



**NOVA**

**IMS**

Information  
Management  
School

**MAA**

---

**Mestrado em Métodos Analíticos Avançados**

Master Program in Advanced Analytics

## **Customer Review Analysis**

**Philipp Tueschen**

Internship Report presented as the partial requirement  
for obtaining a

Master's degree in Data Science and Advanced  
Analytics

**NOVA Information Management School**  
**Instituto Superior de Estatística e Gestão de Informação**

Universidade Nova de Lisboa

NOVA Information Management School  
Instituto Superior de Estatística e Gestão de Informação  
Universidade Nova de Lisboa

# Customer Review Analysis

by

**Philipp Tüschen**

Internship Report presented in partial fulfillment of the  
requirements for the degree of  
Master of Science in Data Science and Advanced Analytics

**Supervisor:** Flávio Luís Portas Pinheiro

**External Supervisor:** Patrick Dubovy

November 2021

Copyright © Philipp Tueschen, Information Management School, NOVA University Lisbon.

The Information Management School and the NOVA University Lisbon have the right, perpetual and without geographical boundaries, to file and publish this dissertation through printed copies reproduced on paper or on digital form, or by any other means known or that may be invented, and to disseminate through scientific repositories and admit its copying and distribution for non-commercial, educational or research purposes, as long as credit is given to the author and editor.

## ACKNOWLEDGEMENTS

First of all I would like to thank cerascreen for giving me the opportunity to work for them for nearly a year. This time gave me valuable insights and experiences in and around the area of advanced analytics while being fortunate to work remotely for cerascreen. A special thanks goes to Patrick Dubovy, he was the main driver to create this opportunity and giving me the room to grow in my own project. Working together on many different side projects I could further grow my knowledge in Data Science. Next to Patrick I would like to thank Jonas Leible for the close collaboration on some parts of the project. Together Jonas and Patrick were always approachable and open for discussion to help improve the quality of the project. Lastly, I would like to thank my academic thesis advisor Flávio Pinheiro of NOVA IMS. Before and during the internship he was there to give me great feedback and advice to keep me challenged, also during the project.

## ABSTRACT

---

Over the last years, Cerascreen has grown rapidly and expanded into more than 20 countries, always focusing on offering more diverse products, supplements, and services. Unfortunately, it collected a lot of data during these years, which was not yet stored, losing valuable insights. In a new initiative Cerascreen wants to be the most trusted digital predictive health platform. Therefore, it intends to utilize its data to understand its customers better and offer superior products and services according to the customer's needs. The focus of this internship report was to find a way to analyze Cerascreen's customers' reviews to understand its customers better and respond to properly the given feedback. In addition, since the reviews have not been stored before, this report also deals with review retrieval. An exploratory data analysis of the reviews' ratings and texts was conducted to find the first significant insights. The investigation found that although the overall review consensus was positive, it differed by country, while the reviews' length was related to their ratings. A topic model was developed to find more information on what customers are talking about. The Model was able to find several different topics, including product-, supplement-, and service-specific reviews. Lastly, a newly created key performance indicator about customers satisfaction uses the new insights about the ratings and the review topics, which a dashboard partially visualized through a dashboard.

**Keywords:** BERTopic, Sentence Embeddings, Text Mining, Topic Modeling, Unsupervised Learning

---

# CONTENTS

|                                                     |            |
|-----------------------------------------------------|------------|
| <b>List of Figures</b>                              | <b>v</b>   |
| <b>List of Tables</b>                               | <b>vi</b>  |
| <b>Acronyms</b>                                     | <b>vii</b> |
| <b>1 Introduction</b>                               | <b>1</b>   |
| 1.1 Company Overview . . . . .                      | 2          |
| 1.2 The Team and Activities . . . . .               | 2          |
| <b>2 Theoretical Framework</b>                      | <b>3</b>   |
| 2.1 Online Reviews . . . . .                        | 3          |
| 2.2 Text Mining . . . . .                           | 4          |
| 2.2.1 Basics of Text Mining . . . . .               | 4          |
| 2.2.2 Text Mining Applications . . . . .            | 5          |
| 2.2.3 Data Preprocessing . . . . .                  | 5          |
| 2.3 Topic Modeling . . . . .                        | 6          |
| 2.3.1 Vectorization . . . . .                       | 7          |
| 2.3.2 Dimensionality Reduction . . . . .            | 9          |
| 2.4 Unsupervised Learning . . . . .                 | 10         |
| 2.4.1 K-Means: A Centroid-Based Technique . . . . . | 10         |
| 2.4.2 Hierarchical Clustering . . . . .             | 12         |
| 2.4.3 Density-Based Clustering . . . . .            | 14         |
| <b>3 Project</b>                                    | <b>18</b>  |
| 3.1 Business Understanding . . . . .                | 18         |
| 3.2 Data Understanding . . . . .                    | 19         |
| 3.2.1 Data Retrieval . . . . .                      | 19         |
| 3.2.2 Data Sets . . . . .                           | 20         |
| 3.3 Data Preparation / Exploration . . . . .        | 21         |
| 3.3.1 Overall Data Set . . . . .                    | 21         |
| 3.3.2 Text Corpora Preparation . . . . .            | 23         |
| 3.3.3 Final Text Corpora . . . . .                  | 25         |

## CONTENTS

---

|                                            |           |
|--------------------------------------------|-----------|
| 3.3.4 Customer Reviews . . . . .           | 26        |
| 3.4 Topic Modeling . . . . .               | 30        |
| 3.4.1 General Information . . . . .        | 30        |
| 3.4.2 Comparison of Topic Models . . . . . | 30        |
| 3.4.3 BERTopic . . . . .                   | 32        |
| 3.5 Application . . . . .                  | 37        |
| <b>4 Conclusion and Discussion</b>         | <b>38</b> |
| <b>Bibliography</b>                        | <b>40</b> |
| <b>Appendices</b>                          | <b>43</b> |
| <b>A Different Topic Models Comparison</b> | <b>43</b> |
| <b>B BERTopic whole topic table</b>        | <b>48</b> |
| <b>C Topic Translations English</b>        | <b>51</b> |

## LIST OF FIGURES

|    |                                                                                                                                                                        |    |
|----|------------------------------------------------------------------------------------------------------------------------------------------------------------------------|----|
| 1  | Text Mining Process . . . . .                                                                                                                                          | 4  |
| 2  | S-BERT for sentence similarity (Reimers and Gurevych, 2019) . . . . .                                                                                                  | 9  |
| 3  | K-means algorithm. The training examples are the dots, and the centroids are the dots with lines around them (Page et al., 2014). . . . .                              | 11 |
| 4  | The elbow plot, where the best amount of clusters seems to be three based on the drop in change of explained variation (total within-cluster sum of squared) . . . . . | 12 |
| 5  | Agglomerative and divisive hierarchical clustering on data objects (Halkidi, 2009) . . . . .                                                                           | 14 |
| 6  | HDBscan output the stronger the membership, the more saturated is the color. (Leland McInnes, 2016) . . . . .                                                          | 17 |
| 7  | The rating distribution per country . . . . .                                                                                                                          | 22 |
| 8  | The total rating for the DACH region and the United Kingdom . . . . .                                                                                                  | 22 |
| 9  | The total rating for the DACH region . . . . .                                                                                                                         | 23 |
| 10 | The total rating for the United Kingdom . . . . .                                                                                                                      | 23 |
| 11 | The average rating per country per year . . . . .                                                                                                                      | 24 |
| 12 | The distribution of the text length of all reviews . . . . .                                                                                                           | 27 |
| 13 | The distribution of the text length of all reviews per rating . . . . .                                                                                                | 27 |
| 14 | Uni-gram for positive and negative ratings . . . . .                                                                                                                   | 28 |
| 15 | Bi-gram for positive and negative ratings . . . . .                                                                                                                    | 28 |
| 16 | The elbow plot for all reviews where the distortion score is the within-cluster sum of squared . . . . .                                                               | 31 |
| 17 | BERTopic clusters . . . . .                                                                                                                                            | 35 |
| 18 | K-Means with different dimensionality reductions . . . . .                                                                                                             | 44 |
| 19 | Agglomerative Clustering with different dimensionality reductions . . . .                                                                                              | 45 |
| 20 | HDBScan with different dimensionality reductions . . . . .                                                                                                             | 47 |

## LIST OF TABLES

|    |                                                                           |    |
|----|---------------------------------------------------------------------------|----|
| 1  | Numbers of reviews per source and its connection . . . . .                | 19 |
| 2  | Number of reviews per country and source . . . . .                        | 21 |
| 3  | Frequency of the review length . . . . .                                  | 27 |
| 4  | Most important words from all ratings . . . . .                           | 29 |
| 5  | The cluster models' silhouette score . . . . .                            | 31 |
| 6  | The topic models' silhouette score and amount of topic clusters . . . . . | 33 |
| 7  | BERTopic most meaningful clusters . . . . .                               | 36 |
| 8  | Sentence embeddings with Umap and K-mean . . . . .                        | 44 |
| 9  | Sentence embeddings with Umap and Agglomerative Clustering . . . . .      | 46 |
| 10 | Sentence embeddings with Umap and HDBscan . . . . .                       | 46 |
| 11 | BERTopic whole topic table a . . . . .                                    | 48 |
| 12 | BERTopic whole topic table b . . . . .                                    | 49 |
| 13 | BERTopic whole topic table c . . . . .                                    | 50 |
| 14 | Most important words from all ratings . . . . .                           | 51 |
| 15 | Sentence embeddings with Umap and Kmean . . . . .                         | 52 |
| 16 | Sentence embeddings with Umap and Agglomerative Clustering . . . . .      | 52 |
| 17 | Sentence embeddings with Umap and HDBscan . . . . .                       | 52 |
| 18 | BERTopic most meaningful clusters . . . . .                               | 53 |
| 19 | BERTopic whole topic table a . . . . .                                    | 54 |
| 20 | BERTopic whole topic table b . . . . .                                    | 55 |
| 21 | BERTopic whole topic table c . . . . .                                    | 56 |

## ACRONYMS

**API** Application Programming Interface

**BERT** Bidirectional Encoder Representations from Transformers

**BOW** Bag-of-words

**DACH** The region of Germany, Austria and Switzerland

**DBscan** Density-Based Spatial Clustering of Applications with Noise

**GmbH** Gesellschaft mit beschränkter Haftung

**HDBscan** Hierarchical Density-Based Spatial Clustering of Applications with Noise

**KPI** Key Performance Indicator

**NLP** Natural Language Processing

**PCA** Principle Component Analysis

**S-BERT** Sentence-BERT

**t-SNE** t-Distributed Stochastic Neighbor Embedding

**TF-IDF** Term Frequency Inverse Document Frequency

**UMAP** Uniform Manifold Approximation and Projection



## INTRODUCTION

Over the last decades, new technologies have enabled people to measure their fitness and track their health routines more carefully. It made them more conscious about their wellness behavior, like their diet, and more willing to share health-related data (Betts, 2020). One of the companies that began to fulfill this demand in Germany is Cerascreen GmbH (hereafter Cerascreen). As a result, Cerascreen, founded in 2012, rapidly grew, becoming the market leader in home health testing with the idea of making it available to everyone. Especially in the last years, Cerascreen rapidly expanded and operates in more than 20 countries, analyzing over 250,000 tests per year. However, to stay competitive in the market and become more profitable, Cerascreen is constantly trying to offer new, innovative solutions for its customer's health.

By offering new and innovative solutions to its customers, Cerascreen collected a lot of data over the years. To reach its goal of becoming the most trusted digital predictive health platform worldwide, it started to analyze all its data. Besides analyzing laboratory data to develop new insights and products, Cerascreen wants to understand its customers better. For this reason, Cerascreen started a project to understand what customers are saying about the company's products and services on different online platforms. Cerascreen currently has access to its customer reviews, but these are only available in various sources and other data formats. Therefore, the data cannot be processed or analyzed. Accordingly, Cerascreen is losing valuable insights into what customers think about the products, the company, and services. This unused knowledge means that Cerascreen does not recognize the customers' problems, ideas, and wishes and therefore cannot respond to them and improve itself. For these reasons, the text mining project aims to better understand customers and respond to the given feedback.

### 1.1 Company Overview

Cerascreen is a young health company that develops easy-to-use home health tests, carefully selected nutritional supplements, coaching, and courses around health nutrition and fitness. This broad offer in services and products underlines Cerascreen's belief that everyone should be able to monitor their health and understand their own body's needs. This way, they can take precise preventative actions to live healthy and happy instead of reacting to illness. Since Cerascreen's launch, its customer base has grown to over 200,000 customers across Europe, the US, and Canada. Although most customers live in Germany, the international market continuously increases, since nowadays more and more people are taking their health into their own hands.

Cerascreen's business concept mainly focuses on a business-to-consumer relationship. The process starts with the customer's purchase of a test online. The customer receives the test a few days later by mail and performs the test according to the instructions. Afterward, the samples are sent by mail to an independent laboratory for analysis. The results are then enhanced with clear explanations to better understand the necessary steps towards fitter health before everything is sent to the customer. The customer can then take action, for example, by purchasing supplements supporting the body in case of any deficiency.

### 1.2 The Team and Activities

Within Cerascreen, I am part of the newly created data science team within the science department. This team consists of one lead data scientist and me. The job functions of the team members vary between data analytics, data engineering, and advanced analytics. The team is operating in two areas with different scopes. One part deals with research studies using the accumulated health data and trying to find new patterns, for example, concerning Epigenetics. The second area focuses on supporting several departments with knowledge through actionable analysis and dashboards to monitor key performance indicators (KPI). Since the department is very young and many data sources haven't been used before, retrieving and storing data in a structured manner is also part of many analyses. Although I was the project owner, the lead data scientist assisted with external expertise to obtain the required data, validate the gained insights, and decide on the next steps.

## THEORETICAL FRAMEWORK

This section provides a relevant theoretical background on text mining, data cleaning techniques, and natural language processing (NLP) methods applied to the problem. Text mining is the process of discovering and extracting relevant information for a current issue from unstructured text (Kao and Poteet, 2007). Data cleaning is the process of dealing with inconsistent, incomplete, and noisy real-world data by altering or deleting respective data points to create a more accurate data set (Han et al., 2011). NLP utilizes techniques of different areas in artificial intelligence, computational linguistics, mathematics, and information science to help the computer understand natural language (Zhang et al., 2019). In NLP, topic modeling is a clustering technique to find clusters of words that are frequently used together that can be defined as topics across documents (Hassani et al., 2020).

### 2.1 Online Reviews

In the digital era, the significance of online reviews steadily grows. People shop online on multiple devices on every imaginable e-commerce platform. The same people give feedback, packed with their expectations and desires, about the products and services they consume, thereby feeding a review culture that influences other people in their buying behavior (Kumar, 2021). The strong influence of the review culture is mainly because customer reviews and ratings are considered as equally or more trustworthy compared to information on official websites. This collaboration of people sharing information online is called electronic word of mouth (Verma and Yadav, 2021).

Furthermore, positive and negative thoughts have different weights as many factors can lead to customer dissatisfaction, which is often even more diverse than stated

in the online reviews themselves. The phenomenon that no variable guarantees satisfaction but many variables that can cause dissatisfaction is called the Anna Karenina principle (Kirilenko et al., 2021). Hence, analyzing these reviews is of great interest for companies nowadays to find as many insights as possible on users' concerns and complaints to improve products and services further. It is essential to find the customer's dissatisfaction since it has a higher weight, amplified by a star rating. Finding and dealing with customer dissatisfaction is why it is crucial to conduct a topic analysis to understand, classify and generalize all topics customers talk about concerning the company (Kirilenko et al., 2021; Verma and Yadav, 2021).

## 2.2 Text Mining

The potential of text mining steadily grows since its ability to extract knowledge from unstructured textual data and find significant patterns and correlations is continuously improving. The use of data and especially text mining to drive decision-making within companies is of increasing importance to stay competitive (Hassani et al., 2020). The same applies to Cerascreen and this project, where text mining techniques are used to construct solutions to the problem at hand.

### 2.2.1 Basics of Text Mining

The names *text mining* and *data mining* are terms used to define the field that extracts knowledge from data. Since data mining tools are for the process of discovering patterns from structured data, they can deal with semi-structured or unstructured textual data, such as texts on the web or locally stored full-text documents. Thus, text mining is the process of finding previously unknown information from a diverse set of written sources. Structured data resides in predefined formats, like a relational database or spreadsheets. On the other hand, unstructured data usually does not reside in a traditional relational database and is stored in its natural form until its application in analysis. Finally, semi-structured data is a combination of the before mentioned data forms (Vijayarani et al., 2015). Figure 1 provides an overview of the text mining process with its primary activities.

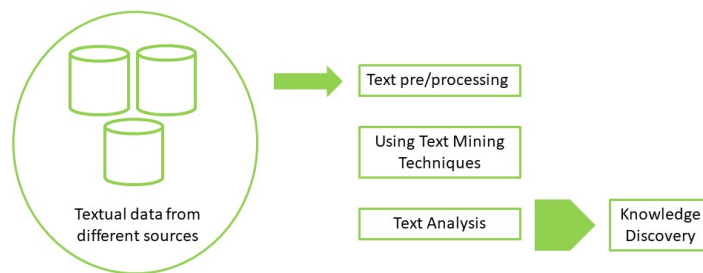


Figure 1: Text Mining Process

### 2.2.2 Text Mining Applications

Information retrieval, information extraction, categorization, and NLP are the main application areas of text mining. Information retrieval is the concept of retrieving information from multiple text-based documents relevant to a specific information need. Therefore, the idea of information retrieval is applied in software systems, also known as document management systems, and web search engines. Information extraction methods identify essential information within text through pattern matching and thus enable finding entities to classify and store them in a database. Further, extraction methods also find their application in feature engineering through sentiment analysis (Al-Twairesh and Al-Negheimish, 2019). It is beneficial for large amounts of text and information only available in natural language.

Categorization identifies the main topics of documents through a predefined set of themes. For example, it counts the words that appear the most in a document, which are then labeled as the document's topics and grouped according to these topics (Vijayarani et al., 2015). NLP is a research area that explores computers' ways of understanding and analyzing any product of natural human communication, such as semantic meanings and grammatical structures. The goal is that computers can understand natural language like a human, for example, understanding the feelings described, to perform functional tasks. Hence, NLP is an interdisciplinary field, which is a combination of computational linguistics and artificial intelligence, which finds its typical applications in areas of text processing, summarization, cross-lingual information extractions, speech recognition, etc. (Vijayarani et al., 2015; Zhang et al., 2019).

### 2.2.3 Data Preprocessing

An essential part of any text mining process is checking for correct and valid data. Several data preprocessing steps are applied to obtain correct, valid data. Usually, preprocessing consists of transformation steps such as standardization, normalization, one-hot encoding, binning, or aggregation (Han et al., 2011). The data transformation methods are the main emphasis during the preprocessing of the data. Data transformations convert data into a format, so that machine learning models can use it to draw relevant conclusions from data analysis. Since these steps are applied to numerical data, text mining preprocessing consists of different steps to clean and transform the textual data to reduce noise in the text and enhance the topic modeling performance. (Zhang et al., 2019). Common steps to normalize text are:

**Lower case and punctuation removal:** This transformation turns all words into lower case, eliminating the same word's variation. For example, Work and work are treated equally as the same word. Furthermore, since punctuations increase the size of the data set and do not add any value to the text analysis, they are removed (Zhang et al., 2019).

**Stopwords removal:** Stopwords are ubiquitous words in a language with no value and often affect the analysis negatively. Therefore, stopwords, such as "the" or "is", are removed from the text before the analysis. The classical way to remove stopwords is to use compiled lists published by different providers since stopwords can vary across domains, especially languages. Hence, experts can add some field-specific words to the compiled lists if through the acquired domain knowledge deemed necessary (Jianqiang and Xiaolin, 2017; Zhang et al., 2019).

**Reverting words:** There is the possibility that in informally written reviews or tweets, words contain repeated letters, which are ways for people to express their sentiments. For the algorithms to identify them as identical words, they have to be reverted to the original word (Jianqiang and Xiaolin, 2017).

**Tokenization:** Tokenization is the process of splitting a defined document into smaller pieces, called tokens, while characters like punctuation are filtered out. Further, the tokens can be classified into word, character, and n-gram types. N-grams are specific sequences of text that occur together within a predefined window size of uni-, bi-, and trigrams; for example, a trigram is "the brown dog" (Zhang et al., 2019).

**Stemming and Lemmatization:** In a text, a word can be written in multiple different forms, e.g., "run," "runs," "running." Further, words with the same root can be used in different syntactic categories, like "fun" and "funny." These natural and common word forms, also called inflection, make it harder for a program to understand the differences. Therefore, stemming reduces the inflection in words to their root forms by cutting off prefixes, suffixes, or infixes even if they do not make sense. Lemmatization, on the other hand, typically uses a vocabulary and morphological analysis to remove the inflectional endings and return the base or the dictionary form of a word, also known as the lemma (Samir and Lahbib, 2018; Zhang et al., 2019).

## 2.3 Topic Modeling

Topic modeling is a practical approach to examine large amounts of unclassified text in today's era of increasing interest in big data analytics among academics and corporations (Alghamdi and Alfalqi, 2015). Topic modeling has gotten much attention and much interest among scholars in a range of domains since its introduction (Liu et al., 2016). At the moment, topic modeling is one of the most efficient techniques in text mining and is utilized across different fields for analyzing various kinds of textual data; for example, for processing legal texts or for analyzing tourist reviews and research papers (Kirilenko et al., 2021; Silveira et al., 2021; Westerlund et al., 2018). It discovers topics in documents by associating words with similar meanings and differentiating between words with multiple meanings and their different applications. Thus, a topic is a cluster of frequently together occurring words.

Further, since documents consist of words, one document can be associated with more than one topic depending on the strength of the related terms. Therefore, the

user can utilize a topic model to detect hidden topics in a selection of documents generating a list of topics and a list of strongly related documents for each detected topic. In contrast, each topic should be distinct (Westerlund et al., 2018).

### 2.3.1 Vectorization

Before text mining algorithms can process the data, it has to be made readable by the computer through vectorization. Thus, text embeddings are paramount to converting text into numerical vectors before any model can be built. Therefore, different ways of vectorizing the data have developed during the last decades. Since each vectorization approach is different, choosing the right approach is paramount to the outcome of the later analysis.

#### 2.3.1.1 Bag-of-Words

The bag-of-words (BOW) model is a commonly adopted and effective approach to turn arbitrary text into fixed-length vectors. Therefore each text piece or document is vectorized as  $d = (x_1, x_2, \dots, x_l)$ , where  $x_i$  is the normalized term frequency of the  $i$ -th word and  $l$  is the size of the document. In other words, the BOW model neglects grammar and the order of words; however, it keeps the multiplicity. The mapping function in the BOW model, on the other hand, is hard or binary, that is, a crisp binary relation that solely expresses the presence or absence of words in the document. The hard mapping function has several drawbacks, among them that the BOW representation as a vector is extremely sparse, and since it treats all keywords equally, it does not consider the rareness of a word. Hence, as BOW only counts word occurrences, the model cannot accurately capture the semantics of words and documents. So, texts that are semantically similar but use different words are assigned to different vectors. Thus, a BOW model has a possible reduced performance for topic modeling since it cannot capture semantics (Zhao and Mao, 2017).

#### 2.3.1.2 Term Frequency Inverse Document Frequency

Term Frequency Inverse Document Frequency (TF-IDF) has the same basics as the BOW model as it counts how often a term occurs in a document (TF). However, instead of treating all keywords equally, the inverse document frequency assigns lower values to the most occurring words and higher values to rarer words. Hence the more infrequent a word is within a document, the higher its importance (Qaiser and Ali, 2018).

$$\text{tfidf}(t, d, D) = \text{tf}(t, d) \times \text{idf}(t, D) \quad (1)$$

where  $t$  are the terms,  $d$  represents each document, and  $D$  represents the collection of documents.

