TUHH
Hamburg University of Technology

ComNets
Institute of Communication Networks

Master's Thesis

Term-based and Embedding-based Similarity Search in Large Unknown Text Datasets

Mattes Ruckdeschel

17.08.2020

Term-based and Embedding-based Similarity Search in Large Unknown Text Datasets

Mattes Ruckdeschel
Matriculation number 21154968
Computer Science and Engineering M.Sc.

Hamburg University of Technology
Institute of Communication Networks
First examiner: Prof. Dr.-Ing. Timm-Giel
Second examiner: Prof. Dr. Chris Biemann
Supervisor: Dr-Ing. Gregor Wiedemann

Hamburg, 17.08.2020

Declaration of Originality

I hereby declare that the work in this thesis was composed and originated by myself and has not been submitted for another degree or diploma at any university or other institute of tertiary education.

I certify that all information sources and literature used are indicated in the text and a list of references is given in the bibliography.

Hamburg, 17.08.2020

Mattes Ruckdeschel

Abstract

Working with large, unstructured data sets is a an increasingly important task in investigative journalism. However, the search engines of journalists' tools are still term-based, although there were large improvements in many information retrieval tasks in the last years.
These improvements are based on deep-neural language models such as Google-BERT, which are able to generate vector representation which capture the semantic characteristics of terms. This thesis is researching, whether the novel breakthroughs in Natural Language Processing that these context-embeddings brought can also provide a benefit for the task of searching for documents in large, unstructured data sets.
For this, test data sets containing heuristically similar document pairs were constructed from two publically available data sets. The test data sets were indexed by Elasticsearch, a widely-used term-based search engine and different context-embeddings creating language models. Then their capabilty to find the semantically similar document pair and their resource-consumption was measured and evaluated.
The findings show, that context-embedding based language models could be useful in searching for information in large, unstructured text data sets. However, their ressource consumption is far greater than that of the tested term-based search engine.
The results were able to showcase the strengths and weakeness of term-based and context-embedding based search engines for the task of searching for documents in large data sets. There are cases, where context-embeddings provide useful results which term-based search engines are not able to find. In order to make context-embeddings useful in investigative journalism, more research and engineering will be necessary.

Contents

1. Introduction

Today's journalistic work faces new challenges as well as opportunities in handling an insurmountable amount of data, which is too large for individuals to comprehend. The problem with that is, that although there is way more information from which a journalist can draw, the news story that this information is telling might never see the light of day due to the sheer amount of data that could draw attention from a viewer and the complex relationship documents might have between one another. For this reason, the field of data-journalism, which combines traditional journalistic work with data science and visualization is on the rise [1].
Journalistic media, as other business ventures handling lots of data, has to incorporate data tools in order process data for investigation. However, journalistic work is distinctive in the way that information is processed, since the end product is a body of text which was crafted by an individual (or a group of individuals), who interprets information in order to find a story worth telling. This is reflected in the requirements that journalistic data processing tools may have.
Investigative journalism is especially prone to rely on a large collection of data, since the incentive to investigate a subject often comes from data leaks. Two such data leaks are the noteworthy *Football Leaks*[1], a leak of over 18.6 Million documents, including contracts and documents containing secret agreements from professional football clubs from all over Europe, and the *Panama Papers*[2], a leak containing over 11.5 Million financial and legal documents which detailed corruption and other white-collar crimes revolving around secretive offshore companies which are in parts linked to 140 politicians from over 50 countries. These two examples showcase the scope of such leaks, not only in size but also in societal impact and importance.
Several tools for handling large, unstructured data in order to make them digestible by journalists already exist. They provide search functionalities based on Elasticsearch indexes in form of a simple key search, as well as batch searches. Although these search options already are very useful, more sophisticated tools for evaluating and sorting big data sets may prove beneficial.
Simple search queries are a helpful tool, but they can only find relationships between documents if these documents share the same vocabulary. There may be documents covering the same topic but differ in their language. Searches based on vocabulary alone therefore only tackle the problem of searching for information in large data sets in a limited way.
The motivation behind the research described in this thesis is the goal to find potential

[1] `https://eic.network/projects/football-leaks`
[2] `https://www.icij.org/investigations/panama-papers/`

improvements for *new/s/leak*, an open-source software developed by the Language Technology Group of the University of Hamburg in cooperation with *DER SPIEGEL*, a large German news organization [2]. *New/s/leak* is a tool for investigative journalists, which aims to make large amounts of text data searchable. Therefore, the research of novel search functionalities is important to bring improvements to the tool.
A feature, which may fulfill the need to provide a context between documents while searching for related information is the *semantic similarity search*, which aims to find documents which are "similar" to a given input document or query, based on the topic that they cover. Text documents must be represented in a way that make semantic similarity search possible. The chosen approach of *language modeling* has a great impact on the way that semantic similarity search can be conducted, because different language models capture different aspects of language.
Semantic similarity search can be achieved with newly developed language models based on neural networks, particularly with the transformer-based language model "BERT", which is a relatively new language model which was released in 2018 [3] with very promising capabilities in lots of Natural Language Processing (NLP)-tasks. Moreover, due to the fine-tuning capabilities and the many possible tweaks in of neural networks, it has lead to a number of language models which are powerful and fast in their respective domain. These language models however were not primarily developed for the task of finding similar documents in large data sets.
The language models render so called *contextual embeddings* or context-embeddings from textual input, usually creating one such vector per word in an input sequence of words. Context-embeddings are dense vector representations of language, which aim to preserve the contextual information of the word they represent.
Since context-embeddings are such a powerful tool for NLP, incorporating them in a useful matter into an already established tool such as *new/s/leak* may have great benefits for investigative journalism. Since *new/s/leak* is an open-source-tool, all interested journalists can benefit from the software, which is especially important for small teams with little resources. As resources may be a limiting factor for smaller news agencies, it is also desirable to develop tools which can run on a wide spectrum of hardware.
The aim of this work is to figure out, whether a semantic similarity search based on these burgeoning language models provides a benefit to journalists and how different metrics of similarity compare in the scope of journalism.
The following research question was formulated and will be discussed in this work.
Comparing a term-based and an embedding-based approach for a similarity search of documents, which approach fulfils the needs of (investigative) journalists searching for information in large, unknown sets of text documents better? What differences do the approaches have in resource consumption, performance and scalability?

To answer these questions, this thesis will conduct an analysis of the following steps. First, the state of the art in search capabilities for large text corpora will be investigated in chapter two.

Afterwards, in chapter three, the theoretical principles of used term-based and novel context-embedding based language models will be presented. Also the question of how these models can be used for similarity search will be answered. Using the findings of chapter two as a comparable standard, and the findings of chapter three as a theoretical foundation, a test setup for testing similarity of documents using publicly available data sets will be described in chapter four. This test setup was used to test the performance in semantic similarity search of several context-embedding based language models. The results were then analyzed quantitatively as well as qualitatively in order to come to an assessment of the capabilities to use the tested language models for an application in semantic similarity search with regards to the domain of large, unstructured data sets. The results and the evaluation will be presented in chapter five.
Finally, a conclusion will be drawn in chapter six, which will close with an outlook on potential future work arising from this thesis.

Hence, the main contributions of this thesis are the assessment of context-embeddings based language models in their potential capabilities to be used in semantic similarity search and the development of a setup for a similarity search engine for large text corpora. Further the capabilities and shortcomings of traditional approaches regarding semantic similarity search are highlighted and compared with the findings of the novel approaches. With these findings, the thesis contributes to research regarding novel search capabilities which improve journalists possibilities when conduction investigative research in large, unstructured data sets.

2. Related Work

This chapter describes the search capabilities of open-source tools which are currently used for data processing in investigative journalism and provide search functionality. Further, an analysis of the use of context-embeddings in search applications in general is conducted.

2.1. Search capabilities of open-source tools

Several tools for processing large collections of unstructured text data exists nowadays. Some of them, such as the forensic tool *Intella*[1] are closed-sourced products with many analytical capabilities. But there are also many open-source software projects which aim to give structure to large text data sets and make them searchable. As they are more analyzable and more comparable to *new/s/leak* and its targeted applications, the focus lies on other open-source solutions.
Since this work deals with a novel way of searching for relevant information in text data, we investigate the search capabilities of already established open-source tools. It has to be mentioned though, that these tools provide a number of additional task such as data wrangling or other NLP-tasks. The tool *alpeh*[2], *Overview*[4] and *DocumentCloud*[3] make the users' data searchable by indexing them, and also let the user search in other stored data sets which may be related to the users' data. However, these tools can also be run on a private server without interaction with the tools' respective search environment. Two other tools which are used specifically for investigative journalism are *Datashare*[4] by the *International Consortium of Investigate Journalism - ICIJ* and *Hoover Search*[5] by the *European Investigative Collaboration- EIC*, which is also part of the data wrangling pipeline of *new/s/leak*.
Although there are many open-source tools for working with large unstructured text data sets, further examination shows, that all of these tools are using the same search mechanism. All of these tools are using *Elasticsearch*[6] for indexing their documents and providing search capabilities to their users. Elasticsearch is a server-based full-text search engine which creates its own search index. Since it provides a REST-API and is

[1] `https://www.vound-software.com/solutions`
[2] `https://github.com/alephdata/aleph`
[3] `https://github.com/documentcloud`
[4] `https://github.com/ICIJ/datashare`
[5] `https://github.com/liquidinvestigations/hoover-search`
[6] `https://github.com/elastic/elasticsearch`

distributed, it is commonly used as a search engine for websites as well as for enterprises. In summary, the tools which are nowadays used for giving structure to unstructured text data sets differ in the way they preprocess and store their data, as well as in their functionality to visualize data and provide NLP-related information but they do not differ in their search capabilities.
Moreover, the search functionality of Intella is - as is elasticsearch - based on *Apache Lucene*, which is an open-source software library for the development search engines based on full text indexing of documents [5].
This leads to the conclusion, that Lucene-based software, such as Elasticsearch or Intella are a de facto standard for searching in large, unstructured text data sets.

2.2. Application of context-embeddings in semantic similarity

There are several models for creating context-embeddings of words. As we are only working with embeddings which come from transformer-based models (such as BERT) [3], we are focusing on related work using these context-embeddings.
Since these models were only beginning to being established in 2018, we found that there is little application outside of academia yet. *Google*, as the company responsible for BERTs development is incorporating the semantic information of BERT-embeddings in order to get a better understanding of natural language search queries.[6].
The General Language Understanding Evaluation (GLUE) benchmark [7] is a collection of tasks for evaluation of natural language systems. It includes a task about sentence similarity based on the STS-data set for textual similarity [8] which consists of English sentence pairs labeled with a similarity score. Transformer based models are on the top ranks regarding that task[7].

[7] for reference, see `https://gluebenchmark.com/leaderboard`

3. Theory

In this chapter, we will first describe the task, which a similarity search engine needs to fulfill, give a brief definition of language models and describe the two general language models which we are comparing in our experiments. Furthermore, we will investigate, how each language model can be used to evaluate the similarity of textual data.

3.1. Task Description

Journalists want to make sense of collection of documents, a *Corpus* C, which is generally too big to fully comprehend. When a journalist has found an interesting document d_i, it is helpful to present a collection of documents, which are relevant to the document. A similarity search system can accomplish this by applying a similarity function $s(d_i, d_j)$, which scores the relevance between two documents $d_i, d_j \in C$, to all other documents $d_j \in C$, in order to find the documents, which are most similar to d_i, according to the metric. The used ranking function differs depending on the way we are numerically representing our documents, the language model. The goal in application is to find for each document $d_i \in C$ the k most similar documents $(d_{i,1}, d_{i,2}, ..., d_{i,k}), d_{i,k} \in C \,\forall\, k$ according to the similarity metric $s(d_i, d_j)$.

3.2. Modeling Language with vectors

The purpose of a language model is to map text data to a numerical representation, so that it can be mathematically processed. Due to the large vocabulary of natural language and the fact that the set of possible sentences is infinitely large, the complexity of high dimensional vectors is necessary for representing language numerically. The advantage of representing text as vectors is, that there are well known and easily calculated metrics for vector spaces, so that the representation makes the comparison of text documents easier and well-formulated. The similarity of two vectors can be calculated with the *cosine similarity* $s(\vec{V_1}, \vec{V_2})$. Given two vectors $\overrightarrow{V_1}, \overrightarrow{V_2}$ the cosine similarity can be calculated as follows [9]:

$$s(\vec{V_1}, \vec{V_2}) = \frac{\vec{V_1} \cdot \vec{V_2}}{|\vec{V_1}||\vec{V_2}|} \tag{3.1}$$

[9] If vectors are already normalized, the normalization in the denominator can be omitted:

$$s(\vec{v_1}, \vec{v_2}) = \vec{V_1} \cdot \vec{V_2} \tag{3.2}$$

In the following sections, two general approaches for creating vector-representations of text will be described.

3.3. Term-based vectors: The Vector Space Model

The Vector Space Model VSM is a well established language model which is used since the 1970s [10]. It is widely used in information retrieval and is the underlying model of the Lucene search engine [11], which Elasticsearch is based on. In the VSM, vectors of documents are based on the terms that the documents contain. Therefore the model is also referred to as the *term vector model*.

3.3.1. Vector representation of text

VSM can be used to create vector representations for documents in a corpus. These *document vectors* have as many dimensions, as there are distinct term in the corpus. This is called the *vocabulary* of the corpus M. Each vector dimension represents a term in M. If the term is present in the document, the value of the vector in the dimension is a non-zero value, which depends on a chosen weighting of the term [9].
A simple way of weighting these terms is the term frequency $tf_{t,d}$, which is the number of times t occurs in the document d. Table 3.1 shows how a Corpus of three sentences can be represented by vectors using the number of occurrences of the word in the sentence in each dimension. Notice how adding the short third sentence to the corpus adds a dimension to all vectors. The third vector representation is also very sparse. Adding a sentence with little shared vocabulary to the corpus adds a dimension for each new word to all vectors in the corpus. If the corpus would be expanded by the sentence *a hare leaps out of a bush*, each document vector would be increased by six dimensions with zero as a value.

Table 3.1.: Example for representing a corpus of three documents with the VSM using tf, example created from [12]

	the	quick	brown	fox	jump	over	lazy	duck	never	say
The quick Brown Fox jumps over the lazy Duck	2	1	1	1	1	1	1	1	0	0
Never jump over the lazy Duck quickly	1	1	0	0	1	1	1	1	1	0
never say never, duck	0	0	0	0	0	0	0	1	2	1

In the case of Lucenes scoring [11], the weighting is not simply a word count but the *term frequency-inverse document frequency - tf-idf*-value $tfid_t = tf_t \cdot idf_t$. Term frequency

Table 3.2.: *Tf-Idf* of terms in example corpus

Term t	$tf\text{-}idf_t$
the	$\log \frac{3}{2} = 0.18$
quick	0.18
brown	0.47
fox	0.47
jump	0.18
over	0.18
lazy	0.18
duck	0
never	0.18
say	0.47

alone is not a good weighting factor, as it considers every term as equally relevant. The inverse document frequency of a term t is defined as follows:

$$idf_t = \log \frac{N}{df_t} \tag{3.3}$$

where N denotes the number of documents in the corpus, and df_t denotes the number of documents $d \in C$ which contain t. The $tf - idf$ is high for rare terms, which have a high discriminating power for documents and low for terms, which are present in most documents and thus do not help to distinguish between documents [9]. This can be seen from table 3.2, which depicts the $tf - idf$-values for the terms in the example corpus. The term *duck* is present in each document. Therefore it is not suitable to distinguish documents. Ente. The terms with the highest $tf - idf$ are the ones which are only present in one of the documents. These are the best terms to find the documents in a search, as they are best characterizing the documents relative to other documents in the corpus.

3.3.2. Similarity scoring of term vectors

The similarity of a document d to a query q can then be calculated using the document vector [9] :

$$s(q, d) = \sum_{t \in q} tf_{t,d} \cdot idf_t \tag{3.4}$$

If document vectors and query vectors are normalized, this is equal to the cosine-similarity. Only terms with non-zero values in the document vector contribute to the calculation.
An important refinement to the weighting function and default similarity measure of Lucene and Elasticsearch [13] is *BM25* [14]. The similarity score of a document d given a query q is calculated as follows [14]:

$$s(q, d) = \sum_{t \in q} idf(t) \cdot \frac{f(t, d) \cdot (k_1 + 1)}{f(t, d) + k_1 \cdot (1 - b + b \cdot \frac{|d|}{avgdl})} \tag{3.5}$$

where $f(t, d)$ denotes the frequency of the term t in the document d, $|d|$, denotes the length of the document d in words and $avgdl$ denotes the average length of documents in the collection C in words. k_1 and b are free parameters which are heuristically chosen. The default values of the BM25 similarity module of Elasticsearch are $k_1 = 1.2$ and $b = 0.75$ [13]. BM25 generally achieves better retrieval results but adds the disadvantage of new parameters that have to be chosen [14].

Figure 3.1 shows an example for how the most similar document to a query q can be found using normalized vectors of a set of documents d_1, d_2, d_3. For visual purposes, a two-dimensional example was chosen. This means, for the representation of the documents, only the values from two dimensions, in this case of the terms *gossip* and *jealous* are illustrated. The vector $\overrightarrow{v}(d_1)$ contains the word *gossip* more often and the vector $\overrightarrow{v}(d_3)$ the word *jealous*.
First, the normalized vector representation of the query $\overrightarrow{v}(q)$ is constructed. The most similar document from the set can then be found by using cosine-similarity. The result is the the document which has the vector representation with the smallest angle with respect to the vector representation of the query. In the example in figure 3.1, it is the vector $\overrightarrow{v}(d_2)$, representing document d_2.

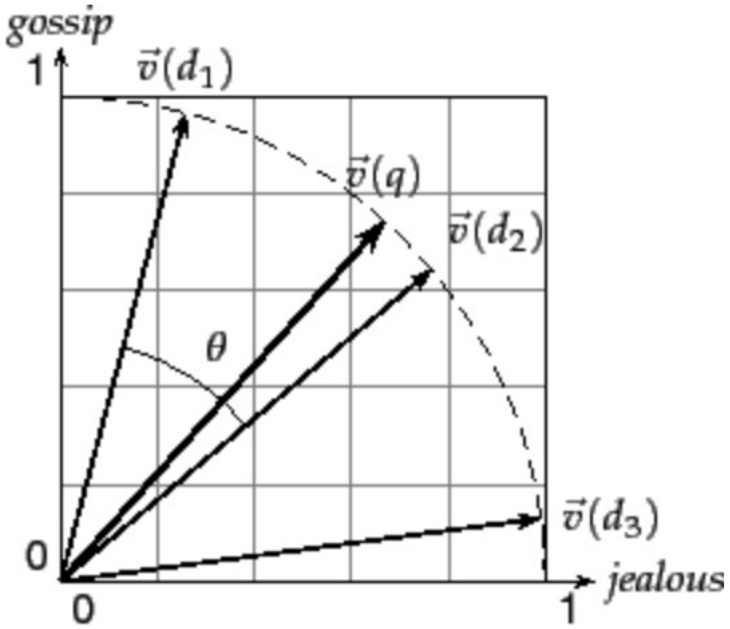

Figure 3.1.: Example of cosine similarity in a normalized vector space containing document vectors $\overrightarrow{v}(d)$, and a query vector. In this two-dimensional example the vocabulary consists of only two words. [9]

3.3.3. More-Like-This-Query

With the *More-Like-This-Query - MLT*, Elasticsearch provides a search query for finding similar documents to a query document in an index, which is based on BM25. The query works either with input text, as shown in figure 3.2, or with documents which are already indexed by Elasticsearch as input, as shown in figure 3.3. From the input, a search query is constructed by selecting terms to describe the input. Based on term selection parameters, like the number of terms to be selected from the input or the minimum term frequency below which terms will not be considered for the query, the terms with the highest *tf-idf* are selected. The constructed search query is then used to search for the highest scoring documents in the index using BM25. Thus, the MLT-Query provides a term-based approach for searching for similar documents.

```
GET /_search
{
  "query": {
    "more_like_this" : {
      "fields" : ["title", "description"],
      "like" : "Once upon a time",
      "min_term_freq" : 1,
      "max_query_terms" : 12
    }
  }
}
```

Figure 3.2.: Example of a more-like-this-query using a text as a query. The document fields "title" and "description" are evaluated for the similarity with the query.

3.3.4. Advantages and disadvantages of term-based vectors

Term-based vectors are an integral part of a lot of information retrieval systems and generally achieve good retrieval results. The construction of term-based vectors is easily implemented. Their disadvantage is that they are high dimensional and sparse, since every term in the vocabulary adds a dimension to the vector, even if it only occurs once in the corpus. Further, they represent language very crudely, since they do not consider word order and carry no semantic information about words. This infers that they cannot asses nuanced similarity, as any form of context, coming from using semantically similar words or word ordering is not carried over into the representation.

```
GET /_search
{
  "query": {
    "more_like_this": {
      "fields": [ "title", "description" ],
      "like": [
        {
          "_index": "imdb",
          "_id": "1"
        },
        {
          "_index": "imdb",
          "_id": "2"
        },
        "and potentially some more text here as well"
      ],
      "min_term_freq": 1,
      "max_query_terms": 12
    }
  }
}
```

Figure 3.3.: Example of a more-like-this-query using a text as a query. The document fields "title" and "description" are evaluated for the similarity with the query.

3.4. Vector representation using context-based embeddings

The disadvantages of term-based vectors are addressed by context-based embeddings. Context-embeddings are vectors which are created by language models which aim to create lower-dimensional vector representations of language. These vectors also try to capture the semantic meaning of the represented piece of text. Usually the vectors are generated for words. The idea is that semantically similar words have a similar vector representation. Thus, the semantic similarity of words can be calculated using cosine-similarity. Moreover, algebraic operations on the generated embeddings lead to conclusions about the semantic relationship of the words that the embeddings represent [15]. These Word embeddings are based on the *distributional hypothesis* [16], which states that semantically similar words occur in a similar context, thus have a similar probability distribution. The representation of a word is based on the context in which the word occurs. The development of models for generating predictive word embeddings was possible by advanced of machine learning techniques [15].
In recent years, deep learning techniques were applied to the task of generating word embeddings [3] [17]. Although pre-trained vectors from earlier models, based on shallow neural networks like *word2vec* [15], were already able to generate word embeddings which

capture the semantic relationships between words, they did not address polysemy, since they generate an embedding per word in the vocabulary regardless of the context [17]. For example, the word *Apple* may refer to the fruit or the Company. Deep neural language models are able to distinguish between references to either one.
Another recent advantage is the pre-training of language models based on very large text corpora, which is necessary to generate good language representation, with deep learning models. This makes incorporating these models into various downstream NLP-tasks possible. Since learning these features is costly in terms of time and resources it would be prohibitive to most users to train these models themselves.

3.4.1. Generation of embedding using BERT

In 2018, researchers at Google released BERT (**B**idirectional **E**ncoder **R**epresentations from **T**ransformers), a pre-trained, deep unsupervised learning language model which outperformed earlier language models in various NLP-Tasks at its inception [3]. Moreover, it lead to the development of many language models based on its architecture such as *Albert* [18] and *RoBERTa* [19]. Google released pre-trained language models which are trained on large corpora. The features of the pre-trained model can be fine-tuned in order to enhance performance in downstream tasks which may have corpora with vastly different vocabulary distributions. A key concept of BERT is that it works bidirectionally, which means that in order to learn the vector representation of a word, BERT looks at the words context in both - the preceding and following - direction [3].

3.4.2. Architecture

BERTs architecture is based on the encoder of the *Transformer*. The Transformer is an encoder-decoder-based neural model for sequence transduction using a concept called *attention* [20]. In encoder-decoder based models for sequence transduction, a vector representation is computed by the encoder, which is used as the input for the decoder in order to generate an output. Since BERT is learning a language representation model in order to generate embeddings, the vector representation of the encoder already is the desired output.
The encoder consists of L stacked layers, which each consist of two sub-layers, a *multihead-attention*-layer and a fully connected feed-forwards network. This is depicted in figure 3.4.

Between the two sub-layers normalization of the sum from the layers' input and its output is applied, so that the output is $Layernorm(x + Sublayer(x))$. This is illustrated in 3.4. The encoder for $BERT_{Base}$ stacks $L = 12$ such layers (For $Bert_{Large}, L = 24$).
The input of each layer is the output of the previous layer. As the layers operate solely on vectors, the input sequence has to be embedded into a vector input as well. BERT uses vectors containing embeddings from Wordpiece [21] as input tokens.

