Universiteit Utrecht

Comparing Text Representations: In Search Of Caring Communities

Master Thesis, Applied Data Science

Kalee Said
Student number: 6721664

Supervisors: Dr. Dong Nguyen & Marielle Zondervan
Second examiner: Dr. Pablo Mosteiro Romero

July 2024

Abstract

Research about caring communities in the Netherlands has become of greater importance as caring communities are a growing movement. By making use of automation and machine learning techniques on Chamber of Commerce data, caring communities can be found at a faster rate. This data contains textual data that can be represented in several ways. Therefore, this thesis aims to determine which text representation technique yields the highest classification performance for identifying 'caring communities' among registered organizations in the Dutch Chamber of Commerce. To answer this question an experiment was performed where six different text representation techniques were used: Word2Vec, TF-IDF, LDA topics, BoW, BERT, and BERTopic. These representations were combined with the Random Forest algorithm to classify instances as either a caring community or not. Two test datasets were used, one dating from 2022 and the other from 2023. Out of all representations, the classification model utilizing Word2Vec reached the highest recall, F1-score, and AUC value. Despite achieving the highest scores, it reached a precision of 0.46 for both datasets indicating a high rate of false positives among identified caring communities. Additionally, it reached a recall of around 0.5 on both test datasets, indicating it does not perform well at finding the existing caring communities in the data. The results could be explained due to the training data not being representative enough of the organizations and their activities as well as due to the class imbalance that was present in the data.

Table of contents

1 Introduction

Caring communities are resident-driven groups that manage neighborhood initiatives. These initiatives provide services, products, and activities that address specific community needs in areas like welfare, healthcare, housing, and participation. They are designed to be by and for those living in the same neighborhood. Examples include providing meeting spaces in community houses or resident-led activities like cooking classes or communal sports (van Zoest et al., 2021).

Research about caring communities in The Netherlands has become of greater importance as caring communities are a growing movement, with initiatives emerging across the country (Nederland Zorg Voor Elkaar et al., 2019). There is a need for a better understanding of the nature, scale, and value of the caring community movement. More research into caring communities can provide insights into the different types of initiatives, their activities, and their potential impact on communities. Expanding research on the topic will also enable local and national policymakers to make more informed decisions about how to support and facilitate the growth of caring communities. Furthermore, it aids in identifying the enabling factors and barriers to these initiatives. Finally, research can help in knowledge sharing and collaboration among initiatives, allowing them to learn from each other's experiences and best practices.

Former research on caring communities has explored several aspects of the movement in The Netherlands. One of the analyses was a survey conducted by knowledge organizations Nederland Zorgt Voor Elkaar, Vilans, and Movisie. This online survey aimed to define the scope, background, goals, and needs of caring communities, reaching 322 respondents (van Zoest et al., 2021). In 2022 Vilans, a knowledge organization focused on healthcare, conducted an internal analysis on the size of the caring community movement. This analysis aimed to identify the optimal method to identify caring communities within Chamber of Commerce data by making use of automation (Vilans, 2022). By attempting to automate this task, the caring communities can be found at a faster rate. Automation can manage large volumes of data more quickly compared to using human evaluators. It also ensures that the criteria for identifying caring communities are applied consistently across all instances of the data. In contrast to having multiple human evaluators with differing standards during the labeling process. The reason that Vilans focused on the Chamber of Commerce data was because of the survey report previously mentioned. This report indicated that the majority of the surveyed caring communities had registered themselves at the Chamber of Commerce. The Chamber of Commerce data was therefore used to potentially find more than those already identified from the survey. The data contains both numerical and textual data, with the textual data being business names and business descriptions organizations provide themselves. In Vilans' analysis, Latent Dirichlet Allocation (LDA) topics were used for text representation. LDA is a generative probabilistic model used for topic modeling. This representation was followed by a Random Forest classifier to classify the instances.

However, LDA text representation might not be the most effective method. Other research suggests that text representations such as TF-IDF and BERT can yield superior performance scores (Velthorst, 2019). Therefore, in this thesis, in collaboration with Vilans, I will expand their former analysis. Their approach will be further built upon by additionally testing five other text representations: Bag-of-Words, TF-IDF, BERT, Word2Vec, and BERTtopic. By adding these representations, I will investigate if the performance of the classification can be improved. Therefore, the research question of this thesis is:

"*Which text representation technique yields the highest classification performance for identifying 'caring communities' among registered organizations in the Dutch Chamber of Commerce?*"

As this project is also on behalf of Vilans, one of their questions regarding the analysis is the number of caring communities that can be identified once the best-performing text representation is found. Therefore, the data science question for this project is 'How many caring communities are present in the 2022 and 2023 Chamber of Commerce data?'

The structure of this thesis is as follows: Section 2 discusses previous research on caring communities as well as the use of text representations for classification purposes. Section 3 describes the data used in this project. Section 4 describes the methodology used for the experiment. Section 5 presents the results of the experimental set-up and classification task. Section 6 interprets the results and lastly, section 7 gives the conclusion.

2 Background

In this section, I will further define the concept of caring communities and expand on previous research in the field to give an insight into the current knowledge. I will also review the research on various text representations and their performance in classification. This review aims to identify the text representations that have previously proven effective, which then will be utilized in my experiment. In selecting the research articles, I have focused on short text and Dutch text. This is because the Chamber of Commerce's textual data is short in length, as further expanded on in section 3.4, and in addition to that written in the Dutch language.

2.1 Caring communities

This section expands on the definition and prior research on caring communities in the Netherlands. As briefly mentioned in the introduction, caring communities, as defined by van Zoest et al. (2020), are resident groups that independently manage their neighborhood initiatives. These initiatives provide services, products, and activities addressing community needs regarding welfare, healthcare, and housing. Other needs include participation, poverty reduction, and integration. They operate on the principle of reciprocity, with members of the communities giving and receiving support, and aim to be sustainable over the long term. This definition excludes private projects without a community function, social enterprises dependent on an entrepreneur, and community houses focused solely on rental or catering services.

Prior research about caring communities has led to several insights in the field. The first attempt to find the size of the caring community movement in the Netherlands was through a survey conducted by Nederland Zorgt Voor Elkaar, Vilans, and Movisie. This survey aimed to define the scope, background, goals, and needs of caring communities. Data was collected using an online survey covering general characteristics and detailed questions about collaboration, financing, development, and learning. A total of 323 self-identified caring community initiatives participated The survey identified several key findings, such as nearly half of the initiatives needing additional funding, one in twenty experiencing a continuous shortage, and one in five facing a volunteer shortage. The survey also explores subtopics such as the potential impact, effects, and financing of caring communities. (van Zoest, et al., 2021).

Vilans (2022) conducted further analysis to develop a classifier for identifying caring communities within Chamber of Commerce data. This analysis focused on two aspects: exploring the number of caring communities that are identifiable with the classifier and the representativeness of these identified instances. Latent Dirichlet Allocation (LDA) topics were used for text representation, and a Random Forest classifier was applied. About 71% of the instances identified as caring communities were correctly classified. Despite this achievement, several areas require further exploration and improvement in future work. Such as if the chosen text representation was the optimal approach as well as the choice of classifier. Additionally, a more precise estimation of the total number of caring communities is necessary. The current estimations made by the classifier, were validated using a list containing caring communities which was created by NZVE which was used to verify if the found organizations by the classifier were caring communities. This list was problematic in usage as it contained inconsistencies and was difficult to automatically process due to noise in the data. Additionally, the caring communities in the list were self-reported through a survey and might not

comply with the definition of caring communities. By improving data quality and the validation process, more known initiatives are likely to be discovered within the Chamber of Commerce data.

2.2 Text representations in classification of short text documents

In my experiment, I will compare various text representations to identify caring communities based on the textual data provided by the Chamber of Commerce. The textual data included are typically short documents as further expanded upon in section 3.4 of the data section. The documents average from 15 to 17 words per dataset and form the basis of the classification task. Given the importance of choosing appropriate the text representations for short documents, several studies have investigated different methods.

Shao et al. (2018) compared various text representation methods for classifying short texts. The researchers evaluated Word2Vec, a word embedding method; Doc2Vec, a document embedding method; and Bag-of-Words (BoW), which generates word frequency vectors. The BoW representation was used in both unigram and bigram forms. The experiment focused on classifying documents from electronic medical records into various classes of medical use of patients. The documents in the dataset averaged around 60 words in length. A Support Vector Machine classifier was used for the task. Results showed that Word2Vec features outperformed BoW unigram features. However, when BoW bigrams were compared with Word2Vec they yielded comparable results. Doc2Vec features exhibited the poorest performance among all representations.

Other text representations are compared in the study by Singla et al. (2020). The researchers compared TF-IDF, which is a weighted word importance measure, and the word embedding methods GloVe and Word2Vec. They paired these with six different classifiers and compared these methods with a BERT model which is a transformer-based model used for text classification. The classification task was a sentiment classification on a dataset consisting of tweets. BERT reached an accuracy of 0.40 which was the highest among all methods. Word2vec and Glove reached comparable results, with an accuracy of 0.35. TF-IDF reached the lowest accuracy of 0.25 but also the highest recall score when it was used in combination with the RNN classifier.

Lastly, Sayed et al. (2024) conducted a study to evaluate the effectiveness of using topics by BERTopic as features for classification, comparing its performance with GloVe embeddings. BERTopic is a topic modeling technique that uses BERT embeddings to cluster and summarize textual data. For the classification, both RoBERTa and SetFit were used as classifiers. The experiment focused on news classification tasks using three datasets with text lengths varying from short abstracts of around 50 words to longer texts of approximately 500 words. Results indicated that the BERTopic representation led to the highest accuracy across all datasets, with an accuracy of 0.85 for short texts and 0.90 for longer texts. The Glove embedding reached an accuracy of 0.83 on short text data and 0.89 on long text. Despite the small difference in accuracies, the researchers deemed BERTopic preferable as it improved accuracy across all datasets with both the BERT and SetFit classifiers.

2.3 Text representations in classification of Dutch text

The Chamber of Commerce data consists of Dutch text. However previous literature mentioned in section 2.2 focuses solely on English text, Therefore, exploring the performance of different text representations on Dutch text for classification tasks is a valuable step.

The research by Velthorst (2019) investigated the effectiveness of various text representation methods for predicting housing market trends using Dutch social media data, specifically focusing on tweets. They compare three main approaches for text representation: BoW, TF-IDF, and LDA topics. For the classification, they used Naive Bayes, Logistic Regression, and Support Vector Machine algorithms to predict the upward or downward trend of housing prices. The findings demonstrate that TF-IDF, which considers both word frequency and rarity, outperforms the BoW approach. The LDA topic modeling resulted in lower accuracy compared to both BoW and TF-IDF.

Another study on Dutch tweets by Reusens et al. (2022) compared several text representations: Word2Vec, TF-IDF, and FastText. The researchers used a dataset containing 47,096 tweets which they used to classify each tweet's sentiment correctly. Logistic Regression, XGBoost, Random Forest, Naive Bayes, and the RobBERTa model were employed for classification. The researchers found that TF-IDF consistently achieved the best performance across all models for most metrics. Other models showed more variation. Only Word2Vec outperformed TF-IDF on accuracy when used with the XGBoost algorithm. FastText performed slightly worse than Word2Vec, though results remained close.

2.4 Topic modelling on Dutch text

De Groot et al. (2023) investigated the performance of topic modeling algorithms for analyzing short texts from various domains. They focused specifically on student course evaluation texts from different faculties. Two approaches were compared: LDA and BERTopic. The models were evaluated based on topic coherence, which is the relatedness of words within a topic, and diversity, the variety of topics discovered. Their results demonstrate that BERTopic outperforms LDA. Notably, BERTopic successfully achieved a topic coherence of 0.091 and a topic diversity of 0.880. These scores surpassed those of LDA, which obtained coherence and diversity scores of 0.031 and 0.718, respectively.

2.5 Conclusions

Based on this literature review I selected the representations that will be used for the experiment. The review suggests that both Word2Vec and BoW bigrams perform well on short text (Shao et al., 2018) as well as BERT when comparing it to other representations (Singla & Kumar,2020). Since the data is in Dutch, TF-IDF, known for its success with Dutch text is utilized (Velthorst, 2019; Reusens et al., 2022). And between the topic modeling approaches BERTopic performs better on Dutch text compared to LDA (de Groot et al., 2023). The LDA approach will still be used as it is used in the original approach of the analysis by Vilans and my task is to extend their analysis.

3 Data

3.1 Context and source

In this experiment, I used four distinct datasets that originated from the Dutch Chamber of Commerce. These datasets include records of organizations that had registered with the Chamber of Commerce, providing all information required during the application process.

The first two datasets originated from the 2022 and 2023 Chamber of Commerce data. I have utilized these two datasets as the test data because of the eventual goal of the task that my project is aiding in, which is determining the number of caring communities in this specific 2022 and 2023 data. By ensuring that a model has not priorly trained on these records a fair evaluation can be made as otherwise there would be an overlap between training and test data.

The third dataset was used as training data and consists of organizations that self-identified as caring communities and were identified through the survey conducted by Nederland Zorgt Voor Elkaar, Vilans, and Movisie as previously mentioned in section 2.1 of the literature. With this identification, Vilans was able to request the records from the Chamber of Commerce of the identified organizations that had also registered themselves there. The records were registered under 45 different business activity codes. In the Netherlands, each company receives an SBI-code (Standaard Bedrijfsidentificatie), and they are used by the Chamber of Commerce to classify companies according to their business activities (CBS, 2024). The team at Vilans manually assessed these SBI-codes to determine if they were logically indicative of caring communities. This analysis resulted in the identification of approximately 12 SBI-codes that corresponded with the definition of a caring community. These SBI-codes represented about three-quarters of the initiatives requested by Vilans from the Chamber of Commerce. Examples of the business activities that these 12 SBI codes cover include ‘local welfare work’ and ‘social work’. The identification of these 12 codes was important as Vilans bought the records for the other three datasets based on these codes.

The fourth dataset used in this study consisted of Chamber of Commerce data from 2022 and was a subset of the 2022 test data. To ensure no overlap between training and test data, the records from this training data that occurred in the 2022 test data were deleted. The data was manually labeled by experts in the field of caring communities at Vilans as either a caring community or not. The dataset was used as training data due to having the property of already being pre-labeled. Subsequently, this dataset was combined with the dataset originating from the survey. This combination was logical because both datasets have a pre-assigned label. The survey dataset contains organizations that self-identified as caring communities, and therefore by their inclusion, they inherently possess the caring community label.

