Faculty of Mathematics and Physics

Analysis of Slovene media articles on the topic of social work

Joint project with the "Jožef Stefan" Institute

Paula Dodig
p.dodig@sudent.tue.nl

Supervizors:
Sergio Cabello (FMF)
Senja Pollak (IJS)

Ljubljana, May 2024

Chapter 1

Introduction

This project concerns the subject of topic modeling in Slovene news articles. It was offered by the Institute Jožef Stefan in collaboration with an expert from the Ljubljana Social Work Centre. The main point of interest is to analyze the distribution and change of topics related to social work problems and scandals in Slovenia. Social work experts are interested in comprehending the media coverage and their sentiment towards social work related topics as well as simply confirming their assumptions about the media and their involvement.

To start of, topic modeling is an unsupervised machine learning technique working on textual data by identifying clusters of words and documents from the corpus body to surface hidden semantic patterns that can be interpreted as topics covered in the text. Even though people are good at identifying these underlying topics from text by reading and interpreting, the increasing amounts of big data available for analysis make it impossible to extract such knowledge manually and efficiently on a larger scale. Specifically, the development of online news has brought a whole new interest in deployment of text to understand some of the social processes going on in society. This is why the Ljubljana Social Work Center believes there is benefit to be reaped.

As there have been plenty of developments in the field of topic modeling in the recent years, we will first review the literature in search of the best performing models that would work on Slovene data. Provided the data, we will then delve into necessary preprocessing and exploratory statistics to get a better idea about its nature. Finally, having prepared the data, we can train the selected models and collect the topic modeling results that can later be analyzed from a sociologist perspective so that some conclusions can be made on the problem.

Chapter 2

Literature Overview

In 2015, Alghamdi and Alfalqi [1] outlined some of the main characteristics and limitations of four most notable models in this field. These include Latent Semantic Analysis, Probabilistic Latent Semantic Analysis, Latent Dirichlet Allocation and the Correlated Topic model, all of which belong to the first generation topic models brought up by David M. Blei [2] - arguably the father of topic modeling. Even though the methods are briefly outlined, this work concisely describes the main characteristics, limitations and the development process of these methods, putting into perspective the intended use.

Through the years, there have been a lot of updates of old and developments of new topic models, each better than the other, bringing us to the current state-of-the-art. Nowadays, one of the most deployed algorithms in this field has got the be the BERTopic [4]. Taking a modular approach, this model is very flexible and adjustable in different phases it takes. Firstly, the documents from the corpus have to be embedded into a numerical representation, meaning that instead of lists of words, we get an array of numbers or a vector using a chosen embedding model. Next, the vectors go through a dimensionality reduction algorithm so as to save resources for other steps. After this, in order to group similar words and documents together, BERTopic performs a density based clustering algorithm HDBSCAN [7]. Finally, the retrieved clusters have to be converted back to understandable topic representations with the CountVectorizer and c-TF-IDF calculation [4].

Along with BERTopic's modularity comes the ability to change specific procedures in each of the steps. Since the project data is fully in the Slovene language, one has to create purposeful document embeddings such that the semantics still hold the intended information after the conversion from textual format. On this matter, there have been a couple different embedding models developed that have shown good results in this language. Firstly, the default BERTopic embedding procedure on languages other than English is the "paraphrase-multilingual-MiniLM-L12-v2" model that has shown good performance in a plethora of languages [6]. Next up is the newly developed "BGE-M3" Hugging face model [3] famous for its ability to successfully process input of different granularities which will be useful with news articles of various lengths. Finally, we also explore a model specifically developed for the Slovene language, SloBERTa [8].