### 2.3.1.3 Word Embeddings

Word embeddings are based on the notion of assigning a low-dimensional vector to each word so that semantically similar words are at a small distance to each other in the vector space. Accordingly, the added value of word embeddings is the cosine similarity measure to evaluate the semantic similarity between two words using the widely used *word2vec* word embeddings. The cosine similarity determines how similar two data points are, regardless of their size. It is mathematically defined as the cosine of the angle between two vectors in a multi-dimensional space. The cosine similarity is positive when words have a similar meaning and negative when words are dissimilar, while a zero value indicates two different words. The *word2vec* framework has two approaches: Continuous Bag of Words, which predicts a word with the surrounding words, and Skip-gram, which predicts a window of words based on a single word. Further, its ability to work with a large amount of unannotated text makes it capable of embedding significant linguistic relationships between words (Mikolov et al., 2013; Zhao and Mao, 2017).

$$\cos(w_i, w_j) = \frac{w_i \times w_j}{\|w_i\| \times \|w_j\|} \quad (2)$$

where  $w_i$  and  $w_j$  are the word embeddings of two words.

### 2.3.1.4 Sentence Embeddings

Universal sentence encoders use pre-trained sentence-level embeddings to encode documents. These publicly available, pre-trained models have been trained on vast amounts of data from sources like Wikipedia at high computational costs by big companies like Google. Both transformer and deep averaging network-based universal encoders exceed the performance of word-level embeddings like *word2vec*. The higher performance is due to its more complex structure using a graph neural network with an attention mechanism. In this way, sentences are fully-connected word graphs, where order and identity are taken into account to build a context-aware representation of these words. Afterward, these representations are transformed into fixed-length sentence encoding vectors, which are built by taking each word positions representation and computing an element-wise sum (Cer et al., 2018). This idea developed into a new state-of-the-art performance on sentence embeddings, called Bidirectional Encoder Representations from Transformers or BERT. Nonetheless, both sentences have to be fed to BERT generating a substantial computational cost; for example, 1000 sentences need roughly 5 million computations, which are close to 65 hours, making it unsuitable for unsupervised tasks like clustering.

Therefore, Sentence-Bert (S-BERT) has been developed as an update to solve the problems of BERT. S-BERT uses siamese and triplet network structures to obtain two fixed-sized semantically meaningful sentence embeddings. Semantically meaningful

is defined as semantically similar sentences close in the same vector space that can be compared with similarity measures like the cosine or the Manhattan and Euclidean distance while keeping BERT’s accuracy. The Figure 2 demonstrates the embedding process. First, S-BERT’s siamese network-like architecture produces two sentences as an input. Next, these sentences are transferred to BERT, where a pooling layer builds their representative embeddings. Lastly, the cosine similarity calculation, for example, then uses the embeddings as input to calculate the similarity score (Reimers and Gurevych, 2019).

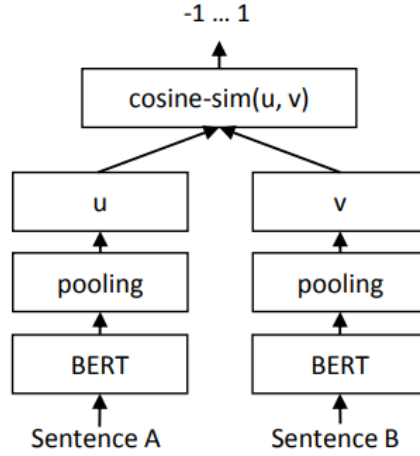


Figure 2: S-BERT for sentence similarity (Reimers and Gurevych, 2019)

### 2.3.2 Dimensionality Reduction

Most machine learning problems entail the curse of dimensionality due to the large number of features slowing down the training phase and making it challenging to find a suitable solution. On the other hand, reducing the dimensionality trades a slight loss of information and slightly declines system performance for a faster training process. Hence it is mainly used for noise reduction and data compression, classification, and visualization (Reddy et al., 2020).

The Uniform Manifold Approximation and Projection algorithm (UMAP) is faster and better scalable than the revolutionary t-Distributed Stochastic Neighbor Embedding (t-SNE), although, at their core, both are very similar to each other. UMAP is part of the nonlinear dimensionality reduction approaches that provide higher quality embeddings and visualizations for more extensive data sets while preserving the global structure with an exceptional run time performance. It reaches this performance because of building a high dimensional graph representation of the present data to use afterward a graph layout algorithm mapping the data points to an optimized, as structurally similar as possible, low dimensional graph. However, although UMAP outperforms t-SNE, it is not as interpretable as the Principle Component Analysis (PCA),

as the dimension of its embedding space has no exact meaning (McInnes et al., 2020). PCA represents the directions of the most significant variance in the source data in the lower-dimensional space. This makes it an interpretable tool for mapping data to a low-dimensional space translating between it and the original feature space since it is a linear mapping of the original features. However, although PCA’s linearity makes the method technically more concise than some nonlinear methods, it happens at the price of variance maximization and the inability to capture nonlinear phenomena in single dimensions (Lovrić et al., 2021).

## 2.4 Unsupervised Learning

Machine learning algorithms that do not need class labeled input samples are called unsupervised learning. In unsupervised learning, clustering algorithms are some of the most often used strategies to find classes within a data set based on the attributes provided in the input data set. For example, an unsupervised learning approach may be used to process a set of revenue data to discover later that there are three interesting data groups. Furthermore, these clusters may correlate to three different revenue levels. In this case, however, we cannot infer the semantic significance of the discovered clusters since the training data is not labeled. Hence, unsupervised learning approaches generally maximize intraclass similarity while decreasing interclass similarity among items in a given class. Clusters of items are created in such a way that the data points inside a cluster are very similar to one another but are distinct to the data points in other clusters (Han et al., 2011).

### 2.4.1 K-Means: A Centroid-Based Technique

K-Means is the most common partitioning methodology for clustering data based on centroid positions. As a starting point, the number of clusters  $k$  and its total number of assigned items  $n$  is specified, where  $k \leq n$  (Han et al., 2011).

A centroid-based partitioning represents a cluster  $C_i$  using its centroid  $c_i$ . The cluster’s centroid is its conceptual center point inside the feature space. The centroid can be defined, for example, by the mean or medoid of the points allocated to the cluster. Further, the function  $dist(p, c_i)$  measures the distance between an object  $p$  in cluster  $C_i$  and the centroid ( $c_i$ ) in the Euclidean space. Further, Han et al. (2011) argues that the intra-cluster variance can determine the overall quality of a cluster, which is the sum of squared errors ( $E$ ) between all of the data points in  $C_i$  and the centroid  $c_i$  defined as

$$E = \sum_{i=1}^k \sum_{p \in C_i} dist(p, c_i)^2 \quad (3)$$

While the Euclidean distance is the most common distance measure for Kmeans, other distance measures, such as the cosine (see above) and the Jaccard similarity coefficient, can be used. The Jaccard similarity coefficient compares and measures the similarity and variety of finite sample sets, defined as the intersection's size divided by the size of the union of the sample sets. The k-means approach for partitioning, represents the centroid of a cluster with the mean value of the points inside the cluster and works as follows:

1. The initial seeding randomly chooses  $k$  data points from the population  $n$  representing the first cluster centroids.
2. The most similar cluster absorbs the remaining data points, based on the data point's euclidean distance from the cluster mean.
3. The intra-variance of the clusters is then enhanced iteratively by generating a new centroid using the data points allocated to the cluster during the current iteration, and all data points are re-assigned to the closest centroid.
4. The process repeats until the data point allocation has little to no change, defining the final centroids.

Figure 3 illustrates the k-means algorithm in action. The graphic shows the algorithm during two iterations. After each iteration, the closest cluster centroid absorbs the remaining data points.

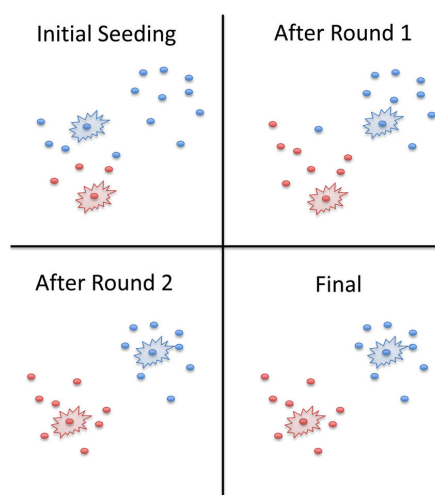


Figure 3: K-means algorithm. The training examples are the dots, and the centroids are the dots with lines around them (Page et al., 2014).

Although k-means is a commonly used algorithm, it does not always converge to an optimal global solution. Instead, it frequently ends up with an optimal local solution, significantly depending on the position of the first centroids. Additionally, k-means

does not perform well with data outliers, significantly distorting the centroid values. Therefore, it is critical to execute k-means with multiple distinct random initializations to reduce the likelihood of discovering suboptimal clusters. A comparison helps to see whether they significantly affect the resultant clusters. Despite its detractors, k-means is still commonly employed because it is time-efficient at a low computational cost.

#### 2.4.1.1 Quality assessment K-means

The k-means method takes as an input the number of clusters  $k$ . The correct number of clusters is significant because a large  $k$  might cause the clustering method to become too sensitive to differences in the training set (variance), impeding the clustering algorithm from generalizing new data points. On the contrary, using a disproportionately small number of clusters may cause the clustering algorithm to overlook important characteristics and produce an extensive segmentation making it useless. The widely used elbow approach is a heuristic determining the sweet spot when the quality does not noticeably increase with a rising number of clusters anymore. Hence, this approach shows the explained variation with the number of clusters in the data set. Therefore, the 'elbow' of the curve determines the number of clusters (Han et al., 2011), which can be seen in Figure 4.

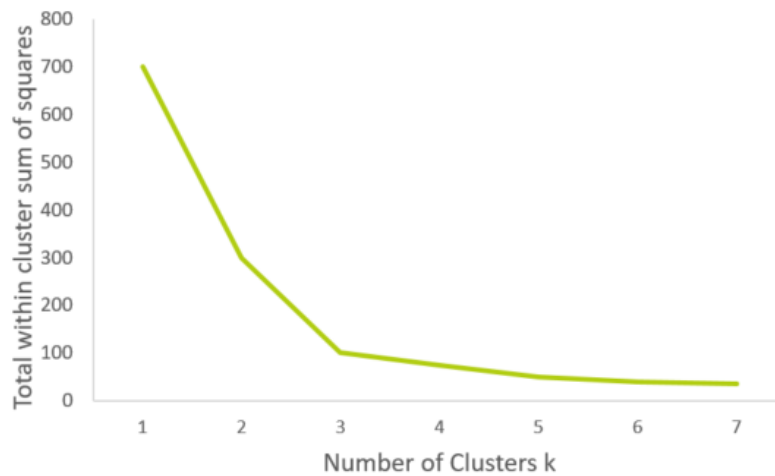


Figure 4: The elbow plot, where the best amount of clusters seems to be three based on the drop in change of explained variation (total within-cluster sum of squared)

#### 2.4.2 Hierarchical Clustering

Although partitioning algorithms satisfy the fundamental clustering criterion of grouping a collection of data points into several exclusive groups, in some situations, it might be of advantage to divide data into groups of hierarchical relationships.

The hierarchical clustering approach organizes groupings of data points into a tree-like structure of clusters, which is how it gets its name. The clusters can be formed on different hierarchical levels, such that on lower levels, clusters are subsets of other clusters on higher levels. In particular, this clustering approach effectively determines the degree of disparity between individual groups of data and which groups share similarities. Based on how the decomposition is generated, the hierarchical clustering approach may be characterized as either agglomerative or divisive, depending on how the hierarchical decomposition is formed. The agglomerative approach begins with each data point producing its own cluster, which is then combined into a larger cluster, also called bottom-up. Once the clusters have been determined to be comparable, they are repeatedly merged until the topmost level of the hierarchy is achieved or a termination condition is activated. The divisive technique begins with a single cluster that comprises all the items and then divides the cluster into smaller groups from the top down. It is then broken into smaller clusters at each iteration until each data point becomes its own cluster (Han et al., 2011). The agglomerative and divisive hierarchical clustering approaches are shown in Figure 5.

Further, indifferent to the hierarchical clustering approach, distance measurements are used to identify the similarity of items in a data set to select data points for the final clusters. Hierarchical clustering can employ multiple, widespread distance measures to construct the proximity matrix. When it comes to distance measurements, the Euclidean Distance is the most common; nevertheless, the Mahalanobis Distance is especially beneficial for non-spherical-shaped distributions, while the Minkowski Distance is applied in a normed vector space acting as the generalization of the Euclidean and Manhattan distance (Han et al., 2011).

Moreover, in both the agglomerative and the divisive methods, a measure has to determine how the distances between clusters are calculated to determine which objects should be grouped. There is not one optimum linking approach; it rather should be selected based on the data type, the desired output, and the observed performance utilizing a variety of various measures (Han et al., 2011). The following are four of the most commonly utilized linking methods:

$$\text{Minimum distance : } dist_{min}(C_i, C_j) = \min_{p \in C_i, p' \in C_j} |p - p'| \quad (4)$$

$$\text{Maximum distance : } dist_{max}(C_i, C_j) = \max_{p \in C_i, p' \in C_j} |p - p'| \quad (5)$$

$$\text{Mean distance : } dist_{mean}(C_i, C_j) = |m_i - m_j| \quad (6)$$

$$\text{Average distance : } dist_{avg}(C_i, C_j) = \frac{1}{n_i n_j} \sum_{p \in C_i, p' \in C_j} |p - p'| \quad (7)$$

Where  $|p - p'|$  is the distance between two objects or points,  $p$  and  $p'$ ;  $m_i$  is the mean for cluster,  $C_i$ , and  $n_i$  is the number of objects in  $C_i$ .

In addition to the previously indicated distance measures, special steps may be taken to aid in the pruning of the tree, such as Ward's approach. It is calculated

similarly to the group average measure, comparing the similarity between two clusters. However, if there is a lot of noise in between clusters, it takes the sum of squares of the distances  $p$  and  $p'$ , which is more stable (Han et al., 2011).

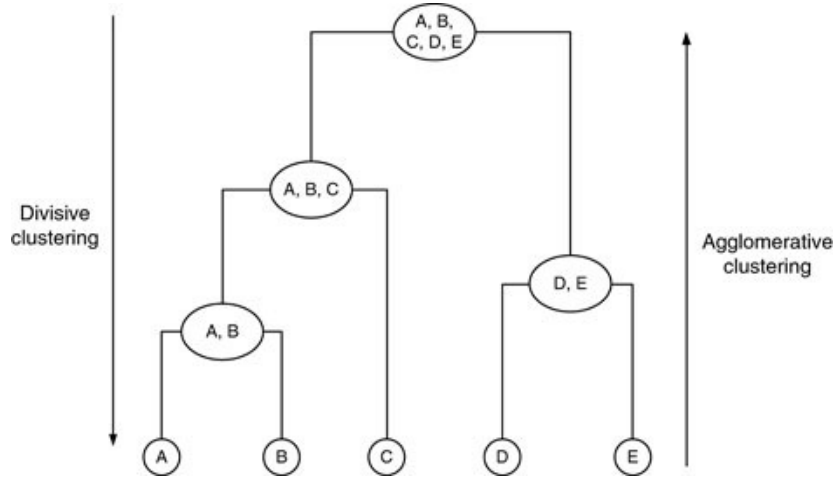


Figure 5: Agglomerative and divisive hierarchical clustering on data objects (Halkidi, 2009)

The use of hierarchical clustering does not necessitate a decision on the number of clusters produced in advance. As a result, hierarchical clustering is an ideal exploratory data analysis starting point for exploring the variations across clusters, particularly when using k-means, which are susceptible to outliers. Hierarchical clustering, on the other hand, has significant limits, most notably in terms of its time-consuming nature (Han et al., 2011).

### 2.4.3 Density-Based Clustering

The majority of partitioning strategies cluster objects according to their distance from one another. However, these approaches are limited to finding clusters with spherical forms and have trouble detecting clusters with irregular shapes. Therefore, other clustering algorithms based on the concept of density have been developed. Their basic premise is that they will continue to develop a given cluster while the density (number of objects or data points) in the "neighborhood" reaches a predefined threshold. For instance, the neighborhood of each data point inside a particular cluster must include a certain minimum number of points. This technique may be used to eliminate noise and outliers and locate clusters of any shape. Density-based approaches can partition a collection of items into numerous distinct clusters or a hierarchical structure of clusters. Typically, density-based approaches focus almost entirely on exclusive clusters and ignore fuzzy ones. Additionally, density-based approaches can be expanded to clustering in subspaces (Han et al., 2011).

**DBscan:** One of the most common density-based clustering algorithms is called Density-Based Spatial Clustering of Applications with Noise (DBscan), identifying core objects, which are objects or data points with dense neighborhoods. Further, the DBscan takes the core objects and links them with their neighborhoods to create dense sectors as clusters. To obtain the clusters, the neighborhood needs to be quantified. Therefore the user needs to determine the parameter  $\epsilon > 0$  that defines the object's neighborhood's radius. Hence, the fixed  $\epsilon$ -neighborhood of an object  $O$  is defined as the area within the radius  $\epsilon$ . Consequently, the density of a neighborhood is measured by the number of objects within the area of the  $\epsilon$ -neighborhood. Moreover, a second parameter *MinPts* determines the density threshold, and thus a core object, which is the pillar of a dense region, is defined as an object that consists of a minimum of *MinPts* points. Further, the DBscan algorithm employs density-connectedness to link core objects and their neighbors in a dense area (Han et al., 2011).

However, the DBscan also has its drawbacks as the user has to determine two input parameters, namely the maximum radius of a neighborhood and the minimum number of points required in the neighborhood of a core object. Generally, these parameters have to be set empirically and are very difficult to define beforehand, especially with high-dimensional, real-world data sets (Han et al., 2011).

**HDBscan:** The solution was introduced by Campello et al. (2013) with the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBscan), converting the DBscan into a hierarchical clustering algorithm and adding a strategy to derive a flat clustering based on cluster stability. Further, HDBscan uses the DBscan's setup to get different  $\epsilon$  values and thus only needs a minimum cluster size as an input parameter. The excess of the mass clustering method returns the most stable clusters over  $\epsilon$ . The HDBscan can detect clusters with varied densities without first selecting an acceptable distance threshold, which saves time. However, the application of an epsilon threshold may be beneficial in some situations, for example, if multiple clusters are spread across the space. The HDBscan steps can be broken down into the following steps:

1. In the beginning, the space has to be altered according to the density and sparsity of the data.
2. Next, the minimum spanning tree of the weighted graph has to be created.
3. The connected data points must be formed into a cluster hierarchy.
4. This cluster hierarchy has to be condensed based on the smallest cluster size
5. The condensed tree then provides the stable clusters

Discovering clusters means finding dense areas of data points. Since a wrongly positioned noisy data point might act as a bridge between dense regions, a more robust

approach is first to decrease the noise, defined as lower density areas. The goal is to make the clustering core more resistant to noise. In reality, this means separating noisy areas from dense areas (Campello et al., 2013; Leland McInnes, 2016). Therefore, the simplest density estimate is to calculate the distances to the  $k$ -th nearest neighbor, for which a density-based hierarchy has to be formulated:

1. The Core Distance is the distance of an object to its  $k$ -th nearest neighbor denoted as  $core_k(x)$
2. The mutual reachability distance functions as a distance metric between two objects to divaricate low density (or high core distance) objects, which is defined as:

$$k_{mreach}(x, y) = \max(k_{core}(x), k_{core}(y), d(x, y)) \quad (8)$$

where  $d(x, y)$  is the initial metric distance.

These metric dense areas remain with the same distance while sparser areas move at least their core distance apart, decreasing the noise level. So now, the mutual reachability metric can find dense areas.

In the next step, the data is handled as a weighted graph, with the data points as vertices and an edge connecting any two points with weight equal to their mutual reachability. In the next step, the Prim algorithm builds the minimum spanning tree one edge at a time, always adding the lowest weight edge that connects the existing tree to a new, not yet existing vertex. It is important to note that the minimum spanning tree for mutual reachability distance is different from the pure distance in the graph (Campello et al., 2013; Leland McInnes, 2016).

Next, the minimal spanning tree is transformed into a hierarchy of linked objects, which can be visualized in a dendrogram. Moreover, pruning this hierarchy at different places is where the HDBscan becomes different from the DBscan. Afterward, large and complex cluster hierarchies are compressed into smaller trees with more data associated with each node. Therefore the minimum cluster size is a parameter for HDBscan defined by the user. This minimum cluster size is then used to traverse down the hierarchy, checking if each split's newly generated cluster has fewer objects than the minimum cluster size. If there are fewer data points than the minimal cluster size, the bigger cluster keeps the parent's cluster identity. A real cluster split occurs when two clusters are separated, and each cluster has at least the minimum cluster size. In the end, a much smaller tree with fewer nodes exists, containing information about the cluster size diminishing at that node (Campello et al., 2013; Leland McInnes, 2016).

In order to extract the correct clusters, a new measure has to evaluate cluster persistence:  $\lambda = \frac{1}{distance}$ . This enables the definition of  $\lambda_{birth}$  and  $\lambda_{death}$ , which are the respective values for when a cluster becomes its own cluster or splits into two

smaller clusters. Further, the value  $\lambda_p$  is a value between  $\lambda_{birth}$  and  $\lambda_{death}$  and is defined as a data point that leaves the cluster due to a cluster split or the lifetime of the cluster. With this definition, stability can now be defined:

$$\sum_{p \in cluster} (\lambda_p - \lambda_{birth}) \quad (9)$$

For the correct cluster selection, all clusters are being considered at first. Then, if the stability sum of the child clusters is greater than the cluster stability, it is updated to the child cluster's stability sum. However, if the cluster stability is greater than the children's stability sum, it is acknowledged as a cluster, and its children are ignored. This process is repeated until the root node is reached and the marked clusters are returned (Campello et al., 2013; Leland McInnes, 2016). Figure 6 shows the output of the returned clusters.

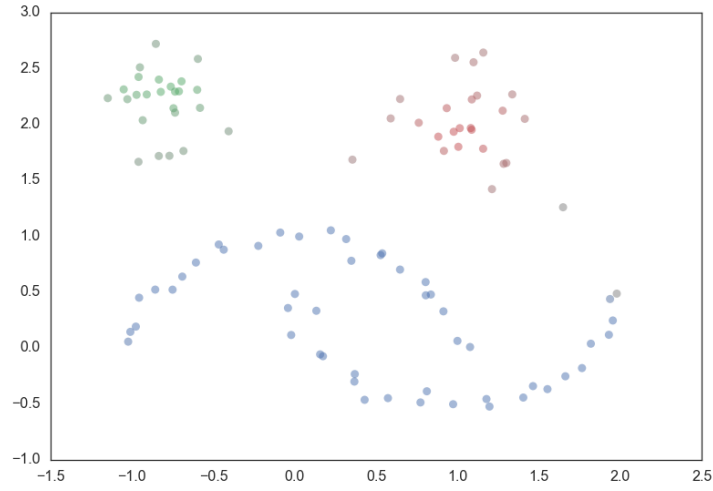


Figure 6: HDBSCAN output the stronger the membership, the more saturated is the color. (Leland McInnes, 2016)

## PROJECT

### 3.1 Business Understanding

Since Cerascreen has not been using all customer-related data, losing valuable knowledge. The non-usage of data can be considered lost knowledge, and Cerascreen currently cannot recognize all the customers' problems, ideas, and wishes expressed online. Therefore it cannot respond to them and improve itself. Of course, customer problems are still handled by the customer support department; however, there is no capacity for analyzing all reviews online next to all the customer complaints by email and other channels.