3.2 Manual labeling of test data

To be able to assess the quality of the classifications on the test data, a fellow student and I manually labeled 8.9% of the 2022 test data and 7.4% of the 2023 test data. This labeling process was conducted under the supervision of an expert on caring communities at Vilans. An initial meeting was held to provide guidance and explain the methodology. The label for each organization was decided based on the business description, the organization's name, and its alignment with the definition of a caring community. If the information was unclear, we conducted an online search to gather more details about the initiative. This was done for all the records that were potentially indicative of having the caring community label. Only records that were certainly not caring communities such as businesses or professionals offering paid services were immediately classified as non-caring communities without further internet searches. Any cases with labeling uncertainty after internet searches were referred to the expert for final classification.

3.3 Features

Each of the four datasets contained a varying number of features. The dataset originating from the survey selection and the manually labeled dataset both had 57 features. The test data from 2022 had 56 features, while the dataset from 2023 had significantly fewer, with only 38 features in total. This reduction occurred because Vilans determined that certain features, such as phone numbers and applicant names, were not useful for the analysis. While there was significant overlap in features across the datasets, some were unique to each specific dataset. Appendix A, Table 1 provides an overview of the features in the datasets, including their definitions and in which dataset(s) each attribute occurs.

Although the initial datasets contained a varying number of features, only eight features were ultimately chosen for the analysis. These include the business name ('HN45', 'H_NAAM_VOL'), the province where the organization is located ('PROV'), the legal form of the organization ('RECHTSVORM'), the number of employees ('W_P_TOTAAL'), the business identification code ('SBI_CODE'), the registration date ('INSCHR_DAT'), and the provided business description (‘Bedrijfsomschrijving’). In section 4.1.1 of the Methods, the reasoning as to why these eight features were chosen is elaborated on. Additionally, the data includes a label column, where a binary class label is assigned. Here a ‘0’ represents a non-caring community and a ‘1’ a caring community. ‘Non-caring communities’ in this case meaning any organization that is not a caring community. Table 1 provides an example of two instances within the dataset; these are fictional for anonymization purposes.

HN45	Prov	SBI_Code	Rechtsvorm	Inschr_Dat	W_P_Totaal	H_Naam_Vol	Bedrijfsomschrijving	Label
Stichting De Haan	A	88102	74	20130324	2	Stichting De Haan	Aanbieden van dagbesteding.	0
Het Buurt Huis	G	88102	61	20140231	0	Het Buurt Huis te Uden	Aanmoedigen van lokale welzijn.	1

Table 1: Two fictional examples to illustrate the dataset.

3.4 Data exploration

To begin with, an overview of the datasets and their respective sizes is provided in Table 2. Additionally, Table 2 in Appendix A presents an overview after merging the datasets that served as training data. An observation that is important to note when it comes to the textual data is that when combined, the maximum number of tokens per document ranges between 105 and 135 tokens. The average amount of tokens per dataset varies between 15 and 17 words. These are the ranges over the three datasets used for the training and testing. Overall, it can be observed that these documents are of short length.

Regarding the outcome label is evident that the label distribution is disproportionate, with the majority of records identified as belonging to a non-caring community as illustrated in Table 2 of the main text. Furthermore, the various features have different relationships with the instances identified as caring communities as well as varying distributions. Appendix B displays a distribution analysis of the features and the output label in Figures 1 to 5, revealing several observations. Firstly, from Figure 1 it can be observed that the legal forms most likely to represent a caring community are code 71 (association), 61 (cooperation), and 74 (foundation). Secondly, in Figure 2 it can be observed that most registrations were after the late 2000's with caring communities having peaks in registrations in both the years 1989-1990 and 2010-2011. In Figure 3 there appears to be an inverse relationship between the number of employees at an organization and the likelihood of it being a caring community. The province where the most registered caring communities are located is Flevoland with a number of 46, which can be observed in Figure 4. And lastly, the SBI code distribution in Figure 5 shows that the code with the highest likelihood of representing a caring community is the code corresponding to 'local wellbeing'.

Dataset	**Description**	**Role**	**Size**	**Records with the caring community label (%)**
Test data 2022	Data from 2022, requested through the Chamber of Commerce based on their SBI code.	Test data 2022	8035 rows	8.9% (out of the 350 hand labelled records)
Test data 2023	Data from 2023, requested through the Chamber of Commerce based on their SBI code.	Test data 2023	11285 rows	7.4% (out of the 350 hand labelled records)
Training data (1)	Data requested through the Chamber of Commerce based on their SBI code. These were manually labelled.	Training data	500 rows	11.3%
Training data (2)	Data selected based on their ID number in the chamber of commerce, identified through a survey.	Training data	448 rows	100%

Table 2: The four datasets, their original sizes as well as their label distribution

3.5 Ethical & legal considerations

The data was acquired through proper channels, with Vilans purchasing it from the Chamber of Commerce, ensuring compliance with legal requirements. To maintain data privacy, personal information such as addresses, names, and emails were deleted. Ethical and legal standards were upheld by overseeing the data responsibly and therefore withholding it from public disclosure and unauthorized sharing.

4 Method

In this experiment, my objective is to compare different text representations with respect to their performance in identifying caring communities within Chamber of Commerce data. This section presents the details of the pre-processing steps conducted, along with the technical workings of the various text representations and the classifier used. Additionally, the performance measures intended to evaluate the text representations are outlined in section 4.5. The code used to perform all the steps is provided in my GitHub repository[1]. The programming language used in this experiment is Python.

4.1 Pre-processing

4.1.1 De-duplication

As previously mentioned, the data initially consisted of four separate datasets. I first merged the two training datasets. This resulted in three datasets, the two test sets and the training dataset. All three datasets underwent the same pre-processing and data-cleaning procedures. First, I removed the duplicates from the training data that resulted from merging these two different sources. The duplicates were identified and removed based on the unique number assigned by the Chamber of Commerce to each branch of a company or organization. In the training data, This resulted in the removal of two duplicates. No duplicaten were found in the 2023 and 2023 datasets. Furthermore, an expert from Vilans had identified a group of initiatives with unique identifiers that were all part of the same organization This insight was based on prior knowledge of working with the data and reduced the training data by 334 records. This same problem did not occur in the test datasets and therefore these two datasets had no records reduced based on this method.

4.1.2 Feature selection

The following step was to delete features not used for analysis. For all three datasets, I reduced the number of features to eight together with the label column. I selected these features based on their potential to indicate the presence of a caring community. Firstly, the business name (HN45, H_NAAM_VOL) plays a role in the analysis because the language used can sometimes be in line with the definition or activities of caring communities. Words like "foundation" or "community house" might be present in the names of caring communities. Similarly, the province in which the organization is located (PROV) is included because of the idea that some areas might have a higher concentration of these organizations. This became clear from the data exploration in section 3.4 where there is a stark difference to be found in frequency of caring communities over the provinces. The legal form of the organization (RECHTSVORM) is another factor considered. This information, provided as a code (for example 62 for cooperation or 74 for foundation), can be indicative of the organization's activities, potentially aligning with those of caring communities. The number of employees (W_P_TOTAAL) is also a relevant factor. Caring communities are typically small-scale and often rely on volunteers, this became clear from the survey study earlier mentioned in section 2.1 of the literature. Out of the caring communities that were questioned about 88% of the initiatives declared to rely on volunteers. And only 14% claimed to have employed workers, therefore, a lower employee count might be in line with

[1] https://github.com/ksaid3/text_representationsZZG

caring communities (van Zoest, et al., 2021). As explained earlier in section 3.1 about the data, the business identification code (SBI_CODE) also plays a role. This code classifies the organization's activities, which can be aligned with the work that caring communities deliver. The registration date (INSCHR_DAT) is included because there might be variations in the number of caring communities initiated across different years. Finally, the business description (Bedrijfsomschrijving) provided by organizations during registration, can potentially describe activities similar to those delivered by caring communities.

In contrast, there were also reasons as to why some features were deleted. In Table 1 of Appendix C, the features that were dropped and the reason why they were dropped are displayed. The reasons for this are varying, firstly based on intuitive knowledge alone it was clear that some features would not be as useful for prediction as others. With examples being the applicant's phone number, name, or email. Secondly, some features would also not be feasible to use as they would result in a large number of features when one-hot encoding, these are features such as the street name or zip code. Furthermore, many of the features in the dataset were redundant representations of one another, examples are 'HN1X30', 'HN2X2X30', and 'HN1X2X30', which are all different lengths of characters given for the business name. Lastly, the 2023 dataset also contained some extra features that did not exist in the other datasets and were therefore consequently removed.

4.1.3 Data cleaning

All the noise in labels or features used in the analysis was deleted, these were values with characters that were not valid to use as a value such as question marks or other characters. This resulted in deleting three records in the training data and none in the two test datasets. Afterwards, the missing data was handled, I deleted records with missing data in any of the features used for analysis because accurate imputation wasn't possible. These features, such as SBI codes and provinces, were categorical and therefore difficult to impute. Features with impossible zero values such as the date of registration, were also deleted. This led to 165 records being deleted from the training data. The test data from 2022 was reduced by 15 records and the test data from 2023 by 18 records.

Lastly, as previously mentioned in section 3.1, to ensure no overlap between the training data and the test data, I deleted all records from the 2022 and 2023 data that also appeared in the training data. The test data from 2022 was reduced by 322 records and the test data from 2023 by 192 records. After these general pre-processing steps, the datasets were significantly reduced in size. Table 3 shows the difference between the original sizes and the sizes after pre-processing. I also converted the datatypes in the datasets to the appropriate datatypes. The registration date of the organization was converted from an integer to a datetime datatype, and other numerical features, such as the SBI-code which were floats but should have been integers, were also converted. Furthermore, all categorical data were one-hot encoded, this applied to the features that represent the province, the SBI-code, and the legal form of the organization. Additionally, the registration data column was split into separate year, month, and day features.

Dataset	Initial size	Size after pre-processing
Training data	949 rows	446 rows
Test data 2022	8035 rows	7686 rows
Test data 2023	11285 rows	11075 rows

Table 3: The sizes of the datasets after pre-processing the data

4.1.4 Preprocessing the textual data

To handle the pre-processing of the textual data, I merged the two features representing the business name and the business description feature into one column named 'combined text', as these three features constitute all of the free format textual data. After the text was combined, the original three features were subsequently deleted. Then to be able to compute the features, the textual data was tokenized, a process that breaks natural language text into smaller units. To facilitate this, I used the word tokenizer function from the NLTK's tokenize package. Then I removed the stop words which are commonly occurring words that typically lack significant meaning. NLTK's stop words corpus provided the stop words for the Dutch language and was used to remove these. URLs and numbers may not contribute to the semantics of the text and can introduce noise in the analysis; therefore, I removed them using regular expressions. Additionally, each token was normalized by converting it to lowercase, ensuring consistency in text processing and preventing different cases of the same word from being treated as separate entities. Lastly, I applied lemmatization, this technique reduces words to their base or dictionary form, also known as lemmas. It helps with normalizing words so that variations of the same word are treated the same. Each token was lemmatized using WordNet's lemmatizer from NLTK (Bird, Klein, & Loper, 2009).

4.2 Feature computation

As previously mentioned, in this thesis six text representations will be compared to each other to find the best-performing one for the task of identifying caring communities. These are: Bag-of-Words, Word2Vec embeddings, TF-IDF, BERT embeddings, LDA topic modeling topics, and BERT topics. These representations are specifically selected because it has been demonstrated in prior literature that these perform well in various classification tasks, as discussed in section 2. In this section the differences between these text representations are highlighted together with their theoretical workings. In table 4, a concrete example of the text representations is given based on the example sentence 'De vrijwilligers van de stichting helpen de buurtbewoners.' to portray how these representations look once applied on the text.

4.2.1 Bag of words

In the Bag-of-Words (BoW) representation, a document is represented as a collection of words, ignoring the grammar and word order of the original text. The focus is on word occurrence and frequency. It creates a sparse matrix where each row corresponds to a document, and each column corresponds to a unique word occurring in the entire corpus of documents. The value in each cell represents the frequency of the word's occurrence in the respective document. BoW has the advantage that it is simple to implement and interpret. However, it also suffers from the loss of semantic meaning and context, it treats each word in isolation, not considering their relationships within the text. This can

potentially be a drawback within certain classification tasks as without capturing semantic meaning, synonyms and related words are not recognized, and neither is the context that the word appears in. Therefore, in my experiment, BoW bigrams are used, as mentioned in section 2.2, bigrams were able to outperform BoW unigrams in classification tasks. Bigrams capture the word order by constructing features based on two adjacent words instead of individual words. In this fashion, the context of the words can be taken into account.

4.2.2 Word2vec

Word2Vec is an embedding technique that in contrast to BoW tries to capture the semantic meanings of words. It represents words in a multi-dimensional vector space, where words with similar meanings or contexts are closer in the vector space. The technique was developed by Mikolov et al. (2013) and works with the idea of distributional semantics. This assumes that words appearing in similar contexts tend to have similar meanings. Word2Vec involves training neural networks on text data to create the representations. There are two main architectures for Word2Vec: Continuous Bag of Words (CBOW) and Skip-gram. CBOW predicts the target word given its context words, while Skip-gram predicts context words given a target word. The resulting word embeddings encode semantic relationships between words. While Word2Vec has the benefit of having the ability to capture semantic information, it has limitations. The paper by Di Gennaro et al. (2021) revealed that Word2Vec might not perform well when it comes to learning syntactic relationships. The performance might also be dependent on the window size, referring to the number of context words considered during training, specifically claiming that larger window sizes lead to better results. For my implementation, I used the Word2Vec model developed by Coosto. This is a model trained on Dutch textual data originating from a large collection of new articles, blogs, forum posts, and social media messages. The model was trained using the CBOW technique (Coosto, 2018). In my implementation, the embedding vectors for all words in each document are averaged to obtain a single vector representation for the entire document. This is done to capture the overall semantic meaning of the text. If a word is not found in the Word2Vec model's vocabulary, a zero vector of the same dimensionality as the embedding vectors is used as a placeholder.

4.2.3 TF-IDF

The TF-IDF (Term Frequency-Inverse Document Frequency) representation is a measure to evaluate the importance of a word in a document relative to the collection of documents. It combines term frequency (TF), which is the frequency of a term in a document, with the inverse document frequency (IDF), the measure of how common or rare a term is across all documents in the corpus. TF-IDF scores are computed by multiplying TF and IDF values for each term, resulting in a score that determines its relevance. The higher the TF-IDF score the more important or relevant the term is. The TF-IDF has as advantage that it is easy to interpret as well as provides a more nuanced understanding of terms compared to simple word counts such as BoW. However, TF-IDF still is not able to capture semantic relationships between words as the Word2Vec representation does.