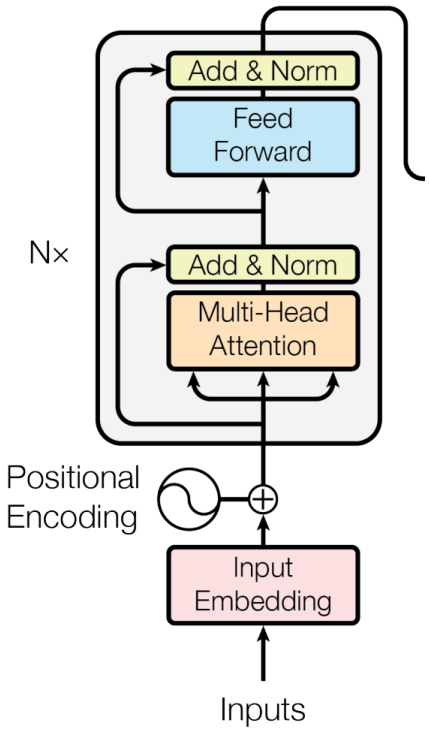

Figure 3.4.: One layer of the transformers encoder, consisting of a sub-layer calculating multi-head attention and a feed-forward neural network. After each layer, the sum of input and output is normalized [20]

From a text input, tokens are created using Wordpiece. The text input does not have to be a single sentence. More sentences together can be tokenized. Sentence endings are tokenized with a special separator token [SEP]. The resulting input for the model is a sequence of tokens.
An additional input is a positional encoding which is needed in order to use the positional information as the self attention does not incorporate positional information. The positional encoding PE is a vector with d_{model} dimensions and is added to the input embedding for each word. It is calculated as follows

$$PE_{(pos,2i)} = sin(pos/10000^{2i/d_{model}}) \tag{3.6}$$

$$PE_{(pos,2i)} = cos(pos/10000^{2i/d_{model}}) \tag{3.7}$$

where pos is the position of the token and i is the dimension. The sinusoidal functions were chosen, because for any offset k, PE_{pos+k} can be calculated as a linear function of PE_{pos} [20]. The first token of a sequence embedded with BERT is always a special [CLS]-token which is used as the aggregate sequence representation for classification tasks [3].

3.4.3. The attention mechanism

In contrast to older sequence transduction models, which are implementing recurrent neural networks, the transformer relies solely on the *attention mechanism*. Attention is a mechanism to assess the importance of various values to a query. The *multi-head Attention* layer of the encoder depicted in figure 3.4, consists of multiple attention heads, which all apply the mechanism of attention on the input simultaneously and independently. Self-attention of a sequence, is used to assess the importance of different positions from the same sequence to each other so we can weight them accordingly when constructing the context-based embedding of the word. This has the advantage that all parts of the input sequence are providing context to the word instead of only previous words, which is the case for language representation models based on Recurrent Neural Networks RNN [3].
Attention can be described as a mapping of the query and key-values pairs to an output, which is a weighted sum of the values [20]. Since all representation are vectors, the attention of multiple queries can be calculated simultaneously with matrices for queries Q, keys K and values V as

$$Attention(Q, K, V) = softmax(\frac{QK^T}{\sqrt{d_k}}V)[20] \tag{3.8}$$

The dot-product function is scaled with the factor $\frac{1}{\sqrt{d_k}}$, in order to avoid small gradients from the softmax function for large query and key dimensions d_k. The elements of V are learned during training. It was found beneficial to not only calculate one attention-value for each position in the input sequence but multiple, in order to create different representation subspaces at different positions [20]. This can be interpreted as learning different perspectives on the importance of other parts of the sequence to a token. An illustration of attention can be seen in figure 3.5. The attention head clearly weights the word *it* to refer to *the animal* from the beginning of the sentence. Thus it gives these word the most attention when encoding the word *it*. Other attention heads will likely give attention to different parts of the sentence.

Reducing the dimension of the multiple attention values and concatenating the results makes it possible to avoid increasing computational cost. Again, computational benefits arise from the fact that the calculations for the *attention heads* are independent from each other, thus making parallel computation possible.
After weighting the relatedness of the tokens using attention, a fully connected feed-forward neural network is applied to each position of the sequence [20]. It has one inner layer with dimension $4H$, where H is the dimension of the model ($Bert_{Base} : H = 768$, $Bert_{Large} : H = 1024$) and uses ReLU as activation [20].

BERT has to be trained on text data in order to learn how to construct term representation from input text. Therefore it was pre-trained on two unsupervised tasks:

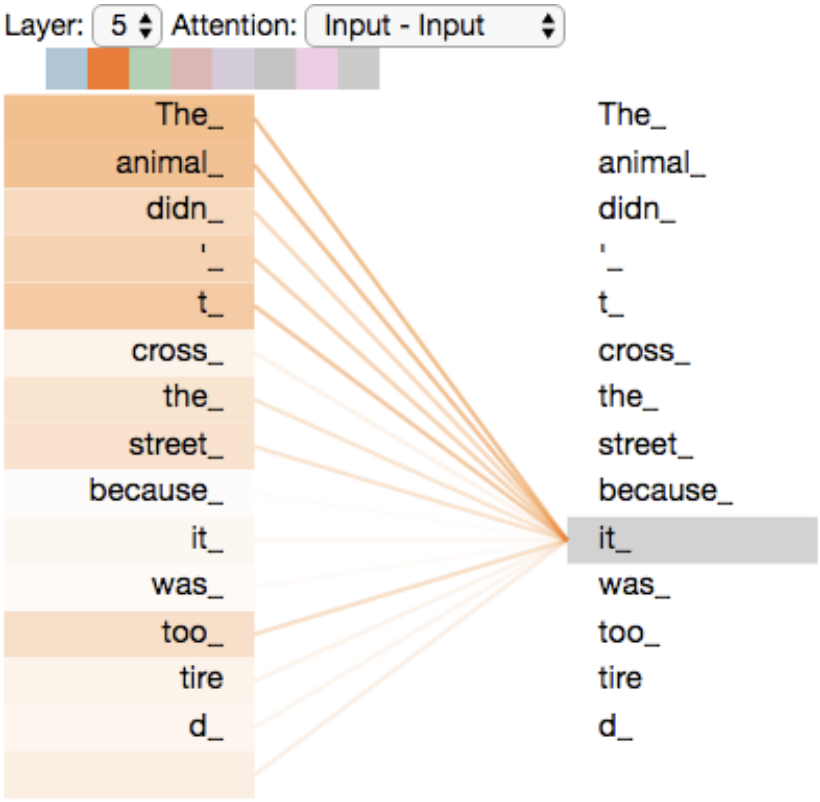

Figure 3.5.: Visualization of attention for the word *it* in a sentence for one attention head in an encoder. The stronger the color, the greater the attention. [22]

1. Masked Language Model ("MLM")
2. Next Sentence Prediction ("NSP")

During the MLM-task, the network gets sequences of tokens as inputs. A random sample of tokens (15% of all tokens in BERTs case) is replaced by a *[MASK]* token. The task is to predict the masked tokens.
NSP trains the language model on understanding the relationship of sentences which is import for many downstream tasks [3]. During pre-training, the model is presented two sentences A and B. In 50% of cases B is the sentence following A. B is labeled accordingly. The model is trained on the task to predict whether B follows A.

3.4.4. Language Models based on BERT

The language models produced by BERT are dependent on the hyperparameters of the network. BERT originally provided two language models which differ in the number

of stacked layers in the encoder. In 2019, there were also smaller models of BERT available [23]. A smaller model was also created by knowledge distillation, a technique, in which a smaller model is trained to reproduce the results of a larger model. The resulting language model *DistilBERT*, has a 40% reduced size and is 60% faster, while retaining 97% of the language understanding capabilities of the original model [24]. Another important parameter for the generation of language models is the training data set which is used. The original BERT models were trained on English data from the BookCorpus [25] and English Wikipedia data [3]. Later, multilingual BERT models were released which use the architecture of $Bert_{Base}$ but were trained on multilingual data sets which includes the articles from the top 100 languages with the largest Wikipedias. Due to the large differences in size, the data set was re-sampled, so that more prevalent languages, such as English, are not as over-represented anymore. For the first released multilingual model $Bert_{Base}$*, multilingual uncased*, lower casing, accent stripping and Unicode normalization were performed on the input. The second released multilingual model $Bert_{Base}$*, multilingual Cased* was trained without those three normalization steps on the input [1]. The developers recommend the Cased model, as it performs better on machine-translation tasks. Today, many language representation models based on BERT which were trained on data sets from various languages exist. The Bavarian State Library has released a German language model based on BERT, which uses a German data set consisting of several crawls and a German Wikipedia dump and has a size of 16 GB [2]. An English model which was not only created with different data sets but also changed other hyperparameters of the original BERT-architecture is RoBERTa (**R**obustly **o**ptimized **BERT a**pproach) [19]. The model was trained longer, and with more data. The Data set contains 160 GB of text data and consists of various openly available data sets. These efforts resulted in improvements across downstream tasks. This validated the theory, that the original BERT-model was under-trained, and that pre-training efforts improve with increasing training data size and diversity [19].

3.4.5. From word embeddings to document embeddings

Since BERT creates word embeddings, the similarity of words can be calculated using cosine-similarity on the vectors. In the test setting however, the similarity of sequences is of interest. There are several approaches to assume a vector representation of sequences based on the vector representation of the containing tokens. The [CLS]-token which already contains an aggregate sequence representation can be used in order to represent the entire sequence. Another approach is to calculate the average embedding of all token embeddings. However, these representations are not trained to have the same characteristic as embeddings, which is that similar vector representation infer semantically similar text [26]. The SentenceBert-model is based on BERT and was trained on labeled

[1] References can be found at `https://github.com/google-research/bert/blob/master/multilingual.md`

[2] References can be found at `https://github.com/dbmdz/berts`

data sets containing similarity scores [26]. Both data sets (SLNI and MultiNLI) only contain English data. Later, a multilingual model was released [27]. Instead of generating document embeddings from the actual output of the language model by taking the [CLS]-token embeddings or aggregating the output embeddings of terms in the sequence, SentenceBert was designed to render document embeddings.

4. Experiments

In this chapter the testing setup is presented. For that the heuristic assumption which this work is based on is described. Afterwards the used data sets are introduced, as well as the pre-processing steps which were applied to the data sets in order to generate test data sets. Further the evaluation metrics and the desired insights from these metrics are established. Then a list of conducted experiments is given in order to navigate through the results.

4.1. Heuristic assumption

The testing setup has to reflect the goal in application to find similar documents to a query document. Since there is no leak data with similarity labels, similar documents had to be defined a priori. The quantitative evaluation of the models regarding capabilities to retrieve similar documents was done with documents pairs, which were constructed from input data sets. The construction was based on the assumption that semantically similar documents are obtained by splitting a document into parts and combining even and odd parts respectively into two test-document pairs as shown in figure 4.1. Since the focus is on document similarity, the assumption was used to create semantically similar documents when creating a testing dataset from the input dataset. Although it is still possible, that more similar document pairs are present in a dataset, the constructed document pairs are considered to be heuristically more similar than random documents from the corpus.

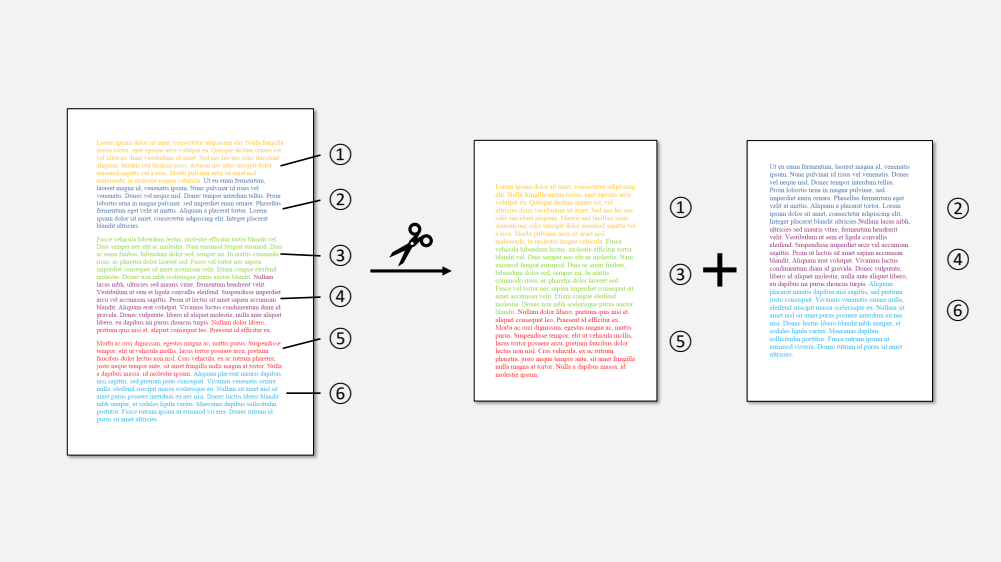

Figure 4.1.: Creation of document pairs from a document. The document is split into paragraphs. Even and odd parts are reconstructed to two documents, which build one document pair

4.2. Data sets

4.2.1. German data set: NSU enquiry board protocols

The NSU enquiry board protocols data set contains 12 documents from enquiry boards of German federal states and the German government. The data set is small compared to the other data sets (12 PDF documents, resulting in 80k document pairs for evaluation) and is in German [2].

4.2.2. Email data set: Enron

The Enron data set is a well known e-mail Data set with corporate mails from Enron employees [28]. The data set was first made public by the American Federal Energy Regulation Commission. It contains roughly half a million mails. It is not well suitable for the task at hand, as it contains many duplicate mails as well as text fragments such as disclaimers and quotes in responses. This invalidated the assumption that the document pairs which are constructed during preprocessing are of high similarity compared to the rest of the corpus. Therefore, it was omitted from in-depths analysis.

4.2.3. Multilingual data set: WW2 Wikipedia-scrape

This data set was scraped from Wikipedia and contains 27.000 documents in four languages, English, German, Spanish and Hungarian. Scraping started with the main article about the second world war, a link network was created by following links from that article for five layers in each respective language [29]. Due to the large number of layers, the data set contains articles on a variety of topics. From the English articles of this data set, a test data set was created. Since the scrape contained many duplicates, which always occupied the top rank positions, the data set was de-duplicated after creation.

4.3. Preprocessing

The input for the tests were folders with text or PDF-files. The files were parsed into strings using Tika [1] and split into text segments using a splitter. The chosen splitter was created individually and splits the input text into paragraphs on an empty line.
It does minimal input cleaning, in the form of deleting non-word characters which occur more than three times in a row. This was done due to the use of special characters for visual purposes in the data sets which did not contain information.
Further, a minimal paragraph length was defined in order to ensure that each paragraph contains processable information. The minimum paragraph length was chosen to be 200 characters. Shorter paragraphs are concatenated. A maximal paragraph length was defined, so that the created paragraphs all have a similar length and because some input documents contained all content on very few lines without empty lines in-between, creating document pairs which are longer than the maximum sequence length of BERT. If a paragraph has to be split due to the maximum length, the splitter tries to split on a sentence ending character (?, ., !), so that a semantically useful unit is not fragmented. The maximum paragraph length was chosen to be 400 characters.
Four paragraphs were added to a pseudo-page. This aligns with a normed page of 1500 characters from the German collection society for print media [30]. One document pair was created per pseudo-page. This was done by concatenating even and odd paragraphs together. Figure 4.1 visualizes the creation of paragraphs for a document. Note that documents in the input data sets tend to be much larger, resulting in a number of document pairs to be created from one input document. In figure 4.1, we can see, that paragraph 2 ends on a naturally occurring paragraph ending. Paragraph 5 however was not split on the paragraph ending, so that it has at least minimal character lengths.
Two documents are created from the input document, one containing paragraphs 1, 3 and 5 and the other containing paragraphs 2, 4 and 6. Based on the aforementioned heuristic these document pairs are regarded as similar.
Figure 4.1 also shows how the algorithms handles trailing paragraphs in documents, from which no document pair can be constructed, as at least four paragraphs are needed in order

[1] `https://tika.apache.org/`

to split an original document into even and odd pairs. The algorithm will concatenate trailing paragraphs and create larger document pairs. This can create problems during embedding creation, as document pairs might exceed the maximum sequence length of the context-embedding based models. As the used data sets contained mostly large original documents, there were little cases where that became a problem. For other data sets the hyperparameters of the preprocessing steps might need some adjustments.
The collection of these document pairs was stored on disk in a text file with one document per line which made reading in the data set as a Python list easy and fast. The index of the python list was used as IDs during indexing of the documents. These preprocessing steps were equal for all test settings.
After pre-processing the used WW2 data set consists of 205,114 documents and the NSU data set consists of 80,954 documents. The WW2 data set has a size of 148,3 MB on disk and the input folder has a size of 163,6 MB. The NSU data set has a size of 46,2 MB and the input folder has a size of 261,8 MB. The large difference in size savings are based on the fact that the NSU input documents were *PDF* files containing formatting and graphics and the WW2 documents were *.txt* files with no formatting. The whole process is depicted in 4.2.
From the test data sets created by this process, document embeddings were rendered and an index was created. This was done for each model which was tested.

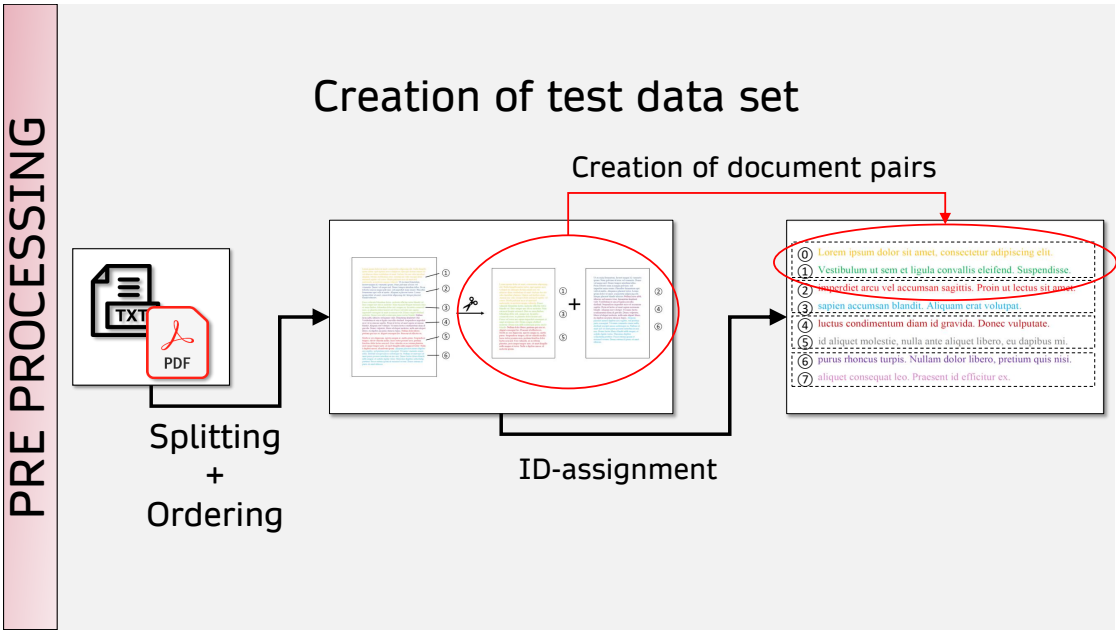

Figure 4.2.: Creation of document pairs from documents in the corpus. From input documents, to pseudo-pages, to document pairs.

4.4. Index creation

4.4.1. Generating an Elasticsearch index

Elasticsearch was running in a docker container. The stored documents contain the raw text and a numerical ID which corresponds to the index of the document in the list. In order to save time when creating the more-like-this-query, the term vectors of all documents were also stored. The documents are stored using the bulk-API of elastic.

4.4.2. Creation of an index for embeddings using Faiss

Embeddings were generated in batches of eight and afterwards stored in a Faiss-index [31]. The index is optimized for storing dense vectors and similarity search. The used index was a flat-index, which does not optimize the stored vectors. It gives the possibility to load the index from disk and retrieve the vectors from it. It calculates the similarity using the metric inner product of the vector, which is the same as calculating cosine-similarity when the vectors are normalized. Therefore, all vectors are normalized before storage. The final index was stored on the hard drive.
Flair [32] was used as a framework to generate word-embeddings for the two tested methods for creating document embeddings, which are based on creating word-embeddings first. Flair also provided the possibility to generate the document-embeddings from the generated word embeddings. The embeddings from the sentence-transformer [26] were generated without Flair, as the model is able to generate the document embeddings directly from input text. All embeddings were generated on a GPU. The used GPU is a Geforce GTX1070.
Although the sequence length of the test-documents was chosen with the maximum sequence length of BERT in mind, due to special characters, such as Arab language or GPS-coordinates which were inefficiently tokenized, generation of embeddings crashed with a reference to memory overflow. This lead to a state in which the GPU became unresponsive. The problem was solved through restarting the generation process. In case of a failed generation, the script stored its current state as well as an incomplete index. It then restarted itself into a secured mode, where the last batch was re-embedded with a batch size of one. Afterwards, the generation continued with the usual batch size parameter. If the generation of an embedding was not possible in the aforementioned secured mode, the test-document pair was excluded from the index. This setup ensured that as many document pairs as possible were embedded.
Afterwards similarity results for each document in the respective index were retrieved from all indices and stored in ranking tables.

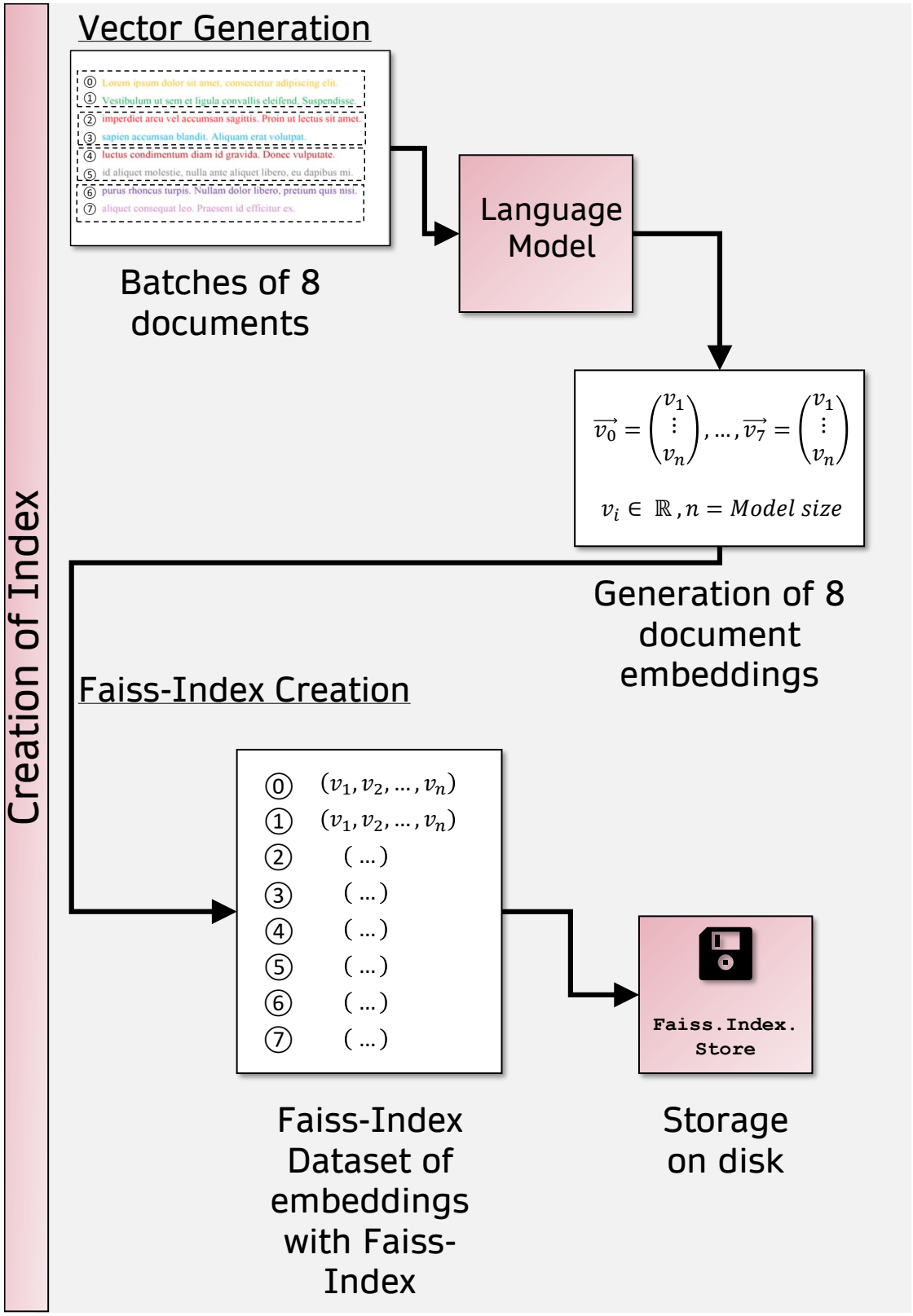

Figure 4.3.: Index creation using a context-embedding based language model. Embeddings are generated in batches of 8, and stored in a Faiss-index.