Chapter 3

Data

The corpus in question has been extracted with the help of Slovenian media monitoring company Kliping, as part of the research project between the 'Jožef Stefan' Institute (hereinafter IJS) and Kliping. The articles have been extracted by the company via keyword extraction where the searched keywords were of the form:

- socialn* del*
- Socialn* del*
- CSD

This ensures that the overall subject is connected to social work in Slovenia. Time of publishing of the articles ranges between the beginning of 2010 and the end of 2023 which equated to 14486 articles. The attributes of the data used are as follows:

- publishing date
- media house (publisher)
- title
- body (main text)

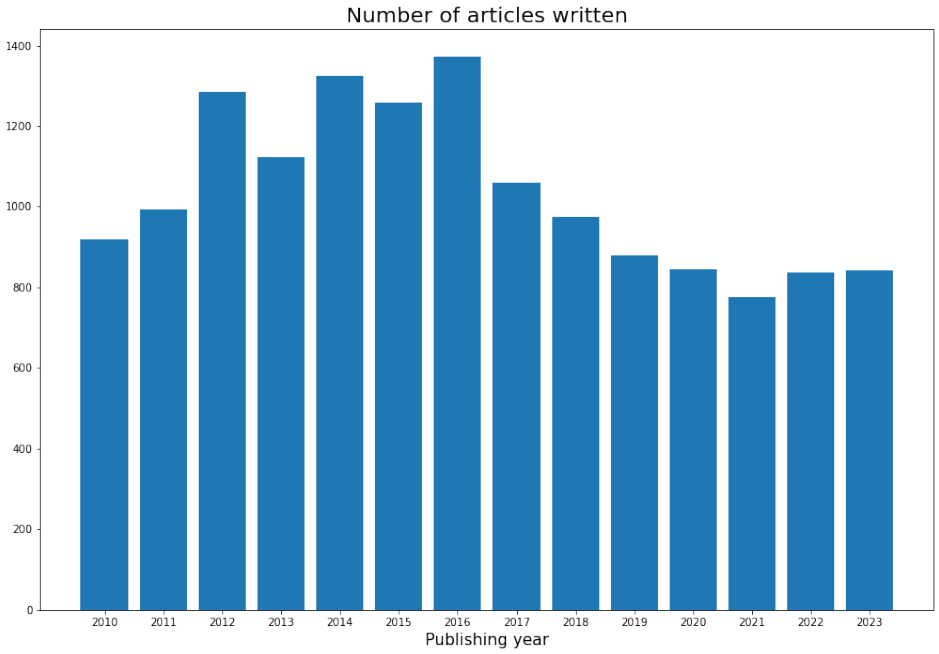
Publishers in the data include 11 of the most famous news sources in Slovenia: Večer, Delo, Dnevnik, Slovenske Novice, Primorske Novice, Svet24, Nedeljski Dnevnik, Mladina, Nedelo, Delo - Sobotna Priloga, Ona plus. Seeing that some of these sources are daily vs weekly publishers, the data contains a wider range of lengths per article, which brings to the diversity and insight of news. In the next chapters, more details on the exploratory analysis and preprocessing of the data will be shared.