As mentioned in the introduction, Cerascreen aims at being a more data-driven company, primarily because it already collects a lot of data from customers that could lead to a lot of new insights. This can lead to better decision-making and thus potentially increase its revenue by increasing its sales or decreasing its costs. Therefore, in a new initiative, Cerascreen wants to look into its many data sources to monetize the potential knowledge that can be retrieved from the data to better understand its customers and offer more excellent products.

Therefore, focusing on customer data, it is crucial that Cerascreen first understands what customers generally say about its products and services across its many different online review platforms. Hence, it serves as an overview of the initial situation to decide on further steps to fix found problems or look deeper into identifying them with a more focused analysis approach. Since there has been no prior customer review analysis, there are no complex requirements, rather than extracting as much knowledge as possible from current data to build a baseline for future projects. Therefore, the scope lies in identifying current data characteristics and topics customers talk about the most, for which textual data sets are required from different sources. Additionally,

the project has the constraint of running a maximum of 2 months. Also, as new data is analyzed, what insights can be drawn from the whole analysis is unknown.

Further, the scope serves as the translation from the business objective into a text mining problem—moreover, the project needed to deliver a presentable output used by different departments. Based on the output, the departments can act upon or request additional future analysis based on newly raised questions.

## 3.2 Data Understanding

At the time of the project, neither has the data been extracted from the different websites required to conduct a text analysis nor has it been stored in any database. Therefore, this section will first discuss how the data has been retrieved from different websites to form a data set for the project. The second part then discusses the data set construction to be used in the analysis.

### 3.2.1 Data Retrieval

Although Cerascreen currently has access to its customer reviews, they are available in different sources, different data formats, and languages. Therefore, before the data can be processed and analyzed, it must be retrieved and stored first. As the data is only available through the various customer review platforms, time constraints demanded a concentration on a selection of websites. The selection process had two parameters: the number of reviews on each platform and the required time for their extraction. The extraction effort had multiple factors, such as if an application programming interface (API) was available to query the data from the website or if a web scrapper had to be built. Connecting to an API can be a quick process; however, every website has its individual API that is either public or has to be purchased. For this project, purchasing an expensive API for any website was ruled out. Therefore, Trustpilot has been tried to be scrapped, being the third-largest review source. Nevertheless, building web scrapers takes a lot of time, and after some attempts, this idea has been discarded for the sake of the project's timeline.

Table 1: Numbers of reviews per source and its connection

| Data Source            | Access                | Number of Reviews | % of total Reviews |
|------------------------|-----------------------|-------------------|--------------------|
| Trustpilot             | Web Scraping          | 247               | 1,85%              |
| Trusted Shop           | API                   | 5804              | 43,45%             |
| Google My Business     | API                   | 83                | 0,62%              |
| Amazon                 | Third Party API       | 2337              | 17,50%             |
| Shopify                | Manual E-Mail exports | 4760              | 35,64%             |
| Google Play Console    | API                   | 84                | 0,63%              |
| Apple AppStore Connect | API                   | 42                | 0,31%              |

According to Table 1, the essential sources based on review numbers are Trusted Shops and Shopify. Trusted Shops offers a public API connection to download and store the data. However, each country's online shop has its own ID and needs its request. Therefore, a Python script loops through a list of all online shops on Trusted Shops and stores them with the correct ID. The requested outcome was one file with all reviews from the last 365 days. Additionally, there was more historical data available on Trusted Shop's databases, and thus, an email contact had to be established to obtain the additional data from the time before mid-2020. This additional data had a different format than the API request necessitating a first data transformation to join both data files appropriately. In contrast, Shopify does not store the customer reviews on its database but employs a third party that does not offer any API. Instead, a registered Shopify administration email has to request a download of all reviews, which are then sent to that email in CSV format.

The other sources with an API were neglected due to the small number of reviews in proportion to the time required to retrieve them. For example, although Amazon has many reviews on its platform about Cerascreen's products, a third-party API had to be purchased to access these reviews. Alternatively, a web scrapping program would have to be written. Hence, it was decided to leave the remaining sources for follow-up projects, keeping in mind that the two sources hold 79% of the reviews.

After deciding on the sources, each country shop had to be selected to extract the reviews. As Cerascreen's biggest markets are in Europe, analyzing all reviews might prove difficult due to the different languages spoken in each country. Therefore, the German-speaking DACH region and the United Kingdom have been selected since they are the biggest markets with the most reviews. Additionally, it is faster to translate one language into the other than five or more. The other countries will follow as their market shares increase and the review numbers with them.

### 3.2.2 Data Sets

The two data sets from Trusted Shops and Shopify carried the same information with slight differences, such as different country shops and specific reviews about the company, products, or services. Therefore, it was still possible to merge both data sets keeping all the needed information for the intended analysis. However, only Shopify offers product-specific reviews, while Trusted Shops offers information on when and what has been responded to the customer comments. Since this information is not required to reach the analysis's goal and both sources have almost an equal amount of reviews, it had to be dropped. Nonetheless, this information will be beneficial in the future follow-up analysis.

For this project, the important information from both sources was the country in which the customer did the review, the used platform, text, rating, and date. This data was vital for the data exploration, gaining first insights about where, when, and

with what rating the customer did the reviews before diving into analyzing the text. Unfortunately, it was later found that the country origin of the review was not part of any of the two original datasets; this information had to be inserted while loading the datasets. Therefore, the Shopify country code has been extracted from the different files previously received by email. In addition, the Trusted Shops country code was extracted from the order reference while downloading the data through the API.

### 3.3 Data Preparation / Exploration

During the data understanding phase, the data was retrieved and merged into one data set while the most important features were identified. Due to the nature of the project, the exploratory phase can already deliver valuable insights about customer reviews. Moreover, the exploratory phase has two steps; the first step focuses on exploring all the features but the text, while the second step specifically analyses the text reviews. Hence, the main findings of both phases are presented in the following sections.

#### 3.3.1 Overall Data Set

This part of the analysis uses the complete data set without looking into the text reviews. After retrieving 10359 reviews since 2014 from the two sources, it is essential first to understand the distribution of the reviews per country and get a first impression about the ratings in each country.

Table 2: Number of reviews per country and source

| Country | Review Platform | Count | Rating | Mean     | Standard Deviation |
|---------|-----------------|-------|--------|----------|--------------------|
| AT      | Shopify         | 596   |        | 4.243289 | 1.214564           |
|         | Trusted Shops   | 3     |        | 5        | 0                  |
| CH      | Shopify         | 521   |        | 4.433781 | 0.988361           |
|         | Trusted Shops   | 337   |        | 4.037151 | 1.475047           |
| DE      | Shopify         | 277   |        | 4.480317 | 0.960130           |
|         | Trusted Shops   | 5007  |        | 4.619193 | 0.897944           |
| UK      | Shopify         | 366   |        | 4.377049 | 1.070101           |
|         | Trusted Shops   | 254   |        | 3.744213 | 1.654806           |

Given that Germany is Cerascreen's primary market, the over-proportionate reviews are not surprising. However, the quick overview in Table 2 already shows the country's preference of review platform and some deviation on the rating the further away the country is from Germany. However, in general, the rating is above four on a scale of one to five, showing that customers giving reviews seem satisfied by Cerascreen.

Looking further into the rating's distributions, shown in Figure 7, individuals only give reviews if they had a very good or an awful experience, which can be translated

into good and bad reviews. This behavior also explains the distribution's violin shape and why there are fewer ratings around the middle with three stars. The plots also show that there are more reviews below a five outside the German market, although the average rating is everywhere around four. Looking closer the good ratings overweight the bad ratings, which also count for the DACH area (see Figures 8 and 9). However, there are still some bad reviews that could indicate some bad experiences customers had. Further, Figure 10 clearly shows the more significant percentage of bad reviews in the United Kingdom. Thus, although the United Kingdom does have a lot of positive reviews, the number of bad reviews also indicates some problems within this market.

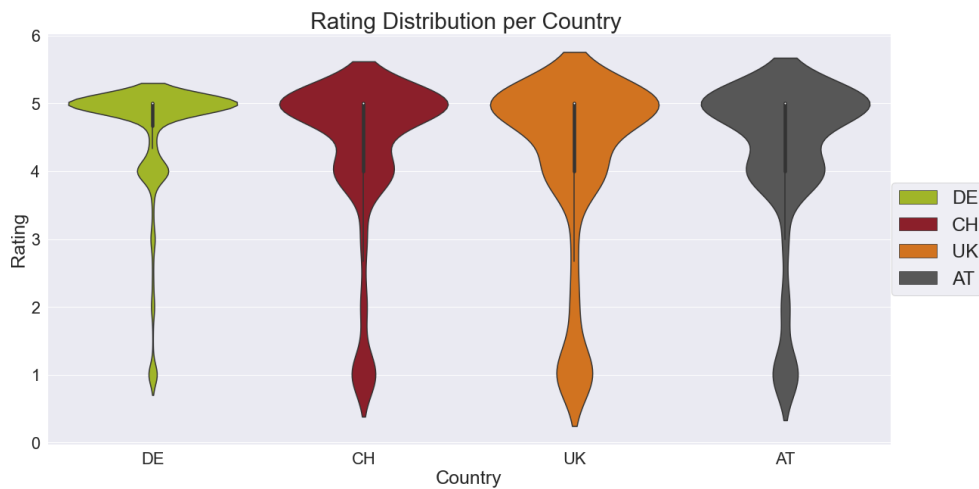


Figure 7: The rating distribution per country

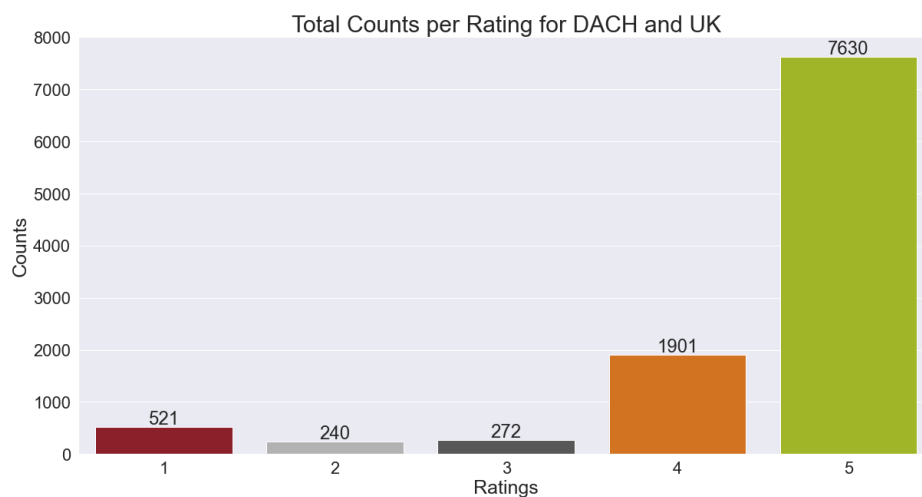


Figure 8: The total rating for the DACH region and the United Kingdom

Considering that Cerascreen was founded in 2012, as the customer base grew over the years, the rating per country per year has also started to vary much stronger. Since 2018, the new markets outside of Germany haven't been as satisfied (Figure 11). It

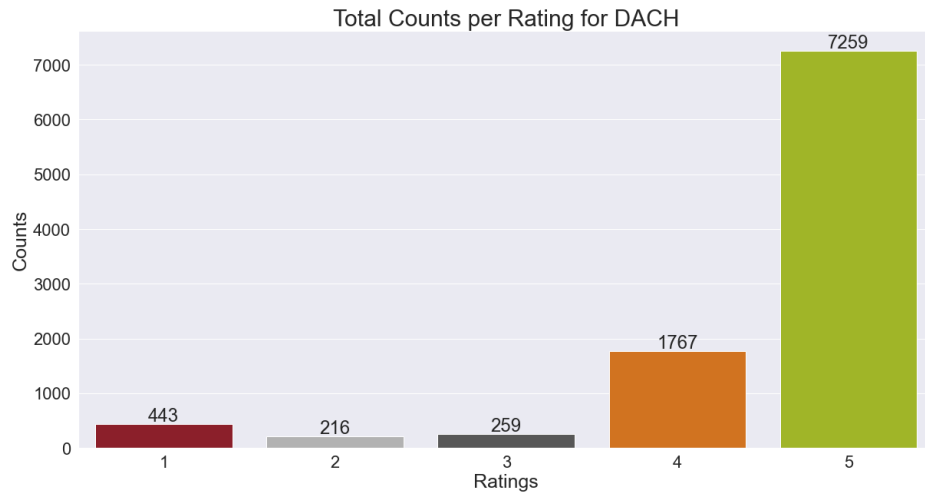


Figure 9: The total rating for the DACH region

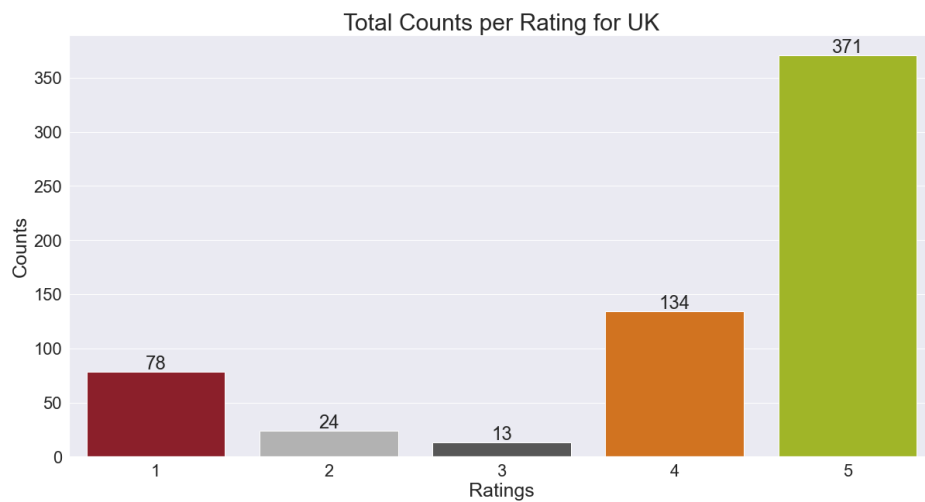


Figure 10: The total rating for the United Kingdom

becomes evident that Germany's ratings keep steady between four and five, while the United Kingdom and Austria have strong dips with bad reviews. Switzerland's reviews start very well but deteriorate with Austria towards the middle of 2021 before recovering to a rating level of 4 and 3, respectively. Especially during the Covid-19 pandemic in 2020 and 2021, customers seemed less satisfied.

### 3.3.2 Text Corpora Preparation

Before the program can translate the text, the review languages must be identified. The country code in the data set indicates the national languages spoken in these countries. However, to be certain, utilizing the language detection model helps to identify whether the review is in English or German. The classification and translation

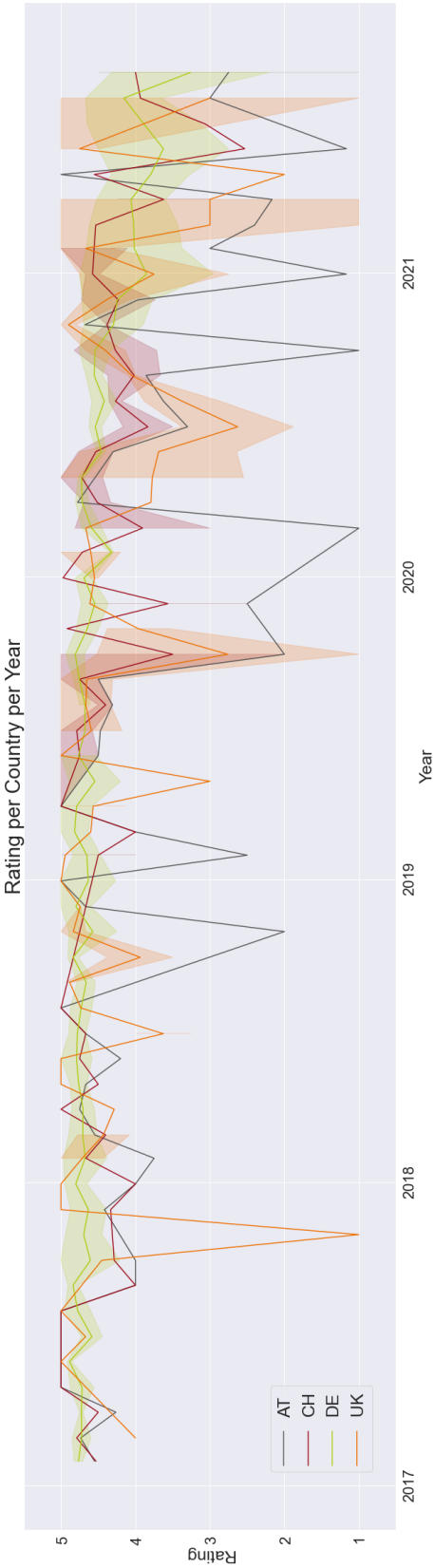


Figure 11: The average rating per country per year

from English into German are necessary because it will keep most of the reviews in the original text, potentially losing less information through translation. Nevertheless, German is the favored language since only about 6% of the reviews are from the United Kingdom, which can be assumed to be primarily English text. Additionally, there might be cases where a person in any country has written a review in another language; however, most reviews written in a country are assumed to be in its national language. There are many different libraries that detect languages. This report uses Langdetect since it was the fastest and easiest to implement, although it became apparent that multi-language detectors misclassified German or English, for example, as French.

Moreover, since the detection models have problems with a very short text (one to three words), more columns are not identified as the German language than in the data set. Therefore, this work finds the most frequent short texts and classifies them manually into German or English, to the best of the author's abilities. This way, after the model has classified the reviews' languages, they can be compared to the manual list, decreasing the classification error. However, there will always be an error margin of wrong language classification, and it is not the scope of this project to increase the language classification accuracy. It is merely a way to incorporate English reviews from the UK, another big market for Cerascreen, into the German text analysis. Hence, the texts reviews that do not match the short text lists or are classified as German or English can be seen as misclassified and, thus, deleted from the data set for simplicity reasons.

After isolating non-German and non-English reviews from the rest of the data set, there is a total of 1168 reviews that are supposedly written in another language. Comparing this subset to the previously established list of short German and English reviews, 921 rows could be found, matching the short review lists. This shows that 247 out of the 1168 reviews are not really another language, which are 2.4% of the whole data set. However, even looking at this 2.4%, most of them can still be classified as German, since they are a mix of another language with German or are spelled wrong. Therefore, this report will consider these 247 reviews as outliers and remove them from the data set. Henceforth, this report will assume that the reviews from the United Kingdom are in English and the reviews from the DACH region are in German.

Although classifying the languages, there might always be some reviews in different languages. So it has to be accounted for that there are a small number of different languages in the text corpora. Further, when translating the rest into German, there will always be a margin of error when translating from one language into another, and some information is lost.

### 3.3.3 Final Text Corpora

In this section, the text corpora is built consisting of the text reviews and their ratings. First, the texts are cleaned, and then all English texts are translated into German.

In general, the shelf translators, like Google translate, are the fastest and easiest to implement. They also provide the most accurate and fastest translations, especially when languages are similar since they have pre-trained models (Vanjani and Aiken, 2020). Hence, this report uses the Google translator to translate English into German.

During preprocessing, there are a few things that have to be managed. First, the sparsity of huge matrices, based on word frequencies, has to be minimized. Therefore, the goal in this section is to remove all noise from the text and normalize it. Noise removal handles redundant text components, like punctuation, tags, URLs, and stopwords. Normalization deals with stemming, which removes suffixes, and lemmatization, which works with the base of the word's root. The outcome is the final text corpora for the analysis and later modeling.

### **3.3.4 Customer Reviews**

This section will first look into the global text characteristics of the reviews before it analyses the content of the reviews' texts. In connection with previous insights, this will shed more light on what customers think about Cerascreen. Afterward, n-grams give a first glance into what has been said in the reviews before topic modeling finds the most important issues expressed by the customers.

#### **3.3.4.1 General Text Characteristics**

Looking purely at the text characteristics, it was first interesting to understand how many words the reviews have, because the longer the reviews the more insights can be drawn from them. Looking at Figure 12, most reviews only contain 1-5 words, which can be considered very short text, while long reviews are very scarce. To better understand, the Table 3 shows the exact distribution of the review lengths in the data set. Here, one can see clearly that 64% of the reviews are shorter than 20 words, the equivalent of a simple sentence or less. As there are many short reviews and only a few long reviews, it mirrors the characteristics of the ratings that have been discussed before. After merging both ratings and reviews in Figure 13 it becomes clear that the better a rating is, the shorter the reviews become. On the other hand, the worst ratings have the highest proportion in long reviews with more than 30 words, and reviews with ten to 20 words take a similar share across all ratings with 25% to 39%. Hence, customers that give a bad rating tend to use more words in a review.

#### **3.3.4.2 Text Analysis**

After the text characteristics had been identified, it was interesting to see what the reviews generally had to say about the company. The first impression of the reviews are that words like "easy, fast, good, delivery, result, and test" dominate in the texts. Further, the reviews were divided into good and bad reviews to get an idea about how

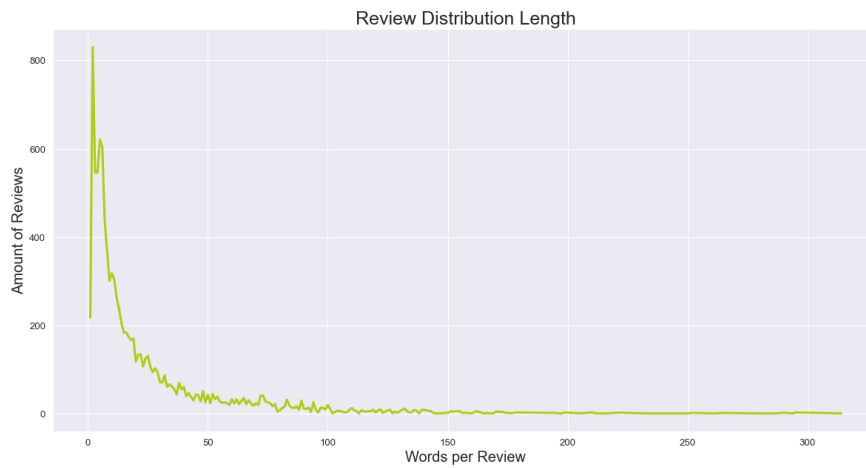


Figure 12: The distribution of the text length of all reviews

Table 3: Frequency of the review length

| Number of words | Amount of Reviews |
|-----------------|-------------------|
| 0 - 5           | 2139              |
| 5 - 10          | 2331              |
| 10 - 20         | 2206              |
| 20 - 30         | 1154              |
| 30 - 40         | 644               |
| 40 - 50         | 406               |
| > 50            | 1479              |

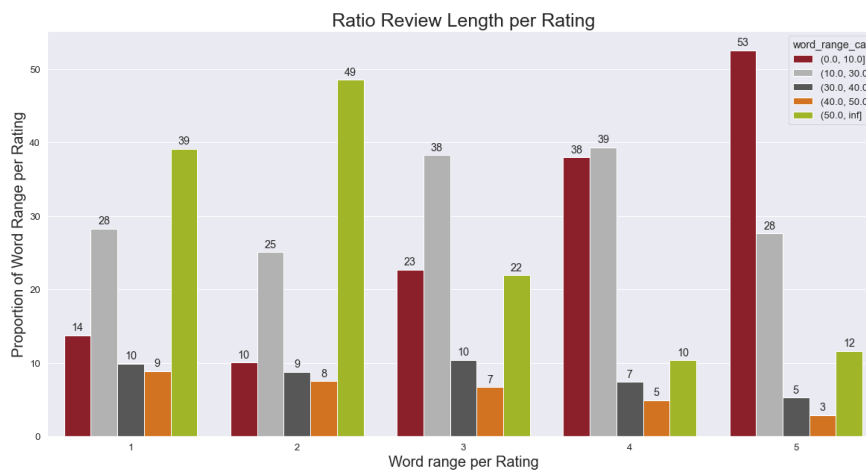
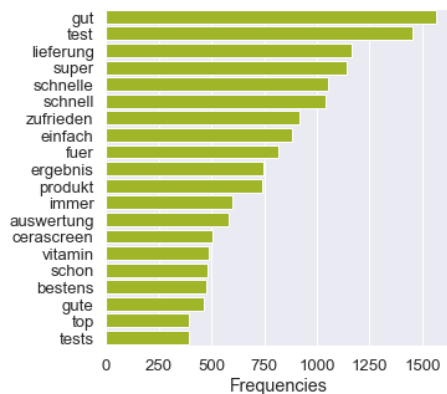


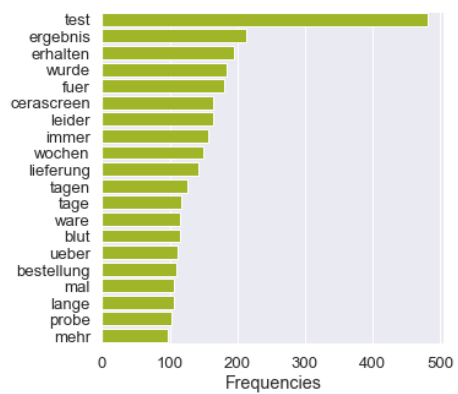
Figure 13: The distribution of the text length of all reviews per rating

this first impression might change. Therefore, it was decided that the good reviews were four and five stars on a rating scale of one to five stars while bad ratings were one to three. This report used a count vectorizer from sklearn to build a simple BOW model producing different sized n-grams to get some first insights into what customers mostly talked about in their reviews.