4.5. Retrieval from index

4.5.1. Generating similarity results from Elasticsearch

For each document, a list of the most similar documents was created using elastics more like this query [2]. Elastic tries to build a query, which is a good representation of the input document by choosing the terms with the highest *tf-idf*. This query is used to search for similar documents. As the default values for the term selection parameters regarding minimum term frequency (the minimum number of times a term has to occur in the document before being considerable for being a representative term) and minimum document frequency (the minimum number of times a term has to occur in the document set before being considerable for being a representative term) lead to situations where only few similar documents were found, these settings were reduced to a value of 1. The results were stored in a .csv file for later analysis. It was also tested, whether removal of stop words had an effect on the results. The quantitative analysis showed no significant difference in the results except the size of the created index on disk. Therefore these results were omitted from the quantitative analysis but are mentioned when comparing the sizes of the models.
In order to be comparable with retrieval from Faiss, the multi-get API of Elasticsearch, which allows multiple search queries to be executed with one request to Elasticsearch was used to retrieve MLT-results in bulks of 1000 queries. The IDs and similarity score of the top k results for each document in the index were stored in a spreadsheet.

4.5.2. Generating similarity search results from Faiss

Faiss is optimized for bulk retrieval, therefore retrieval of similarity results was conducted in bulk. In order to generate the results for an index, the index was loaded into memory from disk which gives the possibility to retrieve the stored embeddings and use them as a search query for the most similar vectors in the index. The similarity results were generated for batches of 1000 embeddings in parallel by issuing a search for the $k + 1$ most similar vector to a vector from the index. $k + 1$ was used, as the most similar vector always is the query vector itself, as it is included in the index. The IDs and similarity score of the top k results for each document in the index were stored in a spreadsheet. This process is depicted in figure 4.4.

[2] `https://www.elastic.co/guide/en/elasticsearch/reference/current/query-dsl-mlt-query.html`

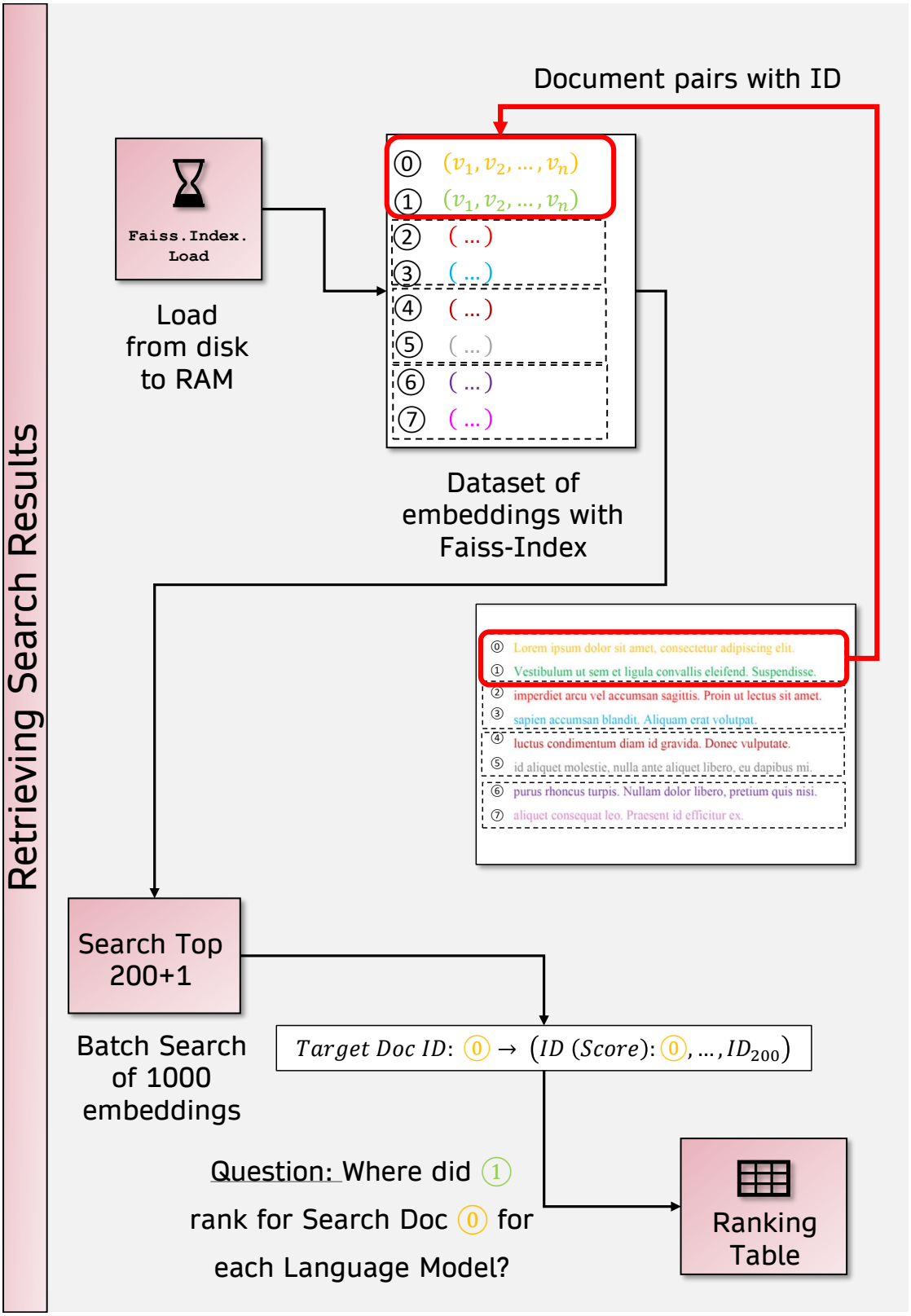

Figure 4.4.: Retrieving Search results from a Faiss index. The index is loaded from disk and in batches of 1000, the search results for embeddings are generated and stored in a ranking table.

4.6. Quantitative evaluation of Results

For both data sets, WW2 and NSU, a ranking of the most similar documents is created for each document. It is then examined, where the other part of the document pair ranks. From that, an average ranking of the other part is calculated. This is the main quantitative evaluation metric. The assumption behind analyzing this metric is that semantically similar document pairs were constructed during the creation of the test setup. Therefore, the term-based approach, as well as the embedding-based approaches can be compared in their ability to find semantically similar documents by analyzing this metric. Additionally the co-occurrence of other important quantitative metrics are the index creation time and the time it takes to generate the similarity results and the disk space that storing the index costs. As all measurements were conducted on the same hardware, a deep analysis into additional resource consumption was omitted. However, the index creation time was measured not only as in-program time but in real time as well, since there were significant differences in the capabilities to create context-embedding for larger document pairs. This lead to a large difference in needed restarts for some models which greatly impacted overall index creation time and thus efficient GPU utilization in real-time.

4.7. Qualitative evaluation of Results

One of the main obstacles in evaluating the capabilities to find semantically similar documents in large, unstructured corpora is that these corpora are not labeled for similarity. Therefore, a qualitative evaluation of some results is needed to assess whether found documents are semantically similar. The document pairs were chosen by several metrics. An analysis was conducted of document pairs, where the targeted document was highly ranked. This was done to find out, whether reasonable semantic similarity between the document pairs can be inferred.
Another analysis was to investigate cases, where document pairs could not be found. This is to find out, whether the top ranked results can be reasonably assessed as being semantically more similar to the query document as the target document.
Finally, an investigation of cases with large and with little co-occurrence between models and Elasticsearch was conducted. The reasoning behind this analysis is the assumption that Elasticsearch as a professional tool is able to provide somewhat useful results. Therefore it is interesting to see, whether vastly different, or more aligned results from models found similarly reasonable results.

4.8. Experiments overview

After pre-processing the two data sets into test data sets, there were several models tested for each test data set. Table 4.1 shows the tested models for the WW2 test data set and table 4.2 for the NSU test data set respectively. There were different models tested for each test data set, as the data sets differed in language and an English language cannot be expected to create useful embeddings for German text.
We can see from the tables that most of the tested models are based on $BERT_{Base}$ and therefore their output vectors have the same number of dimensions.
The word embeddings which are needed for the CLS-method and mean-method for creating document embeddings were generated using Flair. For these models, the quantitative evaluation was conducted. The qualitative evaluation was only conducted for the models, which performed best in the qualitative evaluation.
In order to make results table more readable, all models are abbreviated.

Table 4.1.: Models tested for WW2 test data set

Document Embedding Method	Model Name	Abbreviation	Output Dimension
CLS & Mean	Bert Base Cased English	BC_E	768
CLS & Mean	Bert Base uncased English	BU_E	768
CLS & Mean	Bert Base cased multilingual	BC_{ML}	768
CLS & Mean	roberta-base-openai-detector	R_E	768
SentenceBert	bert-base-nli-stsb-mean-tokens	SBB_E	768
SentenceBert	roberta-base-nli-stsb-mean-tokens	SBR_E	768
SentenceBert	distiluse-base-multilingual-cased	SBD_{ML}	512

Table 4.2.: Models tested for NSU test data set

Document Embedding Method	Model Name	Abbreviation [MB]	Output Dimension
CLS & Mean	Bert Base cased multilingual	BC_{ML}	768
CLS & Mean	Bert Base uncased multilingual	BU_{ML}	768
CLS & Mean	Bert-base-german-dbmdz-cased	BC_G	768
CLS & Mean	Bert-base-german-dbmdz-uncased	BU_G	768
SentenceBert	distiluse-base-multilingual-cased	SBD_{ML}	512

5. Results

In this chapter, the results of the conducted experiments are compiled and analyzed. The chapter is divided in sections for each evaluation criterion described above. The sections consist of the results for both data sets, as well as an evaluation of those results.
First the quantitative analysis is presented, which consists of the measurements regarding resource consumption and the ranking results. The quantitative analysis closes with an evaluation of co-occurrence in rankings results between the different context-embedding based indices and Elasticsearch.
Afterwards, selected examples from the qualitative analysis are presented and evaluated. The chapter closes with an evaluation of the methods that lead to the results.

5.1. Time consumption

5.1.1. Index Creation

First, the vector indices containing the embeddings are created for all tested models. During this, time and other resource consumption metrics were measured. Since all tests were conducted on the same hardware and all models utilized the GPU as much as possible during index creation, the most important difference in resource consumption was the difference in time that index creation took.

Table 5.1 shows the results regarding time consumption during the index creation for the WW2 data set in reference to the time that Elasticsearch took for indexing the data set. The table shows, that index creation is fastest, when the document pair embeddings are created with SentenceBert instead of taking the CLS-token or calculating the mean from the word embeddings in the document pair. The fastest index creation was accomplished by using the SentenceBert Model SBD_{ML}. Comparing only indices with same output vector size, the fastest index creation was accomplished by SBR_E. It is 3.84 times faster than the fastest index creation using a Flair-Model instead of SentenceBert. Regarding Flair models, index creation was faster using mean token weight as a method to create document pair embeddings instead of using the CLS-token. Index creation was slower by a factor between 1.11 for BC_E, which is also the fastest model for both methods, and 1.51 for R_E which is also the slowest model for both methods. The time difference during the creation of indices differs significantly between CLS-based and mean-based models, which need less than 50 % of the time of CLS-based models during index creation.

Table 5.1.: Time consumption during index creation for WW2 data set

Document Embedding	Model	Time consumption in reference to Elasticsearch
CLS	BC_E	107.92
	BU_E	106.56
	BC_{ML}	135.05
	R_E	133.37
Mean	BC_E	49.18
	BU_E	53.18
	BC_{ML}	49.45
	R_E	52.04
SentenceBert	SBB_E	13.36
	SBR_E	12.82
	SBD_{ML}	8.02
Elastic	More-Like-This	1 (127.99s)

The time to create an Elasticsearch index is significantly smaller than the time consumption with any model. Index creation was 8.02 times faster than index creation for the fastest model and 53.18 times faster than index creation for the model with the highest mean ranking result for the target document.

Table 5.2.: Time Consumption during index Creation for NSU data set

Document Embedding	Model	Time consumption in reference to Elasticsearch
CLS	BC_G	358.20
	BU_G	361.95
	BC_{ML}	493.6
	BU_{ML}	447.23
Mean	BC_G	155.93
	BU_G	167.35
	BC_{ML}	159.26
	BU_{ML}	168.75
SentenceBert	SBD_{ML}	27.26
Elastic	More-Like-This	1 (14.50s)

Table 5.2 shows the time consumption during index creation for the NSU data set. The time consumption during index creation for the NSU data set was generally lower than the time consumption during creation of the WW2 data set. The WW2 data set is 2.53

times larger than the NSU data set and index creation took 2.60 times longer for BC_{ML} and 2.74 times longer for DB_{ML}. CLS-based models were again slower than mean-based models. The creation time was 1.94 times larger for BC_G and BU_G and 1.84 times larger for BU_{ML} and BC_{ML}. Creating the index with Elasticsearch took 14.5 seconds, which is 27.3 times faster than index creation with SBD_{ML} and 155.9 times faster than the fastest flair-based model took.

We can see from table 5.1 and table 5.2 that the time consumption between the different document embedding creation methods differed significantly. The CLS-based models' increased time relative to the mean-based models comes from the fact that during index creation, the models failed to create embeddings in batches for significantly more batches. This resulted not only in much more time consumption due to creating embeddings individually, but also due to being forced to reload the code, including loading the large language model into GPU memory. Although loading the model into GPU-memory when using CLS token as a document embedding creation method did not require more GPU-memory than loading the same model with mean token weight as the method, the differences in number of failed batches was very large. The reason for this might be a design decision in the Flair-framework which changed the behavior of some models for longer sequences. This changes of the framework was implemented during the testing phase of this thesis. While mean embeddings were updated with a functionality which allows for longer sequences, CLS embeddings did not receive such an update. This might explain why the CLS method failed more often. Measurements on only the time spent on computation instead of on the absolute time showed, that computational time was also increased for CLS-based models. The difference in computing time between CLS-based and Mean-based models is likely due to the fact that index creation for CLS-based models spent more time creating embeddings individually.
The large difference in time consumption between SentenceBert models and Mean-based models implies that the code for generating SentenceBert embeddings performs embedding generation much more efficiently in this test setting. Since creating document embeddings based on SentenceBert does not use the Flair-framework, which required the construction of Sentence-objects out of input Strings before embeddings, it is conceivable, that the embedding creation process of SentenceBert, which does not require such objects for creating embeddings, is generally faster.
This time difference is crucial when comparing the embedding-based models with Elasticsearch. The larger time difference for the smaller NSU data set implies that the overhead for module loading plays a role in the time difference. Regardless, the time difference between Elasticsearch and even the fastest model SBD_{ML} shows that index creation using context-embeddings comes with a large time cost. Further it has to be mentioned that index creation using context-embeddings comes with the additional cost of powering dedicated hardware in form of a GPU, which Elasticsearch does not require. Therefore considering the cost of acquiring and powering such additional hardware increases the real costs of using context-embedding based approaches.
The conclusion from the index creation time is, that context-embedding based models are

significantly slower in index creation. In order to be feasible, smaller models, such as a DistilBERT-based models are required, as they are significantly faster than their larger counterparts. Also, SentenceBert based models are preferable, as they are able to generate indices faster while having as little errors during index creation as mean-based models. CLS-based models are in the current setup too slow to be considerable as feasible models for large data sets. Considering that the WW2 data set is based on 9717 input documents and the investigative corpus of Football leaks contains 18.6 million documents, which is roughly 1900 times larger, Elasticsearch already requires at least 68 hours ($\sim$ 3 days) for indexing, disregarding potential optimization efforts. The fastest Model SBD_{ML} would require 545 hours ($\sim$ 23 days) and the fastest Flair-based model would require 3365 hours ($\sim$ 140 days) In order to be more competitive, index creation of context-embedding based indices requires more potent hardware than the test hardware.

5.1.2. Retrieval time from index

The retrieval time does not differ between models which share the same vector dimension. Therefore the comparison is only conducted between the average retrieval time of models with the same vector dimension and Elasticsearch.

Table 5.3 shows the time it took the models of different size and Elasticsearch to retrieve similarity information for each document in the test data set. Elasticsearch took 1.98 times longer to retrieve all similarity results from the index models creating 768-dimensional embeddings and 2.7 times longer than SBD_{ML}, which is the only model creating 512-dimensional embeddings.

Table 5.3.: Retrieval time from WW2 indices in reference to Elasticsearch

Size Group	768 Dim Models	512 Dim Model	Elasticsearch
Retrieval Time	50.5 %	36.9 %	2832.80 $s = 100$ %

Table 5.4 shows the retrieval time for the NSU data set. The smaller model SBD_{ML} had the fastest retrieval time, taking only 73.4% of the time that the other models took on average for retrieval. Both models retrieved faster than Elasticsearch, which took almost three times as long to retrieve all results than the larger models and almost four times as long as the smaller model.

Table 5.4.: Retrieval time from NSU indices in reference to Elasticsearch

Size Group	768 Dim Models	512 Dim Model	Elasticsearch
Retrieval Time	35.03 %	25.7 %	715.65 $s = 100$ %

The index retrieval time again shows that smaller models are preferable when considering scaling. The time savings of retrieval will be much more significant when scaling the

results to the size of usual investigative corpora. Elasticsearch seems to be performing much worse in retrieval than the models, it has to be stressed however, that Elasticsearch is not optimized for the task of bulk retrieval. More so, it is able to have fast enough responses to single MLT-requests that these can be queried in production environments. For the context-embedding based models, a similarity search engine would require to pre-calculate and store similarity results for documents inside the corpus in order to have slow response times for single request queries as searching for similar documents to one document is only marginally faster than searching for similar documents to a batch of 1000 documents.
Therefore it cannot be concluded that context-embedding based models perform better than Elasticsearch regarding retrieval time when considering the use-case of a user, who wants to know the top k most similar documents to a document from the corpus.
The conclusion from the results regarding the retrieval times is rather that building a similarity search engine for a corpus using an index based on Faiss containing context-based embeddings, requires much more pre-processing than only creating the index, as the similarity relationships of all documents of the corpus has to be calculated and stored beforehand. In contrast, Elasticsearch only requires the additional storage of term-vectors in order to speed up response time for MLT-queries. Therefore, the effort of a context-embedding based similarity search functionality for a corpus is larger than that of Elasticsearch due to the nature of result generation using context-embeddings and Faiss.

5.2. Disk space

Disk space is an import evaluation criterion, since very large data sets need to be stored as efficiently as possible, so that data set does not waste valuable storgae space When data sets become very large, an efficient scaling of disk space is essential. As all models which share the same embedding vector dimension also share an index size, the tables regarding disk space comparison were constructed for size groups instead of for models. Table 5.5 shows the disk space required to store indices of the WW2 data set. Elasticsearch needs the least disk space to store the index. The small model already requires 1.81 times more disk space while the larger models require 2.72 times as much disk space to store the index.

Table 5.5.: Disk Space required to store a WW2 Index in reference to Elasticsearch

Size Group	768 Dim Models	512 Dim Model	Elasticsearch
Disk Space	2.72	1.81	1(232.4[MB])

Table 5.6 shows the disk space required to store indices of the NSU data set. The WW2 data set is 2.53 times larger, but requires 3.43 times more disk space than the NSU data

Table 5.6.: Disk Space required to store an NSU Index in reference to Elasticsearch

Size Group	768 Dim Models	512 Dim Model	Elasticsearch
Disk Space	3.68	2.46	1(67.7[MB])

set. The proportions between the Elasticsearch-index and the context-embedding-based indices are also larger.

For the smaller data set, Elasticsearch seems to be able to compress the data more than for the larger data set. However, it cannot be concluded from that, how scalable Elasticsearch actually is. From the fact that it is a widely used professional software, it can be assumed that Elasticsearch aims to be optimized in that regard. The proportions between disk space required for storing indices with 768-dimensional and with 512-dimensional vectors are roughly 1.5. The overhead for storing the context-embedding-based indices is very little. However, the raw vectors are stored on disk without any optimization. Therefore, it is to be expected that a larger data set such as Football Leaks would grow approximately linearly. Storing a data set as big as football leaks would require roughly 1.5 Terabyte of disk space.
It has to be mentioned though that the index can theoretically be deleted after acquiring the similarity information and storing only the IDs of the top k most similar documents for each document of a corpus. The disadvantage of this approach however is, that data sets could not be expanded easily. The approach would make a more efficient storage of similarity information rendered from context-embedding possible. Ironically, a more efficient way of storing the information would be in Elasticsearch. It has to be stressed that the created Elasticsearch-index for a data set is capable of various search and analytic functionalities which are desirable for investigative journalism, which explains its prevalence in the available tools.
The Elasticsearch-index is more optimized for long-term storage and provides journalists with more versatile functionality. The context-embedding based indices can theoretically be deleted after rendering similarity information from them. However as it scales so poorly generating context-embeddings for documents in order to extract similarity information about these documents is a very resource-intensive approach regarding disk space.

5.3. Ranking results

The ranking results are the main performance evaluation criterion. This is based on the heuristic that the target document shares a larger similarity with the test document than most documents of the data set, since they are constructed from the same section of an input document. First, it is evaluated, how often the models were able to place the target document into its top 200 most similar documents, as well as into the top 10. The top 10 of search results is a typical result length for the first page of search results which are presented to the user. The top 200 search results were chosen to have a reasonably large

range of results to account for the approximate similarity of the search document and the target document. Ranking all documents in the corpus is computationally not feasible. Afterwards the average ranking of the target document is evaluated for the cases where the target document was found.

Table 5.7 shows how often the different models were able to place the target documents of the WW2 data set into the top 200 and the top 10 of search results respectively. The mean-based model BU_E places the target document into the top 200 and into the top 10 most often. Additionally, in 14.62% of cases it has ranked the target document pair on the first rank, which is also the largest value for all models and is 3.93 times more often than Elasticsearch.
In general, the mean-based models placed the target document into the top 200 and top 10 more often than the CLS-based models. BC_E with mean tokens placed the target document the second most often and the third most placings of the target document were achieved by the smaller, multilingual model SDB_{ML}. This is also the best result of all SentenceBert based models. The worst results were achieved by the RoBERTa-based models with all the document embedding creation methods. In 99.2% of cases, the CLS-based RB_E did not rank the target document in the top 200 most similar documents. Elasticsearch ranked the target document in its top 200 in less than half of test cases. This is significantly lower than the three most highest performing models regarding ranking.

Table 5.7.: Top 200 placings and Top 10 placings of target document for WW2 models

Document Embedding	Model	In Top 200	In Top 10
CLS	BC_E	14.03 %	4.46 %
	BU_E	52.22 %	24.12 %
	BC_{ML}	24.38 %	8.51 %
	R_E	0.08 %	0.002 %
Mean	BC_E	69.52 %	35.78 %
	BU_E	74.97 %	40.6 %
	BC_{ML}	49.39 %	23.04 %
	R_E	0.03 %	0.01 %
SentenceBert	SBB_E	60.94%	26.52 %
	SBR_E	38.42 %	14.64 %
	SBD_{ML}	66.92 %	31.56 %
Elastic	-	45.71 %	17.16 %

Table 5.8 shows the mean ranking of the target document, if it was found in the first place. It shows that the models with the highest overall average placings of the target document into the top 200 from table 5.7, also placed the target document higher on average. The standard deviation σ is very high for all models.

Table 5.8.: Ranking Results of Query Term for WW2 models

Document Embedding	Model	Mean Rank *if found*	σ
CLS	BC_E	52.2	56.1
	BU_E	36.9	49.0
	BC_{ML}	49.2	54.7
	R_E	-	-
Mean	BC_E	31.6	45.3
	BU_E	29.0	43.4
	BC_{ML}	36.3	48.4
	R_E	-	-
SentenceBert	SBB_E	38.2	48.9
	SBR_E	45.8	53.5
	SBD_{ML}	34.5	46.9
Elastic	-	41.3	49.1

Table 5.9 shows how often the different models were able to place the target documents of the NSU data set into the top 200 and the top 10 of search results respectively. Here it can be seen, that SBD_{ML} is the second best performing model regarding the placing of the target document in the top 200 and the third best performing model regarding placing the target document into the top 10. The best performing model is the mean-based monolingual model BU_G. Elasticsearch performs worse than 4 other models but the differences in placement rates are significantly lower than for the WW2 data set. All models perform worse in placing than comparable models for the WW2 data set.

