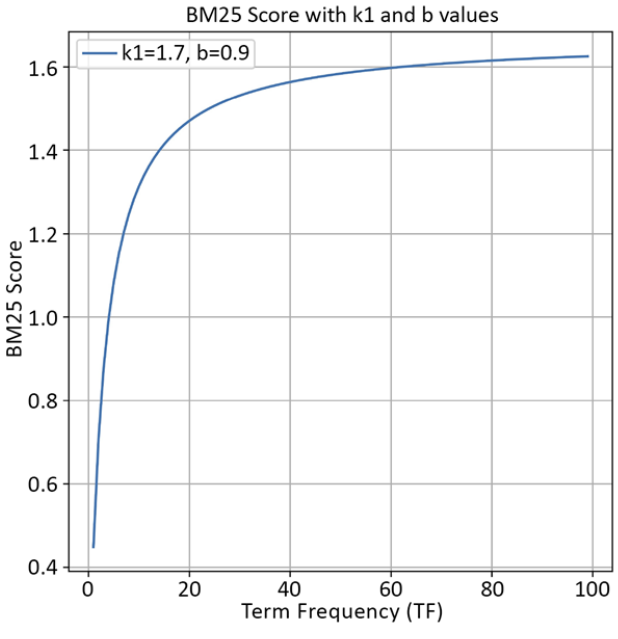
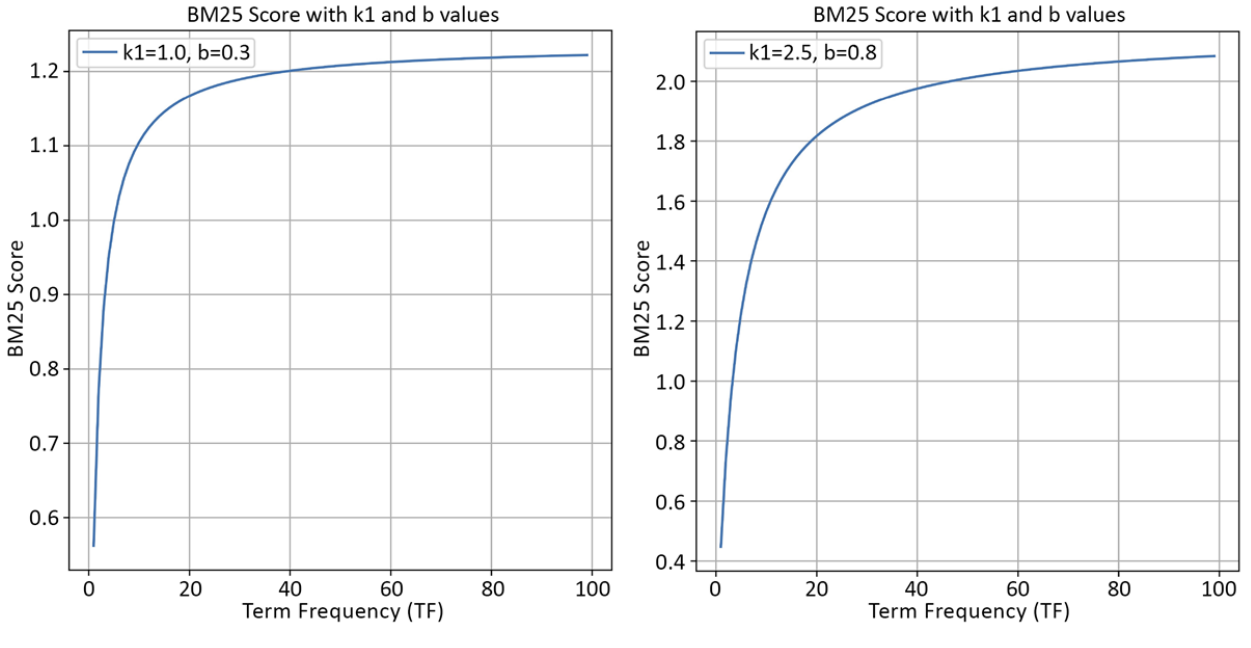


Fig. 2. Chart of the BM25 evaluation function with optimal parameters



a *b*

Fig. 1. Chart of the BM25 evaluation function with priority: *a* – speed; *b* – accuracy

With the given parameter values, it was possible to achieve the best result for the research task. Although b is outside the previously calculated trade-off range, this combination of k_1 and b is well balanced and has a trade-off between speed and accuracy.

Parameter sets were evaluated using cross-validation and performance metrics. At the same time, the model demonstrates an increase in accuracy of up to 14 %, compared to the use of the standard TF-IDF calculation. The results show that the proposed combination best meets the criteria of accuracy and speed of distinguishing real news from fake news.

6. Discussion of results of the research on recognition of fake news based on NLP and balanced parameters of BM25

Research on fake news recognition using natural language processing (NLP) using the BM25 algorithm presents a comprehensive approach to combating disinformation.