Looking at the n-grams in the Figures 14 and 15, satisfied customers are talking about fast delivery, easy-to-use tests, and quick results that are easy to understand. On the other hand, unsatisfied customers talk about long delivery times from days to weeks, faulty tests that must be exchanged, and longer waiting periods until the results arrive. They even mentioned a few times to rather go to the doctor.

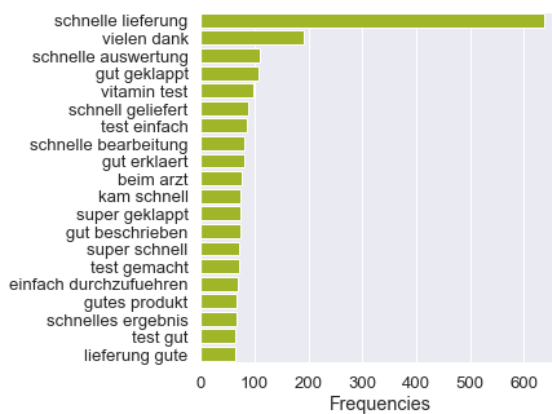


(a) Uni-gram for positive ratings

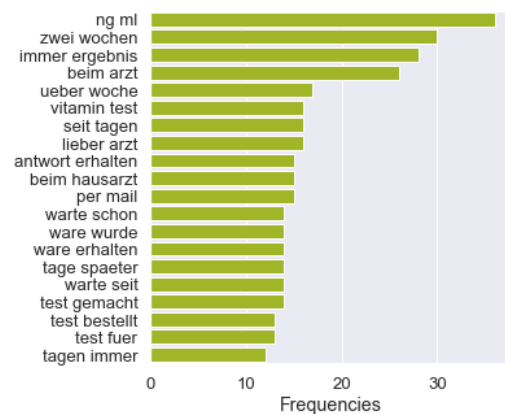


(b) Uni-gram for negative ratings

Figure 14: Uni-gram for positive and negative ratings



(a) Bi-gram for positive ratings



(b) Bi-gram for negative ratings

Figure 15: Bi-gram for positive and negative ratings

The problem with such a simple word count obtained from the CountVectorizer is that large counts of certain common words may dilute the impact of more context-specific words in the corpus. This is solved by the TF-IDF vectorizer, which penalizes words that appear several times across the document. TF-IDF are word frequency scores that highlight more important words to the context rather than those that frequently appear across documents. However, the words from the n-grams almost overlap with the words from the TF-IDF tables in Table 4. They just have the slight difference that the TF-IDF tables are more focused on the delivery time, which supports the previously identified topics .

Table 4: Most important words from all ratings

(a) Words of positive ratings

| Words     | TF-IDF |
|-----------|--------|
| test      | 0.2568 |
| gut       | 0.2559 |
| lieferung | 0.2052 |
| super     | 0.2049 |
| schnelle  | 0.1971 |
| schnell   | 0.1942 |
| zufrieden | 0.1745 |
| einfach   | 0.1716 |
| fuer      | 0.1647 |
| ergebnis  | 0.1527 |

(b) Words of negative ratings

| Words      | TF-IDF |
|------------|--------|
| test       | 0.3138 |
| ergebnis   | 0.1832 |
| erhalten   | 0.1658 |
| wurde      | 0.1644 |
| Cerascreen | 0.1610 |
| fuer       | 0.1583 |
| leider     | 0.1454 |
| immer      | 0.1386 |
| wochen     | 0.1365 |
| lieferung  | 0.1304 |

## 3.4 Topic Modeling

After the data exploration phase already showed some insides about the customer reviews, this section discusses the approaches employed to achieve the project's goal. Additionally, the results are evaluated, and all topic modeling was done on the data set that was developed as a result of the preceding section's data exploration.

### 3.4.1 General Information

It is worth noting that several modeling techniques were employed in this project. We tried to find essential topics customers talk about in their reviews using these modeling techniques. Therefore, unsupervised techniques were compared to build the most applicable model that captures the most diverse topics. Further, since the data was already cleaned during the data understanding, this section only focuses on the model development using different algorithms.

For each model the text is always reduced in dimensionality, sentence encoded and then clustered. Hence, the clustering performance can be affected by dimensionality reduction, sentence encoding, and their parameters. For simplicity reasons, the most widely used sentence transformer for the German language is applied. Afterward, the silhouette score compares the clustering methods and encoding techniques. In addition, the models are compared on their baseline without any parameter adjustment, which would exceed the project's time frame. Thus, the most promising models were chosen to be further developed. However, since the clustering methods deal with text, what is logically correct for the algorithms might not be understandable for a human. Thus, the cluster of topics has to be compared based on common sense. Therefore, a good cluster is defined by finding a diverse set of topics through the noise of commonly used phrases in online reviews. These phrases have been kept in the data set since they represent the view about the company.

### 3.4.2 Comparison of Topic Models

This section is about finding the best-fit clustering algorithm for the topic model. Please refer to the appendix A for a more detailed comparison of all clustering algorithms. Further, this section provides the original topic clusters in German. For an English translation of the topic clusters, please refer to the appendix C.

Before any models could be compared, it was necessary to choose the right amount of clusters. Using the elbow method, Figure 16 indicated the optimal number of clusters is ten since the cluster sum of squared seemed to vary less from that amount of clusters onward. Since ten was the best option for K-Means, the comparison of the models has been conducted with ten cluster topics per model. In the next step, the silhouette scores were compared to find the best reduction technique for each clustering model to produce the most meaningful topics.

The Figures in the appendix A show that UMAP’s reduction enables smaller clusters to be identified throughout the different models. However, although the PCA and the t-SNE find a good range of clusters, they all seem very close to each other (Figures 18-20 in appendix A), allowing for the possibility of overlapping clusters. It is supported by the silhouette score comparison from all models in Table 5. The numbers indicate that UMAP has the smallest intracluster average distance and the highest inter average cluster distance throughout all models. Therefore, UMAP can be established as the best dimensionality reduction technique for these models, and thus this project will also only look at the topics content produced in combination with UMAP.

All in all, compared to the K-Means and Agglomerative Clustering models, HDBscan produces more insights about what customers write in their reviews. However, although the HDBscan is missing the negative annotations, the topic variety and their meaningfulness are much more substantial. Therefore, the HDBscan is used together with the BERTopic modeling approach in the next step.

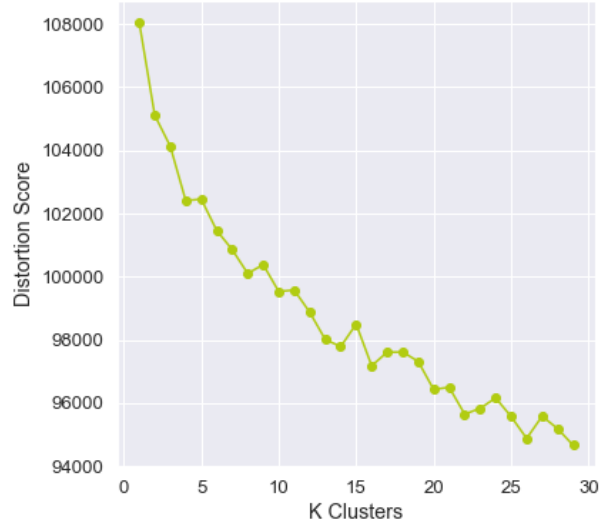


Figure 16: The elbow plot for all reviews where the distortion score is the within-cluster sum of squared

Table 5: The cluster models’ silhouette score

| Cluster Model              | Silhouette Score |
|----------------------------|------------------|
| Bert with PCA and K-means  | 0.40011638       |
| Bert with Tsne and K-means | 0.36476564       |
| Bert with Umap and K-means | 0.4306468        |
| Bert with PCA and HDB      | 0.22121191       |
| Bert with tsne and HDB     | 0.25838166       |
| Bert with Umap and HDB     | 0.27426562       |
| Bert with PCA and AggCls   | 0.36313587       |
| Bert with Tsne and AggCls  | 0.31980598       |
| Bert with Umap and AggCls  | 0.3850319        |

### 3.4.3 BERTopic

#### 3.4.3.1 The algorithm

BERTopic is a topic modeling approach that takes advantage of BERT embeddings and c-TF-IDF to produce dense clusters that allow for easily understandable topics while keeping the keywords in the topic description. Further, it has a similar structure to the set up in the section above. First, it embeds the documents with a preferred embedding technique. In this case, it uses sentence embeddings. Next, UMAP reduces the document's dimension before the HDBscan organizes the documents into semantically similar clusters. The step different from the previous setup is the topic representation creation from the clusters by the c-TF-IDF and maximizing their relevance. The c-TF-IDF is a modification of the TF-IDF, which treats all documents in the cluster as one single document to which it then applies the TF-IDF. Hence, instead of calculating the word's importance in each document of a cluster, TF-IDF now calculates the importance of words across all documents within the clusters. Hence, the more important a word is within a cluster, the more it represents the topic of the cluster. Additionally, it also can further reduce the topic size if it is deemed necessary. As a Python package, it also offers faster and easier navigation through the setup of the model and the retrieval of topics.

Finding an ideal minimum cluster size can only happen through trial and error. Therefore, a total of 39 topics have been created with the following parameters adjusted:

1. The minimum cluster size ranged from 10 to 100.
2. The Manhattan and Euclidean distances were chosen as metrics.
3. The cluster selection methods were leaf or excess of mass.

Afterward, the models were compared against their silhouette score, shown in Table 6, which did not provide a lot of new information since finding the meaningful topics is a hands-on approach as was established with the previous models. Hence, the models were analyzed by manual comparison.

#### 3.4.3.2 The parameters

The parameters metrics and cluster selection method did not affect the general topics because several topics could be found across all clusters, and no extra special topics stood out. Thus, diving into the minimum cluster size, it is important to mention that the higher the number, the lower the number of topic clusters as topics not reaching the minimum number are ignored or merged with another cluster. Further, since most reviews are positive and short, mentioning similar topics a larger minimum cluster size means favoring these topics, while reviews that actually have something to say are

Table 6: The topic models' silhouette score and amount of topic clusters

|    | Model                    | Silhouette Score | Amount of Topic Clusters |
|----|--------------------------|------------------|--------------------------|
| 0  | model_10_manhattan_leaf  | 0.249220         | 359                      |
| 1  | model_10_manhattan_eom   | 0.156600         | 322                      |
| 2  | model_10_euclidean_leaf  | 0.253965         | 362                      |
| 3  | model_10_euclidean_eom   | 0.149808         | 322                      |
| 4  | model_20_manhattan_leaf  | 0.328604         | 180                      |
| 5  | model_20_manhattan_eom   | 0.114349         | 141                      |
| 6  | model_20_euclidean_leaf  | 0.301606         | 180                      |
| 7  | model_20_euclidean_eom   | 0.140242         | 139                      |
| 8  | model_30_manhattan_leaf  | 0.337256         | 119                      |
| 9  | model_30_manhattan_eom   | 0.169409         | 96                       |
| 10 | model_30_euclidean_leaf  | 0.328315         | 119                      |
| 11 | model_30_euclidean_eom   | 0.213829         | 100                      |
| 12 | model_40_manhattan_leaf  | 0.271228         | 89                       |
| 13 | model_40_manhattan_eom   | 0.177054         | 78                       |
| 14 | model_40_euclidean_leaf  | 0.318785         | 90                       |
| 15 | model_40_euclidean_eom   | 0.240437         | 80                       |
| 16 | model_50_manhattan_leaf  | 0.349578         | 73                       |
| 17 | model_50_manhattan_eom   | 0.266453         | 66                       |
| 18 | model_50_euclidean_leaf  | 0.333635         | 68                       |
| 19 | model_50_euclidean_eom   | 0.252544         | 59                       |
| 20 | model_60_manhattan_leaf  | 0.246895         | 57                       |
| 21 | model_60_manhattan_eom   | 0.219200         | 53                       |
| 22 | model_60_euclidean_leaf  | 0.287757         | 57                       |
| 23 | model_60_euclidean_eom   | 0.241368         | 52                       |
| 24 | model_70_manhattan_leaf  | 0.274169         | 49                       |
| 25 | model_70_manhattan_eom   | 0.255960         | 46                       |
| 26 | model_70_euclidean_leaf  | 0.332844         | 51                       |
| 27 | model_70_euclidean_eom   | 0.290193         | 46                       |
| 28 | model_80_manhattan_leaf  | 0.272431         | 44                       |
| 29 | model_80_manhattan_eom   | 0.263582         | 42                       |
| 30 | model_80_euclidean_leaf  | 0.256584         | 43                       |
| 31 | model_80_euclidean_eom   | 0.255908         | 42                       |
| 32 | model_90_manhattan_leaf  | 0.313070         | 42                       |
| 33 | model_90_manhattan_eom   | 0.309555         | 40                       |
| 34 | model_90_euclidean_leaf  | 0.311060         | 42                       |
| 35 | model_90_euclidean_eom   | 0.307003         | 41                       |
| 36 | model_100_manhattan_leaf | 0.361667         | 40                       |
| 37 | model_100_manhattan_eom  | 0.355831         | 38                       |
| 38 | model_100_euclidean_leaf | 0.367273         | 40                       |
| 39 | model_100_euclidean_eom  | 0.362777         | 39                       |

neglected. Although there might be few reviews for a topic does not necessarily mean it is less important or interesting to Cerascreen. Hence, it is essential to find the right amount of topic clusters.

Starting with a minimum cluster size of ten resulted in 359 repetitive topic clusters, which overlapped a lot, meaning they had the same word content. On the other hand, a minimum cluster size of 100 only produced 38 topic clusters. Compared to the 359 clusters before, 38 clusters were very generic, and a lot of information was lost. Hence, it was settled for the model\_40\_manhattan\_eom, which found 78 topic clusters with 14 words each. However, reducing the clusters further was not successful, because of a significant information loss, especially product-specific information, so the original 78 clusters were kept for analysis.

### 3.4.3.3 The analysis's

The analysis focused on the most important clusters that have been picked and displayed in this section in Table 7, the whole table can be seen in the appendix B. A translation of the whole table can be found in the appendix C. The Figure 17 gives

a general overview of the clusters' behavior. One can see right away that many clusters of topics are overlapping, which can be seen again once the analysis dives into the separate topics. From a total of 78 clusters, three topics were meaningless and were ignored, while topics about customers being satisfied had the most occurrences with 27 topics. They are represented by topic ten, and all were generally happy with the product and service but expressed it in different words. Mainly some topics also mentioned nuances of fast deliveries, good results, easy to use products, willingness to purchase again, and even once that the products are a bit too expensive. The second most common topic, with 12 clusters, revolved around fast delivery and fast results. These clusters are represented by topic 7. Further, six clusters, represented by topic 28, mention the products' high quality, Cerascreen being a trustworthy company, and customers would recommend it to others, mentioned in topic nine. Another five clusters, represented by topic 13, are topics about late delivery, varying from waiting days to weeks and comprises the product order as well as receiving the results.

The other 21 clusters offered unique insights into the customer reviews. For example, topic five is about the good taste of the products, and topic 39 details the great taste of the protein powder and bars. Further, in topic six, mentions that the pills for the hair are easy to swallow, while the Omega pills in topic 51 taste like fish. Other powder products in topic 15 help people with their sleep. topic 59 mentions magnesium being perfect for their household, and topic 53 that the curcuma pills are not as good as the drops because these are easier to dispense.

There are also general topics, like 31, where customers just said they did a test at Cerascreen. Additionally, general topics combined also add value, like 52, where the products seem to help the people. This also accounts for 64 and 78, where people feel an immediate effect of the supplements and already feel better after days and months. The Cerascreen concept also seems to be praised by athletes, in topic 73, for better coaching using the supplements after sports. Vegetarians and vegans in topic 65, like the products, because they are especially easy to dispense.

Moreover, customers also like the fast delivery and the customer service in topic eight, while in topic 66, customers praised the excellent instructions and easy-to-use tests. Finally, on topic 77, customers even commend the extensive results report with its good explanations.

Nevertheless, there are also topics, like in 25, where customers complain about the high price of the products for the quality they offer. This is also underlined by topic 69; however, customers say it is a trade-off. The products are more expensive than going to the doctor, but they are also more convenient as they are shipped to the customer. Thus, the topic says the customer pays a premium for convenience, and if a test is purchased in a bundle, it gets cheaper. This convenience is also represented in topic 76, where customers purchased corona tests to take at their home.

All in all, BERTopic has found the most unique topic clusters from all applied models. The detailed topic analysis also shows the alignment with Figure 17, since many topic clusters can be grouped into categories, such as fast and slow delivery or product specifics. Moreover, although bad reviews seem limited, BERTopic found detailed problem topics about the delivery and the high price, just like during the data shown in the exploration phase. Hence, it can be assumed that these are the most significant two sections in the complaints, notably because many similar clusters have been found, which is also the main difference to the other techniques used. Furthermore, many supplements, test, and service-specific topic categories could be found upon which actions can be taken in different forms if needed.



Figure 17: BERTopic clusters

Table 7: BERTopic most meaningful clusters

| Topic_5                           | Topic_6             | Topic_7              | Topic_8                         | Topic_9                | Topic_10             |
|-----------------------------------|---------------------|----------------------|---------------------------------|------------------------|----------------------|
| 0 geschmack                       | kapseln             | schnelle             | service                         | buch                   | wieder               |
| 1 schmeckt                        | die kapseln         | lieferung            | kundenservice                   | das buch               | gerne                |
| 2 lecker                          | kapseln sind        | des                  | guter                           | empfehlen              | wieder alles         |
| 3 der geschmack                   | die kapseln sind    | schnelle lieferung   | guter service                   | kann                   | gerne wieder         |
| 4 gut                             | sind                | tests                | toller                          | mir                    | immer wieder         |
| 5 schmecken                       | schlucken           | auswertung           | lieferung                       | empfehlenswert         | immer                |
| 6 geschmack ist                   | zu schlucken        | des tests            | schnelle lieferung              | weiter                 | wieder gerne         |
| 7 lecker und                      | die                 | test                 | schnelle                        | es                     | immer wieder gerne   |
| 8 der geschmack ist               | haare               | lieferung und        | schneller                       | kann ich               | gerne wieder alles   |
| 9 ist                             | meine               | handhabung           | toller service                  | habe                   | alles                |
| 10 snack                          | meine haare         | schnell              | sehr guter                      | mich                   | gern                 |
| 11 fuer                           | nehme               | bearbeitung          | schneller service               | nehme                  | alles bestens gerne  |
| 12 sehr lecker                    | leicht zu schlucken | schnelle bearbeitung | super service                   | empfehlung             | bestens gerne wieder |
| 13 super                          | leicht              | einfach              | kundenservice sehr              | weiterempfehlen        | bestens gerne        |
| 14 der                            | gute                | einfache             | guter kundenservice             | buch ist               | wieder immer         |
| Topic_13                          | Topic_15            | Topic_19             | Topic_25                        | Topic_28               | Topic_31             |
| 0 lange                           | kissen              | lebensmittel         | preis                           | qualitaet              | Cerascreen           |
| 1 gedauert                        | pulver              | habe                 | teuer                           | gute                   | test                 |
| 2 lieferung                       | schlaf              | test                 | der preis                       | zuverlaessig           | habe                 |
| 3 lange gedauert                  | mir                 | mir                  | preis ist                       | zuverlaessig sehr      | von Cerascreen       |
| 4 lieferzeit                      | das pulver          | dass                 | guter preis                     | sehr                   | den                  |
| 5 etwas                           | das kissen          | meine                | aber                            | sehr zuverlaessig      | schon                |
| 6 lieferung hat                   | habe                | allergien            | der preis ist                   | qualitaet ist          | den test             |
| 7 hat                             | schlafen            | zu                   | ein bisschen                    | sehr gute              | auch                 |
| 8 zu lange                        | mehr                | auf                  | bisschen                        | angebot                | von                  |
| 9 ware                            | es                  | jetzt                | qualitaet                       | sehr gut               | tests                |
| 10 die lieferung                  | schlafe             | nicht                | guter                           | gut                    | bei                  |
| 11 bestellung                     | kann                | eine                 | den preis                       | sehr zuverlaessig sehr | mal                  |
| 12 sehr lange                     | das                 | den                  | ist                             | gutes anbot            | kann                 |
| 13 bis                            | auf                 | mich                 | teuer aber                      | die qualitaet          | bei Cerascreen       |
| 14 hat etwas                      | mal                 | allergie             | finde                           | pure hochwertige       | ein                  |
| Topic_39                          | Topic_51            | Topic_53             | Topic_57                        | Topic_59               | Topic_64             |
| 0 protein                         | omega               | tabletten            | produkt                         | magnesium              | wirkung              |
| 1 proteinpulver                   | fisch               | die tabletten        | das produkt                     | das magnesium          | die wirkung          |
| 2 proteine                        | sind                | sind                 | es                              | nehme                  | wirkung ist          |
| 3 schmeckt                        | omega kapseln       | tropfen              | habe                            | haushalt               | die wirkung ist      |
| 4 es                              | omega ist           | kurkuma              | habe das produkt                | seit                   | wirkt                |
| 5 zu                              | das omega           | die                  | mir                             | dieses magnesium       | schon nach           |
| 6 auf                             | mich                | empfohlen            | geholfen                        | vergangenheit          | es wirkt             |
| 7 mich                            | produkt             | wirkung              | dieses produkt                  | mir                    | noch                 |
| 8 nicht                           | krill               | mich                 | das                             | dieses                 | aber                 |
| 9 riegel                          | da                  | inhaltsstoffe        | dieses                          | es                     | nach                 |
| 10 lecker                         | ist                 | lassen sich          | ich habe                        | hilft                  | gute wirkung         |
| 11 bin                            | epa und             | dosieren             | ich habe das                    | mir gefaellt           | wochen               |
| 12 nur                            | epa                 | hierher              | empfehlung                      | magnesium haushalt     | mal kaufen           |
| 13 proteinriegel                  | habe                | von kurkuma          | diesem produkt                  | magnesium nehme        | noch nicht           |
| 14 bars                           | dha                 | habe                 | bisher                          | gefaellt               | kann                 |
| Topic_65                          | Topic_66            | Topic_69             | Topic_71                        | Topic_73               | Topic_78             |
| 0 vegan                           | anleitung           | test                 | handhabung                      | sport                  | besser               |
| 1 vegan und                       | die anleitung       | leider ist           | einfache                        | training               | fuehle               |
| 2 produkt                         | meine               | zum arzt             | einfache handhabung             | sportler               | nach                 |
| 3 vegetarisch                     | der anleitung       | ist er sehr          | schnelle                        | nach dem sport         | viel besser          |
| 4 auch vegan                      | ueber               | er sehr              | handhabung sehr                 | dem sport              | viel                 |
| 5 veganer                         | ergebnis            | er sehr teuer        | schnelle lieferung              | dem                    | nach einem monat     |
| 6 vegane                          | mich                | leider ist er        | schnelles ergebnis              | nach dem               | winter               |
| 7 fuer                            | bereits             | erspart              | verstaendliche                  | mich                   | haut                 |
| 8 ich jedem                       | durchfuehrung       | sehr teuer           | ergebnisse                      | nach                   | einem monat          |
| 9 der sich                        | das                 | zu hause             | schnelles                       | viel                   | ich mich             |
| 10 sich groestenteils             | das ergebnis        | einfacher            | handhabung und                  | mehr                   | einnahme             |
| 11 ernaeht                        | mir                 | hause                | lieferung                       | dass                   | mich                 |
| 12 guter dosierpipette            | zur post gebracht   | gute                 | ausfuehrlichem ergebnisbericht  | immer nach             | monat                |
| 13 groestenteils vegetarisch oder | der                 | ist er               | ergebnisbericht handhabung sehr | schon                  | besser und           |
| 14 groestenteils vegetarisch      | war                 | sich gut             | ergebnisbericht handhabung      | als sportler           | wie ich mich         |

### 3.5 Application

Creating a topic model that creates many interpretable topic clusters is not where the actual project stopped. After the creation of the topic model, the gained knowledge was organized.