Table 5.9.: Top 200 placings and Top 10 placings of target document for NSU models

Document Embedding	Model	In Top 200	In Top 10
CLS	BC_G	22.9 %	7.3 %
	BU_G	15.9 %	5.0 %
	BC_{ML}	9.5 %	2.8 %
	BU_{ML}	11.9 %	3.7 %
Mean	BC_G	43.3 %	18.3 %
	BU_G	49.2 %	21.3 %
	BC_{ML}	26.6 %	10.2 %
	BU_{ML}	39.3 %	16.0 %
SentenceBert	SBD_{ML}	43.8 %	17.5 %
Elastic	-	39.1 %	16.2 %

Regarding the mean rank of the target document , which is shown in table 5.10, the mean rank is generally lower than for the WW2 data set when comparing the best performing models. The best performing model, which is the mean-based BU_E has an average ranking of 29.0, which is 21.3 ranks lower than Elasticsearch. . The standard deviation is equally high. The best performing embedding-based model shows results which are 10 percentage points higher than Elasticsearchs' results, which ranks the target document lower than all other models.

Table 5.10.: Ranking Results of Query Term for NSU models

Document Embedding	Model	Mean Rank *if found*	σ
CLS	BC_G	53.7	56.5
	BU_G	54.6	57.2
	BC_{ML}	58.5	58.7
	BU_{ML}	56.1	58.0
Mean	BC_G	42.4	56.7
	BU_G	40.8	51.5
	BC_{ML}	47.1	54.8
	BU_{ML}	44.2	53.3
SentenceBert	SBD_{ML}	43.5	52.3
Elastic	-	40.9	50.1

The results show that there is a big difference in performance between methods for creating document embeddings. For every language model, the CLS-based document-embeddings seem to contain less semantic information about the document than the mean-based document-embeddings. The extremely poor performance of RoBERTa-embeddings from the flair-model R_E and the SentenceBert-model SBR_E are very surprising, as the language model is architecturally identical but trained with more data for longer periods than the classic BERT. It outperforms the classic BERT model in other language tasks [19]. The extremely low performance of RoBERTa-based models are most likely due to an implementation error with the Flair-models. But the SentenceBert-based RoBERTa model also performed significantly worse than not only the other SentenceBert models, but also than the multilingual mean-based BERT model. This is very surprising, especially considering that the multilingual Flair-based models generally did not perform very well. This result also does not reflect the results from the author of SentenceBert, which showed a slightly better performance on the STS-benchmark for semantic textural similarity for SBR_E.
The fact that the BERT development team recommends using the cased multilingual model BC_{ML} over the uncased model [1] did not reflect in their performance on this task. The performance of the uncased model was around 50 % better, though both models did

[1] for reference, see `https://github.com/google-research/bert/blob/master/multilingual.md`, visited 12.08.20

perform worse than the monolingual Flair-based models.
When evaluating the Flair-models, unsurprisingly, the monolingual models perform better than the multilingual models. This is due to the decreased language domain for which the language models are trained to create adequate embeddings. Especially the English models are high performing, but relative to the multilingual models, the monolingual German models are also performing better. The better results for the English data are most likely due to better trained English models. BERT was originally a monolingual English language model and for most training scenarios, English data sets outsize data sets in other languages.
Another reason for the difference in performance between the data sets might be the fact that the semantic difference between documents in the WW2 data set is much larger than in the NSU data set. The NSU data set is smaller, and all documents are about the same, very specific topic. This might have made it harder for the language model to identify a semantic difference. The performance of Elasticsearch is also decreased for the NSU data set, though not as drastically. This might be due to the nature of the MLT-query, which incorporates the *tf-idf* of all terms into the decision for what to search. As close paragraphs are not only more likely to be semantically similar, but also to contain the same very specific words, such as locations or names, the difficulty to find the target document might have not increased by much for Elasticsearch.
The large spread for all models is reasonable and underlines that the heuristic nature of the test data sets only give an approximate performance metric about the models.
The most surprising result from the evaluation was, that one of the best performing models, which performed very close to the highest performing model for the German data set, was the smaller multilingual SentenceBert model based on DistilBERT SBD_{ML}. Not only did it perform well on both data sets, it is also the only multilingual model with good performance.

The ranking results from retrieval show that context-embedding models are capable of providing similarity information for small documents. One problem is however, that the sequence length for BERT-based embeddings also limit the document length for which the creation of document embeddings is possible. The very nature of investigative data sets being large and unstructured makes pre-processing of documents as performed in the construction of the here used test data sets necessary. This limits the possibilities of using the powerful context-embeddings for words for large documents.
As for the creation of context-embeddings for documents, it could be shown, they provide vastly different information regarding similarity of documents. As expected the best performing models were monolingual. It was however very surprising to see the significantly smaller and faster multilingual model SBD_{ML} not outperform many of the Flair-based models, but also other monolingual SentenceBert-based models.
The performance of the monolingual mean-based context-embedding models was better, but regarding the requirements of working with large, unstructured text corpora, the results of SBD_{ML} are more promising when aiming to incorporate the findings into the actual work of journalism.

5.4. Co-occurrence

The co-occurrence between Elasticsearch and the models was evaluated in order to see, whether there is a difference between the results of the models and Elasticsearch, independently of the target document.
Table 5.11 shows the co-occurrences between context-embedding ranking results and Elasticsearch. The three models with the three highest co-occurrences with Elasticsearch are the models which also have the highest mean rank. SBD_{ML} has the most co-occurrences with Elasticsearch, followed by the mean-based BU_E and BC_G. These are also the highest performing models regarding the task to find the target document.

Table 5.11.: Co-occurrence of ranking results for WW2 data set

Document Embedding	Model	Mean Co-occurrence	σ
CLS	BC_E	5.95	6.37
	BU_E	17.15	12.90
	BC_{ML}	9.72	9.00
	R_E	-	-
Mean	BC_E	29.06	18.07
	BU_E	32.01	19.10
	BC_{ML}	21.27	16.34
	R_E	-	-
SentenceBert	SBB_E	26.95	17.46
	SBR_E	15.95	12.82
	SBD_{ML}	33.87	20.14

Figures 5.1 and 5.2 show the distributions of co-occurrence values for the Flair-based models used for the WW2 data set. It can be seen, that the distribution shifts to the right for the mean-based models compared to the CLS-based models. There are higher co-occurrences for the mean-based models than for the CLS-based models. The Roberta based models in both cases show almost no co-occurrence with Elasticsearch.

Figure 5.3 shows the co-occurrence between the SentenceBert based models and Elasticsearch. We see that the best performing model regarding the task of finding the target document has a larger co-occurrence that the poorer performing models. The largest co-occurrence was found between the smaller, monolingual model SBD_{ML} and Elasticsearch.

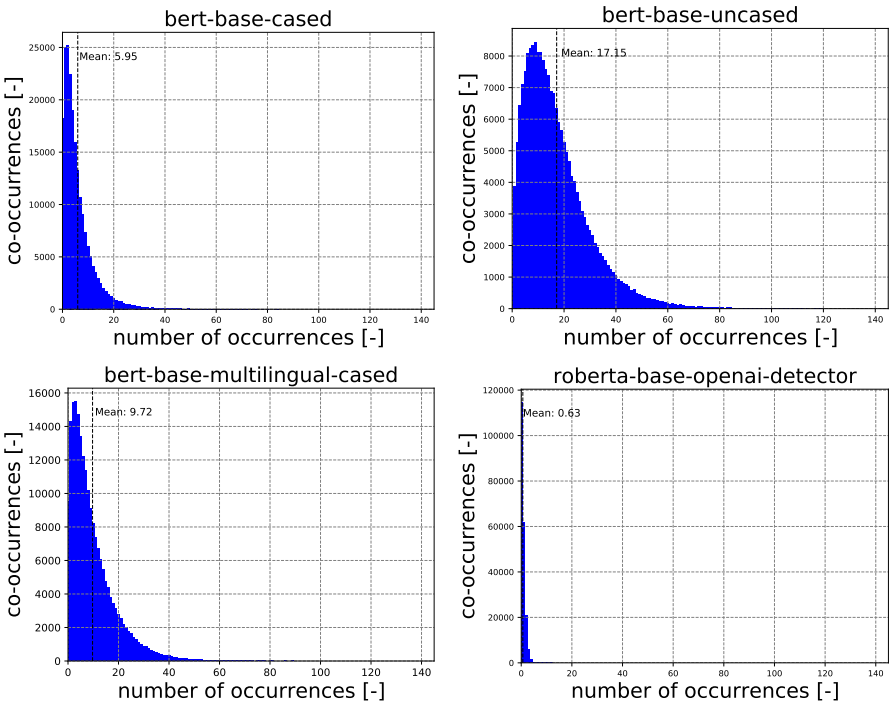

Figure 5.1.: Co-occurrences of CLS-based models for WW2 data set

It can be seen for the WW2 data set, that there is a difference in co-occurrence between the models and that the better performing models have a higher co-occurrence with Elasticsearch than the weaker performing model.

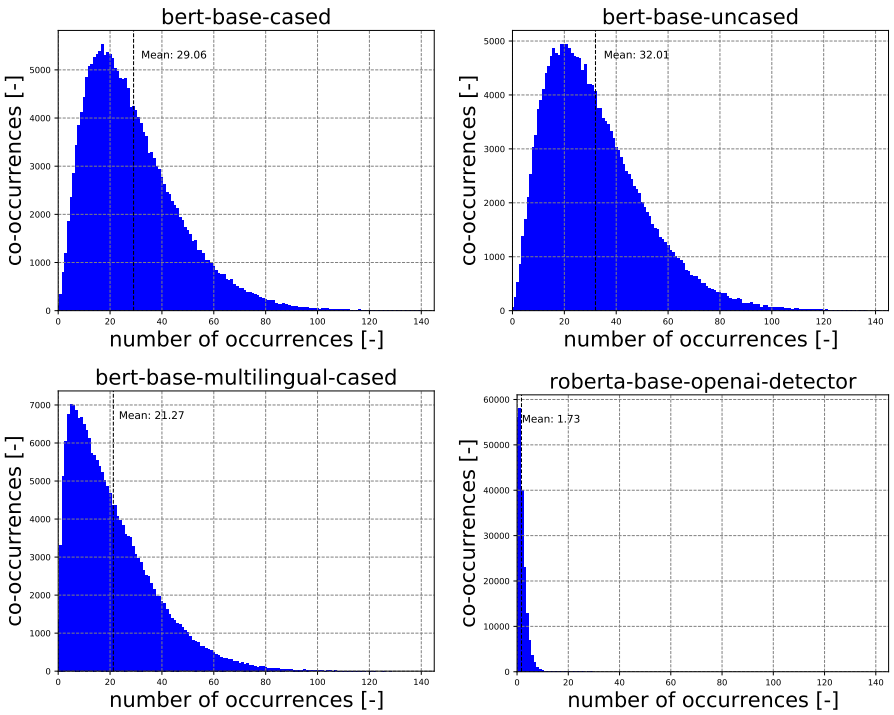

Figure 5.2.: Co-occurrences of mean-based models for WW2 data set

Table 5.12 shows the co-occurrence between ranking results of context-embedding indices and the Elasticsearch index. The model with the highest co-occurrence between itself and Elasticsearch is SBD_{ML}. For the mean-based models, the co-occurrence is significantly larger than the for the CLS-based models. The mean co-occurrence was higher for the monolingual German Flair-models than for the multilingual Flair-models.
The highest co-occurrences between Elasticsearch and a model are that of SBD_{ML} and the mean-based BC_G which are also among the highest performing models regarding the task to find the target document.

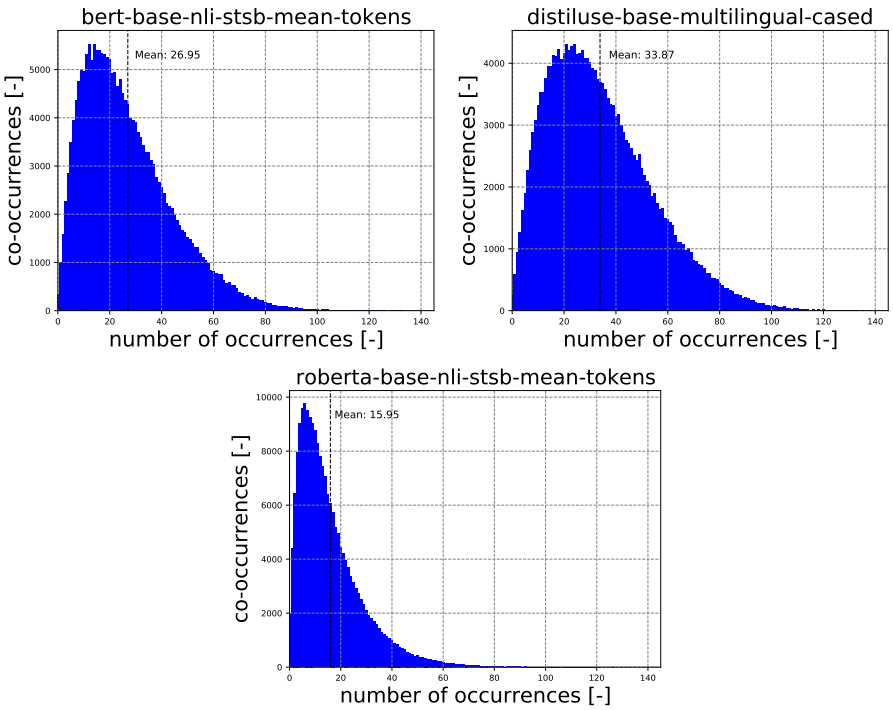

Figure 5.3.: Co-occurrences of SentenceBert-based models for WW2 data set

Table 5.12.: Co-occurrence of ranking results for NSU data set

Document Embedding	Model	Mean Co-occurrence	σ
CLS	BC_G	11.29	16.86
	BU_G	9.12	13.81
	BC_{ML}	6.64	10.98
	BU_{ML}	7.96	12.95
Mean	BC_G	24.41	24.23
	BU_G	23.48	24.75
	BC_{ML}	16.09	21.20
	BU_{ML}	22.17	25.38
SentenceBert	SBD_{ML}	24.99	23.54

The figures 5.4 and 5.5 show the distributions of co-occurrence values for the Flair-based models used for the NSU data set. These figures underline the differences between the document embedding creation strategies. While there is generally less co-occurrence between Elasticsearch and the models for the NSU data set than for the WW2 data set, the CLS-based models for the NSU data set all show little co-occurrence with

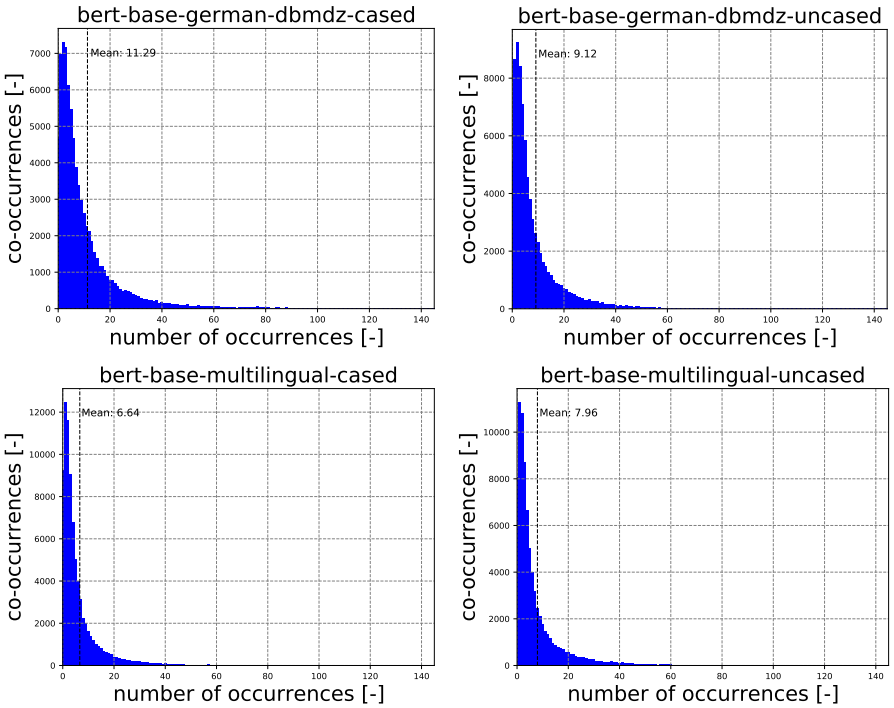

Figure 5.4.: Co-occurrences of CLS-based models for NSU data set

Elasticsearch. Figure 5.6 shows the distribution of co-occurrence values between SBD_{ML} and Elasticsearch. Compared to the other models, the spread is wider shifted to the right, visualizing the larger co-occurrence with Elasticsearch.

The results from the co-occurrence give weight to the assumption that better performing models are somewhat more aligned with Elasticsearch, which we assume is generally capable of finding similar documents. However there are also differences in the results. The average co-occurrence of the best performing models for both data sets is below 20 %. Thus the differences in results between the term-based and the best performing context-embedding-based similarity search is still very large. This might be only partly due to different capabilities regarding semantic similarity search. It will also be partially due to the fact that the top 200 also contains a lot of documents which are not similar to the search document. Since one cannot conclude a priori that the results generated by Elasticsearch are "better", a qualitative analysis of a selection of search results needs to be conducted.

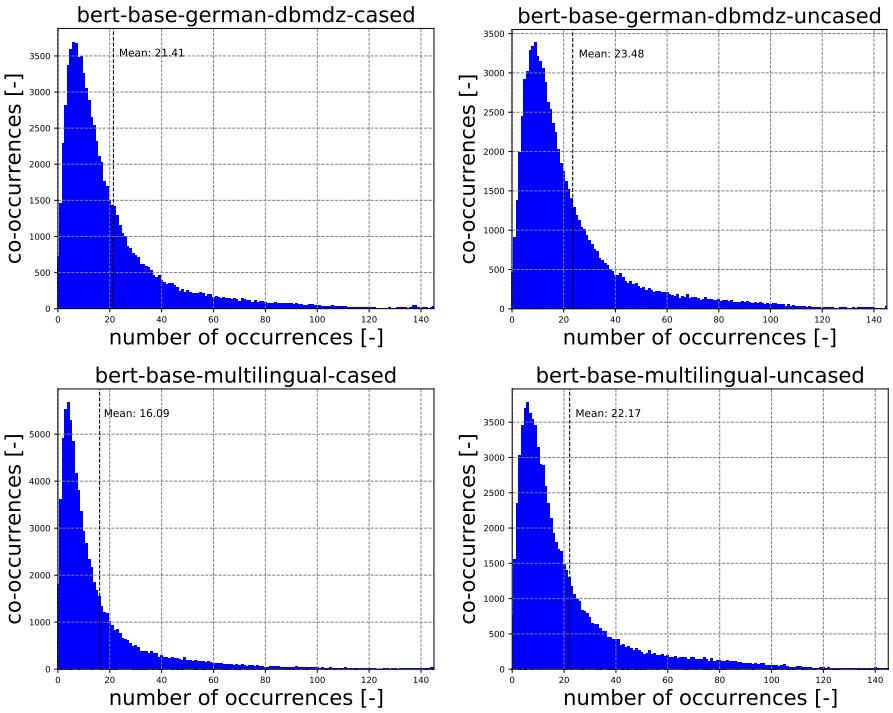

Figure 5.5.: Co-occurrences of mean-based models for NSU data set

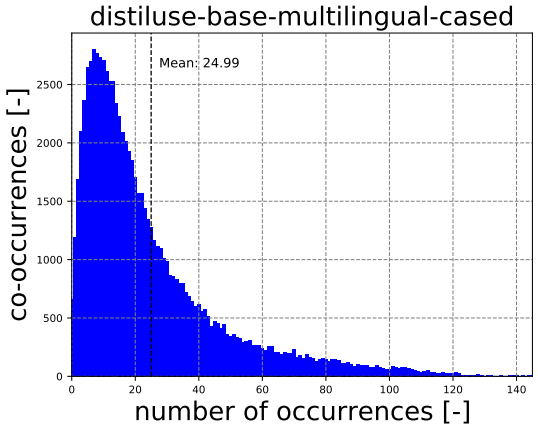

Figure 5.6.: Co-occurrences of SentenceBert-based model SBD_{ML} for NSU data set

5.5. Qualitative analysis

For the selection of test cases, the co-occurrence of top 10 rankings was analyzed. The results can be seen in the appendix A.1. Although the quantitative analysis gives a good estimation about the performance of the model, a qualitative analysis give the possibility to analyze whether there is actually a difference in quality.
To that end many single cases were analyzed, some of which can be found in the appendix. There are many examples were the co-occurrence between good performing models and Elasticsearch is very high. These were easily rankable documents, that mostly did not contain any full text, but extracted data from tables or lists. The more interesting examples were those, where there were large differences between Elasticsearch and the models. The analysis here is narrowed down to a few noteworthy examples from which an interesting insight can be obtained. This is due to the exhaustive size of the data set and the size of the results for single queries.
The qualitative analysis is divided into two parts for each data set. The first part highlights some select examples and showcases some significant differences between Elasticsearch and some models.
The second part is an evaluation of the data set, regarding its suitability for similarity search.

5.5.1. Qualitative analysis of WW2 data set

The WW2 data set used the generally more powerful English language models. Further it covers a large range topically.
The WW2 data set was suitable for the task due to its size and the vast spectrum of topics that it actually contained. This made it possible to have results that were obviously false when analyzing individual search documents. Also, it possibly made it somewhat easier to find the target document, as there are less documents that are semantically similar, when there are lots of topics covered. However, since there are some prevalent topics it still was able to show some of the shortcomings of Elasticsearch, which stem from its term-based similarity search.
An example document which shows the differences in evaluating similarity between context-embeddings and the term-based approach of Elasticsearch is the search document with ID 7762:

Discipline was harsh in the armed forces, and the lash was used to punish even trivial offences—and not used sparingly. Despite the harsh discipline, a distinct lack of self-discipline pervaded all ranks of the British forces. Soldiers had an intense passion for gambling, reaching such excesses that troops would often wager their own uniforms. The British leadership soon discovered it had overestimated the capabilities of its own troops, while underestimating those of the colonists, causing a sudden re-think in British planning. Debate persists over whether a British defeat was a guaranteed outcome. Ferling argues

that the odds were so long, the defeat of Britain was nothing short of a miracle.

The document contains information about the lack of discipline in armed forces and the fear of British leader that a British defeat in the American War of Independence is a possibility. The target document contains some more information about the shortcomings of British troops and the failure of British commanders to achieve a victory over the Americans. It also contains information about Hessian soldiers, which are not mentioned in the target document:

The presence of Hessian soldiers caused considerable anxiety among the colonists, both Patriot and Loyalist, who viewed them as brutal mercenaries. During peacetime, the Army's idleness led to it being riddled with corruption and inefficiency, resulting in many administrative difficulties once campaigning began. Historians such as Ellis and Stewart have observed that, under William Howe's command, the British squandered several opportunities to achieve a decisive victory over the Americans. During planning for the Saratoga campaign, Howe was left with the choice of committing his army to support Burgoyne, or capture Philadelphia, the revolutionary capital. Howe decided upon the latter, determining that Washington was of a greater threat.

The top result of Elasticsearch however, does not revolves around capital punishment in the military:

Capital punishment in Canada was abolished generally in 1976, and for military offences in 1998. Harold Pringle was the last Canadian soldier executed, in 1945, for a military offence. Finland In Finland, the military has jurisdiction over two types of crimes: those that can be committed only by military personnel and those normal crimes by military persons where both the defendant and the victim :§ 2 In crimes where the military has jurisdiction, the military conducts the investigation. In non-trivial cases, this is done by the investigative section of Defence Command or by civilian police, but trivial cases are investigated by the defendant's own unit. The civilian police has always to right to take the case from the military.

Other documents in Elasticsearchs' top results contain words from the search document, but do not share any connection topically. The document on the third rank is about crime in Bangkok, for example.

An estimated 200,000 medical tourists visited Thailand in 2011, making Bangkok the most popular global destination for medical tourism. Crime and safety Bangkok has a relatively moderate crime rate when compared to urban counterparts around the world. Serious offences included 183 murders, 81 gang robberies, 265 robberies, 1 kidnapping and 9 arson cases. Offences against the state were by far more common, and included

54,068 drug-related cases, 17,239 cases involving prostitution and 8,634 related to gambling.