The integration of the BM25 algorithm into the method of recognizing fake news has shown its effectiveness and reliability for text analysis. Its adaptability to documents of different lengths is a notable advantage. Comparative analysis in Table 2 confirms the effectiveness and feasibility of using the BM25 algorithm with high performance requirements. This is explained by the fact that BM25 is adapted to dynamically changing the length of the text and an unbalanced data set. In contrast to the solution from [15], where the tasks of classification and ranking of words in the text are implemented using TF-IDF, the result of involving the BM25 algorithm for evaluating the importance of words in the text in the method of recognizing fake news. This algorithm makes it possible to introduce and balance such a pair of parameters, in which the high accuracy of word estimation does not significantly affect the speed of the method with dynamically changing length of texts.

BM25's ability to consider term frequency, document length, and document inverse frequency offers detailed analysis of textual content. However, difficulties may arise when handling data sets with highly skewed distributions or in texts where the importance of desired terms may not be sufficiently covered. Also, complex linguistic constructions or different languages in the input data can make theoretical expectations unrealistic.

To improve the method of recognizing fake news to increase the efficiency of processing texts with dynamically changing length, as well as unbalanced data and topics, a method for determining the balanced parameters k_1 and b of the BM25 algorithm was developed. To select these parameters of the BM25 function, a number of experiments are conducted within the framework of the stage of classification and assessment of the importance of words. The results of changing the value of the parameters depending on the criteria are described in Table 3, where the optimal limits of fluctuation k_1 and b are given, at which the required level of accuracy is achieved, and computational efficiency is preserved.

In contrast to the experiments [16] on assigning universal values to the parameters of the BM25 algorithm, or the random autocorrelation process method [17], where the BM25 indicator sequence model can be trained to determine the threshold value for each pair of parameters. The result of a balanced selection for given input data makes it possible to create a reliable tool for processing texts of various structures. This is made possible by decreasing the value of b, which minimizes the effect of the document length on the IDF and therefore speeds up the calculation. At the same time, a little higher b also maintains a certain balance in normalizing the length of the document. And increasing the value of k_1, which strengthens the effect of normalizing the frequency of terms.

It is important to note that the defined parameters are balanced for the given input data. The field of application of the research results is media and information resources for automated detection of fake news. The conditions of application of the proposed solution are the use of the same or a similar set of input data for the initial training of the model. Or the correction of parameters in an experimental way with a change in the source, topics, or structure of the texts. For the stability of the solution, it is necessary to perform all stages of pre-processing of the text, especially labeling and cleaning of noise in the text. Under these conditions, the results are adequate and can be reproduced to optimize the recognition of fake news. The disadvantages of the proposed method are the need for a significant number of experiments to check the selected set of parameters of the BM25 algorithm and the English-language set of input data.

Overall, the use of the BM25 algorithm in the context of natural language processing to detect fake news has demonstrated the potential of an effective and reliable tool for working with unstructured data. Further research could be directed at testing the reliability and accuracy of the BM25 in processing the various linguistic features and complex linguistic nuances inherent in fake news recognition.

Future research may focus on comparing BM25 with other state-of-the-art algorithms. Also with the increasing number of experiments with various data sets, complex language constructions. And studying the performance of BM25 in scenarios involving multimedia content or the evolution of language patterns to improve its applicability in social networks.

7. Conclusions

1. The comparative analysis of BM25 with the TF-IDF method theoretically substantiated the advantages of BM25 in the method of detecting fake news with unbalanced data sets and texts of different lengths without losing the accuracy and speed of news analysis. Its adaptability to texts of different lengths and ability to count on the frequency of terms is an important advantage for text analysis.

2. Based on the proposed technique for determining the balanced parameters k_1 and b of the BM25 algorithm, a number of experiments were conducted. They showed that the settings of parameters k_1 and b affect the efficiency of the algorithm. It is important to find optimal values that provide a balance between accuracy and speed of text processing. Therefore, we explored the ranges for each parameter in which the ranking function performs best. So, for faster operation of the algorithm, the coefficient k_1 varies from 1.0 to 1.5, b – from 0.3 to 0.5; to increase the accuracy index, the range changes from 2.5 to 3.5 for k_1 and from 0.6 to 0.8 for b; the balanced value between the two performance evaluation criteria is between 1.5 and 2.5 for k_1 and between 0.5 and 0.8 for b.

Taking into account the peculiarities of the research test data – texts with dynamically changing length and unbalanced data – and the proposed technique, the optimal pair of parameters was selected: k_1=1.7 and b=0.9. Although b is outside the previously calculated trade-off range, this combination of k_1 and b is well balanced and has a trade-off between speed and accuracy. With such values, the model demonstrates an increase in accuracy of up to 14 %, compared to using the standard TF-IDF calculation.