Firstly, a dashboard has been developed in Quicksight that isolates the reviews. The dashboard presents all ratings from all countries and the progress of the countries' national average rating. An overview of the current countries' ratings is a fast indicator showing that customers' happiness with the company is either deteriorating or improving. Since specific sources also allow a per product breakdown, product ratings are displayed on a second dashboard. However, this dashboard cannot capture all reviews, decreasing its overall accuracy. Since Cerascreen wants to maintain an average rating of four out of five, the overall rating dashboard helps, for example, the Amazon team, customer support, and management to take action if necessary.

Although the rating already makes a good key performance indicator on how happy the customers are with Cerascreen, the company is not able to understand, for example, what product problems customers have. Therefore, the created topic model functions as a controlling add-on to the dashboard to deliver a better insight into the current problems. Additionally, the topic model is faster to filter the reviews on their content compared to the employee reading every single review.

The topic model can deliver the actual status quo on what is being said in the reviews, how this state is changing over time, and where the problems lie, helping management take action. The process works as follows: Once the ratings drop, a certain number of clusters talk about a specific topic. Assuming it is a negative topic, the decreasing amount of clusters regarding this topic would imply a decreasing presence of the problem for customers. Further, if, for any reason, the amount of the negative topic cluster were to increase, the topic model can extract the most important reviews from the clusters showing precisely why customers are unhappy. Hence, the topic model is always activated if specific insights about the text reviews are needed. Of course, this also works with the development of positive reviews.

## CONCLUSION AND DISCUSSION

The goal of this project was to identify topics customers talk about the most. A new initiative at Cerascreen wants to monetize on the potential knowledge that can be retrieved from the online reviews to better understand its customers and offer greater products. The analysis should serve as an overview of the initial situation for an executive to decide on what actions to take next. Since there has been no such prior customer review analysis, the data was scattered across different review platforms. Hence, all the data had to be retrieved before serving as input for the general analysis. Nevertheless, the analysis brought up many text characteristics and insights about topics that are important to the customers.

The retrieving step required building a system that retrieves and stores all the data. Afterward, the text analysis compared multiple models to find an approach that is able to produce a variety of meaningful topic clusters. The most effective BERTopic model was able to find 24 significant and meaningful topics presented to the company and made available to the departments that needed them for decision-making and future project planning.

Although the project was an overall success and finds its use across different departments at Cerascreen, there are several points that can be improved in future projects. Unfortunately, these aspects couldn't be incorporated into this project due to time constraints that arose through the extensive review retrieval.

First of all, while this project was running, most reviews were stored on Amazon Web Services. Thus, in order to make it an automated process, the build model will be rolled out on Amazon Web Services and QuickSight for visualization. Additionally, as more reviews will be available, the analysis should also focus on either negative or positive reviews. Further, there should also be an analysis of the reviews on a product and platform level in the future. This will help to understand the specific product

---

problems and the characteristics of each review platform.

Overall, the project has been a success since several departments now know how to take action. It is a good path for Cerascreen's to understand its customers better, and start utilization the company's data more effectively.

## BIBLIOGRAPHY

- Al-Twairesh, N. and H. Al-Negheimish (2019). "Surface and deep features ensemble for sentiment analysis of arabic tweets." In: *IEEE Access* 7, pp. 84122–84131.
- Alghamdi, R. and K. Alfalqi (2015). "A Survey of Topic Modeling in Text Mining." In: *International Journal of Advanced Computer Science and Applications* 6.1. DOI: 10.14569/IJACSA.2015.060121. URL: <http://dx.doi.org/10.14569/IJACSA.2015.060121>.
- Betts Korenda, G. (2020). "Are consumers already living the future of health?" In: *Deloitte's Life Sciences and Health Care Consulting Services*.
- Campello, R. J., D. Moulavi, and J. Sander (2013). "Density-based clustering based on hierarchical density estimates." In: *Pacific-Asia conference on knowledge discovery and data mining*. Springer, pp. 160–172.
- Cer, D., Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar, et al. (2018). "Universal sentence encoder." In: *arXiv preprint arXiv:1803.11175*.
- Halkidi, M. (2009). "Hierarchical Clustering." In: *Encyclopedia of Database Systems*. Ed. by L. LIU and M. T. ÖzSU. Boston, MA: Springer US, pp. 1291–1294. ISBN: 978-0-387-39940-9. DOI: 10.1007/978-0-387-39940-9\_604. URL: [https://doi.org/10.1007/978-0-387-39940-9\\_604](https://doi.org/10.1007/978-0-387-39940-9_604).
- Han, J., J. Pei, and M. Kamber (2011). *Data mining: concepts and techniques*. Elsevier.
- Hassani, H., C. Beneki, S. Unger, M. T. Mazinani, and M. R. Yeganegi (2020). "Text mining in big data analytics." In: *Big Data and Cognitive Computing* 4.1, p. 1.
- Jianqiang, Z. and G. Xiaolin (2017). "Comparison research on text pre-processing methods on twitter sentiment analysis." In: *IEEE Access* 5, pp. 2870–2879.
- Kao, A. and S. R. Poteet (2007). *Natural language processing and text mining*. Springer Science & Business Media.
- Kirilenko, A. P., S. O. Stepchenkova, and X. Dai (2021). "Automated topic modeling of tourist reviews: does the Anna Karenina principle apply?" In: *Tourism Management* 83, p. 104241.
- Kumar, P. (2021). "The Significance of Data Mining Applications in Evaluating Customer Reviews in the Context of E-commerce Transactions." In: *International Journal of Advanced Research in Technology and Innovation* 3.3, pp. 100–110.

- Leland McInnes John Healy, S. A. (2016). *How HDBSCAN Works*. URL: [https://hdbscan.readthedocs.io/en/latest/how\\_hdbscan\\_works.html](https://hdbscan.readthedocs.io/en/latest/how_hdbscan_works.html).
- Liu, L., L. Tang, W. Dong, S. Yao, and W. Zhou (2016). "An overview of topic modeling and its current applications in bioinformatics." In: *SpringerPlus* 5.1, pp. 1–22.
- Lovrić, M., T. Đuričić, H. T. Tran, H. Hussain, E. Lacić, M. A. Rasmussen, and R. Kern (2021). "Should We Embed in Chemistry? A Comparison of Unsupervised Transfer Learning with PCA, UMAP, and VAE on Molecular Fingerprints." In: *Pharmaceuticals* 14.8, p. 758.
- McInnes, L., J. Healy, and J. Melville (2020). "Umap: Uniform manifold approximation and projection for dimension reduction." In: *arXiv preprint arXiv:1802.03426*.
- Mikolov, T., I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). "Distributed representations of words and phrases and their compositionality." In: *Advances in neural information processing systems*, pp. 3111–3119.
- Page, J. T., Z. S. Liechty, M. D. Huynh, and J. A. Udall (2014). "BamBam: genome sequence analysis tools for biologists." In: *BMC Research Notes* 7.1, pp. 1–5.
- Qaiser, S. and R. Ali (2018). "Text mining: use of TF-IDF to examine the relevance of words to documents." In: *International Journal of Computer Applications* 181.1, pp. 25–29.
- Reddy, G. T., M. P. K. Reddy, K. Lakshmana, R. Kaluri, D. S. Rajput, G. Srivastava, and T. Baker (2020). "Analysis of dimensionality reduction techniques on big data." In: *IEEE Access* 8, pp. 54776–54788.
- Reimers, N. and I. Gurevych (2019). "Sentence-bert: Sentence embeddings using siamese bert-networks." In: *arXiv preprint arXiv:1908.10084*.
- Samir, A. and Z. Lahbib (2018). "Stemming and lemmatization for information retrieval systems in amazigh language." In: *International Conference on Big Data, Cloud and Applications*. Springer, pp. 222–233.
- Silveira, R., C. G. O. Fernandes, J. A. M. Neto, V. Furtado, and J. E. P. Filho (2021). "Topic Modelling of Legal Documents via LEGAL-BERT." In: *Proceedings http://ceur-ws.org ISSN 1613*, p. 0073.
- Vanjani, M. and M. Aiken (2020). "A comparison of free online Machine Language Translators." In: *Journal of Management Science and Business Intelligence*, pp. 26–31.
- Verma, S. and N. Yadav (2021). "Past, present, and future of electronic word of mouth (EWOM)." In: *Journal of Interactive Marketing* 53, pp. 111–128.
- Vijayarani, S, M. J. Ilamathi, M. Nithya, et al. (2015). "Preprocessing techniques for text mining-an overview." In: *International Journal of Computer Science Communication Networks* 5.1, pp. 7–16.
- Westerlund, M., S. Leminen, and M. Rajahonka (2018). "A topic modelling analysis of living Labs research." In: *Technology Innovation Management Review* 8.7.
- Zhang, F., H. Fleyeh, X. Wang, and M. Lu (2019). "Construction site accident analysis using text mining and natural language processing techniques." In: *Automation in Construction* 99, pp. 238–248.

Zhao, R. and K. Mao (2017). “Fuzzy bag-of-words model for document representation.” In: *IEEE transactions on fuzzy systems* 26.2, pp. 794–804.



## DIFFERENT TOPIC MODELS COMPARISON

This section contains a detailed comparison of all cluster algorithms deriving why HDBScan has been used.

### A.0.0.1 K-Means

The figure 18c indicated that some clusters are very close to each other. Looking at Table 8 this indication is confirmed by five topics, which all mention that everything with the products and services is good and that customers are satisfied. Furthermore, the other five topics mention fast shipping, fast results, good supplement flavor, good communication, and specifically the vitamin D blood test. Hence, 50% of the topics could be merged into one topic while the other 50% are very generic as no specifics about a service or product are mentioned. Given that most results are positive and probably from very short reviews, this is what K-Means is mostly showing. It misses the smaller clusters of topics where customers give more detailed feedback. Additionally, the model could not identify any negative topics, although, from the previous analysis, it is apparent that customers do have some complaints, given the negative reviews.

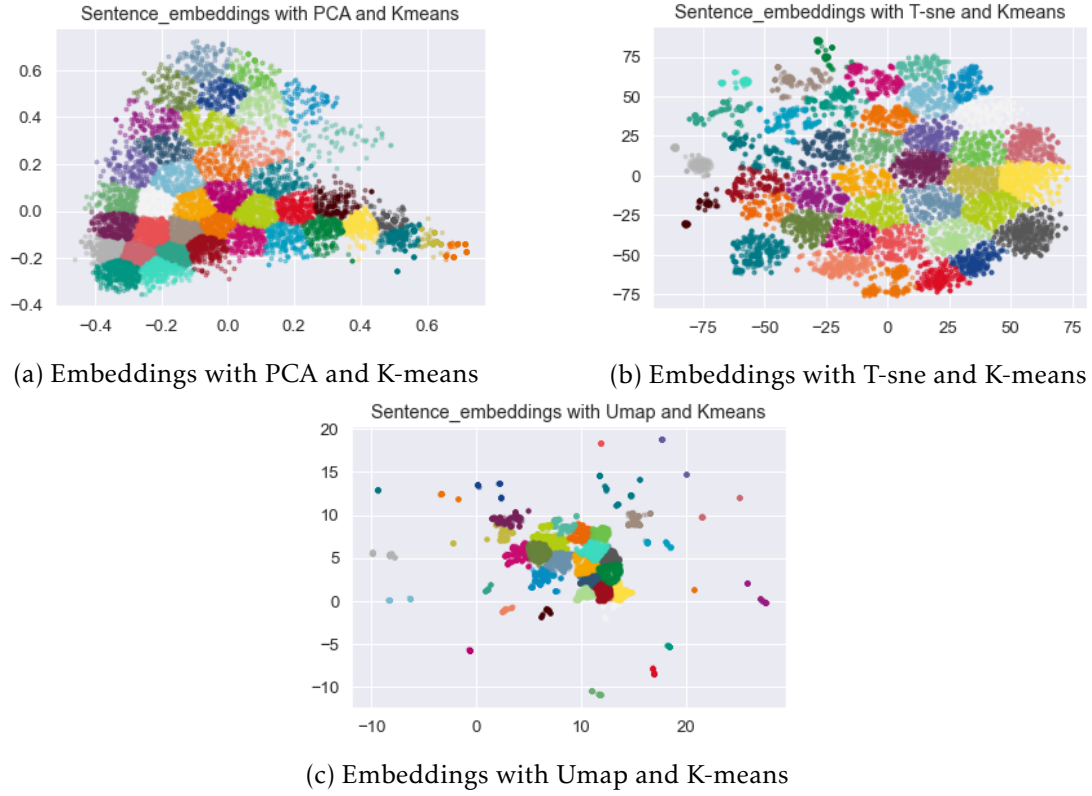


Figure 18: K-Means with different dimensionality reductions

Table 8: Sentence embeddings with Umap and K-mean

| S | Topic_1       | Topic_1_Val | Topic_2       | Topic_2_Val | Topic_3    | Topic_3_Val | Topic_4    | Topic_4_Val | Topic_5      | Topic_5_Val  |
|---|---------------|-------------|---------------|-------------|------------|-------------|------------|-------------|--------------|--------------|
| 0 | test          | 0.071973    | produkt       | 0.442169    | kapseln    | 0.062653    | ok         | 1.584606    | immer        | 0.264539     |
| 1 | ergebnis      | 0.041491    | geklappt      | 0.190803    | fuer       | 0.050702    | prima      | 0.821222    | top          | 0.261874     |
| 2 | lieferung     | 0.036684    | super         | 0.166423    | gut        | 0.047123    | ordnung    | 0.444978    | perfekt      | 0.180431     |
| 3 | erhalten      | 0.035372    | zufrieden     | 0.158469    | super      | 0.038467    | zufrieden  | 0.348159    | funktioniert | 0.171884     |
| 4 | wurde         | 0.035021    | gut           | 0.100704    | produkt    | 0.036371    | super      | 0.346555    | super        | 0.171433     |
| 5 | tests         | 0.034890    | zufriedenheit | 0.086632    | schon      | 0.031645    | tip        | 0.212006    | gut          | 0.155524     |
| 6 | Cerascreen    | 0.034651    | gelaufen      | 0.073771    | geschmack  | 0.031099    | top        | 0.139055    | zufrieden    | 0.140500     |
| 7 | tagen         | 0.031839    | gutes         | 0.073451    | nehme      | 0.027436    | bester     | 0.061175    | danke        | 0.114009     |
| 8 | einfach       | 0.031259    | ganz          | 0.057870    | preis      | 0.027344    | topp       | 0.057088    | bestens      | 0.112503     |
| 9 | schnell       | 0.031107    | danke         | 0.051270    | schmeckt   | 0.027145    | insgesamt  | 0.046269    | gerne        | 0.100289     |
|   | Topic_6       | Topic_6_Val | Topic_7       | Topic_7_Val | Topic_8    | Topic_8_Val | Topic_9    | Topic_9_Val | Topic_10     | Topic_10_Val |
| 0 | bestens       | 3.290803    | gerne         | 0.641547    | gut        | 2.018121    | test       | 0.063056    | schnelle     | 0.204118     |
| 1 | okay          | 0.359111    | top           | 0.545808    | super      | 0.316079    | vitamin    | 0.057803    | lieferung    | 0.197800     |
| 2 | bestes        | 0.065005    | immer         | 0.398179    | far        | 0.058596    | blut       | 0.036976    | schnell      | 0.123564     |
| 3 | alley         | 0.040511    | perfekt       | 0.258503    | gutes      | 0.056478    | ergebnis   | 0.030919    | auswertung   | 0.074528     |
| 4 | alles         | 0.037658    | gern          | 0.230588    | good       | 0.052056    | fuer       | 0.030777    | super        | 0.069347     |
| 5 | alls          | 0.035989    | bestens       | 0.207332    | wirklich   | 0.048011    | Cerascreen | 0.029523    | service      | 0.066614     |
| 6 | kommunikation | 0.025879    | jederzeit     | 0.168200    | guenstig   | 0.038761    | einfach    | 0.024810    | gut          | 0.058602     |
| 7 | besten        | 0.023872    | super         | 0.157967    | eigentlich | 0.036943    | arzt       | 0.022281    | ware         | 0.055971     |
| 8 | ja            | 0.020145    | zufrieden     | 0.117445    | ja         | 0.029139    | gut        | 0.022258    | gute         | 0.055482     |
| 9 | top           | 0.015896    | gut           | 0.084841    | gute       | 0.021922    | schnell    | 0.021969    | zufrieden    | 0.048466     |

---

### A.0.0.2 Agglomerative Clustering

In its default setup, Agglomerative Clustering was also set to 10 clusters for comparison reasons. The figure 19c shows already more spread out clusters, but there also seems to be a center of close clusters. Looking at the cluster's contents in Table 9, just like K-Means agglomerative clustering produces many clusters that have very similar topics. Although there are only four topics revolving around satisfied customers and how everything worked well for them, one topic consists of random words, mostly starting with the letter "g". Nevertheless, it was still able to find two topics about fast delivery of tests and results about the good service. Additionally, there seem to be customers that like the vitamin D test and services, such as a smooth ordering process. The results show that the variety and meaningfulness of the clusters do not differentiate much from the K-Means topics. Furthermore, like K-Means, agglomerative clustering misses more detailed feedback and negative reviews, which the over-proportionate amount of positive reviews can explain in the data set.

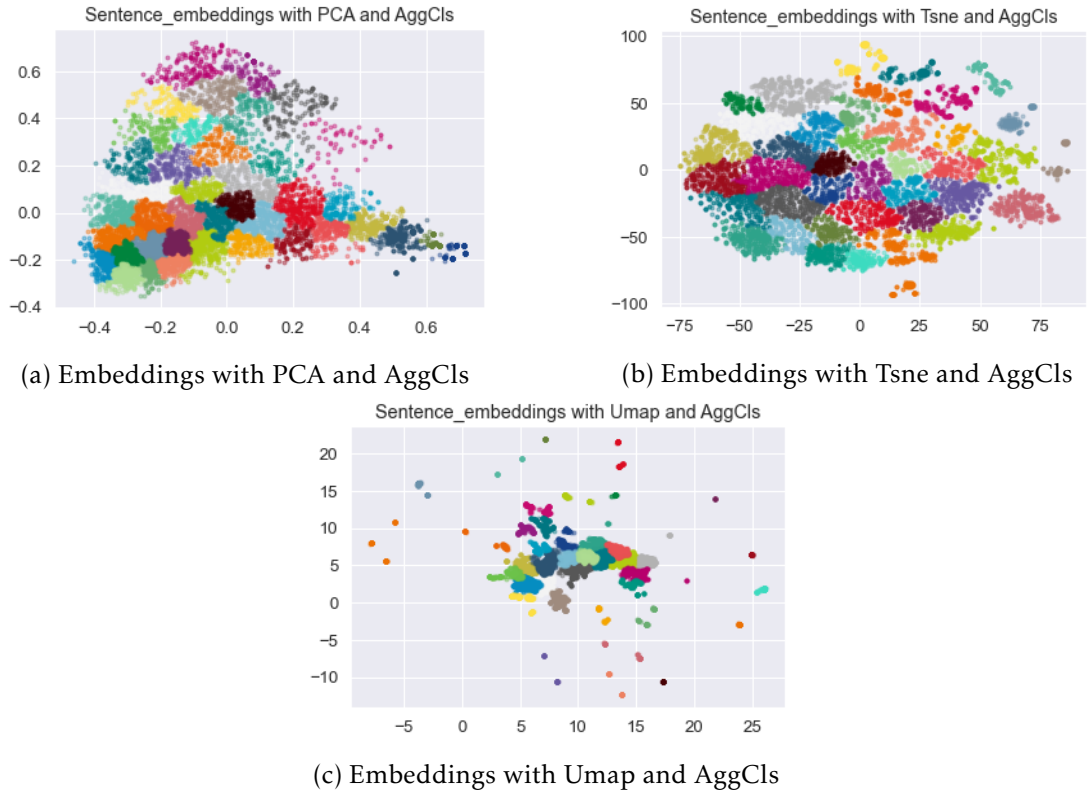


Figure 19: Agglomerative Clustering with different dimensionality reductions

Table 9: Sentence embeddings with Umap and Agglomerative Clustering

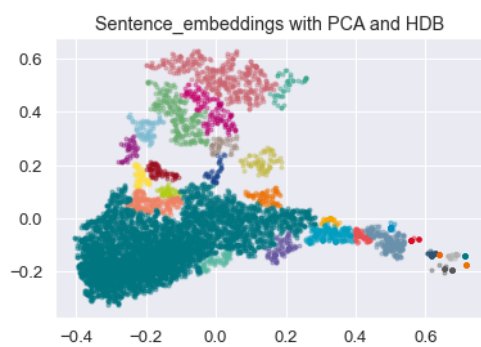
|   | Topic_1       | Topic_1_Val | Topic_2    | Topic_2_Val | Topic_3         | Topic_3_Val | Topic_4   | Topic_4_Val | Topic_5    | Topic_5_Val  |
|---|---------------|-------------|------------|-------------|-----------------|-------------|-----------|-------------|------------|--------------|
| 0 | zufrieden     | 0.447168    | test       | 0.069153    | bestens         | 2.270783    | gut       | 0.967924    | lieferung  | 0.173115     |
| 1 | gerne         | 0.383589    | Cerascreen | 0.040094    | top             | 0.660003    | bestens   | 0.363908    | schnelle   | 0.169788     |
| 2 | top           | 0.243956    | ergebnis   | 0.037671    | perfekt         | 0.319762    | perfekt   | 0.358622    | schnell    | 0.106312     |
| 3 | immer         | 0.231571    | wurde      | 0.033717    | zufrieden       | 0.235402    | okay      | 0.346285    | auswertung | 0.072463     |
| 4 | gut           | 0.227129    | erhalten   | 0.033008    | bestes          | 0.044247    | prima     | 0.310364    | gut        | 0.058761     |
| 5 | zufriedenheit | 0.181233    | tagen      | 0.029620    | einwandfrei     | 0.031268    | danke     | 0.272264    | super      | 0.058514     |
| 6 | danke         | 0.111110    | tests      | 0.027524    | oerfekt         | 0.027574    | vielen    | 0.187310    | service    | 0.052965     |
| 7 | bestens       | 0.109897    | zufrieden  | 0.027163    | alley           | 0.027574    | dank      | 0.178320    | ware       | 0.050425     |
| 8 | gern          | 0.095430    | einfach    | 0.026344    | vollumfaenglich | 0.025633    | zufrieden | 0.107192    | gute       | 0.047698     |
| 9 | super         | 0.085994    | gut        | 0.024834    | alles           | 0.025633    | perfekte  | 0.082138    | einfach    | 0.045975     |
|   | Topic_6       | Topic_6_Val | Topic_7    | Topic_7_Val | Topic_8         | Topic_8_Val | Topic_9   | Topic_9_Val | Topic_10   | Topic_10_Val |
| 0 | produkt       | 0.080234    | super      | 1.113554    | immer           | 0.400451    | ok        | 3.634418    | vitamin    | 0.063112     |
| 1 | kapseln       | 0.064158    | geklappt   | 0.583750    | geklappt        | 0.319185    | super     | 0.487167    | test       | 0.058182     |
| 2 | fuer          | 0.052747    | gut        | 0.476400    | super           | 0.263340    | oki       | 0.083953    | blut       | 0.040359     |
| 3 | gut           | 0.047996    | gelaufen   | 0.383112    | ordnung         | 0.261281    | note      | 0.075547    | fuer       | 0.031972     |
| 4 | super         | 0.046142    | bestens    | 0.157617    | prima           | 0.260012    | gaebe     | 0.000000    | ergebnis   | 0.030775     |
| 5 | geschmack     | 0.032004    | lief       | 0.132883    | bestens         | 0.156093    | gamma     | 0.000000    | Cerascreen | 0.025364     |
| 6 | wirklich      | 0.030151    | ok         | 0.052755    | danke           | 0.147207    | game      | 0.000000    | einfach    | 0.023562     |
| 7 | schmeckt      | 0.028956    | angekommen | 0.042575    | top             | 0.135016    | galt      | 0.000000    | gut        | 0.023135     |
| 8 | einfach       | 0.028425    | toll       | 0.041839    | dank            | 0.124169    | gaesschen | 0.000000    | schon      | 0.022845     |
| 9 | gute          | 0.028108    | heklappt   | 0.038453    | vielen          | 0.117386    | gaengig   | 0.000000    | schnell    | 0.021873     |

