Evaluation of Informatioon Retrieval Systems for Croatian

Kai Garcia, Tin Cvrlje, Qin Wang, Wenyu Li
Uppsala University, Department of Linguistics
Uppsala, Sweden
kai.garcia.2355@student.uu.se,
tin.cvrlje.9198@student.uu.se,
qin.wang.3929@student.uu.se,
wenyu.li.6753@student.uu.se

Abstract

The main goal of our research is to determine what is the best performing ranking algorithm for Croatian language specific information retrieval. We accomplish this by comparing the performance of TF-IDF, BM25, and dual-stage systems using BM25 for initial retrieval and SBERT for re-ranking. Additionally, we evaluate whether utilizing annotated similarity scores for document pairs significantly impacts performance of the ranking algorithms we are comparing. We propose an evaluation method that does not require any keyword extraction or similar annotation by using news article titles as queries and accompanying bodies as the target documents. Our results demonstrate that BM25 outperforms both TF-IDF and the dual-stage systems, and that utilizing similarity scores has minimal impact on the performance of our models. We argue that these results might have been influenced by the small size of our dataset and insufficient computational resources and suggest that the topic remains open for future research.

1 Introduction

The most important part of every Information Retrieval system (IR system) is the ranking algorithm it implements. The main goal of our research is to determine the best performing ranking algorithm for a Croatian language specific IR system and explore whether their performance can be improved upon. We take on this task by comparing the performance of TF-IDF (Term Frequency-Inverse Document Frequency), BM25 (Best Matching 25) and a dual stage IR system incorporating BM25 for the initial retrieval and SBERT (Sentence-BERT) for reranking a specific number of top candidates BM25 returned.

Additionally, to explore a possible improvement of Croatian specific IR systems, we evaluate whether utilizing annotated similarity scores for document pairs significantly impacts the performance of IR systems. We suggest that this type of annotation should prove as an improvement on the keyword matching based ranking algorithms (such as TF-IDF and BM25) because it expands the list of retrieved documents for each query to documents that might be semantically related to the user's query but do not get recognized because of the lexical gap between the query and their content. We suggest that this method might be a useful not only for Croatian, but all medium to low resource languages that the current version of SBERT might struggle with due to the less than sufficient amount of training data.

We also propose a simplified method for IR system evaluation as an alternative to keyword extraction, which is a very labor intensive and overly time consuming task. Since the dataset we use consists of news articles, we utilize their structure by using news article titles as queries and their accompanying bodies as the target documents that should be retrieved by those queries.

To our knowledge, there is no published article that explores these research questions for Croatian language specifically, while exploring the impact of annotating similarity scores between document pairs seems to be an unexplored topic in the field of IR.

2 Related Work

In recent years, a lot of research in the field of information retrieval has focused on building dual stage models capable of retrieving relevant documents, but also reranking them in a more sophisticated manner. A work by Ofori-Boateng et al. (2024) introduced dual-stage information re-

trieval system that solved the challenge of retrieving Spanish medical literature, which us to explore whether a similar model could be adapted to our Croatian news article dataset. In their work, they combined BM25 with several pre-trained language models for reranking to create a complex dual-stage system. We adapted this framework to match with currently available resources for Croatian by using BM25 with cosine similarity for the initial retrieval portion of our model, followed by using a multilingual SBERT for re-ranking.

In their research on creating a semantics only based method for determining short text similarity using word embeddings, Kenter and de Rijke (2015) suggest that utilizing such models can be highly beneficial to most areas of information retrieval. In our research, we evaluated this hypothesis by utilizing hand annotated semantic similarity scores between news article pairs in our dataset to see if it improves each of the IR systems we use.

A work by Karan et al. (2013) introduced a FAQ retrieval system for Croatian that solved the fundamental challenge of lexical gap between queries and documents. Our work draws several parallels with Karan et al.'s research, our document-query structure is similar to how they used FAQ questions and answers as paired texts, we utilize news titles as queries and article bodies as target documents. We also used similar preprocessing techniques for Croatian text, including lemmatization and stopword removal, which is crucial for handling the morphologically rich Croatian language.

3 Method

Our project compares three different ranking algorithms for information retrieval: TF-IDF with cosine similarity, BM25, and a hybrid BM25+SBERT model. Our pipeline processes text data from Croatian news articles, using titles as queries and bodies as the documents. Below, we detail important technical methods for reproducing the key steps.