4.2.4 BERT embeddings

BERT (Bidirectional Encoder Representations from Transformers) embeddings can capture semantic meaning in text in a more advanced fashion compared to earlier techniques like BoW, Word2Vec, and TF-IDF. The technique was developed by Devlin et al. (2018) and is designed to pre-train deep

bidirectional representations. This bidirectional approach allows BERT to understand the context of a word based on all the words surrounding it in both directions, unlike earlier models which can only capture left-to-right or right-to-left context. The embeddings are created through a process of pre-training on a large corpus of texts using two tasks: masked language modeling (MLM) and next-sentence prediction (NSP). In MLM, some words in the input are randomly masked, and the model predicts these masked words, which helps it learn deep bidirectional representations. NSP, on the other hand, involves predicting whether a given pair of sentences appears consecutively in the text, helping the model in understanding sentence relationships. The primary advantage of BERT embeddings is their ability to capture nuanced semantic meanings and context. An important limitation is that BERT's pretraining process learns complex relations between words, but many of these relationships may be irrelevant for certain classification tasks. The irrelevant information may lead away from the essential keywords that are needed in the classification task (Gao, et al., 2021). The BERT model I used for this experiment was the RobBERT model developed by Delobelle et al. (2020). The RobBERT model was trained on a large corpus of Dutch text data, following a similar pre-training approach as BERT but with modifications tailored to the Dutch language. RobBERT training regime omits the NSP task and introduces dynamic masking during pre-training. This is different from MLM as dynamic masking specifically refers to changing the tokens that are masked at each training step. This approach enhances the MLM process by preventing the model from memorizing the positions of masked tokens and encourages learning more generalizable language patterns. In my implementation, the embeddings of the documents are obtained by using the embedding of the CLS token. The CLS token is a token added to the beginning of every input sequence in the BERT model. The embedding corresponding to this token is designed to capture the overall meaning and context of the entire sequence. By extracting and using the CLS token embedding, I obtained a dense representation of the entire document.

4.2.5 LDA topic modeling

Latent Dirichlet Allocation (LDA) is a probabilistic model used for discovering latent topics within a collection of documents. LDA assumes that each document is a mixture of a certain number of topics. Each topic captures a specific theme or concept and words within the document are allocated to topics based on their likelihood of belonging to each topic. By analyzing the word frequencies and co-occurrences, LDA assigns probabilities to each word belonging to a particular topic. Through iterative training, LDA learns the topic-word and document-topic distributions, allowing insights into the main topics present in the corpus. However, LDA may not capture finer-grained semantic relationships between words like Word2Vec or BERT does. Additionally, the performance of LDA can be influenced by factors such as the number of topics chosen and the quality of the input data (Blei, Ng, & Jordan, 2003). To determine the number of topics needed, I made use of the coherence score measure which is a metric used to evaluate the quality and interpretability of topics (Newman, Lau, Grieser, & Baldwin, 2010). It measures the semantic similarity between high-scoring words within each topic, ensuring that the words are related, and the topic is meaningful. Higher coherence scores indicate more meaningful and coherent topics aligning with a better topic quality. In my experiment, I used the C_V coherence measure, which specifically measures topic coherence by combining normalized pointwise mutual information (NPMI) with cosine similarity. This coherence measure was chosen as it has been shown to correlate with human judgment (Röder, Both, & Hinneburg, 2015).

4.2.6 BERTopic

BERTopic is a method designed to extract topics from text data. The method was designed by Grootendorst (2022) and operates through three main stages. Initially, documents are converted into BERT embeddings using a pre-trained language model. Then dimensionality reduction techniques such as UMAP are applied. This step ensures that important features are preserved while reducing computational complexity. Following this dimensionality reduction, documents are clustered using HDBSCAN, a clustering algorithm that handles varying densities within the data. Finally, topic representations are derived from these clusters using a modified TF-IDF approach, allowing for the identification of words associated with each topic. As previously mentioned, BERTopic uses a pre-trained language model to embed documents into dense vector representations, capturing semantic meaning more effectively than LDA's bag-of-words approach. This leads to more nuanced and contextually considerate topic representations. However, a limitation highlighted in the paper by Grootendorst (2022) is that BERTopic assumes single topic documents and overlooks the possibility of multi-topic documents which in turn affects the accuracy of the topic representations. To determine the appropriate number of topics I used the C_V coherence measure, as it was used for the LDA topic modeling as well.

Text Representation	**Example Output**	**Explanation**
BoW	{('de', 'vrijwilligers'): 1, ('vrijwilligers', 'van'): 1, ('van', 'stichting'): 1, ('stichting', 'helpen'): 1, ('helpen', 'buurtbewoners'): 1, ('de', 'de'): 1}	Represents the sentence as a dictionary where each bigram (pair of consecutive words) is a key, and its frequency is the value.
Word2Vec	[0.23, -0.45, 0.12,0.18, 0.56, -0.32, 0.29]	Words are represented as vectors in a high-dimensional space, capturing semantic relationships. Each word in a document represents a vector. Each document in this experiment takes an average of all vectors to represent the document
TF-IDF	{de: 0.023, vrijwilligers: 0.178, van: 0.139, stichting: 0.240, helpen:0.172, buurtbewoners: 1.456}	TF-IDF builds on BoW by weighing terms based on their frequency in a document and their inverse frequency across all documents. The output is a vector with TF-IDF scores for each term.
BERT Embeddings	[0.78, 0.12, 0.56, 0.23, -0.18, 0.45, 0.32, 0.29]	BERT assigns a single vector to the entire sentence, capturing contextual meaning and relationships between words. Similar to Word2Vec, the vector will be a lengthy list of numbers.
LDA topic modelling	Topic 1: (0.7), Topic 2: (0.2) Topic 3: (0.1)	Identifies latent topics within a corpus of documents. In this case, the sentence might be assigned to topic 1, with a higher probability (0.7) than topic 2 (0.2)
BERTopic	Topic 1: (0.7), Topic 2: (0.2) Topic 3: (0.1)	BERTopic provides a distribution of topics with their corresponding probabilities, indicating the document's relevance to each topic.

Table 4: Examples of the text representations based on the example sentence ‘De vrijwilligers van de stichting helpen de buurtbewoners.’

4.3 Classification

The classifier that will be used for the task of classification is the Random Forest algorithm. This algorithm is composed of multiple individual decision trees functioning as an ensemble. Each tree is constructed from different bootstrap samples of the training data, meaning that each tree uses a randomly selected data sample from the same dataset. The effectiveness of the algorithm increases when the individual trees are uncorrelated. Which is why the training data is bootstrapped as well as additional randomness being introduced through feature selection. With this last method each node only considers a subsample of all features to decide to split on. When all individual trees have their prediction for a class, the class that occurs the most will be chosen as the final prediction class (Liaw & Wiener, 2002). The advantage of the algorithm is that it is not prone to overfitting. Another reason to choose this algorithm is the fact that it can handle high-dimensional data well and due to the text representations resulting in large data frames this is beneficial. The classifier also handles continuous variables well when confronted with high-dimensional data. This is an important aspect as most of the features used to build the text representations are continuous. In an experiment by Chen et al. (2020), three popular datasets with a high number of variables were used, employing Random Forest as the primary algorithm for feature selection and the classification task. The researchers found that Random Forest's approach of selecting random subsets of variables for each tree and averaging their importance across multiple trees enables it to effectively handle continuous variables. Lastly, a great benefit of Random Forest is that the feature importance scores can also be calculated and make the classification task more interpretable.

4.4 Evaluation metrics

The evaluation of the text representations' performance in classification was measured using the precision, recall, and the F1-score. The decision to exclude accuracy from the assessment was due to the dataset's imbalance, notably the smaller amount of the caring community class compared to the non-caring community class. Using accuracy as a measure would therefore not accurately reflect the performance of the classification. However, to get a general idea of the total amount of instances that are classified correctly, accuracy is still displayed in the results. The three measures that are used to assess the performances are defined as follows. Precision is the amount of correctly classified instances of caring communities divided by all instances that were assigned to that class by the classifier. The recall is the amount of correctly classified instances of caring communities divided by the number of instances in total belonging to the caring communities.

Furthermore, the F1-score, a harmonic mean of precision and recall, was computed as well. Finally, the ROC curves for all the text representations are plotted. A ROC curve is constructed by plotting the true positive rate against the false positive rate. The true positive rate is the proportion of instances that were correctly predicted to be positive out of all positive classes. Similarly, the false positive rate is the proportion of classes that are incorrectly predicted to be positive out of all negative classes. Each ROC curve also has a corresponding AUC (area under the curve) which is the measure of the ability of a classifier to distinguish between classes and is used as a summary of the ROC curve. The higher the AUC value, the better the classification performance of the model. The AUC values can vary between 0 and 1 (Powers, 2020)

5 Results

In this section, I will present the results of the set-up and the classification. The set-up consists of the test and train split of the data, the number of topics used for the topic representations, and the set-up used for hyperparameter tuning of the classification model. Lastly, the results of the classification task for each of the text representations are presented.

5.1 Set up

5.1.1 Train-test split

After cleaning and pre-processing the data, I prepared it for the classification tasks. The six different text representation methods were applied on the data. And each of them was in turn used to train six separate classification models. To estimate the model's performance on unseen data before the final evaluation, I split the training data into training and test sets. The training set is used to train the model, while the testing set assesses the model's ability to generalize to new data.

In this split 75% of the data was used for training and the other 25% for testing. This split was chosen as it is a balanced approach that provides enough data for both training and testing. It avoids the extremes of having too little training data which can lead to underfitting or too little testing data which can lead to an unreliable evaluation. I used the train_test_split function from Sci-kit Learn for this *(Pedregosa, et al., 2011)*.

5.1.2 Topic modelling results

The number of topics used for each of the topic modeling representations was determined by making use of the coherence score as previously mentioned in section 4.2.5 of the Method chapter. The search range for the number of topics for LDA was between 1 and 50 topics. This search resulted in 11 topics being optimal according to the coherence score. The maximum coherence score reached here was 0.47. A higher topic coherence correlates with higher human interpretability of the topics. Generally, coherence scores in the range of 0.4 to 0.6 are considered sufficient, while scores above 0.6 are considered high performing (Röder, Both, & Hinneburg, 2015). Therefore, the LDA topics should be decently coherent. With the use of human judgment, it becomes clear that some topics are distinct such as topic 4 indicating care services and topic 6 being geared more towards material needs. In table 5 these two topics are displayed. However, there is also a substantial overlap in content to be found between all of the topics. With words like ‘stichting’ and ‘coöperatie’ appearing in nearly every topic. In Table 1 of Appendix D, I displayed all of the 11 topics and the first 15 words of each of the topics.

Topic	First 15 words
4	mensen, stichting, begeleiding, geven, beperking, bv, hulp, coaching, stellen, trainingen, maken, ondersteunen, financiele, maatschap, maatschappelijke
6	leden, cooperatie, ua, buurtcentrum, ten, overeenkomsten, behoeve, stoffelijke, behoeften, bedrijf, uitoefenen, einde, belangen, doet, uitoefent

Table 5: Two topics generated by the LDA topic model algorithm; in the table the first 15 words of each topic are displayed.

For BERTopic, the search range for the number of topics was between 5 and 50, the optimal number of topics according to the coherence score resulted in 5 topics. The method reached a higher score than LDA with a coherence score of 0.57. From human judgment, it becomes clear that the topics displayed here are more intuitive to understand than the LDA topics. The topics are also more distinct, for instance, topic 1 appears to fit into the category of healthcare and wellbeing whereas topic 4 seems to fit into the category of young people and youth centers. In table 6 these two topics are displayed. Furthermore, in Table 2 of Appendix D, all 5 topics are displayed alongside topic -1, which is a topic detected by BERTopic as noise in the data.

Topic	Top n words
1	stichting, zorg, welzijn, begeleiding, mensen, gebied, ondersteuning, ua, ondersteunen, ouderen
4	jongeren, jeugdland, jeugd, jongerencentrum, activiteiten, vereniging, jongerenwerk, bv, stichting, kader

Table 6: Two topics generated by the BERTopic algorithm.

5.1.3 Parameter tuning

Another aspect of the experimental setup involved tuning the Random Forest algorithm's parameters for each text representation. These parameters included the number of trees used for prediction, the maximum depth of the trees, the minimum number of samples required to split an internal node, and the minimum number of samples required at a leaf node. Tuning was performed to ensure optimal model performance, as these parameters significantly influence the classifier's effectiveness. The GridSearchCV function from Sci-kit Learn was used to facilitate this process. The function performs a search over specified parameter values. The exact range of tested values can be found in Table 7, the resulting optimal parameters for each classifier are bolded in the same table.

Text representation	Search range: Number of trees (n_estimators)	Search range: Maximum tree depth (max_depth)	Search range: Minimum samples to split a node on. (min_samples_split)	Search range: Minimum samples required at a leaf node. (min_samples_leaf)
BoW	[15,**50**, 100,150]	[10,20, 30,40,50,**60**,70]	[**2**, 5, 10]	[**1**, 2, 4,5]
Word2Vec	[15,50,100,**150**]	[10,**20**, 30,40,50,60,70]	[2, **5**, 10]	[1, **2**, 4,5]
TF-IDF	[15,50, **100**,150]	[10,20, 30,40,50,60,**70**]	[2, **5**, 10]	[**1**, 2, 4,5]
BERT Embeddings	[**15**, 50, 100,150]	[**10**,20, 30,40,50,60,70]	[2, 5, **10**]	[**1**, 2, 4,5]
LDA topic modelling	[15,50, **100**,150]	[**10**,20, 30,40,50,60,70]	[**2**, 5, 10]	[**1**, 2,4, 5]
BERTopic	[**15**,50, 100,150]	[10,**20**, 30,40,50,60,70]	[**2**, 5, 10]	[**1**, 2, 4,5]

Table 7: The ranges that were used in the grid-search to find the hyperparameters

For the grid search, I employed a five-fold cross-validation strategy to be used on the training data. The F1-score was chosen as the performance metric for the grid search because both precision and recall were initially lower than other metrics before tuning. The two measures are also essential in this context, as finding relevant initiatives is the main goal of this experiment. Therefore, finding each of the caring communities, the recall, and ensuring that this prediction is correct, the precision, is of foremost importance. Once the optimal parameters were found, the classifier was finally used to label the data and I calculated the performance on the 2022 and 2023 test data.

5.2 Classification

The results for each text representation together with the other non-textual features in the dataset can be found in Tables 8 and 9. The two tables represent the results for the dataset of 2022 and 2023, respectively. In the tables, the accuracy, precision, recall, and F1-score are displayed for each class.

The non-caring community class achieved higher scores than the caring community class.