### A.0.0.3 HDBscan Clustering

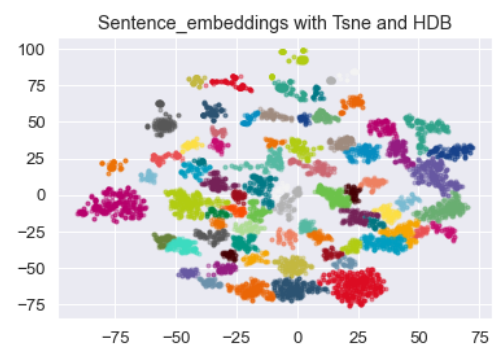
The HDBscan has no input to control the number of clusters, and thus the clusters had to be reduced manually. Further, the clusters in the figure 20c are even closer together than before, but they are also smaller. The other algorithms seemed to have only found one cluster; the HDBscan found more. That HDBscan finds smaller clusters is also represented by the content of the clusters in Table 10 as every topic mentions something different. For example, the topics mention that the products are good for allergic people and that vitamin and iron supplements are good for travel. The testosterone test is used as additional knowledge to have successful training sessions. Further, the tests are also being used for yearly blood level checks. In addition, Cerascreen is also being called a great store, trustworthy, and commendable with quick deliveries. Compared to the previous models, the HDBscan produces much more insights about what customers write in their reviews. However, although the HDBscan is missing the negative annotations, the topic variety and meaningfulness are more substantial.

Table 10: Sentence embeddings with Umap and HDBscan

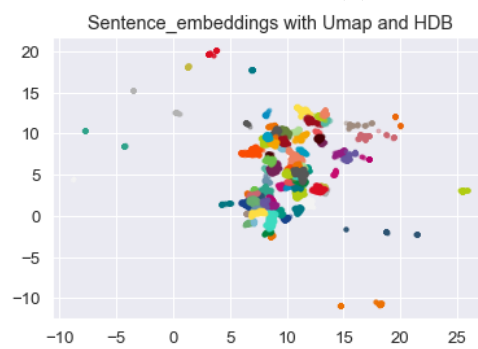
|   | Topic_1       | Topic_1_Val | Topic_2     | Topic_2_Val | Topic_3          | Topic_3_Val | Topic_4            | Topic_4_Val | Topic_5      | Topic_5_Val  |
|---|---------------|-------------|-------------|-------------|------------------|-------------|--------------------|-------------|--------------|--------------|
| 0 | perfekt       | 1.991076    | schnelle    | 1.180348    | tip              | 2.038517    | testosteron        | 0.243989    | super        | 1.185673     |
| 1 | perfekte      | 0.481369    | lieferung   | 1.067121    | top              | 1.337068    | training           | 0.099037    | produkt      | 0.865443     |
| 2 | allergiker    | 0.228932    | lieferung   | 0.103622    | ordnung          | 0.388966    | testosteronspiegel | 0.097301    | superware    | 0.298305     |
| 3 | dankesch      | 0.228932    | lieferung   | 0.103622    | zuverlaessiger   | 0.316716    | test               | 0.097047    | mygreenz     | 0.298305     |
| 4 | droht         | 0.228932    | schneller   | 0.097002    | tipp             | 0.294109    | wissen             | 0.083106    | produkte     | 0.296685     |
| 5 | krise         | 0.203382    | versand     | 0.096669    | bester           | 0.294109    | erfolge            | 0.079179    | betreuung    | 0.244009     |
| 6 | ideale        | 0.196692    | beraus      | 0.089029    | topp             | 0.274462    | tests              | 0.073673    | energie      | 0.209079     |
| 7 | zugehen       | 0.196692    | langsames   | 0.089029    | tiptop           | 0.274462    | schon              | 0.070039    | artikel      | 0.188883     |
| 8 | kapseln       | 0.186613    | rasche      | 0.075116    | shop             | 0.237711    | weitergegeben      | 0.065755    | mehr         | 0.119770     |
| 9 | kochen        | 0.175383    | lieferzeit  | 0.061632    | geklappt         | 0.155343    | durchfuehr         | 0.065755    | fuer         | 0.088979     |
|   | Topic_6       | Topic_6_Val | Topic_7     | Topic_7_Val | Topic_8          | Topic_8_Val | Topic_9            | Topic_9_Val | Topic_10     | Topic_10_Val |
| 0 | ordnung       | 1.444733    | immer       | 1.243886    | immer            | 1.183874    | perfekt            | 1.729622    | zufrieden    | 1.200549     |
| 1 | immer         | 1.285349    | gerne       | 0.876818    | zufrieden        | 0.947802    | gelaufen           | 1.407798    | immer        | 0.964012     |
| 2 | zuverlaessig  | 0.306670    | gern        | 0.475227    | vollsten         | 0.232462    | ablauf             | 0.364571    | jaehrlich    | 0.328135     |
| 3 | tablettten    | 0.281259    | zufrieden   | 0.193637    | zufriedenheit    | 0.196335    | schnuerchen        | 0.298305    | ueberpruefen | 0.274487     |
| 4 | vorbildlich   | 0.215471    | astaxanthin | 0.141056    | gruessen         | 0.160542    | perfekter          | 0.235291    | vorgestellt  | 0.245305     |
| 5 | handling      | 0.210261    | reise       | 0.136416    | kundenbetreuung  | 0.148382    | reibungsloser      | 0.235291    | lasse        | 0.240167     |
| 6 | eisen         | 0.181359    | vitamaze    | 0.132816    | ueberaus         | 0.148382    | erklaeerung        | 0.172278    | zuvoer       | 0.240167     |
| 7 | genau         | 0.137908    | mal         | 0.121056    | note             | 0.144467    | lief               | 0.171808    | abgelaufen   | 0.229987     |
| 8 | unkompliziert | 0.136076    | schon       | 0.112683    | freundlichen     | 0.132307    | empfehlenswert     | 0.153739    | bisher       | 0.169112     |
| 9 | gutes         | 0.135548    | rundum      | 0.099685    | kundenfreundlich | 0.130634    | ergebnisse         | 0.119190    | spiegel      | 0.168001     |



(a) Embeddings with PCA and HDB



(b) Embeddings with Tsne and HDB



(c) Embeddings with Umap and HDB

Figure 20: HDBScan with different dimensionality reductions



## BERTopic WHOLE TOPIC TABLE

Table 11: BERTopic whole topic table a

| Topic_1 | Topic_2              | Topic_3             | Topic_4         | Topic_5                  | Topic_6               |
|---------|----------------------|---------------------|-----------------|--------------------------|-----------------------|
| 0       | sehr                 | vitamin             | blut            | ware                     | geschmack             |
| 1       | einfach              | test                | das             | bestellung               | schmeckt              |
| 2       | der                  | vitamin test        | nicht           | nicht                    | lecker                |
| 3       | test                 | den                 | war             | erhalten                 | der geschmack         |
| 4       | das                  | von                 | den             | noch                     | gut                   |
| 5       | habe                 | habe                | finger          | keine                    | schmecken             |
| 6       | ist                  | zu                  | test            | lieferung                | geschmack ist         |
| 7       | bin                  | mit                 | der             | nach                     | lecker und            |
| 8       | mit                  | mein                | ist             | wurde                    | der geschmack ist     |
| 9       | zu                   | ist                 | man             | die ware                 | ist                   |
| 10      | war                  | der                 | die             | auf                      | snack                 |
| 11      | auch                 | auch                | zu              | die lieferung            | fuer                  |
| 12      | nicht                | spiegel             | es              | wochen                   | sehr lecker           |
| 13      | die                  | das                 | mit             | leider                   | super                 |
| 14      | auf                  | wert                | blutprobe       | antwort                  | der                   |
| 0       | schnelle             | service             | buch            | wieder                   | test                  |
| 1       | lieferung            | kundenservice       | das buch        | gerne                    | der test              |
| 2       | des                  | guter               | empfehlen       | wieder alles             | schnell               |
| 3       | schnelle lieferung   | guter service       | kann            | gerne wieder             | war                   |
| 4       | tests                | toller              | mir             | immer wieder             | der                   |
| 5       | auswertung           | lieferung           | empfehlenswert  | immer                    | einfach               |
| 6       | des tests            | schnelle lieferung  | weiter          | wieder gerne             | tests                 |
| 7       | test                 | schnelle            | es              | immer wieder gerne       | man                   |
| 8       | lieferung und        | schneller           | kann ich        | gerne wieder alles       | test war              |
| 9       | handhabung           | toller service      | habe            | alles                    | sehr                  |
| 10      | schnell              | sehr guter          | mich            | gern                     | test ist              |
| 11      | bearbeitung          | schneller service   | nehme           | alles bestens gerne      | ergebnis              |
| 12      | schnelle bearbeitung | super service       | empfehlung      | bestens gerne wieder     | der test ist          |
| 13      | einfach              | kundenservice sehr  | weiterempfehlen | bestens gerne            | ist                   |
| 14      | einfache             | guter kundenservice | buch ist        | wieder immer             | sind                  |
| 0       | lange                | tagen               | kissen          | schnelle lieferung       | schnell               |
| 1       | gedauert             | test                | pulver          | schnelle                 | alles                 |
| 2       | lieferung            | probe               | schlaf          | lieferung                | schnell und           |
| 3       | lange gedauert       | der                 | mir             | lieferung alles          | alles schnell         |
| 4       | lieferzeit           | nach                | das pulver      | schnelle lieferung alles | alles schnell und     |
| 5       | etwas                | ergebnis            | das kissen      | alles                    | und schnell           |
| 6       | lieferung hat        | der test            | habe            | alles bestens            | geklappt              |
| 7       | hat                  | nicht               | schlafen        | bestens                  | hat alles             |
| 8       | zu lange             | am                  | mehr            | lieferung alles bestens  | alles super           |
| 9       | ware                 | war                 | es              | gerne wieder             | schnell alles         |
| 10      | die lieferung        | in                  | schlafe         | wieder                   | hat                   |
| 11      | bestellung           | das                 | kann            | gerne                    | alles gut             |
| 12      | sehr lange           | ist                 | das             | super schnelle           | super                 |
| 13      | bis                  | das ergebnis        | auf             | wie immer                | unkompliziert         |
| 14      | hat etwas            | auf                 | mal             | ware                     | gut                   |
| 0       | lange                | tagen               | kissen          | schnelle lieferung       | bestens alles bestens |
| 1       | gedauert             | test                | pulver          | schnelle                 | alles bestens alles   |
| 2       | lieferung            | probe               | schlaf          | lieferung                | bestens alles         |
| 3       | lange gedauert       | der                 | mir             | lieferung alles          | alles bestens         |
| 4       | lieferzeit           | nach                | das pulver      | schnelle lieferung alles | bestens               |
| 5       | etwas                | ergebnis            | das kissen      | alles                    | alles                 |
| 6       | lieferung hat        | der test            | habe            | alles bestens            | zum besten alles      |
| 7       | hat                  | nicht               | schlafen        | bestens                  | alles zum besten      |
| 8       | zu lange             | am                  | mehr            | lieferung alles bestens  | ja alles bestens      |
| 9       | ware                 | war                 | es              | gerne wieder             | besten alles          |
| 10      | die lieferung        | in                  | schlafe         | wieder                   | besten alles bestens  |
| 11      | bestellung           | das                 | kann            | gerne                    | bestens ja alles      |
| 12      | sehr lange           | ist                 | das             | super schnelle           | zum besten            |
| 13      | bis                  | das ergebnis        | auf             | wie immer                | bestens alles zum     |
| 14      | hat etwas            | auf                 | mal             | ware                     | bestens ja            |

Table 12: BERTopic whole topic table b

|    | Topic_19                     | Topic_20                  | Topic_21                      | Topic_22                 | Topic_23                 | Topic_24                |
|----|------------------------------|---------------------------|-------------------------------|--------------------------|--------------------------|-------------------------|
| 0  | lebensmittel                 | schnelle                  | nicht                         | super alles              | wie immer                | zufrieden               |
| 1  | habe                         | lieferung und             | sie                           | alles super alles        | wie                      | bin                     |
| 2  | test                         | schnelle lieferung und    | wurde                         | super alles super        | wie immer alles          | sehr zufrieden          |
| 3  | mir                          | lieferung                 | erhalten                      | alles super              | immer                    | bin sehr                |
| 4  | dass                         | schnelle lieferung        | ich habe                      | super                    | immer alles              | test                    |
| 5  | meine                        | auswertung                | habe                          | alles                    | bestens wie              | bin sehr zufrieden      |
| 6  | allergien                    | schnelle auswertung       | keine                         | super alles wunderbar    | alles bestens wie        | ich bin                 |
| 7  | zu                           | und schnelle              | bestellung                    | wunderbar                | bestens wie immer        | ich bin sehr            |
| 8  | auf                          | bearbeitung               | eine                          | wunderbar alles          | alles                    | sehr                    |
| 9  | jetzt                        | und auswertung            | nach                          | alles wunderbar          | wie immer wie            | den test                |
| 10 | nicht                        | gute                      | auf                           | toll alles super         | immer wie                | den                     |
| 11 | eine                         | schnelle und              | test                          | alles toll alles         | wie beschrieben          | tests                   |
| 12 | den                          | lieferung und auswertung  | dass                          | wunderbar alles super    | bestens                  | sehr zufrieden mit      |
| 13 | mir                          | auswertung und            | der                           | alles wunderbar alles    | alles bestens            | mit                     |
| 14 | allergie                     | schnelle auswertung und   | an                            | toll alles               | immer wie immer          | den test zum            |
|    | Topic_25                     | Topic_26                  | Topic_27                      | Topic_28                 | Topic_29                 | Topic_30                |
| 0  | preis                        | schnell                   | produkt                       | qualitaet                | gut gut                  | zufriedenheit           |
| 1  | teuer                        | schnell und               | das produkt                   | gute                     | abc                      | zu meiner               |
| 2  | der preis                    | zuverlaessig              | top produkt                   | zuverlaessig             | gut                      | alles zu                |
| 3  | preis ist                    | und zuverlaessig          | das                           | zuverlaessig sehr        | gut gut gut              | alles zu meiner         |
| 4  | guter preis                  | schnell und zuverlaessig  | fuer                          | sehr                     | supi                     | zu meiner zufriedenheit |
| 5  | aber                         | zuverlaessig schnell      | dieses                        | sehr zuverlaessig        | passt                    | meiner zufriedenheit    |
| 6  | der preis ist                | und zuverlaessig schnell  | produkt ich                   | qualitaet ist            | gut passt                | zufrieden alles         |
| 7  | ein bisschen                 | zuverlaessig schnell und  | produkt hat                   | sehr gute                | xxx                      | zufrieden               |
| 8  | bisschen                     | wie immer                 | dieses produkt                | angebot                  | name                     | alles                   |
| 9  | qualitaet                    | wie immer schnell         | produkten                     | sehr gut                 | gut abc                  | alles zur               |
| 10 | guter                        | wieder schnell und        | das produkt hat               | gut                      | ok gut                   | meiner                  |
| 11 | den preis                    | wieder schnell            | mich                          | sehr zuverlaessig sehr   | supi gut                 | zufriedenheit alles     |
| 12 | ist                          | kompetent                 | produkt ist                   | gutes anbot              | abc gut                  | zur                     |
| 13 | teuer aber                   | gerne wieder schnell      | produkt weiterempfehlen       | die qualitaet            | ok                       | geklappt                |
| 14 | finde                        | immer schnell             | das produkt weiterempfehlen   | pure hochwertige         | ist ok                   | bin                     |
|    | Topic_31                     | Topic_32                  | Topic_33                      | Topic_34                 | Topic_35                 | Topic_36                |
| 0  | Cerascreen                   | zufrieden                 | schnelle                      | gut alles gut            | zufrieden                | top alles               |
| 1  | test                         | sehr zufrieden            | schnelle lieferung            | alles gut alles          | zufrieden ich            | alles top alles         |
| 2  | habe                         | bin                       | lieferung                     | gut alles                | zufrieden ich bin        | alles top               |
| 3  | von Cerascreen               | ich bin                   | lieferung schnelle            | alles gut                | bin                      | top                     |
| 4  | den                          | bin sehr                  | schnelle lieferung schnelle   | gut                      | sehr zufrieden           | top alles top           |
| 5  | schon                        | bin sehr zufrieden        | sehr schnelle lieferung       | alles                    | ich bin                  | ordnung alles           |
| 6  | den test                     | ich bin sehr              | sehr schnelle                 | gut ja alles             | zufrieden bin            | in ordnung alles        |
| 7  | auch                         | zufrieden mit             | schnelle auswertung           | gut ja                   | bin sehr zufrieden       | ordnung alles top       |
| 8  | von                          | sehr                      | lieferung schnelle lieferung  | nichts negatives sagen   | bin sehr                 | alles                   |
| 9  | tests                        | lieferung                 | gute                          | gut eigentlich           | sehr zufrieden ich       | ordnung                 |
| 10 | bei                          | mit                       | preis                         | sagen alles gut          | ich bin sehr             | in ordnung              |
| 11 | mal                          | ware                      | auswertung                    | gut alles ist            | bin zufrieden ich        | alles in ordnung        |
| 12 | kann                         | der lieferung             | lieferung schnelle auswertung | eigentlich alles gut     | bin zufrieden            | alles in                |
| 13 | bei Cerascreen               | mit der                   | schnelle lieferung sehr       | ist gut alles            | zufrieden ich war        | top alles in            |
| 14 | ein                          | mit der lieferung         | lieferung sehr                | ja alles gut             | bin wie                  | top alles tip           |
|    | Topic_37                     | Topic_38                  | Topic_39                      | Topic_40                 | Topic_41                 | Topic_42                |
| 0  | ok alles ok                  | unkompliziert             | protein                       | schnelle                 | top top                  | produkt                 |
| 1  | alles ok alles               | und unkompliziert         | proteinpulver                 | ergebnis                 | top top top              | gutes                   |
| 2  | ok alles                     | unkompliziert schnell     | proteine                      | schnelles                | top                      | gutes produkt           |
| 3  | alles ok                     | einfach                   | schmeckt                      | schnelles ergebnis       | tip top                  | sehr gutes              |
| 4  | ok                           | unkompliziert schnell und | es                            | ergebnisse               | tip                      | sehr gutes produkt      |
| 5  | alles                        | schnell und               | zu                            | schnelle lieferung       | top tip top              | produkte                |
| 6  | alles ok note                | schnell und unkompliziert | auf                           | lieferung                | top tip                  | produkt sehr            |
| 7  | alles oki alles              | schnell                   | mich                          | schnelle ergebnisse      | tip top top              | produkt sehr gutes      |
| 8  | alles oki                    | schnell unkompliziert     | nicht                         | das ergebnis             | tip top tip              | das produkt             |
| 9  | oki alles ok                 | einfache                  | riegel                        | super                    | top top tip              | produkt ist             |
| 10 | ok alles oki                 | einfache handhabung       | lecker                        | ergebnis super           | tiptop top top           | produkt und             |
| 11 | ok note alles                | einfach und               | bin                           | und auswertung das       | die schilddruesefunktion | das produkt ist         |
| 12 | ok note                      | einfach klasse            | nur                           | kurzum schneller versand | top ablauf top           | hochwertiges produkt    |
| 13 | note alles ok                | klasse                    | proteinriegel                 | kurzum schneller         | top ablauf               | sehr                    |
| 14 | note alles                   | und unkompliziert schnell | bars                          | super schnelles          | vertraeglichkeit sues    | hochwertiges            |
|    | Topic_43                     | Topic_44                  | Topic_45                      | Topic_46                 | Topic_47                 | Topic_48                |
| 0  | lieferung schnelle lieferung | bestens alles bestens     | perfekt                       | das produkt ist          | test                     | kommentar               |
| 1  | lieferung schnelle           | alles bestens alles       | ideal                         | produkt ist              | dass                     | kein kommentar          |
| 2  | schnelle lieferung           | bestens alles             | perfekt perfekt               | das produkt              | diesen                   | kein                    |
| 3  | schnelle lieferung schnelle  | alles bestens             | perfekt fuer                  | produkt                  | habe                     | nichts zu               |
| 4  | lieferung                    | bestens                   | fuer mich                     | produkt ist echt         | mein                     | nichts                  |
| 5  | schnelle                     | alles                     | fuer                          | ist echt                 | diesen test              | keine                   |
| 6  | lieferung prompte            | bestes alles              | genau das                     | produkt ist wirklich     | mir                      | nicht guenstig          |
| 7  | prompte                      | alles bestes alles        | ideal fuer                    | ist                      | habe ich                 | einwandfrei             |
| 8  | prompte lieferung            | alles bestes              | perfekt wie immer             | das                      | nicht                    | beanstandungen          |
| 9  | lieferung schneller versand  | bestes                    | perfekt wie                   | echt                     | finde                    | besser geht             |
| 10 | lieferung prompte lieferung  | um alles                  | perfekt fuer mich             | ist wirklich             | den test                 | besser                  |
| 11 | geliefert                    | bestes alles bestens      | mich                          | produkt ist ganz         | schon                    | guenstig                |
| 12 | geliefert schnelle           | bestes alles bestes       | das was mein                  | ist ganz                 | meinen mann gekauft      | zu meckern              |
| 13 | geliefert schnelle lieferung | rund um alles             | mein koerper benoetigt        | wirklich                 | dieser                   | keine beanstandungen    |
| 14 | erhalten schnelle            | um alles bestens          | koerper benoetigt             | spitze                   | mich                     | meckern                 |
|    | Topic_49                     | Topic_50                  | Topic_51                      | Topic_52                 | Topic_53                 | Topic_54                |
| 0  | noch                         | super super               | omega                         | das produkt ist          | tabletten                | dank alles              |
| 1  | bewertung                    | super                     | fisch                         | produkt ist              | die tabletten            | vielen dank alles       |
| 2  | keine                        | super super super         | kapseln                       | das produkt              | sind                     | vielen dank             |
| 3  | kein                         | einfach super             | omega kapseln                 | produkt                  | tropfen                  | danke alles             |
| 4  | ergebnis                     | einfach super super       | omega ist                     | produkt ist wirklich     | kurkuma                  | dank                    |
| 5  | kein ergebnis                | ausgezeichnet             | das omega                     | ist wirklich             | die                      | vielen                  |
| 6  | testergebnis                 | super wunderbar super     | mich                          | produkt ist ganz         | empfohlen                | danke                   |
| 7  | auf                          | wunderbar super           | produkt                       | ist ganz                 | wirkung                  | alles bestens vielen    |
| 8  | noch nicht                   | wunderbar super super     | krill                         | produkt ist echt         | mich                     | bestens vielen dank     |
| 9  | warte                        | super super wunderbar     | da                            | ist echt                 | inhaltsstoffe            | bestens vielen          |
| 10 | kein testergebnis            | super wunderbar           | ist                           | ist                      | lassen sich              | alles                   |
| 11 | bis                          | super einfach             | epa und                       | das                      | dosieren                 | super vielen            |
| 12 | nicht                        | super einfach super       | epa                           | klasse                   | hierher                  | super vielen dank       |
| 13 | eine                         | gemacht super             | habe                          | wirklich                 | von kurkuma              | dank alles super        |
| 14 | meine                        | groartig                  | dha                           | echt                     | habe                     | alles super vielen      |