It is evident, that is was ranked due to the shared vocabulary. However, there is hardly any connection between the documents. Elasticsearch failed to rank the target document. The best-ranked document of mean-based BU_E also is not the target document, but a document about the conflict between British elites and the the American colonists after a war.

This was a successful wartime strategy but, after the war was over, each side believed that it had borne a greater burden than the other. The British elite, the most heavily taxed of any in Europe, pointed out angrily that the colonists paid little to the royal coffers. The colonists replied that their sons had fought and died in a war that served European interests more than their own. Many had never been to Britain, yet they imitated British styles of dress, dance, and etiquette. This social upper echelon built its mansions in the Georgian style, copied the furniture designs of Thomas Chippendale, and participated in the intellectual currents of Europe, such as the Enlightenment. The seaport cities of colonial America were truly British cities in the eyes of many inhabitants.

It is hard to argue for a semantic similarity of the document and the search document, other than a vague connection of a conflict between British and American people. It is equally hard to find explanations for the placements of context-embedding based results. The model ranked the target document on the third place, although it appears to be more similar. The document ranked on the fourth place is dealing with a relatively similar topic, but is from a different input document altogether, dealing with the topic of World War I instead of the American Independence War.

The military leadership of the British Army during the World War I was frequently condemned as poor by historians and politicians for decades after the war ended. Common charges were that the generals commanding the army were blind to the realities of trench warfare, ignorant of the conditions of their men and were unable to learn from their mistakes, thus causing enormous numbers of casualties that they could not control such as a lack of adequate military communications, which was not known before. Furthermore, military leadership improved throughout the war culminating in the Hundred Days Offensive advance to victory in 1918. Some historians, even revisionists, still criticize the British High Command severely, but they are less inclined to portray the war in a simplistic manner

The document also deals with shortcomings of the British army during a war. This is a glimpse into the possibilities which may arise from document-embeddings which are more accurately representing the semantic context of a document. There is a semantic relation between the documents which is not found by merely analyzing the contained terms.
Regarding the top ten results of mean-based BU_E, the best performing model for the query document the conclusion is that the rankings are not reliable yet. However, there

is a great potential for similarity search systems which are capable of overcoming the shortcomings of term-based search systems.
This shortcoming is very evident when analyzing the search ranking results for the document with ID 201. It is an excerpt of the Wikipedia-article about using the term fascist as an insult:

In response to multiple authors Possible explanations for casual uses They employ massive overkill strategy, there are 30, 20 to 30 marshals daily inside the courtroom, it has the atmosphere of an arms camp, the law against us is rigged . . . and our claims that this law violates our constitutional rights and it's the same way that we claim that Mayor Daley didn't have the right to deny us a permit t social structure, though this kind of Maoist or Guevarist analysis often underpinned the rhetorical depiction of Cold War authoritarians as fascists.Some Marxist groups – such as the Indian section of the Fourth International and the Hekmatist groups in Iran and Iraq – have provided analytical accounts as to why the term "fascist" should be applied to groups such as the Hindutva movement, the 1979 Islamic Iranian regime or the Islamist sections of the Iraqi insurgency. Other scholars contend that the traditional meaning of the term fascism does not apply to Hindutva groups and may hinder an analysis of their activities. See also References External links

Both the search document and the target document are dealing with the question whether some group can be described as fascist.

In 2014, with the outbreak of the war in Donbass the Russian nationalists and media returned to the "fascist" rhetoric, frequently describing the Ukrainian government after Euromaidan as "fascist", "Nazi" etc. to march or to assemble in the park. . . . I think it points a direction in the future which is that the government embarked on a course of fascism.Several Marxist theories back up particular uses of fascism beyond its usual remit. For instance, Poulantzas's theory of state monopoly capitalism could be associated with the idea of a military-industrial complex to suggest that 1960s America had a fascist social structure

However, the top ranking result for Elasticsearch is the beginning of the Wikipedia article about Operational research.

Operations research Operations research, or operational research in British usage, is a discipline that deals with the application of advanced analytical methods to help make better decisions. Further, the term 'operational analysis' is used in the British ...

This is most likely based on just a few terms, mainly *analysis* being selected as a term for the search query of Elasticsearch. Most likely, the word fascist did not even make it into the terms which Elasticsearch selected, as the term might be so prevalent in the data set that Elasticsearch did not evaluate it as being useful for identifying similar documents,

due to its low idf. Elasticsearch was not able to place the target document into its top ten, unlike mean-based BU_E which placed it on the ninth rank. The top ranking result is from a different part of the data set (ID 79135), but there is a strong semantic connection to the search document:

... Criticism of theory of a link between Islam and fascism While Islamic Fascism has been discussed as a category of serious analysis by the scholars mentioned above, the term Islamofascism circulated mainly as a propaganda, rather than as an analytic, term after the September 11 attacks on the United States in September 2001 In his diagnosis of this shift he detected a decline in the old liberal consensus of American politics, and what he called the "deliquescence of the Democratic Party"...

It also deals with the discussion of whether describing a group of people as fascist is appropriate. This is again a strong semantic link between the search document and the top ranking document.
Clearly, this result is much more appropriate than Elasticsearchs' result.

5.5.2. Qualitative analysis of NSU data set

The NSU data set contained generally fewer co-occurrences and was topically more focused.
The NSU data set had some issues with the very limited sequence length due to the language used. Nonetheless there are many document pairs which share a semantic similarity. Another issue were the parsed footers and headers of the document which are embedded in many of the documents. It is very likely that this is not an issue for Elasticsearch, since the term-based similarity search is based on *tf-idf*, so that these repeatedly occurring footers and headers will most likely not have affected the scores significantly. The effect on the embeddings may be bigger, as they are generated sequence-by-sequence and therefore do not contain the global information about repeatedly occurring "noise".
In conclusion, the results from the data set were able to underline the results of the WW2 data sets and showed that in order to use context-embeddings for a similarity search of general documents, the maximum sequence length has to be larger, or another approach for generating document embeddings has to be used.
Similarly to the WW2 data set, there were some occasions, where documents did not contain any textual information but filenames or paths. These were the documents where even the worst-performing models were able to find the target document.
An example where there were no co-occurrences between Elasticsearch and the model and where Elasticsearch was able to rank the target document on the highest rank and the highest ranking model, mean-based BU_G, rank the target document in its top 10, whereas the multilingual model SBD_{ML} fails to place the target document in its top 10 search results, revolves around the following test case with the document ID 52096:

Informationssystemen bereits erfolgt und negativ verlaufensei. In einem mit „LB“ – für Ludwigsburg – bezeichneten Unterordner habe er, so der Zeuge KHK J. G., zwei Kontrollstellenlisten, nämlich eine von der Kontrollstelle aus Mundelsheim und eine aus Oberstenfeld, feststellen können, die allerdings denselben Dateinamen gehabt hätten. Die Excel-Dateien seien ansonsten aufgrund eines Programms zur automatisierten Erfassung in die Bearbeitungssoftware „CRIME“ übermittelt worden. Als er, so der Zeuge KHK J. G., die Listen weiter ausgewertet habe, habe er beim Filtern der Gesamtliste „Kennzeichen“ festgestellt, dass im gesamten Fahndungsraum am 25. April 2007 unter den ca. 33.000 Fahrzeugkennzeichen nur sechs Wohnmobile verzeichnet gewesen seien. Darunter habe sich kein Wohnmobil aus Zulassungsbezirken in Thüringen befunden.

The document can be summarized by two statements. First, there was a problem with two files sharing a filename and second, the search for a RV-vehicle with the list of license plate did not give any results.
The target document also deals with the search for an RV and mentions corrupted filenames:

Aufgrund des fehlerhaften Dateinamens sei diese Übertragung für die Ringalarmdaten aus Oberstenfeld zunächst nicht erfolgt. In der Excel-Tabelle und der Halterlistentabelle seien diese Daten jedoch vollständig erfasst gewesen. Aus sächsischen Zulassungsbezirken sei nur das Wohnmobil mit dem Kennzeichen C-PW 87 von C. H. mit ihrem Geburtsdatum, ihrem Geburtsort und einer Chemnitzer Adresse vermerkt gewesen, die restlichen fünf Wohnmobile seien in anderen Bundesländern zugelassen gewesen. Anhand der Halterdaten des Kraftfahrt-Bundesamts sei nicht zu erkennen gewesen, dass dieses Wohnmobil aus Chemnitz auf ein Unternehmen,

The target document also shows another characteristic of the German NSU data set, which consists solely of documents in officialese language. Sentences tend to be longer than the maximum paragraph length. That made is generally more likely that a sentence ending symbol was found during paragraph splitting.
Nonetheless, Elasticsearch was not only able to place the target document in its top 10 but on the highest rank. Moreover, the top three are results from the immediate vicinity, dealing with the same topic. All documents in the top nine contain the word *Wohnmobil (RV)*, which implies that the word was put in the Elasticsearch search query as a document-defining word.
The mean-based BU_G placed not the target document, but the preceding document on the first rank. The document also deals with the same topic, describing the problems with excel-files when searching for a the vehicle. Notably, it does not contain the word *Wohnmobil* and was not placed into the top 10 of Elasticsearch while certainly being relevant. This exemplifies the importance that terms have in the term-based similarity search of Elasticsearch. The absence of the term, which was most likely very relevant for the ranking of Elasticsearch, resulted in a ranking outside of the top 10. The mean-based

BU_G placed the target document on the third rank, therefore it would still have been visible to a potential user. The preceding document is:

Er, so der Zeuge KHK J. G., habe an seinem dritten Arbeitstag bei der Soko „Parkplatz“, dem 9. November 2011, auf deren Laufwerk erstmals die Auswerteergebnisse zur Ringalarmfahndung gesichtet und dabei Folgendes festgestellt: Mit der Erfassung und Auswertung aller Kontrollstellenlisten sei im August 2010 unter der Bezeichnung „Maßnahme 328“ durch den Abschnitt „Operative Auswertung“ der Soko „Par ers „Kontrollstellen“ habe er, so der Zeuge KHK J. G., eine mit „Kennzeichen“ bezeich-nete Gesamtliste im Excel-Format festgestellt. In diesem Ordner sei auch eine Tabelle gespeichert gewesen, in welche bereits die ermittelten Halterdaten der Fahrzeuge eingebunden gewesen seien. Auf Nachfrage sei ihm mitgeteilt worden, dass eine Überprüfung der Halterpersonalien in den polizeilichen Informationssystemen

The multilingual model SBD_{ML} failed to place the target document. The highest ranking document revolve around documents and lists. One can see a connection to the query document, but it is very far fetched. The second highest ranked document for example is about another list which is also referred to as *Garagenliste*. However, the more semantically similar results of the monolingual model were not found, nor was the target document or other documents from its vicinity.

problematik mit GBA bzw. OLG München kämen und was sie „halt vielleicht auch aushalten“ müssten. Um Erklärung gebeten, weshalb die aus dem Jahr 1998 stammende Telefonliste des NSU bzw. „Garagenliste“ erst derart spät, wohl 2012, an Baden-Württemberg übermittelt worden sei, verwies die Zeugin an Thüringen oder das BKA; sie selbst wisse es nicht. nummern stehen würden, diese Personen und Telefonnummern dann überprüftund dem BKA mitgeteilt worden seien, bejahte die Zeugin. Nach Vorhalt der Bewertung des mit der Auswertung der Adressliste befassten Kriminalhauptkommissars B. („Bei den weiterhin aufgefundenen Notizzetteln mit Adressen handelt es sich zum Teil um Adressen bekannter Personen der rechtsextremistischen bundesdeutschen Szene.

This example from the NSU data set shows issues with the data set, as well as with the multilingual model and Elasticsearch.
The data set contains many paragraphs which are slightly malformed, meaning that paragraphs were not split at paragraph ending symbols but in the middle of words. This reduces the semantic similarity of the created documents.
The example also shows that while the mean rank values of the target document did not differ greatly between SBD_{ML} and BU_G, there are still cases, where the monolingual model finds useful information, but the multilingual model does not.
The problem with Elasticsearch which the example highlights is its dependency on terms which makes it miss relevant information.

5.5.3. Conclusion of the qualitative analysis

The qualitative analysis tries to give a glimpse into the results, in order to see whether there are actual semantic links between search results and search documents and whether there is a difference in Elasticsearchs' search results and the ones from the top ranking context-embedding-based models. Due to the large number of results, the qualitative analysis can only give a few examples.
It could be shown, that there are problems with explaining the ranking of documents from context-embeddings-based models. The results of even the top ranking context-embeddings-based models are not always comprehensible, but there are some cases where there were able to find interesting results which may not have been found by term-based search systems. Moreover, it could be shown that Elasticsearchs' results were sometimes off as well. The shortcoming of Elasticsearch can most likely be mitigated by investing more time into fine-tuning term selection parameters, possibly also by customizing the list of stop-words or other words which could be excluded from possibly selected terms. This is however not helpful for the task of dealing with large and unstructured data sets. A system which relies on manual fine-tuning of search parameters is not well suitable for these kind of data sets.
This qualitative analysis showed that there are still lots of possible improvements, not only for the test data sets but also for the tested models.

5.6. Discussion of methods and results

5.6.1. Pre-processing and retrieval

It has to be stressed that the used methods for generating the data sets is based on a heuristic and therefore it cannot be guaranteed that the target document actually shares a significant similarity with the search document in all cases. The results show, that a large number of document pairs were retrievable and therefore the heuristic assumption can be assumed to produce somewhat similar documents relative to the number of documents in the corpus. The large difference in retrieval performance of the models and the results of the quantitative analysis suggest however, that the results from the heuristically created data sets are able to show the differences between the models and provide a indicator about the usefulness of semantic similarity search based on context-embeddings for the task of searching in large data sets.
When analyzing the generated documents more thoroughly, it was revealed that some of the documents were malformed, since the paragraph splitter was not able to find a sentence ending character. This can be seen in some examples in the appendix A.2. A better paragraph splitter might have been useful to generate semantically more similar document pairs.

5.6.2. Resource consumption

All models tried to utilize the resources from the GPU as much as possible and did not utilize other resources such as CPU or RAM extensively. Therefore there was no significant difference in resource consumption between the models. Since all tests were conducted on the same hardware, time and disk space appear to be good enough indicators of resource consumption. Other interesting tests could have been performed on different hardware, such as powerful cloud solutions. This would have made it possible to come up with a "price" for creating embeddings in a more timely manner.

5.6.3. Results

The rankings of the target documents are reasonable, due to the fact that the heuristic assumption cannot be expected to generate similar documents in all cases. They are reliable enough to see a trend in the capabilities of the models, especially since there are significant differences in the results from the models. Although labeled data sets would have produced more insightful results, the results based on the created test data sets allow to make statements about the models and the document creation methods, regarding semantic similarity.
The quantitative analysis was useful in order to see, which documents are ranked highly, regardless of the similarity of the target document. The results from the qualitative analysis confirmed the findings of the quantitative analysis that the documents with the higher average mean rank of the target document were able to find more similar documents. This also gives more weight to the results from the quantitative analysis.

6. Conclusion and outlook

The aim of this work was to evaluate approaches for a similarity search system for documents in large, unstructured data sets. Such data sets are common in investigative journalism. The motivation for this work was to find possible improvements for journalists to extract useful information from such data sets. After reviewing current solutions, it was evident that search tools which are used by journalists today all have the same term-based search engine, Elasticsearch, in common. Although search results of Elasticsearch are useful, their term based nature is not possible to account for the semantic similarity of different terms.
Novel language models which are based on training neural networks with very large data sets in order to produce vector representation of words which incorporate the semantic nature of words produce astonishing results in many NLP-tasks. The goal of this work was to research, whether search systems based on these so called *context-embeddings* can be helpful, by finding semantically similar documents which are not found by term-based approaches to search. Another important consideration was the resource consumption of the search system compared to Elasticsearch, which already is a professional and widely used tool. The problem of context-embeddings is that their creation is computationally expensive. Without a dedicated GPU, it is infeasible to create context-embeddings for even small data sets.
There are vast numbers of different language models for creating context-embeddings. They differ in their size, the size of the output vectors and also in their performance regarding common NLP-tasks. In this research, several models were tested to see, whether there are any differences between the models. An important distinction between models can be the language for which they are trained. Monolingual models, which are trained on the language of the used test data set, as well as multilingual models were tested. Regarding the aim defined at the outset, that the system should be able to process large, unstructured data sets, which might not be monolingual, multilingual models are more desirable as candidates for a search engine which is useful for the task.
An issue with using any of the language models for similarity search of documents is that their output vectors are representing words, not documents. There are several approaches to create document embeddings from word embeddings. Due to the fact that word embeddings are designed to capture only semantic meaning of the words they represent and not the sequence in which they are found, creating document embeddings from them does not necessarily capture semantic meaning about the document. This research tested three document embedding creation strategies.
The conducted research set out from two publicly available data sets, an English data set and a German data set. These data sets were pre-processed into suitable test data

sets. This was done, since there are almost no data sets which are suitable for this kind of similarity research, as manually assigning similarity labels to large data sets is time-consuming and labor-intensive. For testing similarity searches however, it is crucial to have knowledge about what documents from the data set are similar. Therefore a method for automatically creating test data sets consisting of approximately similar document pairs was worked out. The test data sets were constructed by splitting the input data sets into small paragraphs and creating test document pairs from concatenating alternating paragraphs. The work was conducted based on the heuristic that the resulting document pairs are likely semantically similar to a degree, so that a semantic similarity search should be able to place the document into its top 200 search results.
Based on these data sets, context-embeddings were created for each document in the data set and stored into a vector index using the open-source framework Faiss. This was conducted for each combination of language model and document embedding creation method. During this process the time was measured. Additionally an Elasticsearch index was created from the test data set, again measuring time. From these data, it was evaluated whether there is a significant time difference between models, document embeddings strategies and most importantly, Elasticsearch.
The results from this experiment show, that Elasticsearch is significantly faster to a point, where for many models it will not be feasible to create such indices of very large data sets. The difference between models was also significant, with SentenceBert-based models being the fastest in creation. The smallest model, which is based on DistilBERT SBD_{ML} was significantly faster during index creation than any other models. It was still about eight times slower than Elasticsearch. If indices were to be created of large data sets, it would most likely still be feasible to create indices using SBD_{ML}. Therefore regarding time consumption the conclusion is, that as of now, only smaller context-embedding-based models should be realistically used in the context of large, unstructured data sets, unless much more powerful hardware is available to the users.
For each created index, a similarity search was conducted for each document in the index, while measuring the time. For each search document, it was known, which target document should be found. The top 200 search results were queried for each document and it was evaluated, in how many cases the search systems were able to place the target document into its top 200, as well as into its top 10, which is a commonly used first page of search results. Further it was evaluated, what the mean rank of the test document was, if it was found.
The results for the finding rate and the mean target document rank were then compared to the performance of Elasticsearch regarding placing the target document, as well as the time consumption.
It could be shown that there are differences between the tested models. The best performing models were monolingual and were using the mean word embedding of the document as the method to create document embeddings. A model which was in the top three performing models was the smaller, multilingual model SBD_{ML}.
Regarding mean ranking and finding rate of the target document, Elasticsearch did not perform better than the best performing context-embedding based models. The time

consumption during retrieval seem to indicate that Elasticsearch is slower in retrieving data. However that is due to the test environment which is catered to performance improvements for the context-embedding based models which are achieved by batch-searching for similarity results. In real applications, Elasticsearch is fast enough to omit the pre-calculation of similarity results, while context-embedding based models require this pre-calculation in order to serve user requests in a timely manner. Therefore, it has to be stressed that although retrieving *all* results from the index at once is faster for the context-embedding based models, Elasticsearch does not require this calculation of results altogether.
Therefore the calculation of results for the context-embedding based models has to be seen as an extra step in processing data, which is resource-consuming.
Regarding resource-consumption, the disk space required to store the indices was also measured. Here it could be shown, that the used vector indices are stored very inefficiently in comparison to Elasticsearch indices. This is especially problematic when keeping large, unstructured data sets in mind, since required disk space might be well above a Terabyte for an index. However, since the index can be theoretically deleted after rendering similarity information from it, the problem might be manageable.
Since Elasticsearch was used as a *silver-standard*, to which context-embedding based models were compared, a co-occurrence between model results and Elasticsearchs' results was established. The co-occurrence as an evaluation criterion is also independent of the target-document and therefore was assumed to be a useful indicator for the performance of the context-embedding based models in comparison to Elasticsearch.
Since this research is based on a heuristic model of similarity, it is not guaranteed that the target document is actually the most similar document. To this end, some test documents and the results of the similarity search from Elasticsearch and the context-embedding based models regarding the test documents were qualitatively evaluated. The qualitative analysis was done with test documents, where there was no co-occurrence between Elasticsearch and one of the best performing models.
The results from the qualitative analysis show the limits of term-based similarity search, especially when used in large, unstructured data sets where manual fine-tuning to the content is not feasible. It also shows the potential of context-embedding based semantic similarity search by being able to find interesting documents which were not the target document and which were not found by Elasticsearch due to differences in used terms.
Meanwhile the qualitative analysis addressed one of the bigger challenges for context-embedding based search systems which is creating comprehensible results. It is easy to argue about good results but hard to establish why bad results were chosen. This is problematic when investing time and resources into developing a search system for large data sets.

Coming to a conclusion about the results of the conducted experiments regarding the research question formulated in the introduction of this thesis, it can be stated that semantic similarity search using context-embeddings for documents is possible. Moreover with SBD_{ML} a model was found which is a good candidate for creating an experimental

similarity search engine for journalistic work. Due to its smaller size, it scales better when working with large data sets and it performed well compared to other multilingual models.
To answer the question, whether term-based or embeddings-based similarity search fulfill the needs of journalists better, it can be stated that it is more likely that both search methods are able to provide different but useful information from the corpus. However, as semantic similarity search using context-embeddings is more resource hungry and it still in its infancy, its roll will more likely be that of an experimental feature for news agencies which can provide the necessary hardware. The differences in resource consumption are large. It remains to be seen, whether they might decrease in the future, as well as whether there are improvements in the embedding-representation of large documents. The short limit on the processable sequence length was already a limiting factor in designing the used test suites. The limitation will hinder the usefulness in the domain of large, unstructured data sets even more. Regardless, the potential of novel context-embeddings for semantic similarity search could be shown.
The master thesis leads to further interesting research question and potential applications which might prove useful in the coming times.

Outlook

The research described in this master thesis provided promising results regarding the use of context-embedding based search engines. A bottleneck for the assessment of its usefulness is the sparse availability of suitable test data sets.
Therefore it would be interesting to use a semantic similarity search system in investigative journalism in order to have users assess the usefulness of the results. Given enough time, it might be possible to create labeled test data sets from their input by finding out which document pairs are actually considered useful on a larger scale. With better test data sets regarding similarity, models could be fine-tuned to the task of document similarity search.
Document similarity itself is a topic worth researching more. Especially given the very limiting maximal sequence size of the models, this will be necessary to actually use context-embeddings as a semantic similarity search engine for larger documents. Research could be conducted, whether the methods to generate document embeddings from word embeddings can also be applied to the document embeddings of SBD_{ML}, in order to create longer document embeddings from embeddings of shorter documents or paragraphs.
Another aspect of context-embeddings which was not investigate in this work is fine-tuning. It could be worth to investigate, how fine-tuning affects the performance of the models and whether the time and resource investment is worth the potential improvement in retrieval performance.
Also, the retrieval performance of other method to generate documents embeddings can be researched using the same data sets as for the experiments in this thesis. A suggestion would be to not create document embeddings by taking the mean of the word embeddings in the sequence but weighting them by their td-idf.