A few noteworthy observations can be made based on the final results. All models achieved accuracies ranging from 0.83 to 0.91. However, their performance in the caring community's class is notably inferior across the other metrics. Precision for non-caring communities consistently ranged from 0.91 to 0.96, whereas for caring communities, it varied considerably from 0.16 to 0.55. Similarly, recall for non-caring communities was consistently high ranging from 0.95 to 0.99, contrasting with the caring

community's class, which ranged from 0.04 to 0.52. The F1-scores followed the same pattern, with non-caring communities scoring higher, ranging from 0.90 to 0.96, and caring communities scoring significantly lower, ranging from 0.06 to 0.48.

The highest-scoring text representation for the caring community's class is Word2vec.

Word2Vec achieved the highest recall and F1-score on both the 2022 and 2023 datasets. On the 2023 dataset it obtained the highest precision score however, on the 2022 dataset TF-IDF reached the highest score. Figure 1 displays the F1-scores of the models, specifically those calculated for the caring community class. Since finding these communities is a central goal of the project, the F1-score holds particular importance. This metric is valuable because it penalizes models for missing true caring communities and misclassifying non-communities as such. From this figure, it can be observed that Word2Vec has the highest F1-score with the other text representations having a substantially lower score. Word2Vec also appears to be more consistent in performance across the two test datasets.

The lowest-scoring text representation for the caring community's class is BoW.

When looking more into the lowest performance amongst the text representation, I identified BERTopic on the 2023 data as the text representation achieving the lowest precision scores for the caring community's class, with a score of 0.16. This is followed by the LDA representation, which scored 0.21 on the 2022 dataset. BoW on the 2022 and 2023 data exhibited the lowest recall score of 0.04 while LDA achieved the second-lowest recall of 0.08 for the 2023 data. For the F1-score again BoW achieved the lowest scores, on the 2022 data achieved a score of 0.07 and on the 2023 data it achieved a score of 0.07.

Model		Accuracy	Precision	Recall	F1-score
BoW	Non-Caring	0.92	0.91	0.99	0.95
	Caring		0.33	0.04	0.06
TF-IDF	Non-Caring	0.91	0.93	0.98	0.95
	Caring		0.55	0.19	0.29
LDA	Non-Caring	0.88	0.92	0.95	0.94
	Caring		0.21	0.13	0.16
Word2Vec	Non-Caring	0.90	0.95	0.94	0.95
	Caring		0.46	0.52	0.48
BERT	Non-Caring	0.87	0.95	0.92	0.93
	Caring		0.37	0.48	0.42
BERTopic	Non-Caring	0.83	0.92	0.89	0.90
	Caring		0.16	0.23	0.19

Table 8: The performance measures of each text representation by class for the 2022 dataset.

Model		Accuracy	Precision	Recall	F1-score
BoW	Non-Caring	0.92	0.93	0.99	0.96
	Caring		0.25	0.04	0.07
TF-IDF	Non-Caring	0.92	0.93	0.98	0.96
	Caring		0.38	0.12	0.18
LDA	Non-Caring	0.91	0.92	0.98	0.95
	Caring		0.25	0.08	0.12
Word2vec	Non-Caring	0.92	0.96	0.95	0.96
	Caring		0.46	0.50	0.48
BERT	Non-Caring	0.89	0.95	0.94	0.94
	Caring		0.31	0.35	0.33
BERTopic	Non-Caring	0.89	0.93	0.95	0.94
	Caring		0.16	0.12	0.13

Table 9: The performance measures of each text representation by class for the 2023 dataset.

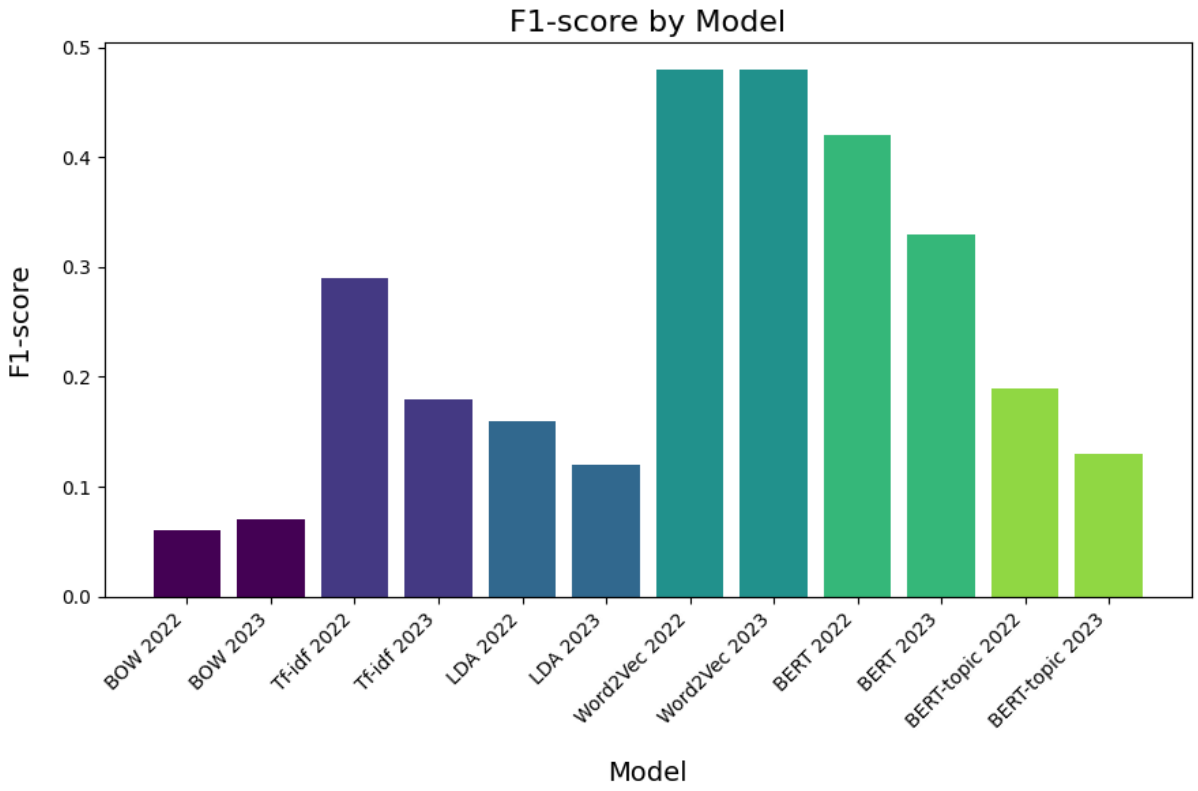

Figure 1: The F1-scores of each text representation for both 2022 and 2023 test datasets

Figures 2 and 3 depict the ROC-curves generated for each text representation for the 2022 and 2023 datasets, respectively. The AUC values ranged from 0.59 to 0.89, with Word2Vec achieving the highest score the 2023 dataset with an AUC of 0.89. The highest score for the 2022 dataset was also achieved by Word2Vec with 0.88. The LDA representation applied to the 2022 dataset exhibited the lowest AUC score of 0.59. On the 2023 dataset BERTopic achieved the lowest AUC score of 0.63.

Lastly, in the appendix D, table 3 the performances on the training data can be found. These are the results of the classifier for each text representation after tuning and finding the hyperparameters.

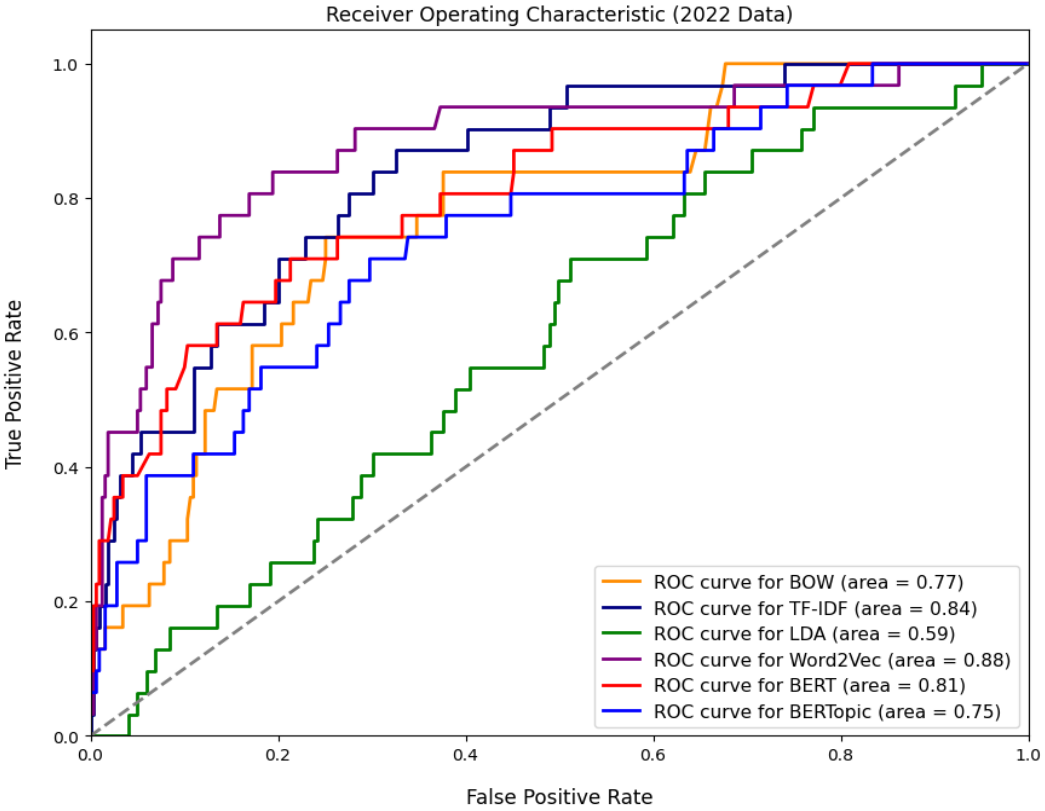

Figure 2: ROC-curves for the 2022 dataset

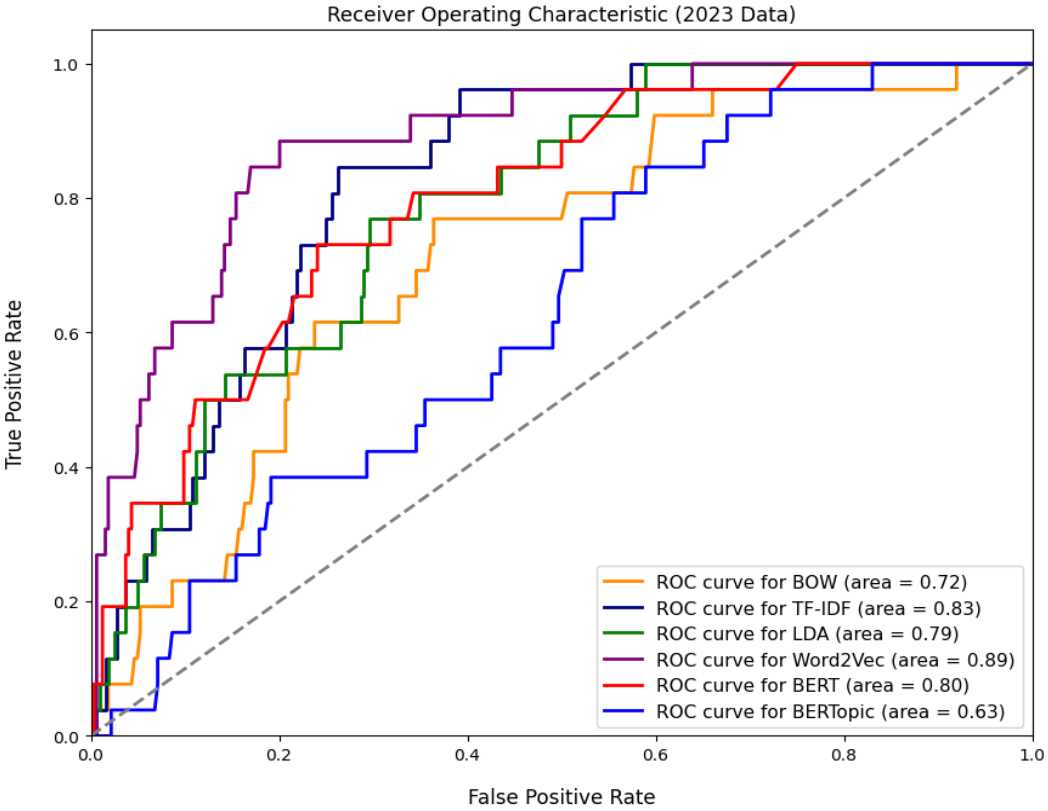

Figure 3: ROC-curves for the 2023 dataset

6 Discussion

The goal of this thesis was to determine which text representation technique yields the highest classification performance in identifying caring communities among registered organizations in Dutch Chamber of Commerce data. The findings indicate that the Word2Vec text representation achieves the highest classification performance in terms of recall, F1-score, and AUC among all six different representations. Only TF-IDF reached a higher precision of 0.55 versus the 0.46 precision score for Word2vec. These scores were with respect to the performance on the caring community class, while for the non-caring community class, all text representations yielded comparable results.

6.1 Performance

Despite reaching the highest overall scores, the results do not necessarily imply that the Word2Vec text representation is ideal for this specific classification task. When entirely focusing on the caring community class the following can be observed.

The precision score of the classification model that utilized Word2Vec was 0.46 in both datasets. The precision, in this case, is the fraction of actual caring communities among all of the instances identified as caring communities. The score of 0.46 suggests that over half of the organizations are incorrectly classified as caring communities. The recall of the caring community for the Word2Vec representation was 0.52 in the 2022 dataset and 0.50 for the 2023 dataset. The recall is the fraction of caring communities the classifier was able to retrieve among all existing caring community instances in the dataset. These scores indicate that the classifier is missing a substantial portion of actual caring communities, with approximately half not being identified. The classifier utilizing Word2Vec also yielded the highest F1-score among all text representations, yet this was merely 0.48 for both datasets. This is expected, as both precision and recall were around that range, and the F1-score balances these two measures. The F1-score of 0.48 suggests that the classifier will often fail to correctly identify caring communities and will frequently misclassify non-caring communities as caring ones. As mentioned in section 5.2 of the results, the performance on the non-caring communities' class was significantly better. Overall, there was a trend where the classifiers struggled to correctly classify the caring communities, exhibiting a bias towards the non-caring communities' class. Another notable discrepancy appeared between the performance scores of the two test datasets. The classifiers performed consistently better on the 2022 test dataset compared to the 2023 dataset. In all cases, the scores for the 2023 dataset were either equal to or lower than those for the 2022 dataset.