3.1 Data Preprocessing

The dataset, loaded via Pandas, contains 2,109 documents with IDs, titles, and bodies. A dictionary maps IDs to titles for query processing. Croatian stopwords are loaded from `stopwords-hr.txt`[1]. Text preprocessing uses the CLASSLA pipeline[2] (`lang='hr', processors='tokenize,pos,lemma'`) to tokenize, lowercase, lemmatize, and filter out punctuation and stopwords. The cleaned data is cached in `lemmatized_data.pkl`, so preprocessing does not need to be repeated every time the model is run (the process takes around 20 minutes).

3.2 Vocabulary and Term-Document Matrix

After preprocessing the data, a vocabulary of 39,954 unique lemmas is built from document bodies. We created a Pandas DataFrame where rows are vocabulary terms, columns are document IDs, and values are term frequencies (TF), computed using `Counter` from the `collections` module.

3.3 TF-IDF Model

The TF-IDF model utilizes 3 values:

- **TF**: How often a term occurs in a document. Equal to f/w, where f = frequency of the term in the document and w = count of all words in the document. We used logarithmically scaled term frequency, equal to log(1 + TF).
- **IDF**: How rare a term is across all documents. Equal to log(n/d), where n = total number of documents and d = documents that contain the term.
- **TF-IDF**: The product of TF and IDF.

We preprocess the query using the same approach as the data (tokenization, lemmatization, lowercasing, and filtering) and convert it to a TF-IDF vector. We then create a TF-IDF matrix and and calculate cosine similarity to rank documents. The top 5 documents are retrieved.

3.4 BM25 Model

The BM25 model uses the BM-25 formula:

$$\text{BM25} = \sum_{\text{term}} \text{IDF} \cdot \frac{\text{TF} \cdot (k_1 + 1)}{\text{TF} + k_1 \cdot (1 - b + b \cdot \frac{\text{DL}}{\text{avgDL}})}$$

where DL is document length, avgDL is average document length, k_1 controls term frequency saturation, and b controls document length normalization. We set k_1=1.5 and b=0.75 for our model, since these are often used as default values that

[1] `https://github.com/stopwords-iso/stopwords-hr/blob/master/stopwords-hr.txt`

[2] `https://github.com/clarinsi/classla`

balance both features. The query is preprocessed using same approach as the data, and documents are ranked by BM25 scores, retrieving the top 5.

3.5 Hybrid BM25+SBERT Model

The hybrid model combines BM25 for initial retrieval and SBERT for reranking. BM25 is used to retrieve k documents, which are then reranked using two SBERT models from the SentenceTransformer library[3]: `distiluse-base-multilingual-cased-v1` and `distiluse-base-multilingual-cased-v2`. The original document texts are encoded into embeddings, and we calculate cosine similarity between query and document embeddings. We use the torch library's topk function to retrieve the top 5 documents after reranking.

3.6 Evaluation

We evaluate the performance of our models based on 4 key metrics:

- Accuracy@5 - percentage of queries where the target document was a part of the five most relevant documents list retrieved by each model
- Average Rank Position - mean position of the target document inside the five most relevant documents list in cases of successful retrieval
- MAP - Mean Average Precision
- Our custom "similarity score usefulness" value which indicates how often a document with a high similarity score to the targeted one wasn't a part of five most relevant documents list retrieved by each model.

4 Data

The data source for our project is the hand-annotated data from the "Article-BERTić: Croatian article semantic similarity dataset" GitHub repository, created by Ir2718 in 2023.[4] The data we use is stored in six .json files, one of them contains a list of fifty dictionaries while the other five contain a list of one hundred dictionaries each. Each dictionary in those lists holds information about a pair of news article documents (their IDs, article titles, article bodies, portals they were scraped from and the date they were published) along with their similarity score annotation ("choice") and information about the annotation/annotator (annotator, annotation id, time it was created at, time it was published at and lead time).

We converted this data to 2 different .json files we used to conduct our experiments. One of them will function as the main dataset used for evaluating the performance of our information retrieval models while the other one will be used to determine whether assigning similarity scores between documents can provide a useful extension to IR systems.