A. Appendix

A.1. Co-occurrences in the top 10 of search results

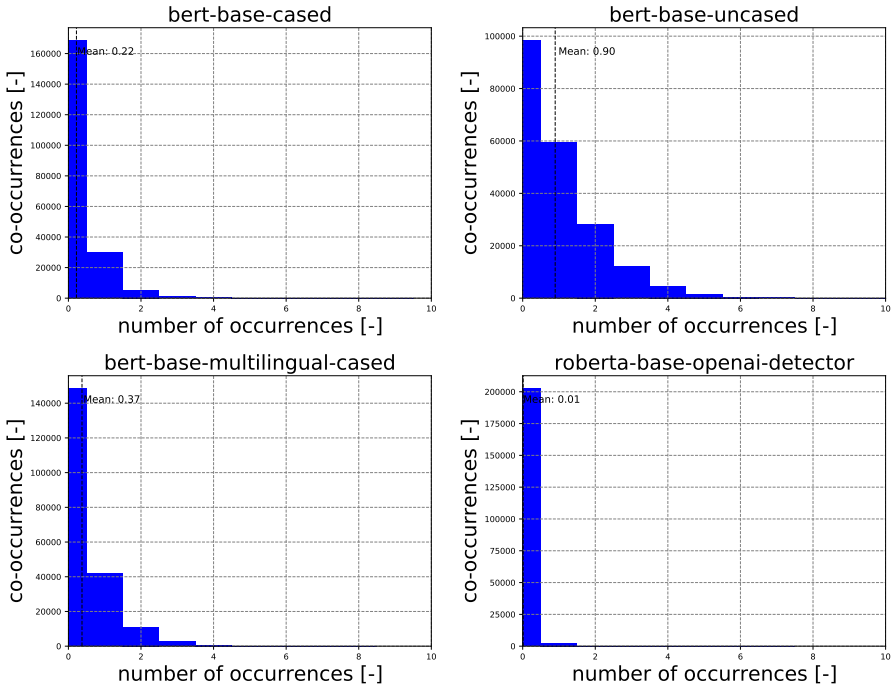

Figure A.1.: Co-occurrences in the top 10 of CLS-based models for WW2 data set

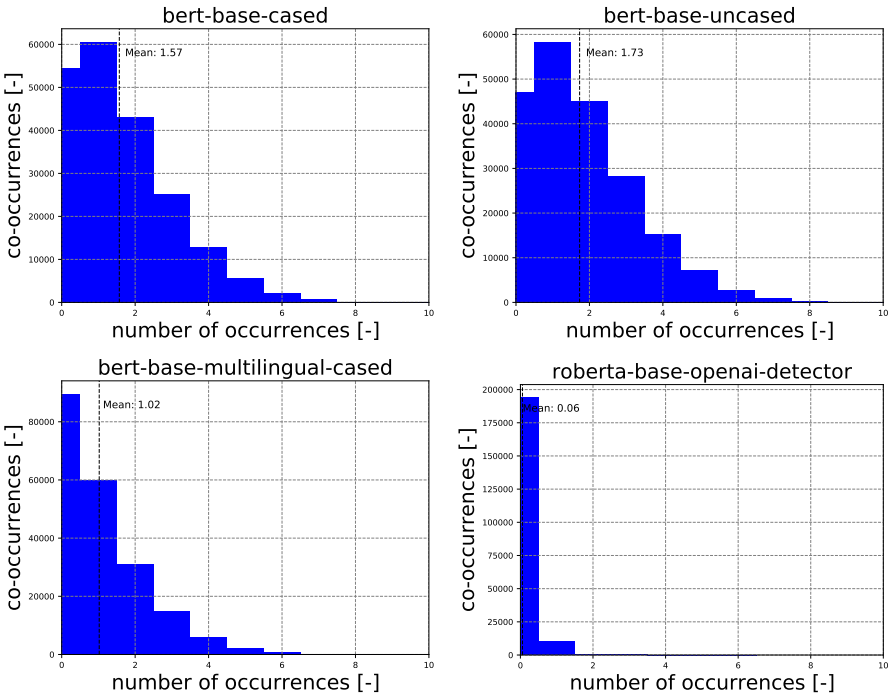

Figure A.2.: Co-occurrences in the top 10 of mean-based models for WW2 data set

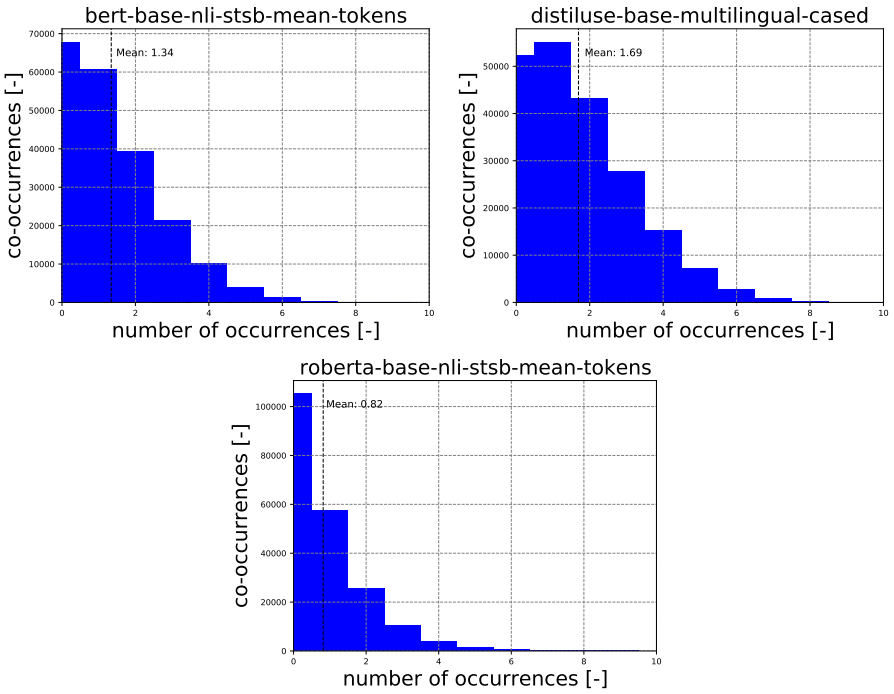

Figure A.3.: Co-occurrences of SentenceBert-based models for WW2 data set

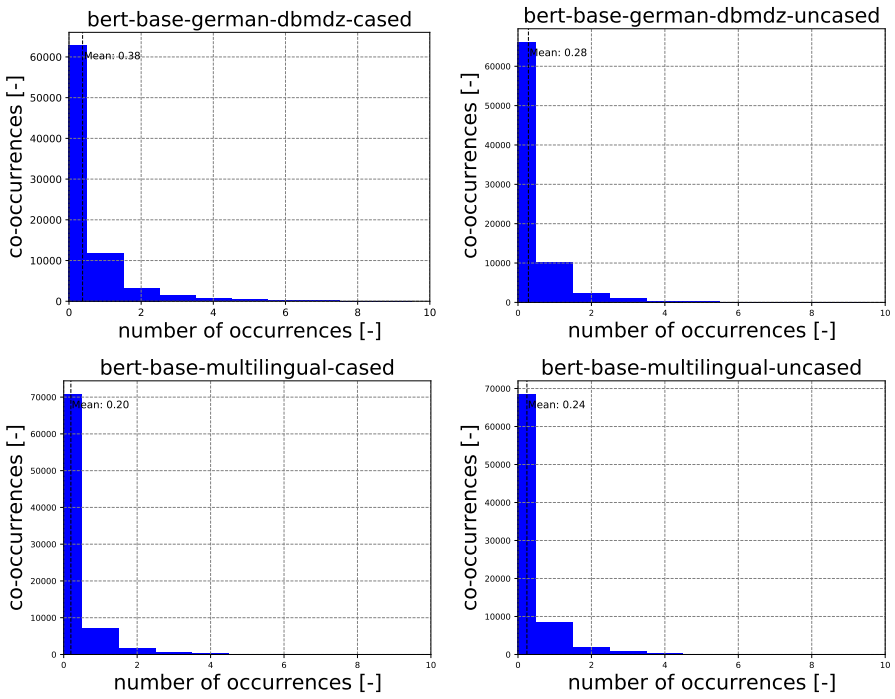

Figure A.4.: Co-occurrences in the top 10 of CLS-based models for NSU data set

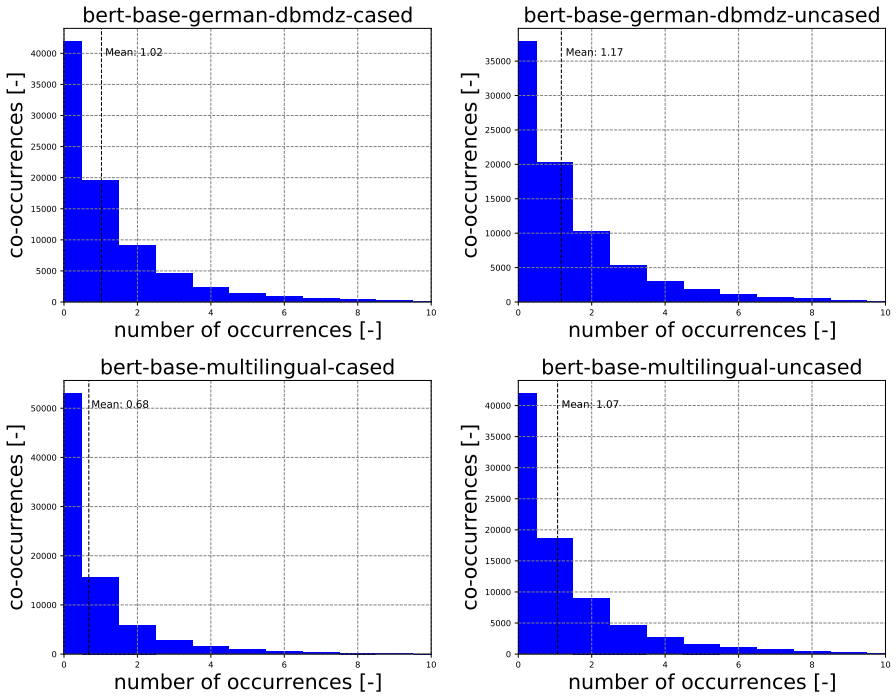

Figure A.5.: Co-occurrences in the top 10 of mean-based models for NSU data set

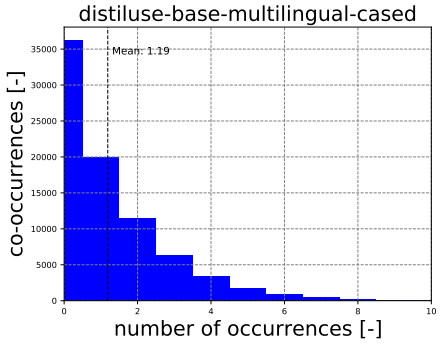

Figure A.6.: Co-occurrences in the top 10 of SentenceBert-based model SBD_{ML} for NSU data set

A.2. Elaborate qualitative analyis

This section describes the qualitative analysis in more detail and without editing. The structure of the examples is as follows:
For each example the search document is presented first. Afterwards the target document is presented. Following this, the search results of the models are presented, one after another. For all presented documents, the id is also given.
More examples are provided digitally with this work. If one desires to search through the data set, the working code is available `https://github.com/MattesR/sentence-bert`.

WW2 examples

Document ID 7762

search document
Discipline was harsh in the armed forces, and the lash was used to punish even trivial offences—and not used sparingly. Despite the harsh discipline, a distinct lack of self-discipline pervaded all ranks of the British forces. Soldiers had an intense passion for gambling, reaching such excesses that troops would often wager their own uniforms. The British leadership soon discovered it had overestimated the capabilities of its own troops, while underestimating those of the colonists, causing a sudden re-think in British planning. Debate persists over whether a British defeat was a guaranteed outcome. Ferling argues that the odds were so long, the defeat of Britain was nothing short of a miracle.

target document
The presence of Hessian soldiers caused considerable anxiety among the colonists, both Patriot and Loyalist, who viewed them as brutal mercenaries. During peacetime, the Army's idleness led to it being riddled with corruption and inefficiency, resulting in many administrative difficulties once campaigning began. Historians such as Ellis and Stewart have observed that, under William Howe's command, the British squandered several opportunities to achieve a decisive victory over the Americans. During planning for the Saratoga campaign, Howe was left with the choice of committing his army to support Burgoyne, or capture Philadelphia, the revolutionary capital. Howe decided upon the latter, determining that Washington was of a greater threat.

Elasticsearch results:
1: (172760):
Capital punishment in Canada was abolished generally in 1976, and for military offences in 1998. Harold Pringle was the last Canadian soldier executed, in 1945, for a military offence. Finland In Finland, the military has jurisdiction over two types of crimes: those that can be committed only by military personnel and those normal crimes by military persons where both the defendant and the victim :§ 2 In crimes where the military has

jurisdiction, the military conducts the investigation. In non-trivial cases, this is done by the investigative section of Defence Command or by civilian police, but trivial cases are investigated by the defendant's own unit. The civilian police has always to right to take the case from the military.
2: (2210):
A similar operation had been conducted on the northern front in Ethiopia but went against theory, for which there were ample forces to execute. Graziani believed the only way to defeat the British was by mass, having overestimated their strength. -General Noel Beresford-Peirse) had two infantry brigades and part of its artillery establishment, the 7th Armoured Division (Major-General Sir Michael O'Moore Creagh) had two armoured brigades composed of two regiments instead of the normal three and fourteen non-brigaded battalions of British infantry. Wavell was to defend Egypt and the Suez Canal against an estimated 250,000 Italian troops based
3: (180372):
An estimated 200,000 medical tourists visited Thailand in 2011, making Bangkok the most popular global destination for medical tourism. Crime and safety Bangkok has a relatively moderate crime rate when compared to urban counterparts around the world. Serious offences included 183 murders, 81 gang robberies, 265 robberies, 1 kidnapping and 9 arson cases. Offences against the state were by far more common, and included 54,068 drug-related cases, 17,239 cases involving prostitution and 8,634 related to gambling.
4: (152239):
The Corinthians, the Spartans, and others in the Peloponnesian League sent more reinforcements to Syracuse, in the hopes of driving off the Athenians; but instead of withdrawing, the Athenians sent another hundred ships and another 5,000 troops to Sicily. Under Gylippus, the Syracusans and their allies were able to decisively defeat the Athenians on land; and Gylippus encouraged the Syracusans to build a navy, which was able to defeat the Athenian fleet when they attempted to withdraw. Following the defeat of the Athenians in Sicily, it was widely believed that the end of the Athenian Empire was at hand. Her treasury was nearly empty, her docks were depleted, and the flower of her youth was dead or imprisoned in a foreign land. They overestimated the strength of their own empire and the beginning of the end was indeed imminent.
5: (75478):
The Islamic Caliphate later guaranteed religious freedom under the conditions that non-Muslim communities accept dhimmi status and their adult males pay the punitive jizya tax instead of the zakat paid by Muslim citizens. Dhimmis were allowed to operate their own courts following their own legal systems in cases that did not involve other religious groups, or capital offences or threats to public order. Despite Dhimmis enjoying special statuses under the Caliphates, they were not considered equals, and sporadic persecutions of non-Muslim groups did occur in the history of the Caliphates.
6: (69844):
judges, which were established in each municipality and city according to the Local Courts and Local Courts Procedure Act of 1875 as special tribunals for minor civil cases. The Royal Court Table in Zagreb was also a jury court for press offences. Judges

were appointed by the king, but their independence was legally guaranteed. Cities (gradovi) and municipalities (općine) were local authorities. Symbols According to the Croatian–Hungarian Agreement in 1868: Croatia, Slavonia and Dalmatia can, within their own frontiers in their internal affairs, use their own combined colours and coat of arms, the latter, however, being surmounted by the Crown of St. Stephen.(Art.
7: (99206):
Luftwaffe preparations Intelligence Faulty intelligence was the component that was mostly responsible for the failure of Adlertag. While the gap between the British and Germans was not yet wide in this regard, the British were starting to gain a decisive lead in intelligence. The breaking of the Enigma machine and poor Luftwaffe signals discipline allowed the British easy access to German communications traffic. Joseph "Beppo" Schmid was commander of the Luftwaffe's Military Intelligence Branch (Abteilung 5 as Chief IC). Throughout this time, Schmid's reports made a series of errors. In July 1940, Schmid grossly overestimated the strengths of the Luftwaffe and underestimated the RAF. The most serious mistakes were made concerning radar, airfield identification, and production sites.
8: (27105):
So marked was their recovery that historians refer to an Italian economic miracle and in the case of West Germany and Austria the Wirtschaftswunder (German for economic miracle). Facing a new power balance between the Soviet East and American West, Western European nations moved closer together. In 1957, Belgium, France, the Netherlands, West Germany, Italy and Luxembourg signed the landmark Trea Between 1945 and 1980, Europe became increasingly socialist. Most European countries became welfare states, in which governments provided a large number of services to their people through taxation. By 1980, most of Europe had universal healthcare and pensions for the elderly. The unemployed were also guaranteed income from the government, and European workers were guaranteed long vacation time.
9: (75742):
The "Edict on Criminal Law Practices against Poles and Jews in the Incorporated Eastern Territories", promulgated 4 December 1941, permitted corporal punishment and death sentences for even the most trivial of offences. After discussion with Hitler, he issued a policy directive to Rosenberg that read in part: The Slavs are to work for us. In so far as we don't need them, they may die. The fertility of the Slavs is undesirable. As to food, they are to not get more than necessary. We are the masters; we come first.
10: (179165):
, Belgians and Australians who were struggling with outmoded aircraft, poor training and weak tactics. As a result, the Allied air successes over the Somme would not be repeated and heavy losses were inflicted by the Germans. During their attack at Arras, the British lost 316 air crews and the Canadians lost 114 compared to 44 lost by the Germans. The offensive proceeded poorly as the French troops, with the help of two Russian brigades, had to negotiate rough, upward-sloping terrain in extremely bad weather. On 3 May the weary French 2nd Colonial Division, veterans of the Battle of Verdun, refused orders, arriving drunk and without their weapons. Lacking the means to punish an entire division,

its officers did not immediately implement harsh measures against the mutineers.

BU_E **results:**
1: (87459):
This was a successful wartime strategy but, after the war was over, each side believed that it had borne a greater burden than the other. The British elite, the most heavily taxed of any in Europe, pointed out angrily that the colonists paid little to the royal coffers. The colonists replied that their sons had fought and died in a war that served European interests more than their own. Many had never been to Britain, yet they imitated British styles of dress, dance, and etiquette. This social upper echelon built its mansions in the Georgian style, copied the furniture designs of Thomas Chippendale, and participated in the intellectual currents of Europe, such as the Enlightenment. The seaport cities of colonial America were truly British cities in the eyes of many inhabitants.
2: (61615):
Russian Army Tsar Nicholas I (reigned 1825–1855) lavished attention on his very large army; with a population of 60–70 million people, the army included a million men. They had outdated equipment and tactics, but the tsar, who dressed like a soldier and surrounded himself with officers, gloried in the victory over Napoleon in 1812 and took enormous pride in its smartness on parade. The Army became the vehicle of upward social mobility for noble youths from non-Russian areas, such as Poland, the Baltic, Finland and Georgia. On the other hand, many miscreants, petty criminals and undesirables were punished by local officials by enlisting them for life in the Army. The conscription system was highly unpopular with people, as was the practice of forcing peasants to house the soldiers for six months of the year.
3: (7763):
The presence of Hessian soldiers caused considerable anxiety among the colonists, both Patriot and Loyalist, who viewed them as brutal mercenaries. During peacetime, the Army's idleness led to it being riddled with corruption and inefficiency, resulting in many administrative difficulties once campaigning began. Historians such as Ellis and Stewart have observed that, under William Howe's command, the British squandered several opportunities to achieve a decisive victory over the Americans. During planning for the Saratoga campaign, Howe was left with the choice of committing his army to support Burgoyne, or capture Philadelphia, the revolutionary capital. Howe decided upon the latter, determining that Washington was of a greater threat.
4: (75222):
The military leadership of the British Army during the World War I was frequently condemned as poor by historians and politicians for decades after the war ended. Common charges were that the generals commanding the army were blind to the realities of trench warfare, ignorant of the conditions of their men and were unable to learn from their mistakes, thus causing enormous numbers of casualties that they could not control such as a lack of adequate military communications, which was not known before. Furthermore, military leadership improved throughout the war culminating in the Hundred Days Offensive advance to victory in 1918. Some historians, even revisionists, still criticise

the British High Command severely, but they are less inclined to portray the war in a simplistic manner with
5: (139306):
Its main result, protecting French borders against all enemies, surprised and shocked Europe. The levée en masse was also effective in that by putting on the field many men, even untrained, it required France's opponents to man all fortresses and expand their own standing armies, far beyond their capacity to pay professional soldiers. Though not a novel idea—see for example thinkers as diverse as Plato and the lawyer and linguist Sir William Jones (who thought every adult male should be armed with a musket at public expense)—the actual practice of a levée en masse was rare before the French Revolution.
6: (191332):
The new developments hoped to exploit the intrinsic bravery of the French soldier, made even more powerful by the explosive nationalist forces of the Revolution. The changes also placed a faith on the ordinary soldier that would be completely unacceptable in earlier times; French troops were expected to harass the enemy and remain loyal enough to not desert, a benefit other Ancien Régime armies did not have. By summer of the following year, conscription made some 500,000 men available for service and the French began to deal blows to their European enemies. Armies during the Revolution became noticeably larger than their Holy Roman counterparts, and combined with the new enthusiasm of the troops, the tactical and strategic opportunities became profound. By 1797 the French had defeated the First Coali left major command positions to those who could be trusted, namely, the aristocracy. The hectic nature of the French Revolution, however, tore apart France's old army, meaning new men were required to become officers and commanders.
7: (130606):
Modern conscription, the massed military enlistment of national citizens, was devised during the French Revolution, to enable the Republic to defend itself from the attacks of European monarchies. Deputy Jean-Baptiste Jourdan gave its name to the 5 September 1798 Act, whose first article stated: "Any Frenchman is a soldier and owes himself to the defense of the nation. The defeat of the Prussian Army in particular shocked the Prussian establishment, which had believed it was invincible after the victories of Frederick the Great. The Prussians were used to relying on superior organization and tactical factors such as order of battle to focus superior troops against inferior ones. Given approximately equivalent forces, as was generally the case with professional armies, these factors showed considerable importance.
8: (139305):
Many individuals who were conscripted by the French deserted, fleeing from their duty of fighting for France without permission, in hopes of never getting caught. There were rough estimates to the number of individuals that deserted during the time of the Levée en Masse, but due to many factors, like the inability to manage and keep track of all the armies or differentiating between men with similar names, the exact number is unclear. Popular reaction For all the rhetoric, the levée en masse was not popular; desertion and evasion were high. However, the effort was sufficient to turn the tide of the war, and there was no need for any further conscription until 1797, when a more permanent system

of annual intakes was instituted. An effect of the levée en masse was the creation of a national army in France, made up of citizens, rather than an all-professional army, as was the standard practice of the time.
9: (32533):
Many eligibles pooled their money to cover the cost of anyone drafted. Families used the substitute provision to select which man should go into the army and which should stay home. There was much evasion and overt resistance to the draft, especially in Catholic areas. The great draft riot in New York City in July 1863 involved Irish immigrants who had been signed up as citizens to swell the vote of the city's Democratic political machine, not realizing it made them liable for the draft. From a tiny frontier force in 1860, the Union and Confederate armies had grown into the "largest and most efficient armies in the world" within a few years. European observers at the time dismissed them as amateur and unprofessional, but British historian John Keegan's assessment is that each outmatched the French, Prussian and Russian armies of the time, and but for the Atlantic, would have threatened any of them with defeat.
10: (16094):
In addition, battles were often made irrelevant by the proliferation of advanced, bastioned fortifications. To control an area, armies had to take fortified towns, regardless of whether they defeated their enemies' field armies. As a result, by far the most common battles of the era were sieges, hugely time-consuming and expensive affairs. The indecisive nature of conflict meant wars were long and endemic. Conflicts stretched on for decades and many states spent more years at war than they did at peace. The Spanish attempt to reconquer the Netherlands after the Dutch Revolt became bogged down in endless siege warfare. The expense caused the Spanish monarchy to declare bankruptcy several times, beginning in 1577.