Lastly, the ROC curves and the AUC values were also calculated. Generally, for any model with an AUC score of 0.5 or below, it can be inferred that the classifier cannot distinguish between positive and negative class points (Powers, 2020). In this experiment, all models achieved AUC scores exceeding 0.5. Once again, the classification model utilizing Word2Vec achieved the highest score with an AUC value of 0.89. It is desirable to have a steep curve as that indicates that there is a high rate of true positives with a lower rate of false negatives. Judging by the curve shapes, Word2Vec has the most desirable outcome for both datasets. In the 2022 dataset, it is also noteworthy to point out that the classification model using LDA is closely aligned with the diagonal line, this line indicates an AUC of 0.5. The score of

0.59 suggests that the model has weaker discriminatory abilities than the models using other representations.

It is also important to note that the performance on the training data was higher than on the test data as can be observed in table 3 of Appendix D, suggesting that the models might be overfitting. These results show higher accuracy, precision, recall, and F1-score values compared to the test data. This indicates that the models perform well on data they have already seen but struggle to generalize to unseen data.

6.2 Error analysis

Based on the literature, the results of the highest-performing representations are not too surprising. In prior research on Dutch short-text classification, classifiers using Word2Vec representations yielded the highest results, with TF-IDF achieving the second-highest score, as seen in the experiment on Dutch short-text tweets by Reusens et al. (2022) discussed in section 2.3 of the literature. Both Word2Vec and TF-IDF also perform better than LDA topics and the BoW representation used for classification, as mentioned in the study by Velthorst (2019) in the same section. However, the classification model using BERTopic scored surprisingly low, considering it outperformed an embedding method as demonstrated by Sayed et al. (2024), and was able to reach a higher coherence score than LDA in this experiment. The performance of the BERT embedding representation was also unexpected. Its recall score came close to the score of Word2Vec and based on the F1-score it can overall be said the BERT was the second-best performing representation after Word2Vec. While BERT came close to the performance of Word2Vec, it was noted in section 2.2 to be likely to reach a higher precision and accuracy compared to TF-IDF and Word2vec (Singla & Kumar, 2020).

To understand why these representations did not perform as expected and why they performed poorly overall, I conducted an error analysis. I examined the instances that were misclassified, specifically the caring communities that were classified as non-caring communities, as the recall score was low across all models for this class. Based on this error analysis, I hypothesized what possibly caused this. There was a total of 43 different instances that were misclassified by all models with varying frequencies. An instance that was misclassified by all models was ‘Stichting DOGMA, a probable reason for this could be that the business description could easily be associated with healthcare services as it mentions care for disabled people and therefore would not meet the definition of a caring community. There were 9 other instances that were misclassified as non-caring communities by all classification models except for the model that used the Word2Vec representation. In an attempt to investigate if any of these had common patterns that were classifier-based, I generated all the feature importance plots, which can be found in Appendix E, figures 1 to 6. Feature importances provide insights into which features are most influential in the decision-making process of the Random Forest models. By analyzing these plots, I could identify which specific features were driving the misclassification. For instance, certain words, topics, or other features might have been strongly associated with the non-caring community class across different models, leading to the incorrect classification of caring communities. An important

observation from the plots is that while the Random Forest classification model uses the Word2Vec representation, and mostly uses the embeddings as a feature to make decisions, the other classification models make more use of the non-textual features to classify. One common feature in the top 20 across all other five models was 'rechtsvorm_74,' which is a characteristic defining the legal form of organizations. It is commonly associated with non-caring communities as 78% of the records in the test data belonged to this category. Going back to the 9 misclassified instances, 7 of these had the legal form of 'rechtsvorm_74', more commonly associated with non-caring communities and therefore a possible reason for this misclassification. The other two instances could have been more text representation-specific misclassifications as other patterns could not be detected in their characteristics. Both instances only had business descriptions of one word, not providing enough context about the initiative and therefore less textual data to create a classification upon.

Examining the 15 initiatives that the Word2Vec-based classification model misclassified reveals the following observations. Firstly, the model made most decisions based on embeddings instead of other non-textual features. Only 'SBI_OVERIG' appears in the top 20 feature importances, as it did for all models. This feature is convenient to split on by the Random Forest model as there are zero instances of caring communities having this SBI-code feature. As none of the misclassified instances by Word2Vec have it as their SBI-code, further analysis will be performed on the textual data only. The high feature importances suggest that semantic information is essential for the Random Forest classifier that used Word2Vec to make accurate predictions. The most likely reason that the classifier failed to correctly classify these might be due to the misclassified records containing terms or phrases that are not well-represented in the training data of the Word2Vec model. With this model, I allude to the Word2Vec model used to create the Word2Vec embeddings as previously mentioned in section 4.2.2. This model was trained on a large corpus of Dutch text data to learn the semantic relationships between words, and was in turn utilized to create the embeddings for the datasets in this experiment. The training corpus used might not have been suitable for the domain represented here or more likely not adequately trained to comprehend the uncommon terminologies used in the misclassified records. The misclassified instances, for example, contain uncommon spellings of words such as 't'dorpshuus,' phrases in the Frisian language such as 'oan e feart,' or simply had little textual data and consisted of the initiative's name and a maximum of one word explaining their business operations. Some instances also did not have names or business descriptions in line with caring communities but were instead classified as a caring community based on internet searches made during the manual labeling process. Most of these 15 instances were also misclassified by the classifiers making use of the BoW, TF-IDF, and LDA topic representations. They likely misclassified these for the same reason of having to handle uncommon phrases or too little textual data available.

Notably, the classification model that made use of BERTopic topics achieved the lowest precision on both test datasets despite the method reaching relatively high performing scores in the mentioned literature of section 2.2 and 2.4. A possible reason for this is that the optimal number of topics based on the coherence score was relatively small. By making use of only six topics it can be difficult to

express the complexity of the characteristics of the minority caring community class. The amount of topics chosen were potentially too little to accurately identify the caring community class.

Lastly, the datasets also had a significant class imbalance, where the minority caring community class was underrepresented compared to the majority non-caring community class. The caring communities class composed under 10% of all instances in both test datasets as previously mentioned in section 3.2. The model may underperform due to this limited data as it has few instances to learn to distinguish the minority class well. It might be the case that the few instances presented are not distinct enough to adequately differentiate them from the majority class. This imbalance is also a potential explanation as to why the 2023 test dataset had a consistently lower performance compared to the 2022 test dataset. The 2023 dataset had a smaller percentage of labeled caring communities and therefore fewer instances to be representative of that class.

6.3 Number of caring communities

While the Word2Vec representation yielded unsatisfactory results, as has already been discussed, it still generated the highest performance and was therefore used to estimate how many caring communities there are to be found in the 2022 and 2023 datasets. In the 2022 dataset, 1106 caring communities were identified, constituting 14,4% of the dataset. In the 2023 dataset, 1214 caring communities were identified, constituting 11,0% of the dataset. This however does not imply a decrease as the 2023 dataset is larger than the 2022 dataset. It is also important to keep in mind that these identified instances are likely not entirely accurate estimations judging by the performance scores observed in the model. The recall on the 2022 dataset was 0.52 and on the 2023 dataset, this was 0.50. Meaning that around half of the caring communities are not found. This does not necessarily mean that the number of caring communities is larger than the number found here by the classifier. The precision, the fraction of actual caring communities among all the found instances, indicates that most of the instances classified as a caring community are in fact not. With the precision score being 0.46 on both datasets most found instances are false positives. In summary, the precision and recall scores suggest that the actual number of accurately identified caring communities could be significantly lower than reported.

6.4 Limitations and future work

Several issues occurred in this experiment. One of the core problems was that the textual data in the Chamber of Commerce data was short in size or too ambiguous to create a clear classification. Therefore, during the manual labelling of the data, a fellow student and I had to resort to the internet to find more information on the initiatives in the dataset to make the definite decision. Because of this, many instances were labeled as caring communities with the help of implicit knowledge not displayed in the data. This may have been a potential reason for the misclassifications made. Furthermore, the test data used to measure the performance of the classifier was manually labeled, and due to the numerous amounts of records in both of the datasets, only a small portion could be labeled. This could have introduced skewed results to a certain extent as precision and recall scores can easily be influenced if there are few positive instances to be found. Lastly, when wanting to estimate how many caring communities there are, as defined in the data science question, it also needs to be considered that numerous communities have ceased to exist after the years 2022 and 2023. Through internet

searches it became clear that there were several cases of initiatives declaring on their website and/or social media pages that they are inactive. Therefore, the counts represented are not reliable both due to the performance of the model as well as the possibility of many organizations not being active anymore.

From the analysis, it became clear that to find an accurate estimate of the communities the process needs to be expanded. Future implementations of this project could attempt to use a different dataset or an attempt to enrich the current dataset by searching for additional features about the nature of the initiatives. Furthermore, the experiment could be extended by also addressing the classification task specifically. An analysis of multiple classifiers and testing them against each other instead of only making use of the Random Forest classifier could create valuable insights. And lastly, ensuring that there is a bigger pre-labeled dataset with more representation of caring communities can aid in creating a model that is better prepared to learn the characteristics of caring communities.

7 Conclusion

This thesis aimed to answer the research question: "Which text representation technique yields the highest classification performance for identifying 'caring communities' among registered organizations in the Dutch Chamber of Commerce?". To answer this question an experiment was performed where six different text representation techniques were used to identify caring communities in Dutch Chamber of Commerce data. The representations used were: Word2Vec, TF-IDF, LDA topics, BoW, BERT, and BERTopic. Out of these six, Word2Vec turned out to be the highest-performing representation. In combination with the Random Forest classification algorithm, it was able to reach the highest recall, F1-score, and AUC value.

Despite achieving the highest scores, the classification model utilizing Word2Vec reached a precision of 0.46 for both test datasets, indicating a high rate of false positives among identified caring communities. Additionally, the classification model showed a low recall of 0.50 to 0.52 for the two test datasets, suggesting it missed a substantial portion of actual caring communities. Not only the classification model utilizing Word2Vec struggle to find the caring community class, but all other models had this problem, showing a bias towards the non-caring community class when classifying.

Potential reasons for these scores varied, overfitting was observed as performance on training data exceeded that on test data, highlighting a struggle to generalize to new instances. The dataset also contained a class imbalance, with caring communities comprising less than 15% of the data. This left few instances to train the models and be a representative of the caring community class. Furthermore, reliance on implicit knowledge not known to the classifier when manually labeling the data possibly contributed to misclassifications. The Chamber of Commerce data was short in size and some cases too ambiguous to create a clear classification alone. These challenges highlight the need for a new or improved dataset that is enriched with more information about the initiatives and their activities. A dataset with a better balance between classes could also aid in bettering the performances. Furthermore, experimenting with a wider variety of classifiers could also help in achieving higher performances on the classification.

Bibliography

CBS. (2024, May). *Sbi-codes*. Retrieved from Business.gov.nl: https://business.gov.nl/running-your-business/business-management/administration/sbi-codes/

Al Sayed, M., Braşoveanu, A. M., Nixon, L. J., & Scharl, A. (2023, September 30). Unsupervised Topic Modeling with BERTopic for Coarse and Fine-Grained News Classification. *International Work-Conference on Artificial Neural Networks*, (pp. 162–174). Ponta Delgada, Portugal.

Bird, S., Klein, E., & Loper, E. (2009). *Natural language processing with Python: analyzing text with the natural language toolkit.* Sebastopol: O'Reilly Media.

Blei, D., Ng, A., & Jordan, M. (2003, January 3). Latent Dirichlet Allocation. *Journal of Machine Learning Research*, pp. 933-1022.

Chen, R. C., Dewi, C., Huang, S. W., & Caraka, R. E. (2020, July 23). Selecting critical features for data classification based on machine learning methods. *Journal of Big Data*.

Coosto. (2018, July 5). *Dutch Word Embeddings*. Retrieved from https://github.com/coosto/dutch-word-embeddings

De Groot, M., Aliannejadi, M., & Haas, M. R. (2022, December 16). Experiments on Generalizability of BERTopic on Multi-Domain Short Text. *arXiv:2212.08459*.

Delobelle, P., Winters, T., & Berendt, B. (2020, January 17). RobBERT: a Dutch RoBERTa-based Language Model. *arXiv:2001.06286*.

Devlin, J., Chang, M.-W., Lee, K., & Kristina, T. (2018, October 11). BERT: Pre-training of Deep Bidirectional Transformers for. *arXiv:1810.04805*.

Di Gennaro, G., Buonanno, A., & Palmieri Francesco, A. N. (2021, April 6). Considerations about learning Word2Vec. *The Journal of Supercomputing* , pp. 12320–12335.

Gao, S. O., Alawad, M., Young, M. T., Gounley, J. O., Schaefferkoetter, N., Yoon, H. J., . . . Coyle. (2021, September 1). Limitations of Transformers on Clinical Text Classification. *Journal of Biomedical and Health Informatics*.

Grootendorst, M. (2022, March 11). BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv:2203.05794*.

Liaw, A., & Wiener, M. (2002, December). Classification and regression by randomForest. *R News*, pp. 18-22.

Mikolov, T., Chen, K., Corrado, G., & Dean, J. (2013, January 16). Efficient Estimation of Word Representations in Vector Space. *arXiv*.

Nederland Zorgt Voor Elkaar, Vilans, Movisie. (2019, September). De Organiserende Burger: Leerprogramma en monitor.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010, June). Automatic evaluation of topic coherence. *In Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*.

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., . . . Prettenhofer, P. (2011). Scikit-learn: Machine Learning in Python. *Journal of Machine Learning Research*, pp. 2825-2830.

Powers, D. (2020). Evaluation: from precision, recall and F-measure to ROC, informedness, markedness and correlation. *arXiv preprint arXiv:2010.16061*.

Reusens, M., Reusens, M., Callens, M., vanden Broucke, S., & Baesens, B. (2022, October 18). Comparison of Different Modeling Techniques for Flemish Twitter Sentiment Analysis. *Analytics*, pp. 117-134.

Röder, M., Both, A., & Hinneburg, A. (2015, February 5). Exploring the Space of Topic Coherence Measures. *WSDM '15: Proceedings of the Eighth ACM International Conference on Web Search and Data Mining*, pp. 399-408.

Shao, Y., Taylor, S., Marshall, N., Morioka, C., & Zeng-Treitler, Q. (2018, December 10). Clinical text classification with word embedding features vs. bag-of-words features. *IEEE International Conference on Big Data* . Seattle, WA, USA.