To create the first of these datasets, we restructured the original .json file so that each dictionary contains information about a single news article. We also deleted all the unnecessary dictionary keys we do not utilize in our research – all of them apart from the IDs, titles and bodies. To achieve that, we split up each original dictionary into a couple of separate ones, in a way that each new dictionary consists of a single article ID, title and body while also adjusting the "id2", "title2" and "body2" key names to match the simple "id", "title" and "body" ones. Finally, we removed the duplicate instances of all documents that appeared more than once and then assigned new IDs to each document starting from 1. This file ended up being 2109 documents long.

The second file we created contains information about news article pairs (IDs, titles and bodies) along with their similarity scores, but only if the similarity score indicates high similarity between the document pair. Based on the labeling guidelines provided (listed below) in the GitHub repository we decided to only include document pairs with a similarity score of 4 or higher.

- 5 - The articles are completely equivalent and express the same meaning
- 4 - The articles are mainly equivalent; some unimportant details are different or missing
- 3 - The articles are mainly equivalent, but some important details are different or missing
- 2 - The articles are not equivalent, but they do

[3] `https://huggingface.co/sentence-transformers`

[4] `https://github.com/ir2718/article-bertic`

```
{
  "id": 298,
  "title": "Potpisan je ugovor: za vrtić 1,45 milijuna eura",
  "body": "SLATINA - Grad Slatina sklopio je s Ministarstvom obrazovanja te Središnjom agencijom za financiranje i ugovaranje programa i p
  "portal": "glas-slavonije.hr",
  "date_published": "2023-03-19 00:00:00",
  "id2": 11479939,
  "title2": "Đakovu 1,7 mil. eura za vrtiće Sjever i Vila",
  "body2": "Đakovu 1,7 mil. eura za vrtiće Sjever i Vila\nĐAKOVO\nUgovore o dodjeli nepovratnih sredstava za projekte koji se financiraju
  "portal2": "glas-slavonije.hr",
  "date_published2": "2023-03-15 00:00:00",
  "choice": "3",
  "annotator": 1,
  "annotation_id": 312,
  "created_at": "2023-03-23T15:17:27.207679Z",
  "updated_at": "2023-03-23T15:28:23.636930Z",
  "lead_time": 133.243
},
```


Figure 1: A single dictionary in the original .json file from the GitHub repository


```
{
    "id": 1,
    "title": "60 tisuća eura za komunalne prioritete",
    "body": "Dan Mjesnog odbora Josipovac i crkveni god svetog Josipa obilježeni su u subotu polaganjem vijenaca na spomen-obilj
  }
```


Figure 2: A single dictionary containing information about a single news article from our processed .json file


```
{
  "id": 7,
  "title": "Hrvatice nisu mogle izići nakraj s Norvežankama",
  "body": "Mlada hrvatska ženska nogometna U-17 reprezentacija poražena je jučer od Norveške s 2:0 u drugom nastupu u okviru elitnog
  "id2": 8,
  "title2": "'Bili tići' izazvali su delirij, ali niste još vidjeli borbenost naših nogometašica, pogledali smo utakmicu u Hrvacama"
  "body2": "Mlada hrvatska ženska U-17 reprezentacija poražena je od Norveške (0:2) u drugom nastupu u okviru elitnog kvalifikacijsk
  "choice": "4"
},
```


Figure 3: A single dictionary containing a pair of news articles and their similarity score from our processed .json file

share some details

- 1 - The articles are not equivalent, but parts of their contents are similar
- 0 - The articles are completely different[5]

Additionally, we assigned the new IDs of each document in the first file to the appropriate documents in this one by mapping them by their titles to keep the IDs consistent. We again removed all of the unnecessary keys. This file contains 313 news article pairs along with their similarity scores.

Considering none of our models require a test/train split, there is no need to split the dataset.

5 Experiments

We used the TF-IDF model as our baseline, since it is a well-established retrieval method and the simplest of our models computationally. For the SBERT model we tested two variants, one using the original distiluse-base-multilingual-cased-v1 and another using the expanded distiluse-base-multilingual-cased-v2. For each model, we preprocessed the individual article data by removing whitespace and stopwords, lowercasing everything, and lemmatizing using a CLASSLA pipeline for Croatian. We used the data to build a term-document matrix. We then iterated over the individual articles, preprocessed each title, and entered it into the model as the query. Each model returned a list of 5 document IDs ranked in order of relevance, along with 5 metrics.

For the scoring function in our BM25 and SBERT models, we used hyperparameters k1 = 1.5 and b = 0.75. These values are frequently used as defaults in BM25 scoring, and we found them sufficient for getting strong results.