Conflicts of interest

The authors declare that they have no conflicts of interest in relation to the current study, including financial, personal, authorship, or any other, that could affect the study and the results reported in this paper.

Funding

The research was carried out with the financial support of RebelRoam OÜ, Tallinn, Estonia.

Data availability

The manuscript has associated data in the data warehouse.

Use of artificial intelligence

The authors used artificial intelligence technologies within acceptable limits to provide their own verified data, which is described in the research methodology section.

References

1. Sharifani, K., Amini, M., Akbari, Y., Aghajanzadeh Godarzi, J. (2022). Operating Machine Learning across Natural Language Processing Techniques for Improvement of Fabricated News Model. International Journal of Science and Information System Research, 12 (9), 20–44. Available at: https://ssrn.com/abstract=4251017
2. Yoo, J.-Y., Yang, D. (2015). Classification Scheme of Unstructured Text Document using TF-IDF and Naive Bayes Classifier. Advanced Science and Technology Letters. doi: https://doi.org/10.14257/astl.2015.111.50
3. Fan, H., Qin, Y. (2018). Research on Text Classification Based on Improved TF-IDF Algorithm. Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018). doi: https://doi.org/10.2991/ncce-18.2018.79
4. Dai, W. (2018). Improvement and Implementation of Feature Weighting Algorithm TF-IDF in Text Classification. Proceedings of the 2018 International Conference on Network, Communication, Computer Engineering (NCCE 2018). doi: https://doi.org/10.2991/ncce-18.2018.94
5. Izzah, I. K., Girsang, A. S. (2021). Modified TF-Assoc term weighting method for text classification on news dataset from twitter. IAENG International Journal of Computer Science, 48 (1), 142–151. Available at: http://www.iaeng.org/IJCS/issues_v48/issue_1/IJCS_48_1_15.pdf
6. Wang, S., Jiang, L., Li, C. (2014). Adapting naive Bayes tree for text classification. Knowledge and Information Systems, 44 (1), 77–89. doi: https://doi.org/10.1007/s10115-014-0746-y
7. Alammary, A. S. (2021). Arabic Questions Classification Using Modified TF-IDF. IEEE Access, 9, 95109–95122. doi: https://doi.org/10.1109/access.2021.3094115
8. Dogan, T., Uysal, A. K. (2019). On Term Frequency Factor in Supervised Term Weighting Schemes for Text Classification. Arabian Journal for Science and Engineering, 44 (11), 9545–9560. doi: https://doi.org/10.1007/s13369-019-03920-9
9. Ketola, T., Roelleke, T. (2023). Automatic and Analytical Field Weighting for Structured Document Retrieval. Advances in Information Retrieval, 489–503. doi: https://doi.org/10.1007/978-3-031-28244-7_31
10. Liu, T., Xiong, Q., Zhang, S. (2023). When to Use Large Language Model: Upper Bound Analysis of BM25 Algorithms in Reading Comprehension Task. 2023 5th International Conference on Natural Language Processing (ICNLP). doi: https://doi.org/10.1109/icnlp58431.2023.00049
11. Mishchenko, L., Klymenko, I. (2023). Method for detecting fake news based on natural language processing. The VI International Scientific and Practical Conference «Modern ways of solving the problems of science in the world», Warsaw, 375–378. Available at: https://eu-conf.com/ua/events/modern-ways-of-solving-the-problems-of-science-in-the-world/
12. Introduction to Information Retrieval BM25, BM25F, and User Behavior Chris Manning and Pandu Nayak. Available at: https://web.stanford.edu/class/cs276/handouts/lecture12-bm25etc.pdf
13. Lv, Y., Zhai, C. (2012). A Log-Logistic Model-Based Interpretation of TF Normalization of BM25. Advances in Information Retrieval, 244–255. doi: https://doi.org/10.1007/978-3-642-28997-2_21
14. Vo, N., Lee, K. (2020). Where Are the Facts? Searching for Fact-checked Information to Alleviate the Spread of Fake News. Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP). doi: https://doi.org/10.18653/v1/2020.emnlp-main.621
15. Seitz, R. (2020). UNDERSTANDING TF-IDF AND BM-25. Available at: https://kmwllc.com/index.php/2020/03/20/understanding-tf-idf-and-bm-25
16. Liu, C., Sheng, Y., Wei, Z., Yang, Y.-Q. (2018). Research of Text Classification Based on Improved TF-IDF Algorithm. 2018 IEEE International Conference of Intelligent Robotic and Control Engineering (IRCE). doi: https://doi.org/10.1109/irce.2018.8492945
17. Liu, T., Zhang, S., Xiong, Q. (2023). Separated Model for Stopping Point Prediction of Autoregressive Sequence. 2023 IEEE 12th Data Driven Control and Learning Systems Conference (DDCLS). doi: https://doi.org/10.1109/ddcls58216.2023.10167110