Document ID 201

search document
In response to multiple authors Possible explanations for casual uses They employ massive overkill strategy, there are 30, 20 to 30 marshals daily inside the courtroom, it has the atmosphere of an arms camp, the law against us is rigged . . . and our claims that this law violates our constitutional rights and it's the same way that we claim that Mayor Daley didn't have the right to deny us a permit t social structure, though this kind of Maoist or Guevarist analysis often underpinned the rhetorical depiction of Cold War authoritarians as fascists.Some Marxist groups – such as the Indian section of the Fourth International and the Hekmatist groups in Iran and Iraq – have provided analytical accounts as to why the term "fascist" should be applied to groups such as the Hindutva movement, the 1979 Islamic Iranian regime or the Islamist sections of the Iraqi insurgency. Other scholars contend that the traditional meaning of the term fascism does not apply to Hindutva groups and may hinder an analysis of their activities. See also References External links
target document

In 2014, with the outbreak of the war in Donbass the Russian nationalists and media returned to the "fascist" rhetoric, frequently describing the Ukrainian government after Euromaidan as "fascist", "Nazi" etc. to march or to assemble in the park.... I think it points a direction in the future which is that the government embarked on a course of fascism.Several Marxist theories back up particular uses of fascism beyond its usual remit. For instance, Poulantzas's theory of state monopoly capitalism could be associated with the idea of a military-industrial complex to suggest that 1960s America had a fascis

Elasticsearch results:

1 (31370):
Operations research Operations research, or operational research in British usage, is a discipline that deals with the application of advanced analytical methods to help make better decisions. Further, the term 'operational analysis' is used in the British (and some British Commonwealth) military as an intrinsic part of capability development, management and assurance. Employing techniques from other mathematical sciences, such as mathematical modeling, statistical analysis, and mathematical optimization, operations research arrives at optimal or near-optimal solutions to complex decision-making problems.
2 (122878):
Directorate of Analysis The Directorate of Analysis, through much of its history known as the Directorate of Intelligence (DI), is tasked with helping "the President and other policymakers make informed decisions about our country's national security" by looking "at all the available information on an issue and organiz There is an office dedicated to Iraq; regional analytical offices covering th This Directorate was created in an attempt to end years of rivalry over influence, philosophy and budget between the United States Department of Defense (DOD) and the CIA. In spite of this, the Department of Defense recently organized its own global clandestine intelligence service, the Defense Clandestine Service (DCS), under the Defense Intelligence Agency (DIA).
3 (148480):
Military intelligence Military intelligence supports the combat commanders' decision making process by providing intelligence analysis of available data from a wide range of sources. To provide that informed analysis the commanders information requirements are identified and input to a process of gathering, analysis, protection, and dissemination of information about the operational environment, Most militaries maintain a military intelligence capability to provide analytical and information collection personnel in both specialist units and from other arms and services. Personnel selected for intelligence duties, whether specialist intelligence officers and enlisted soldiers or non-specialist assigned to intelligence may be selected for their analytical abilities and intelligence before receiving formal training.
4 (163879):
Two years later, the city hosted the tumultuous 1968 Democratic National Convention, which featured physical confrontations both inside and outside the convention hall, with

anti-war protesters, journalists and bystanders being beaten by police. Richard M. Daley, son of Richard J. Daley, was elected in 1989. His accomplishments included improvements to parks and creating incentives for sustainable development, as well as closing Meigs Field in the middle of the night and destroying the runways. After successfully running for re-election five times, and becoming Chicago's longest-serving mayor, Richard M. Daley declined to run for a seventh term.
5 (168547):
Cecil Cook, the Northern Territory Protector of Natives, noted that: generally by the fifth and invariably by the sixth generation, all native characteristics of the Australian Aborigine are eradicated. The problem of our half-castes will quickly be eliminated by the complete disappearance of the black race, and the swift submergence of their progeny in the white. During this period many Aboriginal activists began to embrace the term "black" and use their ancestry as a source of pride. Activist Bob Maza said: I only hope that when I die I can say I'm black and it's beautiful to be black.
6 (5827):
In the century after Tracy, the term ideology moved back and forth between positive and negative connotations. (Perhaps the most accessible source for the near-original meaning of ideology is Hippolyte Taine's work on the Ancien Régime (the first volume of "Origins of Contemporary France"). He describes ideology as rather like teaching philosophy by the Socratic method, but without extending the v) The term "ideology" has dropped some of its pejorative sting, and has become a neutral term in the analysis of differing political opinions and views of social groups. Analysis There has been considerable analysis of different ideological patterns. This kind of analysis has been described by some as meta-ideology—the study of the structure, form, and manifestation of ideologies.
7 (167784):
The Americans also occasionally used the French term corvette. History In the Royal Navy, the sloop evolved into an unrated vessel with a single gun deck and three masts, two square rigged and the aftermost fore-and-aft rigged (corvettes had three masts, all of which were square-rigged). During the War of 1812 sloops of war in the service of the United States Navy performed well against their Royal Navy equivalents. The American ships had the advantage of being ship-rigged rather that brig-rigged, a distinction that increased their maneuverability. They were also larger and better armed. Cruzier-class brig-sloops in particular were vulnerable in one-on-one engagements with American sloops of war.
8 (122954):
Several investigations (e.g., the Church Committee, Rockefeller Commission, Pike Committee, etc.) have been conducted about the CIA, and many documents have been declassified. Influencing public opinion and law enforcement Drug trafficking Two offices of CIA Directorate of Analysis have analytical responsibilities in this area. The Office of Transnational Issues applies unique functional expertise to assess existing and emerging threats to U.S. national security and provides the most senior U.S. policymakers, military planners, and law enforcement with analysis, warning, and crisis support.
9 (148675):

Most current communist groups descended from the Maoist ideological tradition still adopt the description of both China and the Soviet Union as being "state capitalist" from a certain point in their history onwards—most commonly, the Soviet Union from 1956 to its collapse in 1991 and China from 1976 to the present. Use by liberal economists Murray Rothbard, an anarcho-capitalist philosopher, uses the term interchangeably with the term state monopoly capitalism and uses it to describe a partnership of government and big business in which the state intervenes on behalf of large capitalists against the interests of consumers.
10 (45648):
To provide an analysis, the commander's information requirements are first identified, which are then incorporated into intelligence collection, analysis, and dissemination. Areas of study may include the operational environment, hostile, friendly and neutral forces, the civilian population in an area of combat operations, and other broader areas of interest. Intelligence activities are conducted Personnel performing intelligence duties may be selected for their analytical abilities and personal intelligence before receiving formal training. Contents Levels of intelligence Intelligence operations are carried out throughout the hierarchy of political and military activity.

BU_E **results:**
1: (79135):
fed straight into the formulations of Hassan al-Banna, the leader of the Muslim Brotherhood, and the essays of Sayyid Qutb to influence modern Islamic radicalism. Euphemisms must be dropped, Herf counseled, and killing civilians, Muslim or otherwise, defined as a "war crime", so that Mahmoud Ahmadinejad should be indicted for incitement to genocide. Criticism of theory of a link between Islam and fascism While Islamic Fascism has been discussed as a category of serious analysis by the scholars mentioned above, the term Islamofascism circulated mainly as a propaganda, rather than as an analytic, term after the September 11 attacks on the United States in September 2001 In his diagnosis of this shift he detected a decline in the old liberal consensus of American politics, and what he called the "deliquescence of the Democratic Party". Many former left-liberal pundits, like Paul Berman and Peter Beinart having no knowledge of the Middle East or cultures like those of Wahhabism and Sufism on which they descant authoritatively, have, he claimed, and his view was
2: (79130):
The American journalist and former Nixon speechwriter William Safire wrote that the term fulfilled a need for a term to distinguish traditional Islam from terrorists: "Islamofascism may have legs: the compound defines those terrorists who profess a religious mission while embracing totalitarian methods and helps separate them from devout Muslims who want no part of terrorist means. George Orwell, it has been noted in this connection, observed as early as 1946 that " Walter Laqueur, after reviewing this and related terms, concluded that "Islamic fascism, Islamophobia and antisemitism, each in its way, are imprecise terms we could well do without but it is doubtful whether they can be removed from our political lexicon.

3: (180223):
But clearly this is not true of, say, Portugal or the various South American dictatorships. Or again, antisemitism is supposed to be one of the distinguishing marks of Fascism; but some Fascist movements are not antisemitic. Learned controversies, reverberating for years on end in American magazines, have not even been able to determine whether or not Fascism is a form of capitalism. The word fascist is sometimes used to denigrate people, institutions, or groups that would not describe themselves as ideologically fascist, and that may not fall within the formal definition of the word. As a political epithet, fascist has been used in an anti-authoritarian sense to emphasize the common ideology of governmental suppression of individual freedom. g, the 1922 Committee, the 1941 Committee, Kipling, Gandhi, Chiang Kai-Shek, homosexuality, Priestley's broadcasts, Youth Hostels, astrology, women, dogs and I do not know what else . Except for the relatively small number of Fascist sympathisers, almost any English person would accept 'bully' as a synonym for 'Fascist'. That is about as near to a definition as this much-abused word has come.
4: (122687):
Accusations of racism Several critics, including David North, Goldhagen has said that there is no racist or ethnic argument about Germans in his text. Some of his critics have agreed with him that his thesis is "not intrinsically racist or otherwise illegitimate", including Ruth Bettina Birn and Norman Finkelstein (A Nation on Trial). ". The book had a "mostly scathing" reception among historians, Steve Crawshaw writes that although the German readership was keenly aware of certain "professional failings" in Goldhagen's book, Crawshaw further asserts that the book's critics were partly historians "weary" of Goldhagen's "methodological flaws", but also those who were reluctant to concede that ordinary Germans bore responsibility for the crimes of Nazi Germany. There were regional variations in anti-Semitism even within Germany. But Hitler's exemplified and brought to an apotheosis the particular form of eliminationist anti-Semitism that came to the fore in the latter part of the nineteenth century. Whatever the variations, I think Austrian and German anti-Semitism can be seen of a piece, where there was a central model of Jews and a view that they needed to be eliminated. In 2006, the American columnist Jonah Goldberg argued that "Goldhagen's thesis was overstated but fundamentally accurate. There was something unique to Germany that made its fascism genocidal. Around the globe there have been dozens of self-declared fascist movements (and a good deal more that go by different labels), and few of them have embraced Nazi-style genocide. Indeed, fascist Spain was a haven for Jews during the Holocaust" he said. See also References Notes Bibliography External links
5: (59088):
An additional 23 individuals have been the subject of contempt proceedings. Criticism Skeptics argued that an international court could not function while the war in the former Yugoslavia was still going on. This would be a huge undertaking for any court, but for the ICTY it would be an even greater one, as the new tribunal still needed judges, a prosecutor, a registrar, investigative and suppor Response to criticism Supporters of the work of the ICTY responded to critics in various publications. In a response to David

Harland's Selective Justice, Jelena Subotić, an assistant professor of political science at Georgia State University and author of Hijacked Justice: Dealing with the Past in the Balkans, responded that the critics of the Tribunal miss the point, "which is not to deliver jus " Marko Hoare claims the accusations of the tribunal's "selective justice" stem from Serbian nationalist propaganda. He wrote: "This is, of course, the claim that hardline Serb nationalists and supporters of Slobodan Milosevic have been making for about the last two decades.
6: (180222):
Writing for the Tribune in 1944, Orwell stated: .It is not easy, for instance, to fit Germany and Japan into the same framework, and it is even harder with some of the small states which are describable as Fascist. It is usually assumed, for instance, that Fascism is inherently warlike, that it thrives in an atmosphere of war hysteria and can only solve its economic problems by means of war preparation or foreign conquests. But still, when we apply the term 'Fascism' to Germany or Japan or Mussolini's Italy, we know broadly what we mean. Some have argued that the terms fascism and fascist have become hopelessly vague since the World War II period, and that today it is little more than a pejorative used by supporters of various political views to insult their opponents. In this sense, the word fascist is intended to mean oppressive, intolerant, chauvinist, genocidal, dictatorial, racist, or aggressive. George Orwell wrote in 1944: . the word 'Fascism' is almost entirely meaningless. In conversation, of course, it is used even more wildly than in print. I have heard it applied to farmers, shopkeepers, Social Credit, corporal punishment, fox-hunting, bull-fightin Linda and Morris Tannehill Anarcho-capitalist authors Linda and Morris Tannehill claim in their 1970 self-published work The Market for Liberty that "Fascism is a system in which the government leaves nominal ownership of the means of production in the hands of private individuals but exercises control by means of regulatory legislation and reaps most of the profit by means of heavy taxation. In effect, fascism is simply a more subtle form of government ownership than is socialism". See also Notes References External links
7: (89464):
Examples Some argue that military drafts, and other forms of coerced government labour. constitute "state-operated slavery." "Slavery" has been used by some anti-psychiatry proponents to define involuntary psychiatric patients due to there are no unbiased physical tests for mental illness, yet the psychiatric patient must follow the orders of his/her psychiatrist. The labor market, as institutionalized under today's market economic systems, has been criticized by mainstream socialists and by anarcho-syndicalists, who utilise the term wage slavery as a pejorative or dysphemism for wage labour. Historians agree that films have largely shaped historical memories, but they debate issues of accuracy, plausibility, moralism, sensationalism, how facts are stretched in search of broader truths, and suitability for the classroom. See also References Bibliography and further reading Surveys and reference External links Historical Modern
8: (6231):
ed in reverse would yield some interesting results regarding German, American, and British competence.Sadkovich also states that such a fixation on Germany and such denig-

rations of Italians not only distort analysis, they also reinforce the misunderstandings and myths that have grown up around the Greek theater and allow historians to lament and debate the impact of the Italo-Greek conflict on the British and German war efforts, yet dismiss as unimportant its impact on the Italian war effort. Alan Levine even goes most authors one better by dismissing the whole Mediterranean theater as irrelevant, but only after duly scolding Mussolini for 'his imbecilic attack on Greece'. Anti-Italianism after World War II Libya and Yugoslavia have shown high levels of anti-Italianism since WWII, as illustrated by the following manifestations: Italian-American organizations National organizations which have been active in combatting media stereotyping and defamation of Italian Americans are: Order Sons of Italy in America, Unico National, National Italian American Foundation and the Italic Institute of America. See also References Further reading
9: (200):
In 2014, with the outbreak of the war in Donbass the Russian nationalists and media returned to the "fascist" rhetoric, frequently describing the Ukrainian government after Euromaidan as "fascist", "Nazi" etc. to march or to assemble in the park.... I think it points a direction in the future which is that the government embarked on a course of fascism.Several Marxist theories back up particular uses of fascism beyond its usual remit. For instance, Poulantzas's theory of state monopoly capitalism could be associated with the idea of a military-industrial complex to suggest that 1960s America had a fascis
10: (80325):
They neither emerge nor are being given attention. No one is explaining to students what colonization has been. We have prevented students from understanding why the decolonization took place. In metropolitan France in 1963, 43% of French Algerians lived in bidonvilles (shanty towns). " But, as it did during wartime, the French state claimed torture were isolated acts, instead of admitting its responsibility for the frequent use of torture to break the insurgents' morale and not, as Aussaresses has claimed, to "save lives" by gaining short-term information which would stop "terrorists". Other publications Translations may be available for some of these works. See specific cases. Films See also References He also argues that the least controversial of all the numbers put forward by various groups are those concerning the French soldiers, where government numbers are largely accepted as sound. Most controversial are the numbers of civilians killed.

NSU examples

Document ID 52096

search document
Informationssystemen bereits erfolgt und negativ verlaufensei. In einem mit „LB“ – für Ludwigsburg – bezeichneten Unterordner habe er, so der Zeuge KHK J. G., zwei Kontrollstellenlisten, nämlich eine von der Kontrollstelle aus Mundelsheim und eine aus Oberstenfeld, feststellen können, die allerdings denselben Dateinamen gehabt hätten. Die

Excel-Dateien seien ansonsten aufgrund eines Programms zur automatisierte Landtag von Baden-Württemberg Drucksache 15 / 8000 195 Als er, so der Zeuge KHK J. G., die Listen weiter ausgewertet habe, habe er beim Filtern der Gesamtliste „Kennzeichen“ festgestellt, dass im gesamten Fahndungsraum am 25. April 2007 unter den ca. 33.000 Fahrzeugkennzeichen nur sechs Wohnmobile verzeichnet gewesen seien. Darunter habe sich kein Wohnmobil aus Zulassungsbezirken in Thüringen befunden.
target document
nErfassung in die Bearbeitungssoftware „CRIME“ übermittelt worden. Aufgrund des fehlerhaften Dateinamens sei diese Übertragung für die Ringalarmdaten aus Oberstenfeld zunächst nicht erfolgt. In der Excel-Tabelle und der Halterlistentabelle seien diese Daten jedoch vollständig erfasst gewesen. Aus sächsischen Zulassungsbezirken sei nur das Wohnmobil mit dem Kennzeichen C-PW 87 von C. H. mit ihrem Geburtsdatum, ihrem Geburtsort und einer Chemnitzer Adresse vermerkt gewesen, die restlichen fünf Wohnmobile seien in anderen Bundesländern zugelassen gewesen. Anhand der Halterdaten des Kraftfahrt-Bundesamts sei nicht zu erkennen gewesen, dass dieses Wohnmobil aus Chemnitz auf ein Unternehmen,

Elasticsearch results:
1: (52097):
nErfassung in die Bearbeitungssoftware „CRIME“ übermittelt worden. Aufgrund des fehlerhaften Dateinamens sei diese Übertragung für die Ringalarmdaten aus Oberstenfeld zunächst nicht erfolgt. In der Excel-Tabelle und der Halterlistentabelle seien diese Daten jedoch vollständig erfasst gewesen. Aus sächsischen Zulassungsbezirken sei nur das Wohnmobil mit dem Kennzeichen C-PW 87 von C. H. mit ihrem Geburtsdatum, ihrem Geburtsort und einer Chemnitzer Adresse vermerkt gewesen, die restlichen fünf Wohnmobile seien in anderen Bundesländern zugelassen gewesen. Anhand der Halterdaten des Kraftfahrt-Bundesamts sei nicht zu erkennen gewesen, dass dieses Wohnmobil aus Chemnitz auf ein Unternehmen,
2: (52099):
Er habe, so der Zeuge KHK J. G., keine Anhaltspunkte dafür, ob die Ermittler vor dem 4. November 2011 die Tabellen im Hinblick auf Wohnmobile ausgewertet hätten, auch im Hinblick darauf, dass am 2. April 2009 der Zeuge J. L. bekundet habe, dass ihm einen Tag vor der Tat am Pumpenhäuschen ein Wohnmobil aufgefallen sei, das nicht zu den Schaustellern gehört habe und dieses am Abend des 25. Er habe, so der Zeuge KHK J. G., nicht überprüft, von welcher PD überhaupt Daten vorhanden seien. Er habe lediglich den Namensfehler bei den Daten aus Oberstenfeld und Mundelsheim erkennen können. Er habe einen Gesamtdatenbestand von 33.037 Fahrzeugen vorgefunden. Die Angaben der Sachverständigen Aust und Laabs in ihrem Buch habe er gelesen, könne sie aber nicht nachvollziehen.
3: (52094):
Auch sei es ihnen nicht gelungen, in dieser Zeit überhaupt jedes Kennzeichen zu erfassen. Vielleicht hätten sie genannte Kennzeichen nur deshalb erfasst, weil es ein auswärtiges Kennzeichen gewesen wäre. Aber an ein Wohnmobil hätten sich beide überhaupt nicht

mehr erinnern können. kplatz“ beim LKA Baden-Württembergbegonnen worden. Zuvor hätten dem LKA nur einzelne Kontrolllisten entweder elektronisch im Excel-Format oder handschriftliche Listen vorgelegen. Durch zwei Unterstützungskräfte der Bereitschaftspolizei seien zu diesem Zeitpunkt alle Kontrollstellenlisten in einheitliche Excel-Dateien übertragen und dann in eine Gesamtliste kopiert worden. Bei der Sichtung des Ordn
4: (33377):
5640 Bei der Kontrollstelle Oberstenfeld, die ca. 25 bis 30 Minuten beziehungsweise 20 Kilometer vom Tatort entfernt ist, wurde unter anderem ein Wohnmobil registriert. Eine Kontrollliste zum Kontrollpunkt LB 3 bestätigt, dass delte. 5642 Dem Ermittlungsbericht lässt sich entnehmen, dass das Wohnmobil die Kontrollstelle zwischen 14.30 Uhr und 14.37 Uhr passierte. Eine detaillierte Weg-ZeitBerechnung des LKA ergab, dass es im Zeitfenster zwi
5: (52002):
Ein weiterer Be zu bringen seien, sei die Feststellung des Wohnmobils am Tattag um 14:37 Uhr in der Kontrollstelle in Oberstenfeld. Oberstenfeld sei ungefähr 20 Kilometer von Heilbronn entfernt, befinde sich also auch in einer Entfernung, die in einer Weg-Zeit-Berechnung mit der Tatzeit gegen 14:00 Uhr korrespondiere. Ein Foto der Dienst worden. Diese Datei trage den Dateinamen „Aktion Polizeipistole“. Zu dem Aufnahmeort der Dienstwaffe hätten sie sehr umfangreiche Ermittlungen angestellt. Die Dienstwaffe sei vor einem relativ auffälligen blauen Hintergrund, einem Teppich, fotografiert worden. Sie hätten, so der Zeuge KOR A. K., die Hoffnung gehabt, beim Nachmieter in der Polenzstraße in Zwickau, in der das Tri
6: (52116):
usrüstungsge-genstände sowie die Flucht aus dem Tatortnahbereich mit Mountainbikes, das Verladen der Bikes in ein Wohnmobil und das Verlassen des Parkplatzes verblieben. Die wahrscheinlichste Fluchtroute sei von der Neckartalstraße Richtung Landturm, dann vor Lauffen links ab nach Ilsfeld und über Flein über einen Hof. Man fahre dann unter der Autobahnbrücke durch, während oben ein Streifenwagen d Die Theresienwiese sei ein beliebter Pausenplatz auch für Raucher gewesen. Dies alles habe seines Erachtens eine entsprechende Planung ermöglicht. Auf die Frage, ob das Wohnmobil auf der von ihm rekonstruierten Fluchtstrecke vor der Kontrollstelle in Oberstenfeld keine weitere Kontrollstelle hätte passieren müssen, gab der Zeuge KHK J. G. an, jedenfalls sei das Kennzeichen sonst nirgends festgeste
7: (52115):
uf der Landstraße 1100 entfernen, weil auf dieser Route der Verkehrsfluss nicht ständigdurch Ampeln, wie in östlicher Richtung durch das Stadtgebiet, unterbrochen werde. Die Landstraße 1100 führe weiter über Flein, Ilsfeld, Beilstein direkt nach Oberstenfeld. Von dort seien es nur wenige Kilometer bis zum Landkreis Rems-Murr. Anhand der durchgeführten Fluchtweganalysen, die er, so der Zeuge KHK J. der Strecke gewesen sei, hätten sie nicht geprüft. Anhand der erfassten Uhrzeiten habe dasLandtag von Baden-Württemberg Drucksache 15 / 8000 198 Wohnmobil am 25. April 2007 nach 14:30 Uhr, aber vor 14:38 Uhr, aus Beilstein kommend die Kontrollstelle in Oberstenfeld passiert. Somit wären den Tätern etwa sechs bis zehn Minuten für die eigentliche Tatausführung,

den Raub der Dienstwaffen und Polizeia
8: (52093):
Im Rahmen dieser beschriebenen Fahndungsmaßnahmen seien am 25. April 2007 im Zeitraum von etwa zweieinhalb bis zweidreiviertel Stunden über 33 000 Fahrzeugkennzeichen von den eingesetzten Kräften handschriftlich notiert worden. Später habe er, so der Zeuge KHK J. G., die beiden eingesetzten Beamten befragt und darüber nur einen Aktenvermerk verfasst, weil beide gesagt hätten: „Wir wissen, wir waren bei dem Ringalarm. Wir haben aber, wie gesagt, 410 Kennzeichen." Beide seien wohl mit dem Notieren beschäftigt gewesen, weil starker Verkehr geherrscht habe, es sei ihnen nicht möglich gewesen, die passierenden Fahrzeuge näher zu beobachten.
9: (22083):
iche Feststellungen, dass er je-mals in Baden-Württemberg gelebt hat. Also wir gehen davon aus, 3446 Hißlinger, Protokoll-Nr. 35 I der 35. Sitzung am 20. Oktober 2016, S. 19. 3447 Hißlinger, Protokoll-Nr. 35 I der 35. Sitzung am 20. Oktober 2016, S. 14. 3448 Hißlinger, Protokoll-Nr. 35 I der 35. Sitzung am 20. Oktober 2016, S. 15. Das von der Terrorgruppe „NSU" angemietete Wohnmobil wurde am 25. April 2007 in einer Kontrollstelle im Bereich Oberstenfeld gesichtet. Im Nahbereich der Kontrollstelle befindet sich der Wohnsitz des Andreas G.,
10: (64034):
schen Gründen nicht angebracht. Aus denselben Gründen sollte nicht an den Stefan Apel (Cousin von Beate Zschäpe) herangetreten werden. Schlussbemerkung: Es wird davon ausgegangen, dass die drei Gesuchten im Bereich Chemnitz untergetaucht Teilnehmer: EKHK J., KHK Tra. - beide LKA Sachsen EKHK'in Lipprandt, KHK Dressler, KHK Kleimann - TLKA. Das hiesige Dezernat 22 wird in den nächsten Wochen in Zusammenarbeit mit dem LKA Sachsen bzw. der Außenstelle in Chemnitz die bisher offen gebliebenen Ermittlungen und