Our retrieval function for the SBERT model contained the hyperparameter top k, the number of documents returned in retrieval. We ran four tests on both SBERT variants, setting top k = 5, 25, 50, and 100, respectively, to observe the effect of adding more documents on the reranking process. In addition, we examined whether similarity scores could potentially aid in retrieval. For each of our models' retrieved documents, we checked if it appeared in our dataset of paired documents with high similarity scores. If it was present, we then checked if the paired document also appeared in the five retrieved results. We calculated the percentage of times the paired document did not appear, calling this metric "similarity score usefulness." A higher score implies the model frequently misses the second document when retrieving the first, and thus similarity scores would be a useful feature.

[5] https://github.com/ir2718/article-bertic

6 Results

The results we got are summarized in table 1. BM25 achieved the best performing metrics overall, with the highest accuracy and MAP, and the lowest average rank. It slightly outperformed the TF-IDF model baseline. SBERT v1 and v2 matched BM25's Accuracy@5 at $k = 5$, although their average rank was slightly higher and their MAP was noticeably lower. Increasing k significantly worsened performance, with v1 suffering the greatest drop across all metrics. SBERT v2 consistently outperformed v1 across all values of k. Similarity score usefulness was low across all models, slightly rising when k was increased for SBERT.

7 Discussion

Our results show that the simple BM25 model performs the strongest. BM25 is more robust than TF-IDF because it incorporates normalization for document length. Normalization accounts for the fact that higher term frequencies for longer articles might be caused by the vocabulary being larger. BM25 also implements saturation for term frequency, since the relationship between frequency and relevance is nonlinear and eventually tapers off. Both features help the BM25 model better handle long news articles that may skew TF-IDF values.

We assumed the addition of SBERT reranking would further improve BM25's performance; however, this was not the case. The SBERT variants perform nearly as well as simple BM25 when the number of documents returned in the retrieval stage is small (k = 5). Increasing this parameter worsens performance. Future studies could switch the roles in our hybrid model, using SBERT for initial retrieval and BM25 for reranking. It is important to note that this would require more computational resources than we had at our disposal.

As expected, the v1 SBERT variant performed noticeably worse than v2. v1 was trained on 14

Method	k	Accuracy@5 (%)	Avg Rank	MAP	Sim. Score
TF-IDF	-	94.22	1.39	0.8061	0.0002
Simple BM25	-	**95.78**	**1.31**	**0.8466**	0.0001
SBERT v1	5	95.78	1.78	0.7183	0.0001
	25	75.44	1.74	0.5796	0.0002
	50	69.46	1.72	0.5348	0.0003
	100	64.58	1.71	0.5009	0.0003
SBERT v2	5	95.78	1.62	0.7570	0.0001
	25	84.07	1.59	0.6741	0.0002
	50	80.65	1.57	0.6521	0.0002
	100	78.24	1.55	0.6375	0.0002

Table 1: Summary of results of all our information retrieval systems using titles as queries

languages that did not include Croatian, while v2 was trained on 53, including Croatian. This highlights the downsides of using a multilingual model for a language-specific task when that language was not in the training.

Similarity scores proved to be irrelevant to our systems' performance. For all our tests, the usefulness metric fell into a negligible range, 0.0001-0.0003. We suggest this was caused by our use of news article titles as queries, which are often long and descriptive. The query vector likely contained enough information for the system to retrieve both the target document and the paired document (a document with a high similarity score to the target document) in the top 5 almost all the time. Similarity scores would likely be more useful for shorter queries, such as 2-to-3 word phrases or proper nouns. The query would not contain enough keywords to retrieve all relevant articles, so linking articles that share similar content would be useful. Our dataset was also small, limiting the usefulness of similarity pairings compared to a large dataset with more candidates for retrieval.

Use of news article titles as queries and accompanying bodies as the target documents proved to be a viable method for evaluating our IR systems, producing high accuracy scores and excellent average rank of the target document. However, this method has its drawbacks, and these should be accounted for when conducting future research.

8 Limitations

One major limitation is the small size of our dataset. Our data consists of 2109 documents, when ideally we should be working with tens of thousands. This contributes to the poor accuracy of the SBERT model when increasing k because the impact of irrelevant documents becomes stronger, decreasing precision. Additionally, only 313 document pairs have a relevant similarity score. We suspect that we would have a larger amount of document pairs with high similarity scores. A larger dataset would also provide more ambiguity and create a larger lexical gap between document pairs covering the same topics and thus make similarity scores more relevant to our research. A "backup" system of paired similar documents could then be useful for the user if they want a document that did not make the top 5.