Singla, S., & Kumar, V. (2020, November 19). Multi-Class Sentiment Classification using Machine Learning and Deep. *International Journal of Computer Sciences and Engineering*.

van Zoest, F., Stouthard, L., van Schaijk, R., Sok, K., van der Heijden, N., & Smelik, J. (2021). *Monitor Zorgzame Gemeenschappen.*

Velthorst, M. (2019, January). Predicting The Dutch Housing Market Trends Using Twitter. *6th Swiss Conference on Data Science (SDS).* Bern, Switzerland: IEEE.

Vilans. (2022, Oktober). Datagedreven monitor Zorgzame Gemeenschappen.

Wenying, D., ORCID, C. G., Shuang, Y., Nengcheng, C., & Lei, X. (2023, April 7). Applicability analysis and ensemble application of BERT with TF-IDF, TextRank, MMR, and LDA for topic classification based on flood-related VGI. *ISPRS International Journal of Geo-Information*, p. 240.

Appendix

Appendix A

Feature name	Description	2022 Data	2023 Data	Monitor Data	Label Data
RGL	Register letter	X	X	X	X
DOSSIER	File number	X	X	X	X
VGNUMMER	Establishment number	X	X	X	X
HN1X30	Trade name 1 x 30 positions	X	X	X	X
STRVA	Street /house number/addition of the establishment address	X	X	X	X
PCPLVA	Postal code and city of the establishment address	X	X	X	X
STRCA	Street/house number/addition of the correspondence address	X	X	X	X
PCPLCA	Postal code and city of the correspondence address	X	X	X	X
HN1X2X30	Trade name 1st line 2 x 30 positions	X	X	X	X
HN2X2X30	Trade name 2nd line 2 x 30 positions	X	X	X	X
HN45	Trade name 45 positions	X	X	X	X
PCVA_CIJF	Postal code of the establishment address		X		
PCVA_LTRS	Postal letters of the establishment address		X		
PCCA_CIJF	Postal code of the correspondence address		X		
PCCA_LTRS	Postal letters of the correspondence address		X		
BEHKN	Managing chamber number		X		
GEOKN	Geographical chamber number		X		
GEMK_VA	Municipality code of the establishment address	X	X	X	X
GEMK_CA	Municipality code of the correspondence address	X	X	X	X
GEMNAAM	Municipality name	X	X	X	X
PROV	Province	X	X	X	X
TEL_NRS	Telephone number	X		X	X
MOB_TEL_NR	Mobile phone number	X		X	X
FUNCTIE	Function	X		X	X
VOORLETTER	First letter	X		X	X
VOORVOEGSE	Prefix	X		X	X
ACHTERNAAM	Last name	X		X	X
SBI_CODE	Standard Business Classification code	X	X	X	X
SBI_OMSCHR	Standard Business Classification description	X	X	X	X
NEVENACT_1	Secondary activity code (1st)	X	X	X	X
NEVENACT_2	Secondary activity code (2nd)	X	X	X	X
HFD_N_VEST	Head/branch office indication	X		X	X
CD_EC_ACT	Indication of economic activity	X	X	X	X

KL_WP_TOT	Classes of total employees	X	X	X	X
KL_WP_FULL	Classes of full-time employees	X	X	X	X
PEILDAT_WP	Reference date of employees at the entity	X		X	X
PEILDAT_WP_OND	Reference date employees at the company			X	
P_DAT_WP_O	Reference date employees at the company	X			X
RECHTSVORM	Registered legal form	X	X	X	X
INS_REDEN	Reason for registration	X		X	X
UITS_REDEN	Reason for deregistration	X		X	X
REDEN_OPH	Reason for discontinuation	X		X	X
RSIN	Identification number for legal entities and partnerships	X		X	X
VENN_NM_DM	Name of the entity	X	X	X	X
NMI	Non-Mailing Indicator	X		X	X
BOEKJAAR	Fiscal year	X		X	X
DAT_OPRI_A	Date on which the entity was officially established	X		X	X
DAT_DEP_JS	Date of filing of annual accounts	X		X	X
INSCHR_DAT	Registration date	X	X	X	X
OPHEFF_DAT	Dissolution date	X		X	X
DAT_OPRICH	Date of establishment	X		X	X
DAT_VEST	Starting date of establishment	X		X	X
DAT_UITSCH RP	Date of deactivation of the registration			X	
DAT_ONTB_RP	Date of dissolution of the registration			X	
VEST_DATUM	Date the entity moved to its current establishment	X		X	X
DAT_VOORTZ	Date of appointment as chairperson	X		X	X
URL	Website URL	X		X	X
DOMEIN	Website URL		X		
P_W_FULLT	Full-time employees	X		X	X
W_P_FULLT	Full-time employees		X		
W_P_TOTAAL	Total number of employees	X	X	X	X
W_P_PARTT	Part-time employees	X		X	X
WP_TOT_OND	Total number of employees at the company	X	X	X	X
H_NAAM_VOL	Full registered trade name	X	X	X	X
IND_OPHEFF	Indicator for discontinuation	X		X	X
Label	Label of caring communities	X	X	X	X
Bedrijfsomschrijving	Textual description of business activities	X	X	X	X

Table 1: Overview of the features in the datasets, including the feature name, a description and in which dataset the attribute occurs in.

Dataset name in code.	Description	Role	Size	Records with the caring community label (%)
Df_KVK2022	Data from 2022, requested through the Chamber of Commerce based on their SBI code.	Test data 2022	8035 rows	8.9%
Df_KVK2023	Data from 2023, requested through the Chamber of Commerce based on their SBI code.	Test data 2023	11285 rows	7.4%
Df_labelled	Merged dataset consisting of the manually labelled dataset and the dataset with records identified through the survey	Training data	949 rows	47.2%

Table 2: The datasets after merging the datasets intended for training.

Appendix B

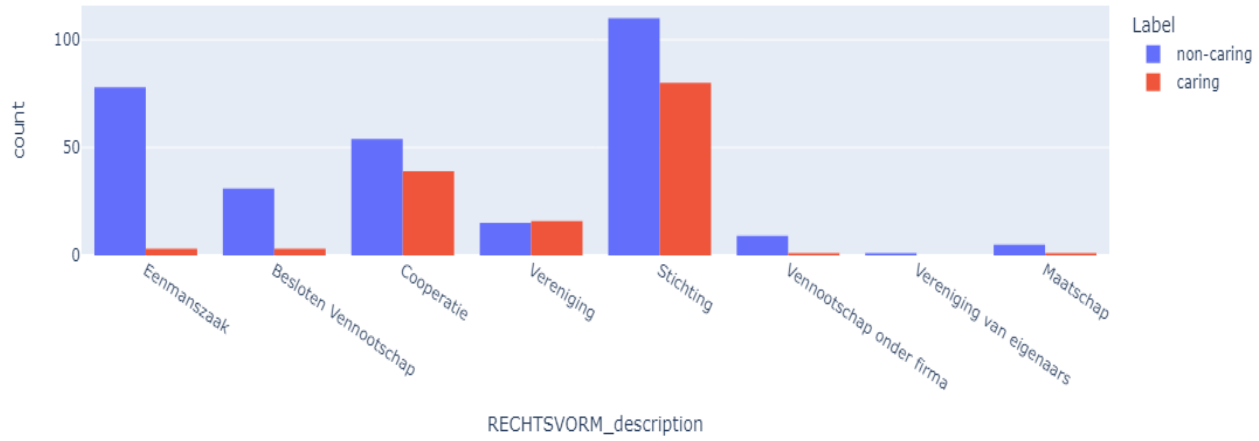

Figure 1: Distribution analysis of 'RECHTSVORM'

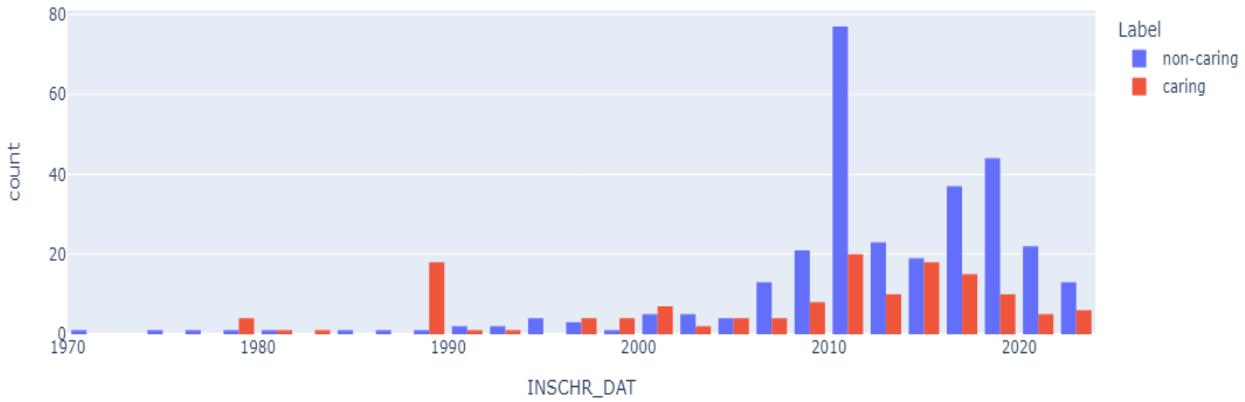

Figure 2: Distribution analysis of 'INSCHR_DAT'

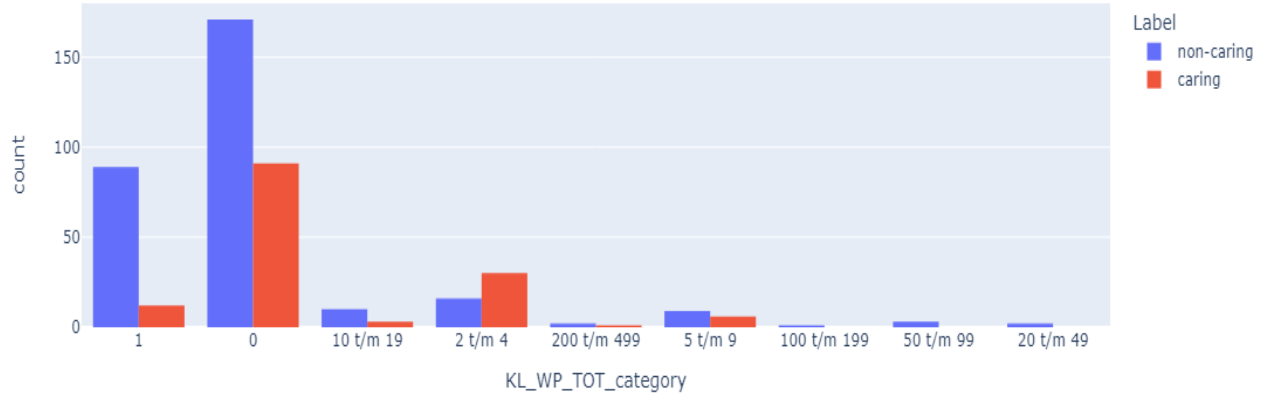

Figure 3: Distribution analysis of 'KL_WP_TOT'

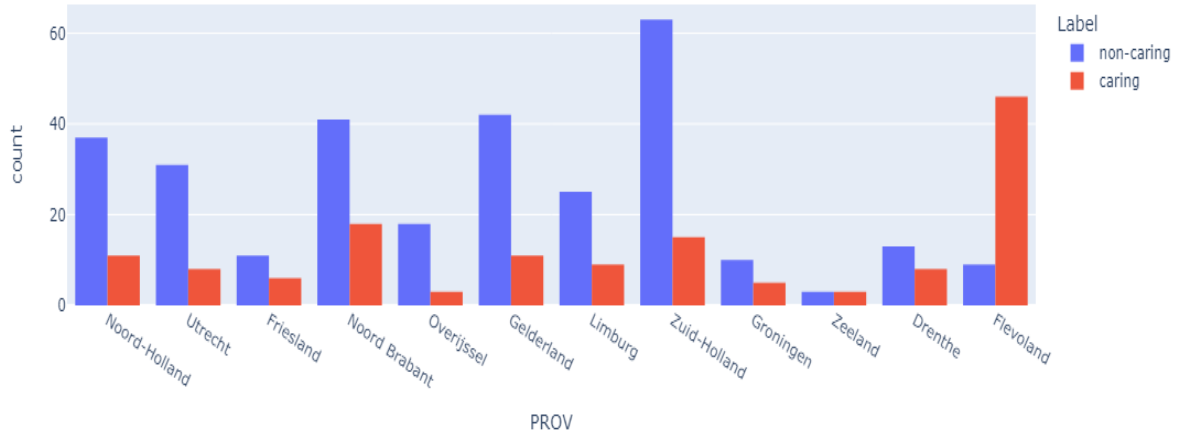

Figure 4: Distribution analysis of 'PROV'

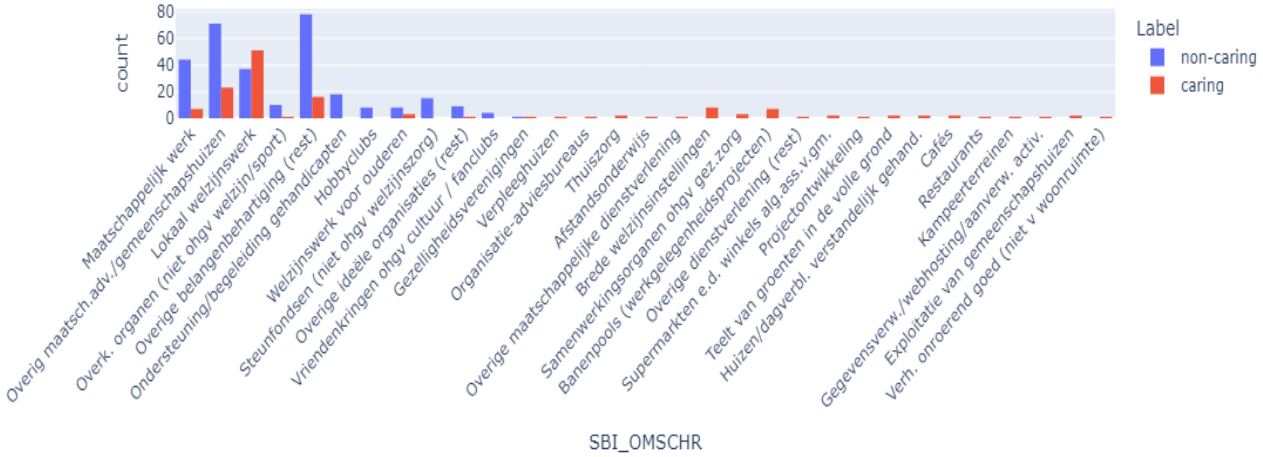

Figure 5: Distribution analysis of 'SBI_OMSCHR'