BU_G **results:**
1: (52095):
Er, so der Zeuge KHK J. G., habe an seinem dritten Arbeitstag bei der Soko „Parkplatz", dem 9. November 2011, auf deren Laufwerk erstmals die Auswerteergebnisse zur Ringalarmfahndung gesichtet und dabei Folgendes festgestellt: Mit der Erfassung und Auswertung aller Kontrollstellenlisten sei im August 2010 unter der Bezeichnung „Maßnahme 328" durch den Abschnitt „Operative Auswertung" der Soko „Par ers „Kontrollstellen" habe er, so der Zeuge KHK J. G., eine mit „Kennzeichen" bezeich-nete Gesamtliste im Excel-Format festgestellt. In diesem Ordner sei auch eine Tabelle gespeichert gewesen, in welche bereits die ermittelten Halterdaten der Fahrzeuge eingebunden gewesen seien. Auf Nachfrage sei ihm mitgeteilt worden, dass eine Überprüfung der Halterpersonalien in den polizeilichen Informationssy
2: (52933):
feldvernehmungen nachgefragt,aber letzten Endes sei nichts dabei heraus gekommen, und sie hätten auch keine auffälligen Rufnummern gefunden. Landtag von Baden-Württemberg Drucksache 15 / 8000 339 Bezüglich der achtmaligen Kontaktaufnahme durch eine Service-

SMS-Nummer erklärte die Zeugin PK'in N. K., sie wisse noch, dass die Rufnummern in dieser Verbindungsdatentabelle gewesen seien und sie abgeklä valides Bild. Auf Nachfrage erklärte die Zeugin, die Daten, wann welche SMS eingegangen seien, hätten aber schon vorgelegen. Alle Daten seien in einer Gesamttabelle, die als „Verbindungsdaten Opfer" bezeichnet worden sei, erfasst worden. Diese Tabelle habe schließlich sämtliche Verbindungsdaten im Zeitraum vom 13. Juni 2005 bis zum 28. April 2007 umfasst. An diese Erfassung habe sich dann die Fest
3: (52097):
nErfassung in die Bearbeitungssoftware „CRIME" übermittelt worden. Aufgrund des fehlerhaften Dateinamens sei diese Übertragung für die Ringalarmdaten aus Oberstenfeld zunächst nicht erfolgt. In der Excel-Tabelle und der Halterlistentabelle seien diese Daten jedoch vollständig erfasst gewesen. Aus sächsischen Zulassungsbezirken sei nur das Wohnmobil mit dem Kennzeichen C-PW 87 von C. H. mit ihrem Geburtsdatum, ihrem Geburtsort und einer Chemnitzer Adresse vermerkt gewesen, die restlichen fünf Wohnmobile seien in anderen Bundesländern zugelassen gewesen. Anhand der Halterdaten des Kraftfahrt-Bundesamts sei nicht zu erkennen gewesen, dass dieses Wohnmobil aus Chemnitz auf ein Unternehmen,
4: (53310):
lbezogen auszuwerten und später dann even-tuell eine Gesamtauswertung vorzunehmen. In Einzelspuren seien diese Videoauswertungen aber schon einbezogen worden. Die Videoaufzeichnung der Gaststätte „Bukowski" habe man gleich zu Beginn priorisiert ausgewertet. Dort seien auf einer Straße, die Fluchtstrecke oder Anfahrt zum Tatort hätte sein können, Fahrzeuge aufgenommen worden, deren Kennzeichen man f) Kriminalhauptkommissar a.D. H. T. Der Zeuge Kriminalhauptkommissar H. T., der von Februar 2009 bis zur Pensionierung am 26. April 2012 als Hauptsachbearbeiter in der SOKO „Parkplatz" tätig war, antwortete auf die Frage, ob, nachdem die SOKO „Parkplatz" im August 2010 die Listen aus der Ringalarmfahndung erstmals zusammengeführt habe, gezielt nach kontrollierten Wohnmobilen gesucht worden sei, z
5: (76552):
Der Ermittlungsständen die abgleichbaren Daten der identifizierten Altfälle miteinander und mit über 600 Dateien im BKA abzugleichen. Das Prüfverfahren habe neu entwickelt werden müssen und dementsprechend seien Dauer und Arbeitsaufwand der einzelnen Prüfschritte nicht zu prognostizieren gewesen. Von den 240 bundesweit übersandten Treffermeldungen seien 34 auf Baden-Württemberg entfallen. ziffer – wobei der Fall „W. W.", bei dem die Täter bekannt seien, sehr offensichtlich anzuerkennen gewesen wäre – und ob die Untersuchung tatsächlich so ordentlich abgelaufen sei wie gewünscht, wenn bei 745 Altfällen keiner davon einen PMK-rechts-Bezug habe, antwortete der Zeuge D., dass nicht allein 745 Fälle, sondern durch die Kollegen in allen Ländern über 3 ersten bis zur letzten Blattseite de
6: (52092):
sich bei der polizeilichen Fahndung gemäß der PDV 384.1 um eine „Verschluss-sache – Nur für den Dienstgebrauch" handele. Bei den Ringalarmstellen habe sich um sogenannte Durchfahrtskontrollen und keine Anhaltekontrollen gehandelt, weil es zu diesem Zeitpunkt

keine konkreten Fahndungshinweise auf Personen und Fahrzeuge gegeben habe. Das bedeute, dass lediglich die durchfahrenden Fahrzeuge erfasst worden seien. Zwischen 14:30 und 14:37 Uhr sei an einem Kontrollpunkt in Oberstenfeld, aus Beilstein kommend, ein Wohnmobil mit dem amtlichen Kennzeichen C – für Chemnitz – PW 87 als 20. Fahrzeug schriftlich festgehalten worden. Dieser Kontrollpunkt sei von 14:30 Uhr bis 16:53 Uhr von zwei Beamten des PP Großbottwar besetzt gewesen, welche dort in diesem Zeitraum 410 vorbeifahrende Fahrzeuge notiert hätten.
7: (51266):
Sie hätten, so der Zeuge KOR A. M., dann einen Beamten von der Soko „Parkplatz“ nur zur Asservatenauswertung im Hinblick auf mögliche Bezüge zu Baden-Württemberg dorthin geschickt. Dabei sei z. B. ein am 26. Juni 2003 aufgenommenes Bild gefunden worden, das wahrscheinlich Uwe Böhnhardt in der Nordbahnhofstraße in Stuttgart zeige. Im Brandschutt habe man verschiedene Stadtpläne gefunden, zwei Stadtpläne von Stuttgart mit Markierungen sowie Stadtpläne von Heilbronn und Ludwigsburg ohne Markierungen. In der Wohnung des Trios habe man verschiedene Adresssammlungen festgestellt, daraus habe man eine sogenannte „10.000er-Liste“ zusammengestellt. Einige dieser Markierungen auf dem Stuttgarter Stadtplan, 11 – wenn er sich richtig
8: (53907):
Auf weiteren Vorhalt berichtete der Zeuge KOK M. G., auf der Bundesautobahn A 6 sei am 25. April 2007 im Bereich Heilbronn um 13:05 Uhr ein Pkw BMW mit dem amtlichen Kennzeichen S-KI 2750 geblitzt worden, wobei dieses Kennzeichen als Behördenkennzeichen für die amerikanische Zulassungsstelle ausgegeben gewesen sei. Eine Halterabfrage habe ergeben, dass das Fahrzeug auf das US-Militär zugelassen ge Im Zuge der LKA Baden-Württemberg Videoaufnahmen und Bildmaterial im Umkreis des Tatorts, auch Aufnahmen aus Geschwindigkeitsüberwachungsmaßnahmen, erhoben worden. Auf die Frage, ob überprüft worden sei, was der Fahrzeughalter zum Zeitpunkt der Geschwindigkeitsmessung in der Heilbronner Gegend gemacht habe, wiederholte der Zeuge KOK M. G., dass bei der US-Botschaft nachgefragt worden sei, ob es i
9: (75143):
Weiter gab er zu bedenken, dass eine Überprüfung sämtlicher Ferienhäuser, Pensionen und dergleichen aufgrund deren Anzahl grenz- und überschaubar. Für die Überprüfung von Campingplätzen habe man sich entschieden, da deren Anzahl „enger umgrenzbar“ sei und man zudem „auf unheimlich viele Campingplatzanmietungen schon vorher einen Hinweis“ gehabt habe. Landtag von Baden-Württemberg Drucksache 16 / 5250 92 M. S. M.-F. B. vermittelt. Dieser habe mithilfe seines Personalausweises und seiner Geburtsurkunde unterstützt. Man habe bei der Stadt Chemnitz einen Reisepass mit falschem Lichtbild beantragt, also ein Echtdokument mit einem falschen Lichtbild erstellen lassen. Ob dieses Echtdokument in Baden-Württemberg benutzt worden sei, habe man Sachbearbe
10: (53684):
Er habe, so der Zeuge KOR F. H., am 31. Mai 2007 die Meldung erhalten, dass eine Spur am Opferfahrzeug festgestellt worden sei, die Bezüge zu 21 Kriminalfällen aufweise. Diese DNA-Spur sei von den anderen Dienststellen in Deutschland und Österreich und dann auch von ihnen als tatrelevant eingestuft worden. Als Konsequenz dieser DNA-Spur sei

die Sonderkommission neu strukturiert und das LKA Baden-W Im Februar 2009 sei die Sonderkommission wegen der Dauerbelastung innerhalb der Polizeidirektion Heilbronn in das LKA Baden-Württemberg verlagert worden, um die Direktion zu entlasten. Es seien gemischte Teams gebildet worden, um einen Wissenstransfer zu gewährleisten. Am 17. März 2009 sei die DNA-Spur auf einem Fingerabdruckblatt festgestellt worden. Ab diesem Zeitpunkt sei klar gewesen, dass kei

BU_G results:
1: (31470):
kehrt auch. Das nicht, aber eine Abstimmung, dass man hier nicht etwa mehrfach vertreten ist, das hat es schon gegeben.“ 4218 Der Zeuge Schmalzl, der von August 2005 bis Dezember 2007 Präsident des LfV Baden-Württemberg war, hat 4216) Schreiben des Innenministeriums Baden-Württemberg vom 27. August 2012, MAT A BW-8/3, Bl. 6, 7. 4217) Schreiben des Innenministeriums Baden-Württemberg vom 27. August 2012, MAT A BW-8/3, Bl. 6, 7.
2: (79419):
Landtag von Baden-Württemberg Drucksache 16 / 5250 844 problematik mit GBA bzw. OLG München kämen und was sie „halt vielleicht auch aushalten“ müssten. Um Erklärung gebeten, weshalb die aus dem Jahr 1998 stammende Telefonliste des NSU bzw. „Garagenliste“ erst derart spät, wohl 2012, an Baden-Württemberg übermittelt worden sei, verwies die Zeugin an Thüringen oder das BKA; sie selbst wisse es nicht. nummern stehen würden, diese Personen und Telefonnummern dann überprüftund dem BKA mitgeteilt worden seien, bejahte die Zeugin. Nach Vorhalt der Bewertung des mit der Auswertung der Adressliste befassten Kriminalhauptkommissars B. („Bei den weiterhin aufgefundenen Notizzetteln mit Adressen handelt es sich zum Teil um Adressen bekannter Personen der rechtsextremistischen bundesdeutschen Szene.
3: (2837):
Reißmann in Bezug genommenen „Abstimmungen“, die angeblich stattfanden. Die vorliegenden Unterlagen enthalten keinerlei Hinweise auf solche Abstimmungen. 3015 3. UA, Protokoll Gunter Rechenberg v. 22.10.2013, S. 12. 3016 3. UA, Protokoll Uwe Reißmann v. 22.10. schau aller Unterlagen und Aussagen auch unbestrittenen – engen Zusammenarbeit der einzelnen Dienststellen nicht als notwendig angesehen worden sei: „Stellv. Vors. Kerstin Köditz: Ist denn mal überlegt worden, eine Sonderkommission
4: (51680):
Landtag von Baden-Württemberg Drucksache 15 / 8000 122 ren. Ob weitere Unterlagen in dieser Richtung existiert haben sollten, entziehe sich ihrer Kenntnis. Zunächst habe das LfV gar nicht genau gewusst, worin der Hinweis von Amtsrat a.D. G. S. genau bestanden habe, da seine Vernehmung lediglich beim BKA vorgelegen habe. Auf der rat a.D. G. S. gleich an das BKA und nicht zuerst an das LfV als seine frühere Behörde gewandt habe, wenn er rückblickend gemeint habe, irgendetwas Sachdienliches beitragen zu können. Sie, so die Zeugin Präsidentin Beate Bube, würde es nach aller Lebenserfahrung als das Normalste der Welt empfinden, wenn man in einem solchen Fall zunächst erst noch einmal mit den früheren Kollegen oder Vo

5: (29863):
Drucksache 17/14600 – 328 – Deutscher Bundestag – 17. Wahlperiode dass diese von einem BKA-Beamten verfasst wurden. 2855 Der Zeuge Brümmendorf hat diesen Umstand damit erklärt, dass die im Schreibprogramm des LKA Thüringen 2857 e) Praktikum der Beamtin Beischer-Sacher beim LKA Thüringen im Frühjahr 1997 Bereits im Jahr 1997 hatte die Beamtin Beischer-Sacher für vier Wochen beim LKA Thüringen hospitiert. 2858 Aus diesem Zeitraum waren ihr bereits die Namen von Zschä
6: (78450):
Nochmals angesprochen auf die vom Zeugen R. geschilderte Drohung, Kontakte bei der Presse auszuspielen, beteuerte der Zeuge K., dass dies nicht stimme. Er habe über einen längeren Zeitraum Kontakte zu zwei Redakteuren des FOCUS in Frankfurt gehabt; sie hätten intensiv und extensiv die „IJU-/Sauerland-Geschichte“ diskutiert und durchgekaut, aber weitere Kontakte habe es da nicht gegeben. Landtag von Baden-Württemberg Drucksache 16 / 5250 675 heute noch existent seien. Dies habe ihm 2004/2005 sein damaliger Einsatzoffizier erzählt, zu ges Verhältnis gehabt habe. Auf Nachfrage, was er im Zusammenhang mit der Gefährderliste als „hart an der Legalität“ gemeint habe, erläuterte der Zeuge, dass diese als „Streng Geheim“ eingestuft gewesen sei, weshalb man sie als Deutscher ohne entsprec
7: (77634):
Es lohne sich nicht wirklich, sich für welche einzuset Landtag von Baden-Württemberg Drucksache 16 / 5250 532 zen, die das te sie dies. Es sei zutreffend, dass sie deswegen das jetzt nicht mehr mitmache. Zur Wahl gehe sie indes. nd besten Freundinnen. Die habe damals schon weiter weggewohnt; da fahre man halt abends nicht mehr heim, wenn man schon mal da sei. Die sei vorher bei der „Deutschen Stimme“ beschäftigt und auch in der NPD gewesen. Dann gebe es noch K. F. aus Bayreuth, die aber nicht mehr in der Szene sei. Gefragt, ob auch Leute aus Ostdeutschland dabei gewesen seien – H. [wohl R. H. gemeint] komme ja aus Ostdeut
8: (56808):
er E-Mail abgestellt. Die Tabelle ist mitdem Staatsministerium und dem Justizministerium abgestimmt. Im Ergebnis ist festzustellen: 91 Landtag von Baden-Württemberg Drucksache 15 / 8000 63 1. Beweisbeschlüsse eines Untersuchungsausschusses des Bundestages werden von den Ländern im Wege der Amtshilfe beantwortet (§ 18 Abs. Soweit der Landes Fristen gesetzt wurden, wurden sie eingehalten (vgl. Beweisbeschlüsse BW-13 [Akten zu Personen auf „Telefonliste Mundlos“] und BW-15 [Regelungen über Auswahl von V-Personen]). 2. Die erstmalige Auswertung und Vervielfältigung von Akten nimmt naturgemäß mehr Zeit in Anspruch als spätere Recherchen in demselben Aktenbestand.
9: (79329):
aus aber auch Ab-sprachen auf Sachbearbeiterebene. Insbesondere was den Komplex Ludwigsburg angelangt sei, habe ihre Stellvertreterin, Frau R., die auch den Komplex federführend bearbeitet habe, eigenständig mit dem BKA abgestimmt. Sie selbst habe natürlich Informationen davon gehabt. Frau R. habe aber über die alten Kontakte von damals verfügt und sie [die Zeugin] habe gewusst, dass Frau R. das gewissenhaft mache. m Landtag von Baden-Württemberg Drucksache 16 / 5250 829 Grundsatz her müsse

man sagen – so seien sie mit dem BKA verblieben und das sei nach ih–, dass die StPOErmittlungen des BKA und GBA vor den polizeirechtlichen Ermittlungen ihrer Stelle zu sehen seien. Das heiße, wenn beide jetzt an einer Person dran gewesen seien, dann hätten sie ihre Ermittlungen zurückgestellt, bis das BKA ihnen grünes Lic
10: (22581):
stellt werden. Im Rahmen des Prozesses vor dem Oberlandesgericht München ließ sich Beate Drucksache 18/12950 – 1048 – Deutscher Bundestag – 18. Wahlperiode Zschäpe im Dezember 2015 schriftlich dahingehend ein, es sei Uwe Böhnhardt gewesen, der ren Ausschnitte dieser Videoaufzeichnungen abgespeichert, die im Rahmen der Fahndung veröffentlicht worden waren. Beim Abspeichern auf dem Computer waren für die einzelnen Ausschnitte Dateibezeichnungen mit den Spitznamen von Böhnhardt („Gerri“) und Mundlos

List of Figures

List of Tables

Bibliography

[1] Nathalie Henry Riche, Christophe Hurter, Nicholas Diakopoulos et al. *Data-Driven Storytelling*. USA: A. K. Peters, Ltd., 2018.

[2] Gregor Wiedemann, Seid Muhie Yimam and Chris Biemann. 'New/s/leak 2.0 – Multilingual Information Extraction and Visualization for Investigative Journalism'. In: *Social Informatics*. Ed. by Steffen Staab, Olessia Koltsova and Dmitry I. Ignatov. Cham: Springer International Publishing, 2018, pp. 313–322.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee et al. 'BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding'. In: 2018. arXiv: 1810.04805.

[4] M. Brehmer, S. Ingram, J. Stray et al. 'Overview: The Design, Adoption, and Analysis of a Visual Document Mining Tool for Investigative Journalists'. In: *IEEE Transactions on Visualization and Computer Graphics* 20.12 (2014), pp. 2271–2280.

[5] Mark Senn. *Intella Connect™ User Manual*. 2017. URL: https://www.vound-software.com/docs/connect/1.7.2/reviewer/03_02_keyword_search.html?highlight=search#search-query-syntax (visited on 15th June 2020).

[6] Pandu Nayak. *Understanding searches better than ever before*. 2019. URL: https://www.blog.google/products/search/search-language-understanding-bert/ (visited on 15th June 2020).

[7] Alex Wang, Amanpreet Singh, Julian Michael et al. 'GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding'. In: *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*. Brussels, Belgium: Association for Computational Linguistics, November 2018, pp. 353–355. DOI: 10.18653/v1/W18-5446. URL: https://www.aclweb.org/anthology/W18-5446.

[8] Daniel Cer, Mona Diab, Eneko Agirre et al. 'SemEval-2017 Task 1: Semantic Textual Similarity Multilingual and Crosslingual Focused Evaluation'. In: *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)* (2017). DOI: 10.18653/v1/s17-2001. URL: http://dx.doi.org/10.18653/v1/S17-2001.

[9] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze. *Introduction to Information Retrieval*. USA: Cambridge University Press, 2008.

[10] G. Salton, A. Wong and C. S. Yang. 'A Vector Space Model for Automatic Indexing'. In: *Commun. ACM* 18.11 (November 1975), pp. 613–620. DOI: 10.1145/361219.361220. URL: https://doi.org/10.1145/361219.361220.

[11] Apache Software Foundation. *Lucene Class Description for TFIDFSimilarity*. 2020. URL: https://lucene.apache.org/core/8_5_2/core/org/apache/lucene/search/similarities/TFIDFSimilarity.html (visited on 16th June 2020).

[12] Abhishek Dubey, Ayush Gupta, Nitish Raturi et al. 'Item-Based Collaborative Filtering Using Sentiment Analysis of User Reviews'. In: *Applications of Computing and Communication Technologies*. Ed. by Ganesh Chandra Deka, Omprakash Kaiwartya, Pooja Vashisth et al. Singapore: Springer Singapore, 2018, pp. 77–87.

[13] Elasticsearch B.V. *Elasticsearch similarity Documentation*. 2020. URL: https://www.elastic.co/guide/en/elasticsearch/reference/current/similarity.html (visited on 16th June 2020).

[14] Stephen Robertson, S. Walker, S. Jones et al. 'Okapi at TREC-3'. In: *Overview of the Third Text REtrieval Conference (TREC-3)*. Gaithersburg, MD: NIST, 1995, pp. 109–126. URL: https://www.microsoft.com/en-us/research/publication/okapi-at-trec-3/.

[15] Tomas Mikolov, Kai Chen, Greg Corrado et al. 'Efficient Estimation of Word Representations in Vector Space'. In: (2013). arXiv: 1301.3781.

[16] Zellig S. Harris. 'Distributional Structure'. In: WORD 10.2-3 (1954), pp. 146–162. DOI: 10.1080/00437956.1954.11659520.

[17] Matthew Peters, Mark Neumann, Mohit Iyyer et al. 'Deep Contextualized Word Representations'. In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics, June 2018, pp. 2227–2237. DOI: 10.18653/v1/N18-1202.

[18] Zhenzhong Lan, Mingda Chen, Sebastian Goodman et al. 'ALBERT: A Lite BERT for Self-supervised Learning of Language Representations'. In: (2019). arXiv: 1909.11942.

[19] Yinhan Liu, Myle Ott, Naman Goyal et al. 'RoBERTa: A Robustly Optimized BERT Pretraining Approach'. In: (2019). arXiv: 1907.11692.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar et al. 'Attention is All you Need'. In: *Advances in Neural Information Processing Systems 30*. Ed. by I. Guyon, U. V. Luxburg, S. Bengio et al. Curran Associates, Inc., 2017, pp. 5998–6008. URL: http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf.

[21] M. Schuster and K. Nakajima. 'Japanese and Korean voice search'. In: *2012 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*. 2012, pp. 5149–5152.

[22] Jay Alammar. *The Illustrated Transformer*. 2018. URL: http://jalammar.github.io/illustrated-transformer/ (visited on 12th August 2020).

[23] Iulia Turc, Ming-Wei Chang, Kenton Lee et al. 'Well-Read Students Learn Better: On the Importance of Pre-training Compact Models'. In: (2019). arXiv: 1908.08962v2.

[24] Victor Sanh, Lysandre Debut, Julien Chaumond et al. 'DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter'. In: (2019). arXiv: 1910.01108.

[25] Yukun Zhu, Ryan Kiros, Rich Zemel et al. 'Aligning Books and Movies: Towards Story-Like Visual Explanations by Watching Movies and Reading Books'. In: *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. USA: IEEE Computer Society, 2015, pp. 19–27. DOI: 10.1109/ICCV.2015.11.

[26] Nils Reimers and Iryna Gurevych. 'Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks'. In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, November 2019. URL: http://arxiv.org/abs/1908.10084.

[27] Nils Reimers and Iryna Gurevych. 'Making Monolingual Sentence Embeddings Multilingual using Knowledge Distillation'. In: *arXiv preprint arXiv:2004.09813* (April 2020). URL: http://arxiv.org/abs/2004.09813.

[28] Bryan Klimt and Yiming Yang. 'The Enron Corpus: A New Dataset for Email Classification Research'. In: *Proceedings of the 15th European Conference on Machine Learning*. ECML'04. Pisa, Italy: Springer-Verlag, 2004, pp. 217–226. DOI: 10.1007/978-3-540-30115-8_22.

[29] Gregor Wiedemann, Seid Yimam and Chris Biemann. 'Multilingual Information Extraction Pipeline for Investigative Journalism'. In: *EMNLP - Software demonstrations*. 2018, pp. 78–83.

[30] VG WORT. *Definition of a normed page from VG Wort*. 2020. URL: https://www.vgwort.de/auszahlungen/wissenschaftliche-publikationen/fach-und-sachzeitschriften.html (visited on 16th June 2020).

[31] J. Johnson, M. Douze and H. Jégou. 'Billion-scale similarity search with GPUs'. In: *IEEE Transactions on Big Data* (2019), pp. 1–1.

[32] Alan Akbik, Duncan Blythe and Roland Vollgraf. 'Contextual String Embeddings for Sequence Labeling'. In: *COLING 2018, 27th International Conference on Computational Linguistics*. 2018, pp. 1638–1649.