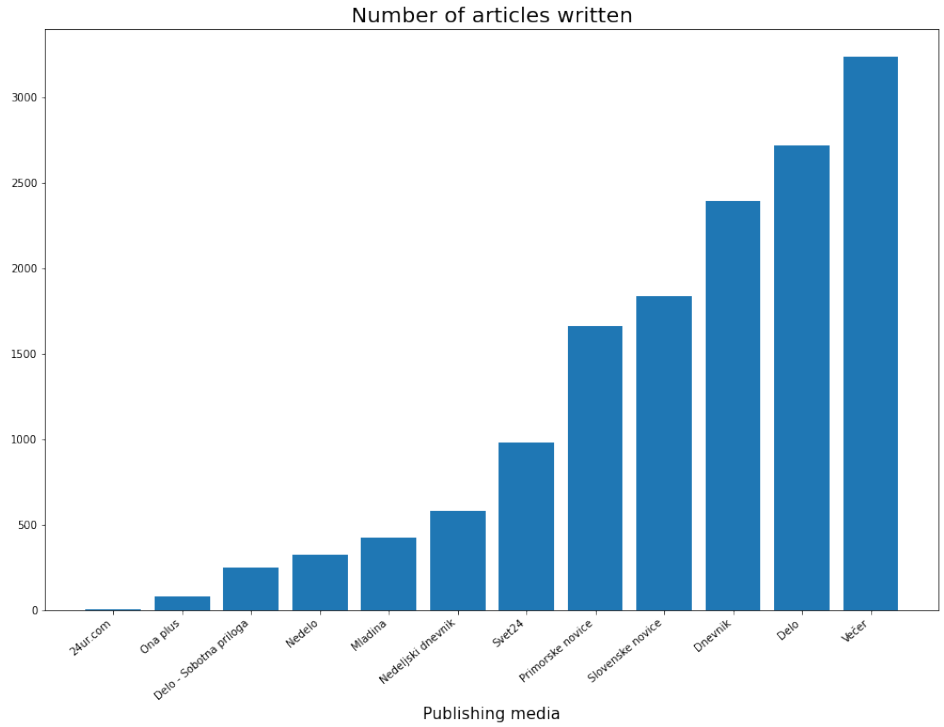
Chapter 4

Methodology

4.1 Preprocessing

To be able to use the proposed models, one needs to get a better idea on the data and its purity. For clarity, we first present some of the basic exploratory statistics prior to modification of text. In the figures bellow, we can see the distribution of the articles per year of publishing as well as per publisher. This type of relatively constant activity plays in favour of our analysis so as to avoid bias. Even though there are some fluctuations in how many articles concerning social work have been written, there are no significant drops of the interest in this topic. On the other hand, the publishers were not all on the same level of activity. This distribution needs to be taken into account if there is interest in distinguishing the topics based on who wrote about them.





Given the inevitable text redundancy, some steps need to be made to ensure the most information dense results from the models. Firstly, the data has to be stripped of all **non-letter characters** such as numbers, punctuation and HTML encoding. Next, since the models do not know how to combine words with upper and lower case spelling, all text has been converted to **lower case**, regardless of the naming rules in natural language. Both these procedures have been done with the help of RegEx library in Python. To further purify the information, words with low semantic meaning (such as 'and', 'I', 'but'...) have to be discarded. These types of words are called **stopwords** which rarely contribute and more often distort the distribution of topics since they present no extra information and appear in almost any cluster by default. Finally, we convert words with a same root into unique **lemmas** since words with different spellings would otherwise be considered different while in reality, we extract the same meaning from them. An example could be seen in words such as *studying, study, studies* all represented by the root lemma *study*.

4.2 Models

Finally, after preparing the dataset, model training can be done using BERTopic. As mentioned before, this procedure will involve fitting three different embeddings of the documents and feeding them to the BERTopic model consecutively, holding other parameters constant. Going through the steps of BERTopic as described in Chapter 2, after the data has been fed to BGE-M3, Multilingual-MiniLM and SloBERTa model, we prepare a dimensionality reduction algorithm based on theory in topological data analysis and manifold approximation called UMAP [5]. The dimensionality of the UMAP model was set to 5 and a cosine metric was used for calculating the distance between points. More specifically, distance between two

embeddings (or vectors) $\boldsymbol{a}$ and $\boldsymbol{b}$ is given by:

$$\frac{\boldsymbol{ab}}{|\boldsymbol{a}||\boldsymbol{b}|} = \frac{\sum_i^n a_i b_i}{\sqrt{\sum_i^n a_i^2}\sqrt{\sum_i^n b_i^2}}$$

UMAP model with these parameters gets fed directly to the BERTopic, which does the rest of the clustering on its' own. Since each of the three embeddings had to be fed separately, there will be three set of results from the BERTopic on our data that can later be combined in one comprehensive solution. For the sake of simplicity, the next chapter 5 shows results for only one of the runs to get an idea about the effectiveness of BERTopic. The models have consistently been called: topic-model-multi, topic-model-bge and topc-model-slo. Since intrepretation of each result individually would take too much space, we will focus on topic-model-bge, the BERTopic with the embedding model BGE-M3.

Chapter 5

Results

As it has been mentioned, we present the results of one of the models up for additional discussion. This is the topic-model-bge. Based on the first impressions from the mentors and the Social scientist, this model performs quite satisfactorily.