Appendix C

Feature name	Reason for removal
'RGL'	The same value is used for each row
DOSSIER/VGNUMMER	Unique values, these are not useful for analysis
'HN1X30', 'HN2X2X30', 'HN1X2X30'	HN45 is already used as it is the most complete business name
'GEMK_CA', 'GEMNAAM', 'GEMK_VA', 'STRVA', 'PCPLVA', 'STRCA','PCPLCA',	These are all proxies for adress, the only location indication that will be employed in the analysis is 'PROV'
MOB_TEL_NR', 'TEL_NRS', 'VOORLETTER', 'VOORVOEGSE', 'ACHTERNAAM,' 'PEILDAT_WP', 'PEILDAT_WP_OND', 'NMI' (email**),'BOEKJAAR', 'URL',**	Not useful for analysis, they logically would not be used for this prediction task
'FUNCTIE,' 'NEVENACT_1', 'NEVENACT_2','HFD_N_VEST' ',**'INS_REDEN',**	These values were missing for most records
'SBI_OMSCHR'	Another way of specifying the 'SBI_CODE' and therefore redundant
'UITS_REDEN', 'REDEN_OPH', 'IND_OPHEFF'	Most organizations in the data are still actively operating so including these features results in numerous empty rows.
'RSIN',	Legal identity number is not relevant for prediction.
,'DAT_OPRI_A,' 'DAT_DEP_JS', 'DAT_OPRI.' 'OPHEFF_DAT','DAT_VEST', 'VEST_DATUM', 'DAT_VOORTZ',	All represent the date the organization was registered or initiated, 'INSCHR_DAT' was the only feature employed in analysis that represents this date.
W_P_TOTAAL, P_W_FULLT, 'WP_TOT_OND', 'W_P_PARTT' en KL_WP_full	All represent the total amount of employees, 'KL_WP_TOT' remained to be used in the analysis.
'CD_EC_ACT'	Specifies the economic activity of an organization, this is defined as 'utilizing a total of more than 15 working hours per week'. This can be seen as an overlap of information with the 'KL_WP_ TOT' value and will therefore be removed.
PCVA_CIJF', 'PCVA_LTRS', 'BEHKN', 'GEOKN', 'PCCA_CI JF', 'DOMEIN', 'PCCA_LTRS', 'W_P_FULLT'	These features only occurred in the 2023 dataset and were therefore removed.

Table 1: All deleted features and the rationale behind their removal

Appendix D

Topic	First 15 words
1	stichting, gemeenschapshuis, adviesorganen, samenwerkings, overkoepelende, organen, gebied, exploitatie, welzijnswerk, delft, activiteiten, thuiszorg, fonds, fondsen, ouderen
2	stichting, ua, zorgcooperatie, dorp, beheren, kwetsbare, werk, wonen, bevorderen, buurt, dagbesteding, blijven, maatschappelijke, welzijn, mensen
3	stichting, buurt, gebied, vereniging, ua, bv, begeleiding, leden, cooperatie, alsmede, ten, diensten, doel, cooperatieve, alle
4	mensen, stichting, begeleiding, geven, beperking, bv, hulp, coaching, stellen, trainingen, maken, ondersteunen, financiele, maatschap, maatschappelijke
5	stichting, ua, leden, cooperatie, cooperatieve, enof, vereniging, belangen, activiteiten, coop, ondersteunen, filosoferen, ter, ondersteuning, sociale
6	leden, cooperatie, ua, buurtcentrum, ten, overeenkomsten, behoeve, stoffelijke, behoeften, bedrijf, uitoefenen, einde, belangen, doet, uitoefent
7	stichting, zorg, welzijn, begeleiding, mensen, buurthuis, individuele, ten, activiteiten, advies, daartoe, exploitatie, dementie, beperking, natuur
8	stichting, bevorderen, welzijn, bv, binnenstad, mogelijk, aanbieden, zorg, arnhem, faciliteiten, diensten, activiteiten, mensen, eigen, creatieve
9	cooperatie, ua, zorg, gebied, werk, diensten, ondersteuning, stichting, jeugdland, maatschappelijk, activiteiten, vereniging, welzijn, bieden, leven
10	vereniging, jongerencentrum, zorg, inwoners, ua, gebied, bv, wijk, voorst, belangen, leden, jongeren, exploiteren, alle, ver
11	stichting, bv, welzijn, nederland, ouderen, welzijnswerk, centrum, begeleiding, jeugd, activiteiten, exploiteren, beheren, organiseren, vrouwen, zoals

Table 1: Topic modelling results for the LDA algorithm

Topic	Top n words
-1	stichting, bv, buurt, dialoog, ontwikkelen, brengen, sport, organiseren, geven, stellen
1	stichting, zorg, welzijn, begeleiding, mensen, gebied, ondersteuning, ua, ondersteunen, ouderen
2	cooperatie, ua, leden, belangen, cooperatieve, ten, behartigen, uitoefenen, stoffelijke, overeenkomsten
3	exploitatie, stichting, gemeenschapshuizen, gemeenschapshuis, buurtcentrum, buurthuis, vitaal, verhuur, dorp, huis
4	jongeren, jeugdland, jeugd, jongerencentrum, activiteiten, vereniging, jongerenwerk, bv, stichting, kader
5	buurt, ouderencentrum, buurthuis, cooperatie, ecodorp, oudehaske, wlz, boekel, stg, eten

Table 2: Topic modelling results for the BERTopic algorithm

Model		Accuracy	Precision	Recall	F1-score
BOW	Non-Caring	0.80	0.77	0.99	0.86
	Caring		0.95	0.50	0.66
TF-IDF	Non-Caring	0.84	0.80	0.99	0.88
	Caring		0.96	0.60	0.74
LDA	Non-Caring	0.79	0.76	0.97	0.86
	Caring		0.91	0.50	0.65
Word2Vec	Non-Caring	0.85	0.82	0.97	0.89
	Caring		0.93	0.64	0.76
BERT	Non-Caring	0.82	0.82	0.91	0.86
	Caring		0.82	0.67	0.74
BERTopic	Non-Caring	0.84	0.83	0.93	0.88
	Caring		0.85	0.69	0.76