## APPENDIX B. BERTOPIC WHOLE TOPIC TABLE

Table 13: BERTopic whole topic table c

|    | Topic_55                | Topic_56                    | Topic_57         | Topic_58                   | Topic_59                        | Topic_60           |
|----|-------------------------|-----------------------------|------------------|----------------------------|---------------------------------|--------------------|
| 0  | schnelle lieferung      | gelaufen                    | produkt          | gut sehr gut               | magnesium                       | Cerascreen         |
| 1  | danke                   | alles                       | das produkt      | sehr gut sehr              | das magnesium                   | zufrieden          |
| 2  | lieferung               | alles super                 | es               | gut sehr                   | nehme                           | bei Cerascreen     |
| 3  | schnelle                | alles super gelaufen        | habe             | sehr gut                   | haushalt                        | bin                |
| 4  | verpackung              | super gelaufen              | habe das produkt | gut                        | seit                            | sehr zufrieden     |
| 5  | dank                    | gelaufen alles              | mir              | sehr                       | dieses magnesium                | weiter             |
| 6  | vielen                  | super gelaufen alles        | geholten         | gut einfach                | vergangenheit                   | empfehlen          |
| 7  | vielen dank             | alles sehr gut              | dieses produkt   | einfach sehr gut           | mir                             | bin mit Cerascreen |
| 8  | lieferung danke         | alles sehr                  | das              | sehr gut einfach           | dieses                          | bei                |
| 9  | verpackt                | abgelaufen                  | dieses           | einfach sehr               | es                              | Cerascreen ist     |
| 10 | gute verpackung         | lief                        | ich habe         | gut wirklich               | hilft                           | empfohlen          |
| 11 | gut verpackt            | super                       | ich habe das     | gut wirklich sehr          | mir gefaellt                    | sehr               |
| 12 | sehr schnelle lieferung | alles reibungslos           | empfehlung       | gut einfach gut            | magnesium haushalt              | mit Cerascreen     |
| 13 | ware                    | alles gut                   | diesem produkt   | einfach gut                | magnesium nehme                 | mit                |
| 14 | schnell                 | sehr gut alles              | bisher           | einfach                    | gefaellt                        | kann Cerascreen    |
|    | Topic_61                | Topic_62                    | Topic_63         | Topic_64                   | Topic_65                        | Topic_66           |
| 0  | alles alles             | lieferung                   | noch             | wirkung                    | vegan                           | anleitung          |
| 1  | alles                   | einfache                    | test             | die wirkung                | vegan und                       | die anleitung      |
| 2  | alles tiptop            | schnelle                    | kein             | wirkung ist                | produkt                         | meine              |
| 3  | tiptop                  | schnelle lieferung          | mein             | die wirkung ist            | vegetarisch                     | der anleitung      |
| 4  | alles supi              | schneller versand           | ergebnis         | wirkt                      | auch vegan                      | ueber              |
| 5  | supi                    | unkomplizierte              | noch kein        | schon nach                 | vegane                          | ergebnis           |
| 6  | alles tiptop alles      | handhabung                  | warte            | es wirkt                   | vegane                          | mich               |
| 7  | alles supi alles        | schnelle lieferung einfache | mein test        | noch                       | fuer                            | bereits            |
| 8  | tiptop alles            | einfache handhabung         | der test hat     | aber                       | ich jedem                       | durchfuehrung      |
| 9  | supi alles              | lieferung einfache          | nicht            | nach                       | der sich                        | das                |
| 10 | passt alles             | versand                     | habe             | gute wirkung               | sich groestenteils              | das ergebnis       |
| 11 | was es                  | schnelle lieferung und      | kein ergebnis    | wochen                     | ernaehrt                        | mir                |
| 12 | supi alles supi         | lieferung und               | mein test ist    | mal kaufen                 | guter dosierpipette             | zur post gebracht  |
| 13 | tiptop alles tiptop     | schneller                   | test hat         | noch nicht                 | groestenteils vegetarisch oder  | der                |
| 14 | macht was es            | und unkomplizierte          | noch nicht       | kann                       | groestenteils vegetarisch       | war                |
|    | Topic_67                | Topic_68                    | Topic_69         | Topic_70                   | Topic_71                        | Topic_72           |
| 0  | geklappt                | produkt                     | test             | zufrieden sehr zufrieden   | handhabung                      | ergebnis           |
| 1  | empfehlen               | schnelle lieferung          | leider ist       | zufrieden sehr             | einfache                        | tagen              |
| 2  | empfehlen alles         | lieferung                   | zum arzt         | sehr zufrieden sehr        | einfache handhabung             | das ergebnis       |
| 3  | bestens                 | schnelle                    | ist er sehr      | zufrieden                  | schnelle                        | zufrieden          |
| 4  | empfehlenswert alles    | produkte                    | er sehr          | sehr zufrieden             | handhabung sehr                 | nach tagen         |
| 5  | alles                   | gute                        | er sehr teuer    | zufrieden zufrieden        | schnelle lieferung              | abgeschickt        |
| 6  | empfehlenswert          | schnelle lieferung produkt  | leider ist er    | rundum                     | schnelles ergebnis              | war das ergebnis   |
| 7  | zu empfehlen            | produkt schnelle lieferung  | erspart          | sehr                       | verstaendliche                  | bin                |
| 8  | alles bestens           | lieferung produkt           | sehr teuer       | rundum zufrieden           | ergebnisse                      | war                |
| 9  | geklappt kann ich       | produkt schnelle            | zu hause         | zufrieden rundum           | schnelles                       | ergebnis           |
| 10 | geklappt kann           | gutes produkt               | einfacher        | sehr zufrieden zufrieden   | handhabung und                  | war das            |
| 11 | hat alles               | gutes                       | hause            | rundum zufrieden sehr      | lieferung                       | ergebnis war       |
| 12 | empfehlen alles bestens | produkte schnelle lieferung | gute             | zufrieden zufrieden sehr   | ausfuehrlichem ergebnisbericht  | ergebnis da        |
| 13 | geklappt sehr           | produkte schnelle           | ist er           | zufrieden rundum zufrieden | ergebnisbericht handhabung sehr | sehr zufrieden     |
| 14 | sehr empfehlenswert     | gute produkte               | sich gut         | sehr zufrieden rundum      | ergebnisbericht handhabung      | total zufrieden    |
|    | Topic_74                | Topic_75                    | Topic_76         | Topic_77                   | Topic_78                        |                    |
| 0  | danke                   | alles prima alles           | corona           | gelaufen                   | besser                          |                    |
| 1  | zufrieden danke         | prima alles                 | wir              | gut gelaufen               | fuehle                          |                    |
| 2  | sehr zufrieden danke    | prima alles prima           | test             | alles gut                  | nach                            |                    |
| 3  | danke danke             | alles prima                 | antikoerper      | alles gut gelaufen         | viel besser                     |                    |
| 4  | sehr zufrieden          | prima                       | war              | war alles                  | viel                            |                    |
| 5  | zufrieden               | alles                       | der              | gelaufen alles             | nach einem monat                |                    |
| 6  | danke ich bin           | weiter so alles             | einen            | gelaufen alles gut         | winter                          |                    |
| 7  | danke ich               | prima nur etwas             | den              | gut gelaufen alles         | haut                            |                    |
| 8  | bin sehr zufrieden      | prima erledigt alles        | dass             | bestens gelaufen           | einem monat                     |                    |
| 9  | vielen dank             | prima erledigt              | hatten           | alles                      | ich mich                        |                    |
| 10 | bin sehr                | prima nur                   | auch             | gut                        | einnahme                        |                    |
| 11 | vielen                  | nur etwas teuer             | da               | alles bestens gelaufen     | mich                            |                    |
| 12 | vielen dank danke       | erledigt alles prima        | dem              | es war alles               | monat                           |                    |
| 13 | dank danke              | topp alles prima            | coronavirus      | war alles gut              | besser und                      |                    |
| 14 | dank                    | alles prima nur             | im               | es war                     | wie ich mich                    |                    |



## TOPIC TRANSLATIONS ENGLISH

This section provides an English translation for all German worded tables in chronological order. It is important to point out that important nuances that exist within the German topic clusters get lost in translation to English. Hence, sometimes words appear twice in the English translation while in German these are two different words that have similar meanings.

Table 14: Most important words from all ratings

(a) Words of positive ratings

| Words     | TF-IDF |
|-----------|--------|
| test      | 0.2568 |
| good      | 0.2559 |
| delivery  | 0.2052 |
| great     | 0.2049 |
| fast      | 0.1971 |
| fast      | 0.1942 |
| satisfied | 0.1745 |
| easy      | 0.1716 |
| for       | 0.1647 |
| result    | 0.1527 |

(b) Words of negative ratings

| Words         | TF-IDF |
|---------------|--------|
| test          | 0.3138 |
| result        | 0.1832 |
| received      | 0.1658 |
| was           | 0.1644 |
| Cerascreen    | 0.1610 |
| for           | 0.1583 |
| unfortunately | 0.1454 |
| always        | 0.1386 |
| weeks         | 0.1365 |
| delivery      | 0.1304 |

## APPENDIX C. TOPIC TRANSLATIONS ENGLISH

Table 15: Sentence embeddings with Umap and Kmean

|   | Topic_1       | Topic_1_Val | Topic_2      | Topic_2_Val | Topic_3     | Topic_3_Val | Topic_4    | Topic_4_Val | Topic_5    | Topic_5_Val  |
|---|---------------|-------------|--------------|-------------|-------------|-------------|------------|-------------|------------|--------------|
| 0 | test          | 0.071973    | product      | 0.442169    | capsules    | 0.062653    | ok         | 1.584606    | always     | 0.264539     |
| 1 | result        | 0.041491    | worked       | 0.190803    | for         | 0.050702    | swell      | 0.821222    | top        | 0.261874     |
| 2 | delivery      | 0.036684    | great        | 0.166423    | good        | 0.047123    | alright    | 0.444978    | perfect    | 0.180431     |
| 3 | received      | 0.035372    | satisfied    | 0.158469    | great       | 0.038467    | satisfied  | 0.348159    | works      | 0.171884     |
| 4 | was           | 0.035021    | good         | 0.100704    | product     | 0.036371    | great      | 0.346555    | great      | 0.171433     |
| 5 | tests         | 0.034890    | satisfaction | 0.086632    | already     | 0.031645    | tip        | 0.212006    | good       | 0.155524     |
| 6 | Cerascreen    | 0.034651    | worked       | 0.073771    | taste       | 0.031099    | top        | 0.139055    | satisfied  | 0.140500     |
| 7 | days          | 0.031839    | good         | 0.073451    | take        | 0.027436    | best       | 0.061175    | thanks     | 0.114009     |
| 8 | easy          | 0.031259    | whole        | 0.057870    | price       | 0.027344    | top        | 0.057088    | splendid   | 0.112503     |
| 9 | fast          | 0.031107    | thanks       | 0.051270    | tastes good | 0.027145    | overall    | 0.046269    | gladly     | 0.100289     |
|   | Topic_6       | Topic_6_Val | Topic_7      | Topic_7_Val | Topic_8     | Topic_8_Val | Topic_9    | Topic_9_Val | Topic_10   | Topic_10_Val |
| 0 | splendid      | 3.290803    | gladly       | 0.641547    | good        | 2.018121    | test       | 0.063056    | fast       | 0.204118     |
| 1 | okay          | 0.359111    | top          | 0.545808    | great       | 0.316079    | vitamin    | 0.057803    | delivery   | 0.197800     |
| 2 | best          | 0.065005    | always       | 0.398179    | far         | 0.058596    | blood      | 0.036976    | fast       | 0.123564     |
| 3 | all           | 0.040511    | perfect      | 0.258503    | good        | 0.056478    | result     | 0.030919    | evaluation | 0.074528     |
| 4 | everything    | 0.037658    | gladly       | 0.230588    | good        | 0.052056    | for        | 0.030777    | great      | 0.069347     |
| 5 | all           | 0.035989    | splendid     | 0.207332    | really      | 0.048011    | Cerascreen | 0.029523    | service    | 0.066614     |
| 6 | communication | 0.025879    | anytime      | 0.168200    | cheap       | 0.038761    | easy       | 0.024810    | good       | 0.058602     |
| 7 | best          | 0.023872    | great        | 0.157967    | actually    | 0.036943    | doctor     | 0.022281    | goods      | 0.055971     |
| 8 | yes           | 0.020145    | satisfied    | 0.117445    | yes         | 0.029139    | good       | 0.022258    | good       | 0.055482     |
| 9 | top           | 0.015896    | good         | 0.084841    | good        | 0.021922    | fast       | 0.021969    | satisfied  | 0.048466     |

Table 16: Sentence embeddings with Umap and Agglomerative Clustering

|   | Topic_1      | Topic_1_Val | Topic_2    | Topic_2_Val | Topic_3    | Topic_3_Val | Topic_4   | Topic_4_Val | Topic_5    | Topic_5_Val  |
|---|--------------|-------------|------------|-------------|------------|-------------|-----------|-------------|------------|--------------|
| 0 | satisfied    | 0.447168    | test       | 0.069153    | splendid   | 2.270783    | good      | 0.967924    | delivery   | 0.173115     |
| 1 | gladly       | 0.383589    | Cerascreen | 0.040094    | top        | 0.660003    | splendid  | 0.363908    | fast       | 0.169788     |
| 2 | top          | 0.243956    | result     | 0.037671    | perfect    | 0.319762    | perfect   | 0.358622    | fast       | 0.106312     |
| 3 | always       | 0.231571    | was        | 0.033717    | satisfied  | 0.235402    | okay      | 0.346285    | evaluation | 0.072463     |
| 4 | good         | 0.227129    | received   | 0.033008    | best       | 0.044247    | swell     | 0.310364    | good       | 0.058761     |
| 5 | satisfaction | 0.181233    | days       | 0.029620    | impeccable | 0.031268    | thanks    | 0.272264    | great      | 0.058514     |
| 6 | thanks       | 0.111110    | tests      | 0.027524    | perfect    | 0.027574    | many      | 0.187310    | service    | 0.052965     |
| 7 | splendid     | 0.109897    | satisfied  | 0.027163    | all        | 0.027574    | thanks    | 0.178320    | goods      | 0.050425     |
| 8 | gladly       | 0.095430    | easy       | 0.026344    | completely | 0.025633    | satisfied | 0.107192    | good       | 0.047698     |
| 9 | great        | 0.085994    | good       | 0.024834    | everything | 0.025633    | perfect   | 0.082138    | easy       | 0.045975     |
|   | Topic_6      | Topic_6_Val | Topic_7    | Topic_7_Val | Topic_8    | Topic_8_Val | Topic_9   | Topic_9_Val | Topic_10   | Topic_10_Val |
| 0 | product      | 0.080234    | great      | 1.113554    | always     | 0.400451    | ok        | 3.634418    | vitamin    | 0.063112     |
| 1 | capsules     | 0.064158    | worked     | 0.583750    | worked     | 0.319185    | great     | 0.487167    | test       | 0.058182     |
| 2 | for          | 0.052747    | good       | 0.476400    | great      | 0.263340    | ok        | 0.083953    | blood      | 0.040359     |
| 3 | good         | 0.047996    | worked     | 0.383112    | alright    | 0.261281    | great     | 0.075547    | for        | 0.031972     |
| 4 | great        | 0.046142    | splendid   | 0.157617    | swell      | 0.260012    | would     | 0.000000    | result     | 0.030775     |
| 5 | taste        | 0.032004    | worked     | 0.132883    | splendid   | 0.156093    | gamma     | 0.000000    | Cerascreen | 0.025364     |
| 6 | really       | 0.030151    | ok         | 0.052755    | thanks     | 0.147207    | game      | 0.000000    | easy       | 0.023562     |
| 7 | tastes good  | 0.028956    | arrived    | 0.042575    | top        | 0.135016    | galt      | 0.000000    | good       | 0.023135     |
| 8 | easy         | 0.028425    | great      | 0.041839    | thanks     | 0.124169    | gasses    | 0.000000    | already    | 0.022845     |
| 9 | good         | 0.028108    | worked     | 0.038453    | many       | 0.117386    | common    | 0.000000    | fast       | 0.021873     |

Table 17: Sentence embeddings with Umap and HDBscan

| Topic_1 | Topic_1_Val      | Topic_2  | Topic_2_Val   | Topic_3  | Topic_3_Val       | Topic_4  | Topic_4_Val         | Topic_5  | Topic_5_Val  |          |
|---------|------------------|----------|---------------|----------|-------------------|----------|---------------------|----------|--------------|----------|
| 0       | perfect          | 1.991076 | fast          | 1.180348 | tip               | 2.038517 | testosterone        | 0.243989 | great        | 1.185673 |
| 1       | perfect          | 0.481369 | delivery      | 1.067121 | top               | 1.337068 | training            | 0.099037 | product      | 0.865443 |
| 2       | allergy sufferer | 0.228932 | delivery      | 0.103622 | alright           | 0.388966 | testosterone levels | 0.097301 | great goods  | 0.298305 |
| 3       |                  | thanks   | 0.228932      | delivery | 0.103622          | reliable | 0.316716            | test     | 0.097047     | mygreenz |
| 4       | threatens        | 0.228932 | fast          | 0.097002 | tipp              | 0.294109 | knowledge           | 0.083106 | products     | 0.296685 |
| 5       | crisis           | 0.203382 | shipping      | 0.096669 | best              | 0.294109 | successes           | 0.079179 | assistance   | 0.244009 |
| 6       | ideal            | 0.196692 | greatly       | 0.089029 | top               | 0.274462 | tests               | 0.073673 | energy       | 0.209079 |
| 7       | approach         | 0.196692 | slow          | 0.089029 | tiptop            | 0.274462 | already             | 0.070039 | article      | 0.188883 |
| 8       | capsules         | 0.186613 | quick         | 0.075116 | shop              | 0.237711 | forwarded           | 0.065755 | more         | 0.119770 |
| 9       | cook             | 0.175383 | delivery time | 0.061632 | worked            | 0.155343 | perform             | 0.065755 | for          | 0.088979 |
| Topic_6 | Topic_6_Val      | Topic_7  | Topic_7_Val   | Topic_8  | Topic_8_Val       | Topic_9  | Topic_9_Val         | Topic_10 | Topic_10_Val |          |
| 0       | alright          | 1.444733 | always        | 1.243886 | always            | 1.183874 | perfect             | 1.729622 | satisfied    | 1.200549 |
| 1       | always           | 1.285349 | gladly        | 0.876818 | satisfied         | 0.947802 | worked              | 1.407798 | always       | 0.964012 |
| 2       | reliable         | 0.306670 | gladly        | 0.475227 | fullest           | 0.232462 | procedure           | 0.364571 | yearly       | 0.328135 |
| 3       | pills            | 0.281259 | satisfied     | 0.193637 | satisfaction      | 0.196335 | perfectly           | 0.298305 | examine      | 0.274487 |
| 4       | commendable      | 0.215471 | astaxanthin   | 0.141056 | regards           | 0.160542 | perfect             | 0.235291 | visualized   | 0.245305 |
| 5       | handling         | 0.210261 | journey       | 0.136416 | customer service  | 0.148382 | smoothly            | 0.235291 | let          | 0.240167 |
| 6       | iron             | 0.181359 | vitamaze      | 0.132816 | overly            | 0.148382 | explanation         | 0.172278 | previously   | 0.240167 |
| 7       | precise          | 0.137908 | once          | 0.121056 | great             | 0.144467 | worked              | 0.171808 | expired      | 0.229987 |
| 8       | uncomplicated    | 0.136076 | already       | 0.112683 | friendly          | 0.132307 | recommendable       | 0.153739 | until now    | 0.169112 |
| 9       | good             | 0.135548 | all around    | 0.099685 | customer friendly | 0.130634 | results             | 0.119190 | levels       | 0.168001 |