In figure 5.1, we can see some of the top most significant and most mentioned topics from the dataset. There are plenty of topics with a regular stable mention, while at the same time, there are a couple that stand out. One of these is the topic in yellow, topic number 3, with representative lemmas: babica, dek, vnuk, volenje. This topic shows a significant spike around the year 2016 suggesting a potential scandal or a popular story of the time. Our social work expert acknowledged that, remembering a story about a custody dillemma with a child and his grandparents. What was interesting to the experts was that this spike is so outstanding that all other topics fell significantly lower on the scale. In light blue, we can also see an interesting movement of topic number 1 with words: begunec, azil, azilen, migrant. One can see three spikes coming up in years 2012, 2016 and 2022 which might correspond to migrant crisis waves in Europe. As the migrant problem in Slovenia has become present only in the most recent years, we can see that the Slovenian media was quite interested in how this issue from other countries would impact themselves, so much that this is the second most mentioned topic overall, even when there were not so many migrants and asylum seekers in Slovenia.

In figure 5.2, we show a snippet of the hierarchical clustering from BGE. As we let the number of topics be a free parameter in BERTopic, we can see all of the clusters that were separated by the algorithm. Regardless of the type of embedding, this number hit around 260 topics which were not all shown for better transparency and clarity. In this snippet from the figure however, we can see the relationships between different topics, showing us how closely related they are. Each of the bigger branches later get their own "name" or representation. The top green branches connected together are topics mostly covering words such as "smrt" and "deklica" while the bottom purple branches cover "umor" and "obtoen".

Finally, in figure 5.3 we can see the word scores for the top 15 topics of the BGE BERTopic model. As each topic is represented by a set of words, we need to be able to distinguish the importance of those representative words. The word scores in this graph represent the L_1 normalized class frequencies calculated from the c-TF-IDF procedure in the following formula:

$$W_{x,c} = ||tf_{x,c}|| \times \log(1 + \frac{A}{f_x})$$

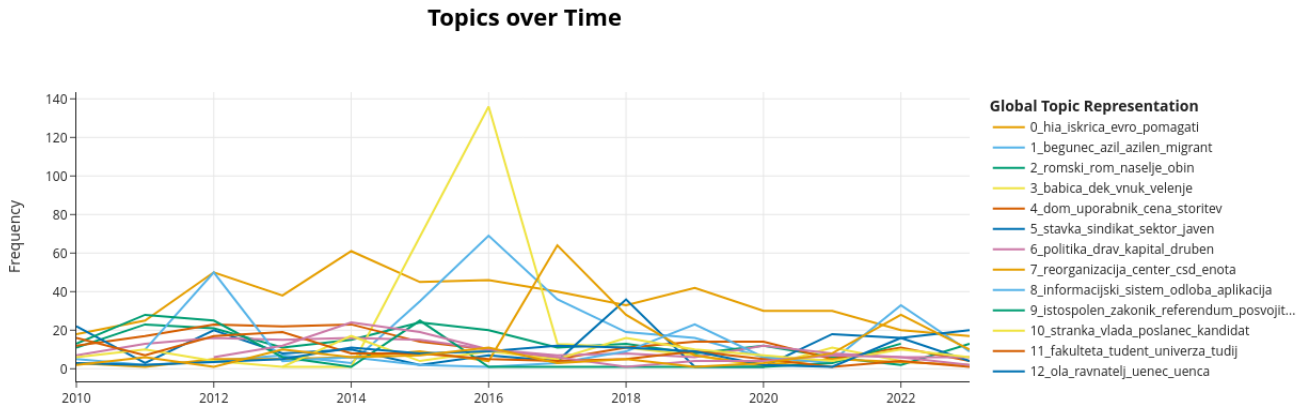


Figure 5.1: Distribution of BGE topics through time

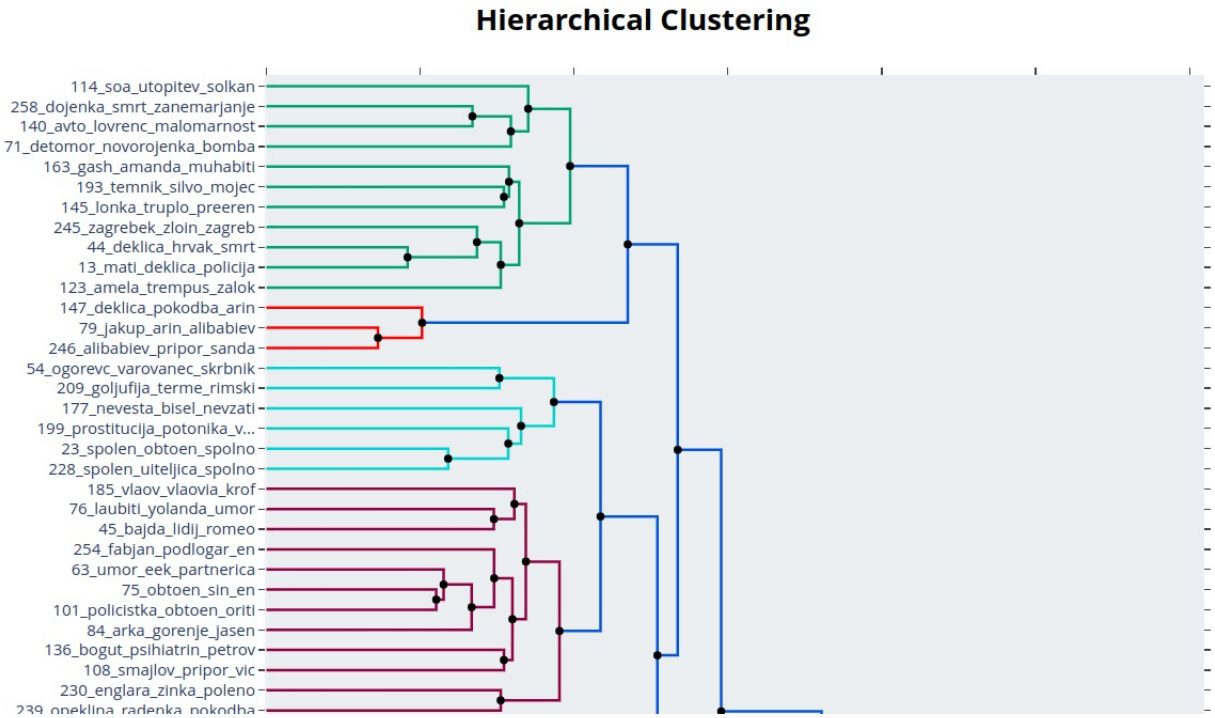


Figure 5.2: Hierarchical representation of BGE topics

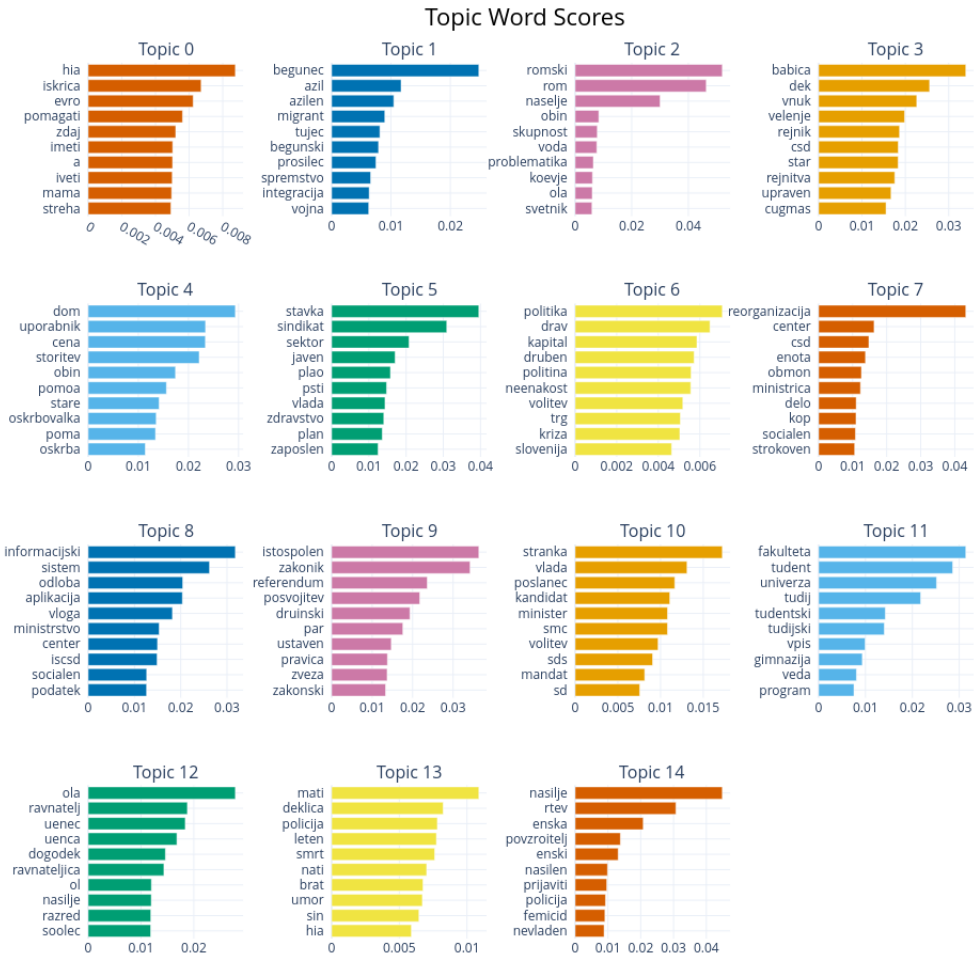


Figure 5.3: An example of word representatives in top 15 BGE topics

where

- $tf_{x,c}$ is the frequency of word x in class c
- f_x is the frequency of word x across all classes
- A is the average number of words per class

Chapter 6

Discussion & Conclusions

As this is still not the end of the project for IJS, we continue to analyze the results of all three topic models and are getting more insights. These models will later be combined into a unique model using the merge-models method with optimal parameters so as to not lose any information. There should be additional metrics for evaluating the models and more interaction with the social work expert for more domain knowledge. We still have to report coherence scores on these topics to also get a quantitative evaluation. As of now, the topics are already comprehensive and insightful, even though there are many. BERTopic has provided a good multilingual bases for the Slovene articles and it dealt with the variations in the data well.