Table 3: Results of the different text representations on the training data

Appendix E

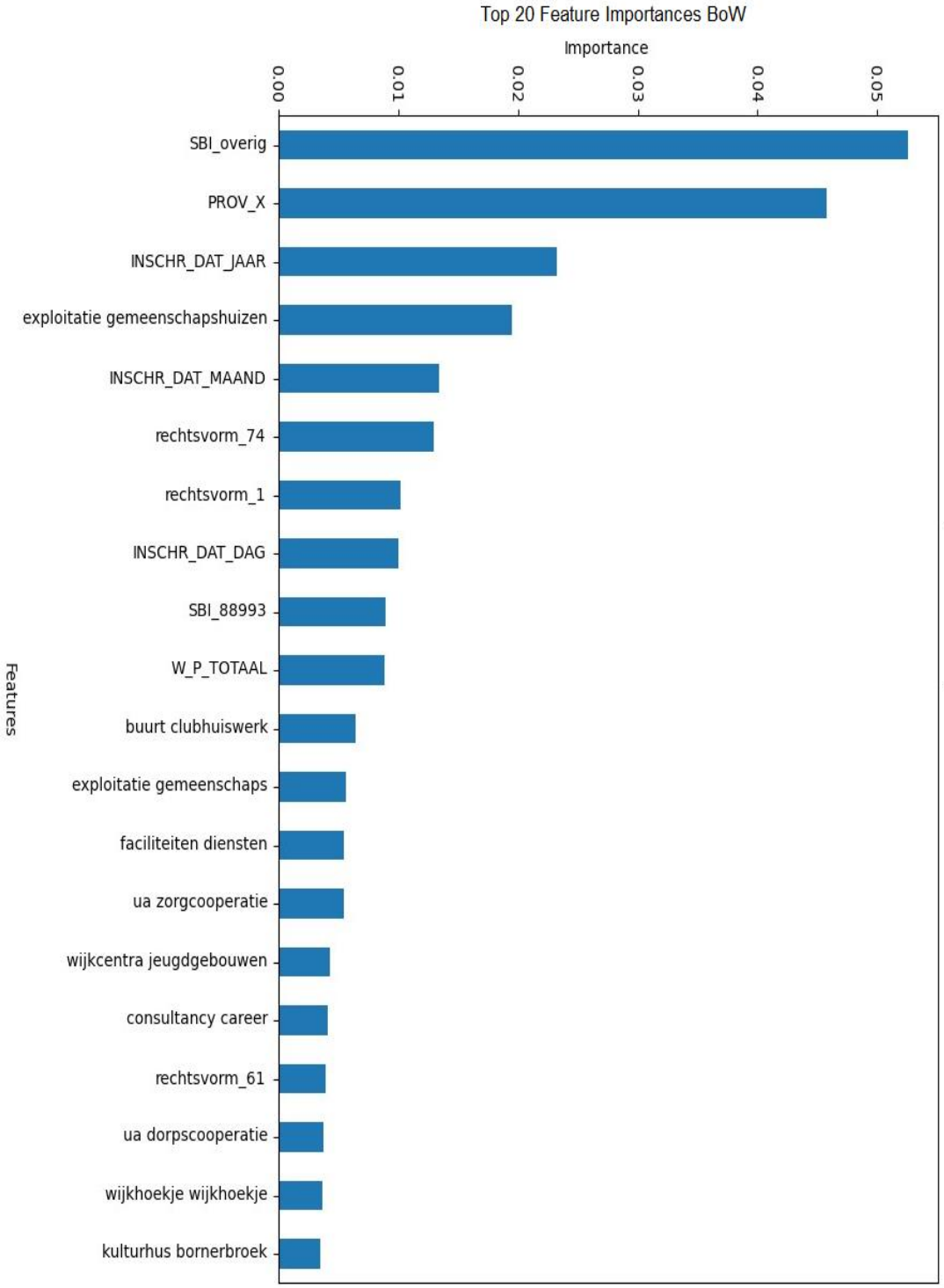

Figure 1: The top 20 feature importances for the BoW model

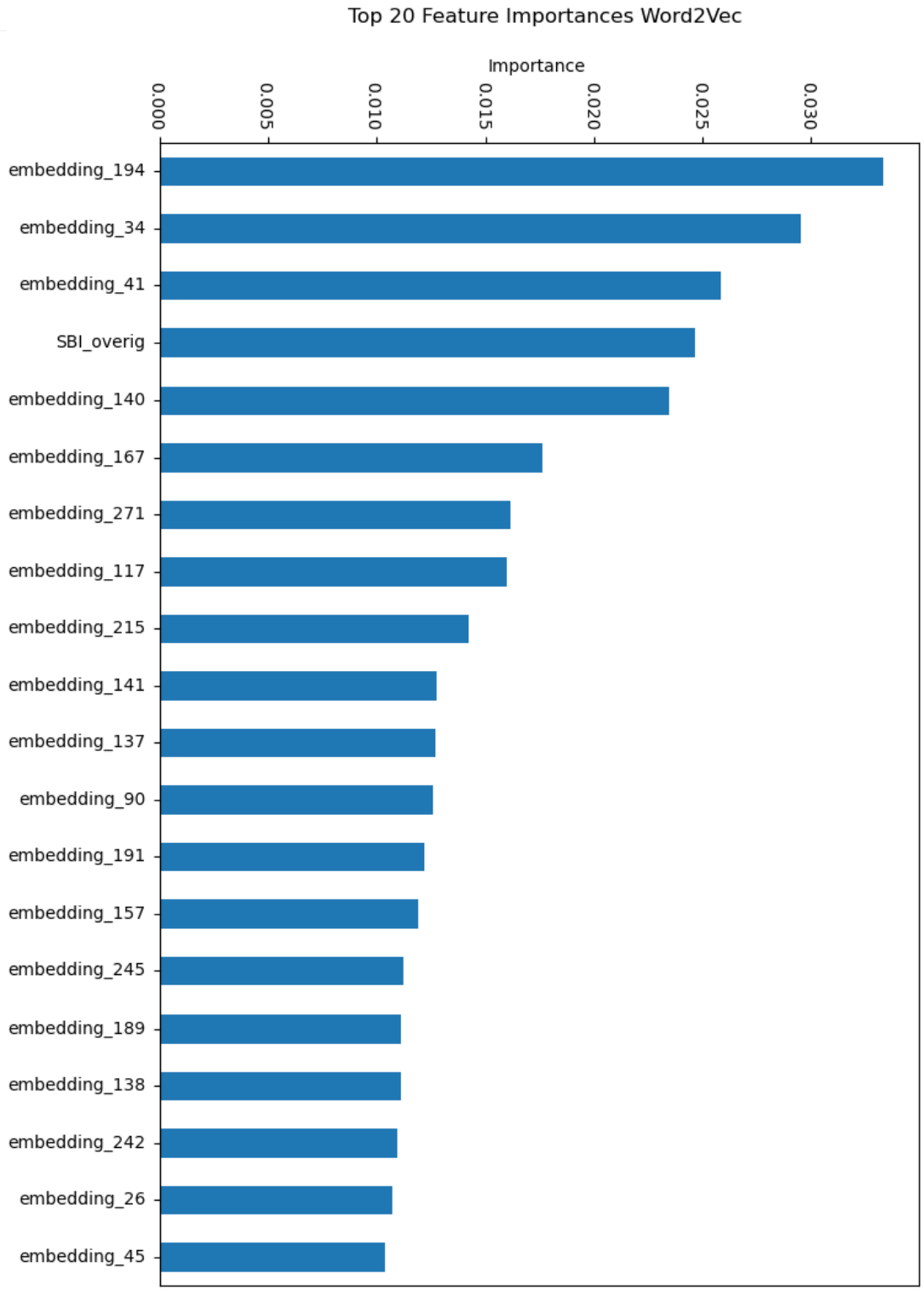

Figure 2: The top 20 feature importance for the Word2Vec model

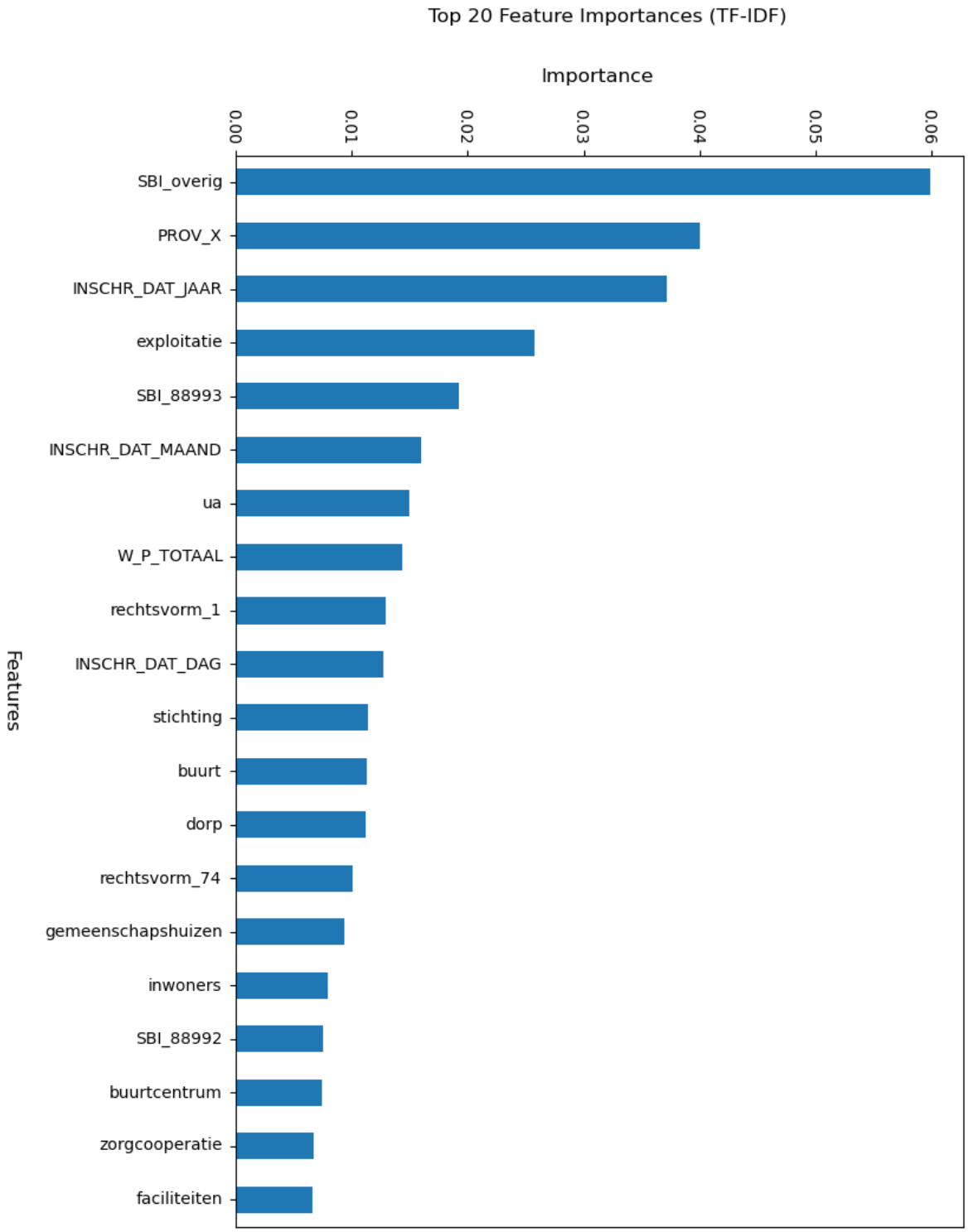

Figure 3: The top 20 feature importances for the TF-IDF model

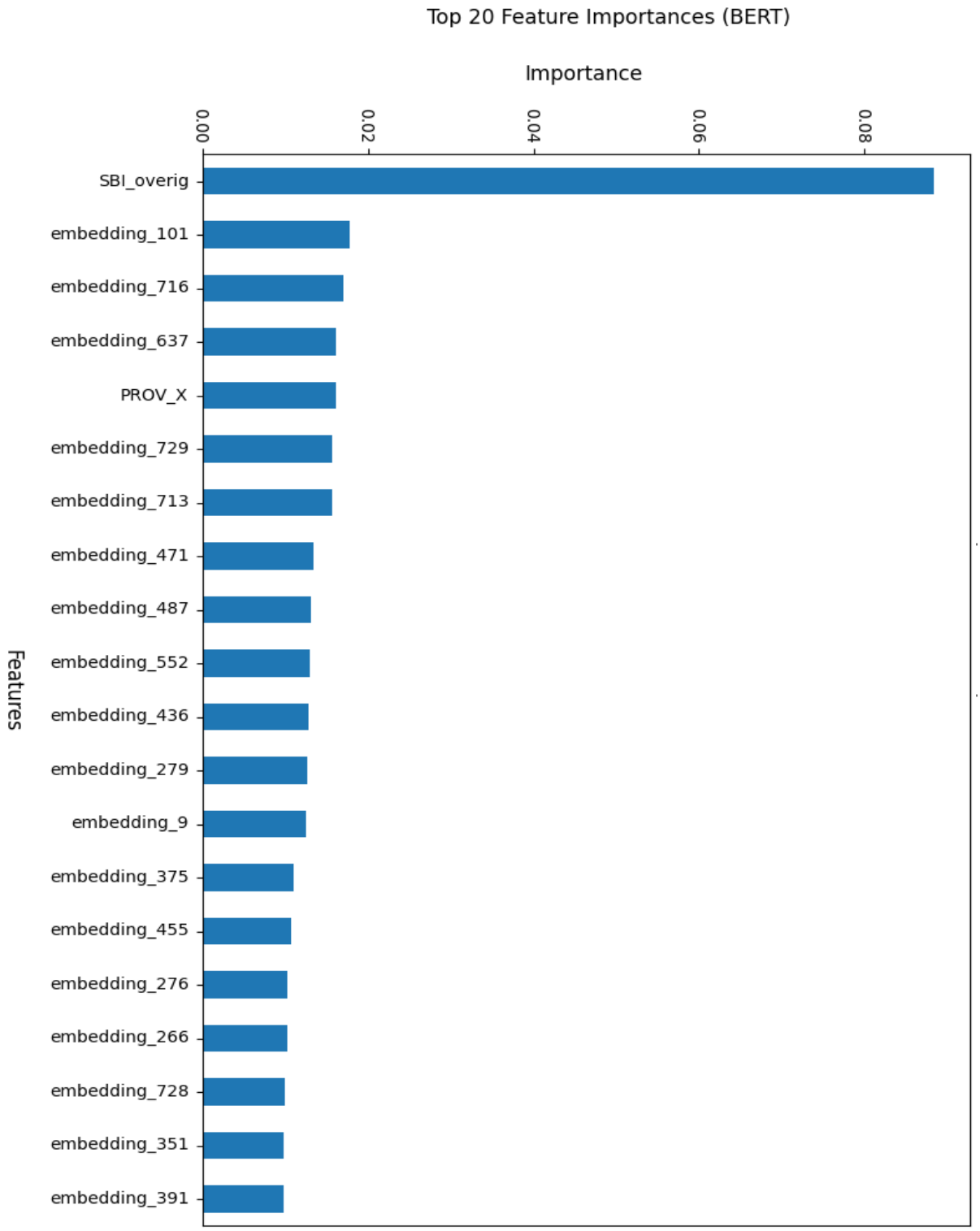

Figure 4: The top 20 feature importances for the BERT model

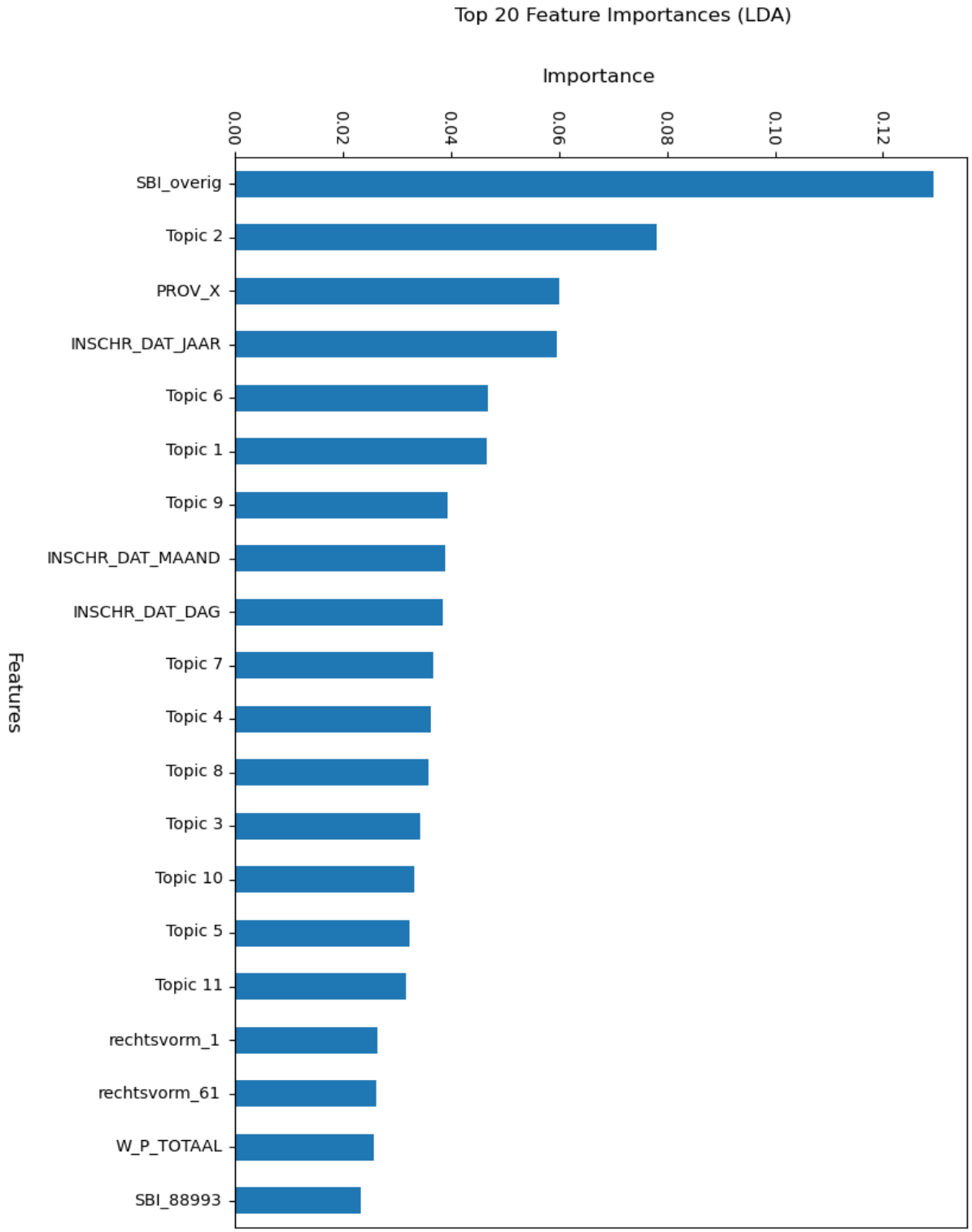

Figure 5: The top 20 feature importances for the LDA model

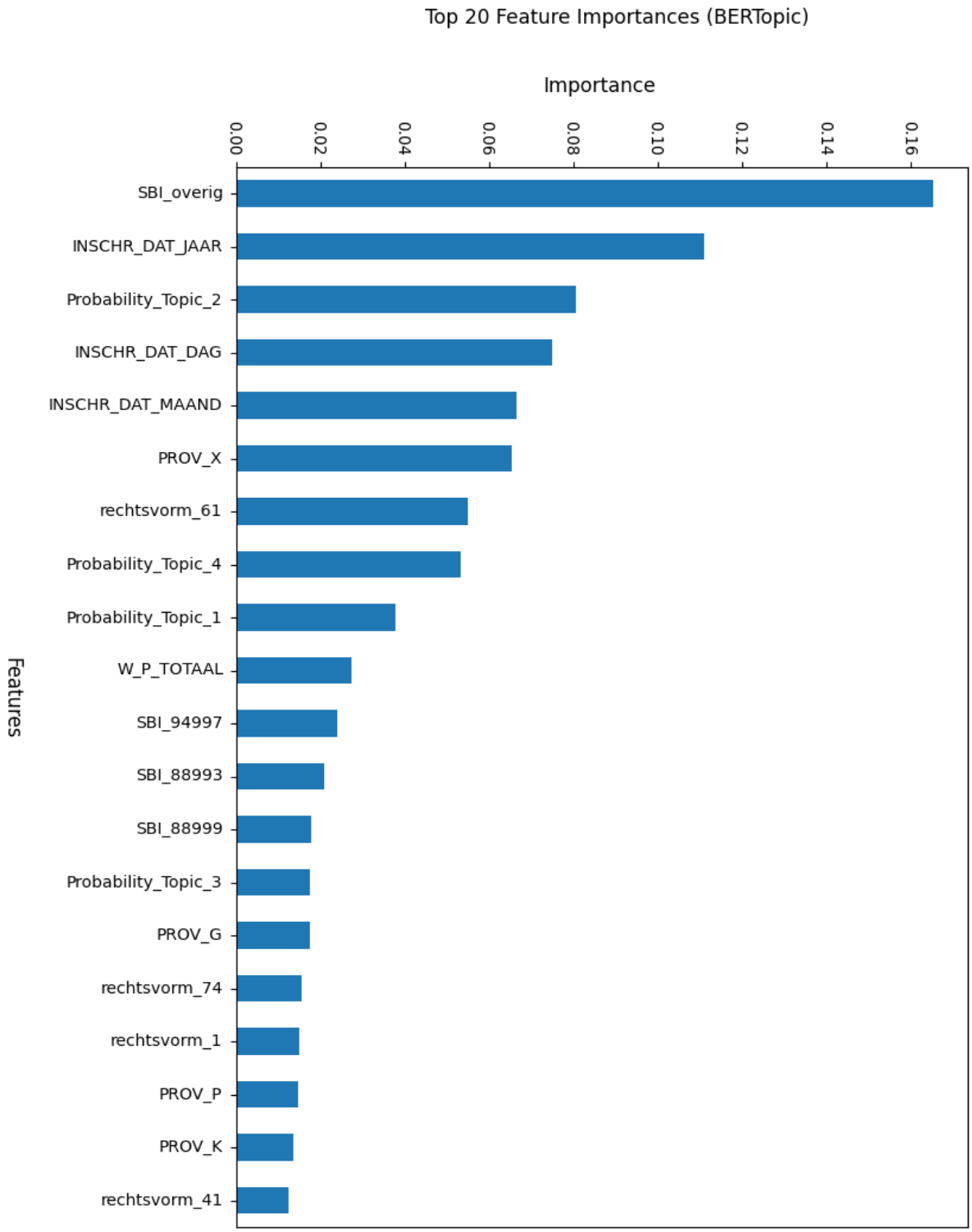

Figure 6: The top 20 feature importance for the BERTopic model