Our approach of using titles as queries made testing our models efficient, but it limited the role of similarity scores because the queries were really long and extensive. It also assumes only a single relevant document per query. While this was sufficient to determine that our IR systems consistently retrieve the most relevant document, it has an obvious downside - inability to evaluate entire output lists produced by our models.

We were also limited by our computing resources. We suggest it would be worthwhile to check how a dual stage IR system that uses SBERT for initial retrieval and BM25 for reranking would perform compared to the ones we

present in this paper. However, since running the SBERT model with k=100 took around 3 hours, having SBERT run on the entire 2109 document long dataset is simply too much for the resources we currently have on our disposure.

9 Ethical Considerations

Privacy and data usage are our primary concerns. News articles contain personal information, and our system must ensure that such information is handled with the care for future use. It is essential that all news content stored and processed by the system is properly licensed. Furthermore, compliance with Croatian data protection regulations and the EU regulations is necessary to safeguard personal data and uphold individuals' rights.

Societal impact of the system cannot be overlooked. The way our system influences how people access and consume Croatian news can affect public opinions. It may lead to certain news stories receiving more visibility than others, thereby shaping public perception. And this may cause consequences for society.

10 Author Contributions

Creating our pipelines was a group effort. All authors worked together in person and decided which data, algorithms, and models to use collectively. We also troubleshooted the code together, but member did take initiative at certain points of our research process listed bellow.

Tin: Formulating the general skeleton for our research, finding and preprocessing the data, deciding on how to evaluate the models.

Kai: Formulating the general skeleton for the data preprocessing pipeline, writing code for the TF-IDF model, writing reusable code for evaluating the four metrics for the models.

Wenyu: Formulating the similarity score metric, writing code for the simple BM25 model.

Wang: Finding related articles concerning BM25 and SBERT, writing code for the dual-stage SBERT model.

11 AI Contributions

We used AI to help us debug our code and suggest ways to implement unfamiliar functionality. We tried to use AI as a troubleshooting tool, rather than having it generate all our code for us. For example, when implementing CLASSLA and SBERT, we first relied on documentation from PyPI and HuggingFace to set up our pipelines. While resolving library version conflicts and integrating data into the pipeline, we used AI to explain why errors were happening and to clarify what variables we needed to pass in.

12 Conclusions

Our study successfully compared the performance of TF-IDF, BM25, and dual-stage BM25+SBERT information retrieval systems on Croatian news data. We found that BM25 is the best performing system. It handles variance in document length better than TF-IDF because it accounts for length normalization and term saturation. SBERT's performance grew worse as we increased the number of retrieved documents due to a precision-recall tradeoff. The small size of our dataset, combined with our approach of using titles as queries, limited the usefulness of similarity scores. For future research, we suggest working with a larger dataset and modifying the dual-stage model to use SBERT for retrieval and BM25 for reranking.

References

Ir2718. 2023. Article-BERTić: Croatian Article Semantic Similarity Dataset. Data set, GitHub. https://github.com/ir2718/article-bertic

Mladen Karan, Lovro Žmak, and Jan Šnajder. 2013. Frequently Asked Questions Retrieval for Croatian Based on Semantic Textual Similarity. In *Proceedings of the 4th Biennial International Workshop on Balto-Slavic Natural Language Processing*, pages 24–33, Sofia, Bulgaria. Association for Computational Linguistics. https://aclanthology.org/W13-2405

Tom Kenter and Maarten de Rijke. 2015. Short Text Similarity with Word Embeddings. In *CIKM'15: Proceedings of the 24th ACM International Conference on Information and Knowledge Management*, pages 1411–1420, Melbourne, Australia. The Association for Computing Machinery. https://doi.org/10.1145/2806416.2806475

Richmond Ofori-Boateng, Monica Aceves-Martins, Nirmalie Wiratunga, and Carlos Moreno-Garcia. 2024. A Zero-Shot Monolingual Dual Stage Information Retrieval System for Spanish Biomedical Systematic Literature Reviews. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, volume 1, pages 3725–3736. Association for Computational Linguistics.