In conclusion, we the project of Slovene news analysis in the topic of social work continues to be a success. We selected BERTopic as the best option to deal with the language offering a variety of embeddings for testing. These included the paraphrase-multilingual-MiniLM-L12-v2, BGE-M3 and SloBERTa model. We cleared the data with classical natural language preprocessing methods including lower case conversion, non-letter filtering and lemmatization. The resulting topic representations were presented in word, time series and hierarchical form and the interpretation was in line with the opinion of experts. More work that could be done will follow in the coming weeks.

Bibliography

[1] R. Alghamdi and K. Alfalqi. A survey of topic modeling in text mining. *Int. J. Adv. Comput. Sci. Appl.(IJACSA)*, 6(1), 2015.

[2] D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of Machine Learning Research*, 3(Jan):993–1022, 2003.

[3] J. Chen, S. Xiao, P. Zhang, K. Luo, D. Lian, and Z. Liu. Bge m3-embedding: Multi-lingual, multi-functionality, multi-granularity text embeddings through self-knowledge distillation, 2024.

[4] M. Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[5] L. McInnes, J. Healy, and J. Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[6] N. Reimers and I. Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 11 2019.

[7] G. Stewart and M. Al-Khassaweneh. An implementation of the hdbscan* clustering algorithm. *Applied Sciences*, 12(5):2405, 2022.

[8] M. Ulčar and M. Robnik-Šikonja. Sloberta: Slovene monolingual large pretrained masked language model. *Proceedings of Data Mining and Data Warehousing, SiKDD*, 2021.