Table 18: BERTopic most meaningful clusters

| Topic_5                  | Topic_6             | Topic_7             | Topic_8                     | Topic_9             | Topic_10                   |
|--------------------------|---------------------|---------------------|-----------------------------|---------------------|----------------------------|
| 0 taste                  | capsules            | fast                | service                     | book                | again                      |
| 1 tastes good            | the capsules        | delivery            | customer service            | the book            | gladly                     |
| 2 delicious              | capsules are        | the                 | good                        | recommend           | again everything           |
| 3 the taste              | the capsules are    | fast delivery       | good service                | can                 | gladly again               |
| 4 good                   | are                 | tests               | great                       | to me               | always again               |
| 5 tastes                 | swallow             | evaluation          | delivery                    | recommendable       | always                     |
| 6 taste is               | to swallow          | the tests           | fast delivery               | further             | again gladly               |
| 7 delicious and          | the                 | test                | fast                        | it                  | always again gladly        |
| 8 the taste is           | hair                | delivery and        | fast                        | can I               | gladly again everything    |
| 9 is                     | my                  | handling            | great service               | have                | everything                 |
| 10 snack                 | my hair             | fast                | very good                   | me                  | gladly                     |
| 11 for                   | take                | processing          | fast service                | take                | everything splendid gladly |
| 12 very delicious        | easy to swallow     | fast processing     | great service               | recommendation      | splendid gladly again      |
| 13 great                 | easy                | easy                | customer service very       | recommend           | splendid gladly            |
| 14 the                   | good                | easy                | good customer service       | book is             | again always               |
| Topic_13                 | Topic_15            | Topic_19            | Topic_25                    | Topic_28            | Topic_31                   |
| 0 long                   | pillow              | groceries           | price                       | quality             | Cerascreen                 |
| 1 took long              | powder              | have                | expensive                   | good                | test                       |
| 2 delivery               | sleep               | test                | the price                   | reliable            | have                       |
| 3 long took long         | to me               | to me               | price is                    | reliable very       | by Cerascreen              |
| 4 delivery time          | the powder          | that                | good price                  | very                | the                        |
| 5 slightly               | the pillow          | my                  | but                         | very reliable       | already                    |
| 6 delivery did           | have                | allergies           | the price is                | quality is          | the test                   |
| 7 did                    | sleep               | to                  | a a little                  | very good           | also                       |
| 8 to long                | more                | of                  | a little                    | offer               | by                         |
| 9 goods                  | it                  | now                 | quality                     | very good           | tests                      |
| 10 the delivery          | sleep               | not                 | good                        | good                | at                         |
| 11 order                 | can                 | one                 | the price                   | very reliable very  | once                       |
| 12 very long             | the                 | the                 | is                          | good offer          | can                        |
| 13 until                 | of                  | me                  | expensive but               | the quality         | at Cerascreen              |
| 14 did slightly          | once                | allergy             | find                        | pure high end       | a                          |
| Topic_39                 | Topic_51            | Topic_53            | Topic_57                    | Topic_59            | Topic_64                   |
| 0 protein                | omega               | pills               | product                     | magnesium           | effect                     |
| 1 protein powder         | fish                | the pills           | the product                 | the magnesium       | the effect                 |
| 2 proteins               | capsules            | are                 | it                          | take                | effect is                  |
| 3 tastes good            | omega capsules      | drops               | have                        | household           | the effect is              |
| 4 it                     | omega is            | curcuma             | have the product            | since               | works                      |
| 5 to                     | the omega           | the                 | to me                       | this magnesium      | already after              |
| 6 of                     | me                  | recommended         | helped                      | past                | it works                   |
| 7 me                     | product             | effect              | this product                | to me               | still                      |
| 8 not                    | krill               | me                  | the                         | this                | but                        |
| 9 bar                    | there               | ingredients         | this                        | it                  | after                      |
| 10 delicious             | is                  | they let            | I have                      | helps               | good effect                |
| 11 at                    | epa and             | to dose             | I have the                  | to me like          | weeks                      |
| 12 only                  | epa                 | hither              | recommendation              | magnesium household | can be bought              |
| 13 protein bar           | have                | by curcuma          | that product                | magnesium take      | not yet                    |
| 14 bars                  | dha                 | have                | until now                   | like                | can                        |
| Topic_65                 | Topic_66            | Topic_69            | Topic_71                    | Topic_73            | Topic_78                   |
| 0 vegan                  | instructions        | test                | handling                    | sport               | better                     |
| 1 vegan and              | the instructions    | unfortunately is    | easy                        | training            | feel                       |
| 2 product                | my                  | to doctor           | easy handling               | athlete             | after                      |
| 3 vegetarian             | the instructions    | is it very          | fast                        | after the sport     | a lot better               |
| 4 also vegan             | over                | it very             | handling very               | the sport           | a lot                      |
| 5 vegan                  | result              | it very expensive   | fast delivery               | the                 | after one month            |
| 6 vegan                  | me                  | unfortunately is it | fasts result                | after the           | winter                     |
| 7 for                    | already             | spares              | understandable              | me                  | skin                       |
| 8 I each                 | implementation      | very expensive      | results                     | after               | one month                  |
| 9 it does                | the                 | to home             | fasts                       | a lot               | I me                       |
| 10 it mostly             | the result          | easier              | handling and                | more                | intake                     |
| 11 eats                  | to me               | home                | delivery                    | that                | me                         |
| 12 good dropping pipette | to the mail brought | good                | extensive report            | always after        | month                      |
| 13 mostly vegetarian or  | the                 | is it               | result report handling very | already             | better and                 |
| 14 mostly vegetarian     | was                 | itself good         | result report handling      | as athlete          | like I me                  |

## APPENDIX C. TOPIC TRANSLATIONS ENGLISH

Table 19: BERTopic whole topic table a

| Topic_1            | Topic_2                 | Topic_3                  | Topic_4                         | Topic_5                  | Topic_6                        |
|--------------------|-------------------------|--------------------------|---------------------------------|--------------------------|--------------------------------|
| 0 very             | vitamin                 | blood                    | goods                           | taste                    | capsules                       |
| 1 easy             | test                    | the                      | order                           | tastes good              | the capsules                   |
| 2 the              | vitamin test            | not                      | not                             | delicious                | capsules are                   |
| 3 test             | the                     | was                      | received                        | the taste                | the capsules are               |
| 4 the              | by                      | the                      | still                           | good                     | are                            |
| 5 have             | have                    | finger                   | none                            | tastes                   | swallow                        |
| 6 is               | to                      | test                     | delivery                        | taste is                 | to swallow                     |
| 7 at               | with                    | the                      | after                           | delicious and            | the                            |
| 8 with             | my                      | is                       | was                             | the taste is             | hair                           |
| 9 to               | is                      | man                      | the goods                       | is                       | my hair                        |
| 10 was             | the                     | me                       | of                              | snack                    | take                           |
| 11 also            | also                    | to                       | the delivery                    | for                      | easy to swallow                |
| 12 not             | levels                  | it                       | weeks                           | very delicious           | easy                           |
| 13 the             | the                     | with                     | unfortunately                   | great                    | easy                           |
| 14 of              | value                   | blood sample             | answer                          | the                      | good                           |
| Topic_7            | Topic_8                 | Topic_9                  | Topic_10                        | Topic_11                 | Topic_12                       |
| 0 fast             | service                 | book                     | again                           | test                     | worked                         |
| 1 delivery         | customer service        | the book                 | gladly                          | the test                 | did everything                 |
| 2 the              | good                    | recommend                | again everything                | fast                     | worked did                     |
| 3 fast delivery    | good service            | can                      | gladly again                    | was                      | worked everything              |
| 4 tests            | great                   | to me                    | always again                    | the                      | worked did everything          |
| 5 evaluation       | delivery                | recommendable            | always                          | easy                     | everything                     |
| 6 the tests        | fast delivery           | further                  | again gladly                    | tests                    | good worked                    |
| 7 test             | fast                    | it                       | always again gladly             | man                      | did                            |
| 8 delivery and     | fast                    | can I                    | gladly again everything         | test was                 | great worked                   |
| 9 handling         | great service           | have                     | everything                      | very                     | everything good worked         |
| 10 fast            | very good               | me                       | gladly                          | test is                  | everything great worked        |
| 11 processing      | fast service            | take                     | everything splendid gladly      | result                   | everything good                |
| 12 fast processing | great service           | recommendation           | splendid gladly again           | the test is              | worked                         |
| 13 easy            | customer service very   | recommend                | splendid gladly                 | is                       | everything great               |
| 14 easy            | good customer service   | book is                  | again always                    | are                      | good worked everything         |
| Topic_13           | Topic_14                | Topic_15                 | Topic_16                        | Topic_17                 | Topic_18                       |
| 0 long             | days                    | pillow                   | fast delivery                   | fast                     | splendid everything splendid   |
| 1 took long        | test                    | powder                   | fast                            | everything               | everything splendid everything |
| 2 delivery         | test                    | sleep                    | delivery                        | fast and                 | splendid everything            |
| 3 long took long   | the                     | to me                    | delivery everything             | everything fast          | everything splendid            |
| 4 delivery time    | after                   | the powder               | fast delivery everything        | everything fast and      | splendid                       |
| 5 slightly         | result                  | the pillow               | everything                      | and fast                 | everything                     |
| 6 delivery did     | the test                | have                     | everything splendid             | worked                   | to best everything             |
| 7 did              | not                     | sleep                    | splendid                        | did everything           | everything to best             |
| 8 to long          | at                      | more                     | delivery everything splendid    | everything great         | yes everything splendid        |
| 9 goods            | was                     | it                       | gladly again                    | fast everything          | best everything                |
| 10 the delivery    | in                      | sleep                    | again                           | did                      | best everything splendid       |
| 11 order           | the                     | can                      | gladly                          | everything good          | splendid yes everything        |
| 12 very long       | is                      | the                      | great fast                      | great                    | to best                        |
| 13 until           | the result              | of                       | like always                     | uncomplicated            | splendid everything to         |
| 14 did slightly    | of                      | once                     | goods                           | good                     | splendid yes                   |
| Topic_19           | Topic_20                | Topic_21                 | Topic_22                        | Topic_23                 | Topic_24                       |
| 0 groceries        | fast                    | not                      | great everything                | like always              | satisfied                      |
| 1 have             | delivery and            | it                       | everything great everything     | like                     | am                             |
| 2 test             | fast delivery and       | was                      | great everything great          | like always everything   | very satisfied                 |
| 3 me               | delivery                | received                 | everything great                | always                   | am very                        |
| 4 that             | fast delivery           | i have                   | great                           | always everything        | test                           |
| 5 my               | evaluation              | have                     | everything                      | splendid like            | am very satisfied              |
| 6 allerg           | fast evaluation         | came                     | great everything wonderful      | everything splendid like | i am                           |
| 7 to               | and fast                | order                    | wonderful                       | splendid like always     | i am very                      |
| 8 on               | handling                | an                       | wonderful everything            | everything               | very                           |
| 9 now              | and evaluation          | after                    | everything wonderful            | like always like         | the test                       |
| 10 not             | good                    | on                       | great everything great          | always like              | the                            |
| 11 an              | fast and                | test                     | everything great everything     | like described           | tests                          |
| 12 the             | delivery and evaluation | that                     | wonderful everything great      | splendid                 | very satisfied with            |
| 13 me              | evaluation and          | the                      | everything wonderful everything | everything splendid      | with the test to               |
| 14 allergy         | fast evaluation and     | an                       | great everything                | always like always       |                                |
| Topic_25           | Topic_26                | Topic_27                 | Topic_28                        | Topic_29                 | Topic_30                       |
| 0 price            | fast                    | product                  | quality                         | good good                | satisfaction                   |
| 1 expensive        | fast and                | the product              | good                            | abc                      | to my                          |
| 2 the price        | reliable                | top product              | reliable                        | good                     | everything to                  |
| 3 price is         | and reliable            | the                      | reliable very                   | good good good           | everything to my               |
| 4 good price       | fast and reliable       | for                      | very                            | great                    | to my satisfaction             |
| 5 but              | reliable fast           | this                     | very reliable                   | suits                    | my satisfaction                |
| 6 the price is     | and reliable fast       | product I                | quality is                      | good suits               | satisfied everything           |
| 7 a a little       | reliable fast and       | product did              | very good                       | xxx                      | satisfied                      |
| 8 a little         | like always fast        | this product             | offer                           | name                     | everything                     |
| 9 quality          | like always fast        | products                 | very good                       | good abc                 | everything to the              |
| 10 good            | again fast and          | the product did          | good                            | ok good                  | my                             |
| 11 the price       | again fast              | me                       | very reliable very              | great good               | satisfaction everything        |
| 12 is              | capable                 | product is               | good offer                      | abc good                 | to the                         |
| 13 expensive but   | gladly again fast       | the quality              | pure high end                   | ok                       | worked                         |
| 14 find            | always fast             | the product recommend    |                                 | is ok                    | at                             |
| Topic_31           | Topic_32                | Topic_33                 | Topic_34                        | Topic_35                 | Topic_36                       |
| 0 Cerascreen       | satisfied               | fast                     | good everything good            | satisfied                | top everything                 |
| 1 test             | very satisfied          | fast delivery            | everything good everything      | satisfied I              | everything top everything      |
| 2 have             | at                      | delivery                 | good everything                 | satisfied I at           | everything top                 |
| 3 by Cerascreen    | I at                    | delivery fast            | everything good                 | at                       | top                            |
| 4 the              | at very                 | fast delivery fast       | good                            | very satisfied           | top everything top             |
| 5 already          | at very satisfied       | very fast delivery       | everything                      | I at                     | alright everything             |
| 6 the test         | I at very               | very fast                | good yes everything             | satisfied at             | in alright everything          |
| 7 also             | satisfied with          | fast evaluation          | good yes                        | at very satisfied        | alright everything top         |
| 8 by               | delivery                | delivery fast delivery   | nothing negative to say         | at very                  | everything                     |
| 9 tests            | at                      | good                     | good actually                   | very satisfied I         | alright                        |
| 10 at              | with                    | price                    | to say everything good          | I at very                | in alright                     |
| 11 once            | goods                   | evaluation               | good everything is              | at satisfied I           | everything in                  |
| 12 can             | the delivery            | delivery fast evaluation | actually everything good        | at satisfied             | everything in                  |
| 13 at Cerascreen   | with the                | fast delivery very       | is good everything              | satisfied I was          | top everything in              |
| 14 a               | with the delivery       | delivery very            | yes everything good             | at like                  | top everything tip             |

Table 20: BERTopic whole topic table b

|    | Topic_37                      | Topic_38                       | Topic_39            | Topic_40                       | Topic_41                    | Topic_42                 |
|----|-------------------------------|--------------------------------|---------------------|--------------------------------|-----------------------------|--------------------------|
| 0  | ok everything ok              | uncomplicated                  | protein             | fast                           | top top                     | product                  |
| 1  | everything ok everything      | and uncomplicated              | protein powder      | result                         | top top top                 | good                     |
| 2  | ok everything                 | uncomplicated fast             | proteins            | fasts                          | top                         | good product             |
| 3  | everything ok                 | easy                           | tastes good         | fasts result                   | tip top                     | very good                |
| 4  | ok                            | uncomplicated fast and         | it                  | results                        | tip                         | very good product        |
| 5  | everything                    | fast and                       | to                  | fast delivery                  | top tip top                 | products                 |
| 6  | everything ok great           | fast and uncomplicated         | of                  | delivery                       | top tip                     | product very             |
| 7  | everything ok everything      | fast                           | me                  | fast results                   | tip top top                 | product very good        |
| 8  | everything ok                 | fast uncomplicated             | not                 | the result                     | tip top tip                 | the product              |
| 9  | ok everything ok              | easy                           | bar                 | great                          | top top tip                 | product is               |
| 10 | ok everything ok              | easy handling                  | delicious           | result great                   | tiptop top top              | product and              |
| 11 | ok great everything           | easy and                       | at                  | and evaluation the             | the thyroid function        | the product is           |
| 12 | ok great                      | easy great                     | only                | in short fast shipping         | top procedure top           | high end product         |
| 13 | great everything ok           | great                          | protein bar         | in short fast                  | top procedure               | very                     |
| 14 | great everything              | and uncomplicated fast         | bars                | great fasts                    | tolerable sues              | high end                 |
|    | Topic_43                      | Topic_44                       | Topic_45            | Topic_46                       | Topic_47                    | Topic_48                 |
| 0  | delivery fast delivery        | splendid everything splendid   | perfect             | the product is                 | test                        | comment                  |
| 1  | delivery fast                 | everything splendid everything | ideal               | product is                     | that                        | no comment               |
| 2  | fast delivery                 | splendid everything            | perfect perfect     | the product                    | this                        | no                       |
| 3  | fast delivery fast            | everything splendid            | perfect for         | product                        | have                        | nothing to               |
| 4  | delivery                      | splendid                       | for me              | product is real                | my                          | nothing                  |
| 5  | fast                          | everything                     | for                 | is real                        | this test                   | none                     |
| 6  | delivery prompt               | best everything                | precise the         | product is really              | to me                       | not cheap                |
| 7  | prompt                        | everything best everything     | ideal for           | is                             | have I                      | impeccable               |
| 8  | prompt delivery               | everything best                | perfect like always | the                            | not                         | complaints               |
| 9  | delivery fast shipping        | best                           | perfect like        | real                           | find                        | better works             |
| 10 | delivery prompt delivery      | over everything                | perfect for me      | is really                      | the test                    | better                   |
| 11 | delivered                     | best everything splendid       | me                  | product is whole               | already                     | cheap                    |
| 12 | delivered fast                | best everything best           | that which my       | is whole                       | my husband bought           | to complain              |
| 13 | delivered fast delivery       | around everything              | my body needs       | really                         | my                          | none complaints          |
| 14 | received fast                 | over everything splendid       | body needs          | great                          | me                          | complain                 |
|    | Topic_49                      | Topic_50                       | Topic_51            | Topic_52                       | Topic_53                    | Topic_54                 |
| 0  | still                         | great great                    | omega               | the product is                 | pills                       | thanks everything        |
| 1  | rating                        | great                          | fish                | product is                     | the pills                   | many thanks everything   |
| 2  | none                          | great great great              | capsules            | the product                    | are                         | many thanks              |
| 3  | no                            | easy great                     | omega capsules      | product                        | drops                       | thanks everything        |
| 4  | result                        | easy great great               | omega is            | product is really              | curcuma                     | thanks                   |
| 5  | no result                     | excellent                      | the omega           | is really                      | the                         | many                     |
| 6  | test result                   | great wonderful great          | me                  | product is whole               | recommended                 | thanks                   |
| 7  | of                            | wonderful great                | product             | is whole                       | effect                      | everything splendid many |
| 8  | still not                     | wonderful great great          | krill               | product is real                | me                          | splendid many thanks     |
| 9  | wait                          | great great wonderful          | there               | is real                        | ingredients                 | splendid many            |
| 10 | no test result                | great wonderful                | is                  | is                             | lets itself                 | everything               |
| 11 | until                         | great easy                     | epa and             | the                            | to dose                     | great many               |
| 12 | not                           | great easy great               | epa                 | great                          | hither                      | great many thanks        |
| 13 | one                           | done great                     | have                | really                         | by curcuma                  | thanks everything great  |
| 14 | my                            | splendid                       | dha                 | real                           | have                        | everything great many    |
|    | Topic_55                      | Topic_56                       | Topic_57            | Topic_58                       | Topic_59                    | Topic_60                 |
| 0  | fast delivery                 | worked                         | product             | good very good                 | magnesium                   | Cerascreen               |
| 1  | thanks                        | everything                     | the product         | very good very                 | the magnesium               | satisfied                |
| 2  | delivery                      | everything great               | it                  | good very                      | take                        | at Cerascreen            |
| 3  | fast                          | everything great worked        | have                | very good                      | household                   | at                       |
| 4  | packaging                     | great worked                   | have the product    | good                           | since                       | very satisfied           |
| 5  | thanks                        | worked everything              | to me               | very                           | this magnesium              | further                  |
| 6  | many                          | great worked everything        | helped              | good easy                      | past                        | recommend                |
| 7  | many thanks                   | everything very good           | this product        | easy very good                 | to me                       | at with Cerascreen       |
| 8  | delivery thanks               | everything very                | the                 | very good easy                 | this                        | at                       |
| 9  | packaged                      | expired                        | this                | easy very                      | it                          | Cerascreen is            |
| 10 | good packaging                | worked                         | I have              | good really                    | helps                       | recommended              |
| 11 | good packaged                 | great                          | I have the          | good really very               | to me like                  | very                     |
| 12 | very fast delivery            | everything smoothly            | recommendation      | good easy good                 | magnesium household         | with Cerascreen          |
| 13 | goods                         | everything good                | this product        | easy good                      | magnesium take              | with                     |
| 14 | fast                          | very good everything           | until now           | easy                           | like                        | can Cerascreen           |
|    | Topic_61                      | Topic_62                       | Topic_63            | Topic_64                       | Topic_65                    | Topic_66                 |
| 0  | everything everything         | delivery                       | still               | effect                         | vegan                       | instructions             |
| 1  | everything                    | easy                           | test                | the effect                     | vegan and                   | the instructions         |
| 2  | everything tiptop             | fast                           | no                  | effect is                      | product                     | my                       |
| 3  | tiptop                        | fast delivery                  | my                  | the effect is                  | vegetarian                  | the instructions         |
| 4  | everything great              | fast shipping                  | result              | works                          | also vegan                  | over                     |
| 5  | great                         | uncomplicated                  | still no            | already after                  | vegan                       | result                   |
| 6  | everything tiptop everything  | handling                       | wait                | it works                       | vegan                       | me                       |
| 7  | everything great everything   | fast delivery easy             | my test             | still                          | for                         | already                  |
| 8  | tiptop everything             | easy handling                  | the test did        | but                            | I each                      | implementation           |
| 9  | great everything              | delivery easy                  | not                 | after                          | the itself                  | the                      |
| 10 | suits everything              | shipping                       | have                | good effect                    | itself mostly               | the result               |
| 11 | was it                        | fast delivery and              | no result           | weeks                          | eats                        | to me                    |
| 12 | great everything great        | delivery and                   | my test is          | once buy                       | good dropping pipette       | to the mail brought      |
| 13 | tiptop everything tiptop      | fast                           | test did            | still not                      | mostly vegetarian or        | the                      |
| 14 | does was it                   | and uncomplicated              | still not           | can                            | mostly vegetarian           | was                      |
|    | Topic_67                      | Topic_68                       | Topic_69            | Topic_70                       | Topic_71                    | Topic_72                 |
| 0  | worked                        | product                        | test                | satisfied very satisfied       | handling                    | result                   |
| 1  | recommend                     | fast delivery                  | unfortunately is    | satisfied very                 | easy                        | days                     |
| 2  | recommend everything          | delivery                       | to doctor           | very satisfied very            | easy handling               | the result               |
| 3  | splendid                      | fast                           | is it very          | satisfied                      | fast                        | satisfied                |
| 4  | recommendable everything      | products                       | it very             | very satisfied                 | handling very               | after days               |
| 5  | everything                    | good                           | it very expensive   | satisfied satisfied            | fast delivery               | sent                     |
| 6  | recommendable                 | fast delivery product          | unfortunately is it | all around                     | fasts result                | was the result           |
| 7  | to recommend                  | product fast delivery          | spares              | very                           | understandable              | at                       |
| 8  | everything splendid           | delivery product               | very expensive      | all around satisfied           | results                     | was                      |
| 9  | worked can I                  | product fast                   | to home             | satisfied all around           | fasts                       | results                  |
| 10 | worked can                    | good product                   | easier              | very satisfied satisfied       | handling and                | was the                  |
| 11 | did everything                | good                           | home                | all around satisfied very      | delivery                    | result was               |
| 12 | recommend everything splendid | products fast delivery         | good                | satisfied satisfied very       | extensive result report     | result there             |
| 13 | worked very                   | products fast                  | is it               | satisfied all around satisfied | result report handling very | very satisfied           |
| 14 | very recommendable            | good products                  | itself good         | very satisfied all around      | result report handling      | total satisfied          |

Table 21: BERTopic whole topic table c

|    | Topic_74              | Topic_75                    | Topic_76    | Topic_77                   | Topic_78        |
|----|-----------------------|-----------------------------|-------------|----------------------------|-----------------|
| 0  | thanks                | everything swell everything | corona      | worked                     | better          |
| 1  | satisfied thanks      | swell everything            | us          | good worked                | feel            |
| 2  | very satisfied thanks | swell everything swell      | test        | everything good            | after           |
| 3  | thanks thanks         | everything swell            | antibody    | everything good worked     | a lot better    |
| 4  | very satisfied        | swell                       | was         | was everything             | a lot           |
| 5  | satisfied             | everything                  | the         | worked everything          | after one month |
| 6  | thanks I at           | further so everything       | a           | worked everything good     | winter          |
| 7  | thanks I              | swell only slightly         | the         | good worked everything     | skin            |
| 8  | at very satisfied     | swell settled everything    | that        | splendid worked            | one month       |
| 9  | many thanks           | swell settled               | had         | everything                 | I me            |
| 10 | at very               | swell only                  | also        | good                       | intake          |
| 11 | many                  | only slightly expensive     | there       | everything splendid worked | me              |
| 12 | many thanks thanks    | settled everything swell    | the         | it was everything          | month           |
| 13 | thanks thanks         | top everything swell        | coronavirus | was everything good        | better and      |
| 14 | thanks                | everything swell only       | in          | it was                     | like I me       |



