Topic modeling for unstructured text survey responses

Exploring methods for processing survey responses from the financial industry

Zhang, Lirong

Award date:
2021

EINDHOVEN UNIVERSITY OF TECHNOLOGY
Department of Mathematics and Computer Science

Topic modeling for unstructured text survey responses

Exploring methods for processing survey responses from the financial industry

Lirong Zhang

Supervisors:
dr. N. (Natalia) Sidorova
dr. V. (Vlado) Menkovski
Msc. X. (Xu) Cao

Eindhoven, Oct 2021

Abstract

It is essential for companies in any sector to ensure a high level of customer satisfaction. Customer experience surveys are widely used to gather information from customers on how satisfied they are about the service and product offering from companies and their overall experience in doing business with the company. Customer experience surveys consist of open-ended and close-ended questions. Compared to the categorized answers collected from close-ended questions, unstructured survey responses are often underutilized for analysis. The traditional way of processing unstructured textual data is done manually, which is labor-intensive and resource-demanding. This project used the textual data of a customer experience survey conducted by a financial service company. As such, the main research goal is addressed as: *"An empirical comparison of state-of-the-art topic models in the context of unstructured survey responses."*

An extensive set of evaluation criteria supports the empirical comparison presented in this work. First, the topic quality is evaluated using a specific topic coherence measure, namely the C_v topic coherence score. It was selected based on results from the literature indicating good agreement between the score and human judgment. The topic quality is also evaluated qualitatively in terms of interpretability by analyzing feedback from domain experts. The second criterion relates to model sensitivity concerning typographical errors because unstructured survey responses are prone to such imperfections. To this end, a novel model sensitivity measure is proposed based on determining the consistency between two topic modeling results. The third criterion aims to understand practical aspects associated with applying the topic models in the context of unstructured survey responses.

With the evaluation criteria defined, the primary methodology used in this empirical study consists of three phases. First, we applied several linguistic cleaning and preprocessing steps such as tokenization, lemmatization, and part-of-speech tagging to the data. Second, we implemented three different topic models to the data and considered a fourth model available in a commercial platform. The four models are the text mining model within Qualtrics Customer Experience platform, the LDA (Latent Dirichlet Allocation) model, the NMF(Non-negative Matrix Factorization) model, and the SBERT and clustering model. After applying the models to the data, the results were evaluated following the evaluation criteria.

The main conclusions that were observed are as follows. First, the domain experts confirmed that the topic models applied to unstructured survey responses are promising since they can recognize some topics identified. Second, none of the four considered models are superior. Different models show different characteristics. For example, the NMF model tends to find more topics than the SBERT model. Furthermore, the LDA model is more sensitive to typographical errors than the others. Third, some observations suggest that the C_v topic coherence score is inappropriate for informal language datasets.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In this chapter, the context of the thesis is first introduced, then the motivation of conducting the thesis is demonstrated. Then, the research questions based on the motivation are formally stated. After this, the approach of how the thesis is undertaken is illustrated. Lastly, the structure of the thesis is introduced.

The financial services industry is concerned with various economic services that the finance industry offers. Financial service companies need to understand customer needs from assets management to equipment leasing services and continue improving their services in a competitive business sector like this.

To understand customer needs, we first need to collect and identify customer feedback or "hearing the voice of the customer" introduced [1]. In work conducted by Griffin and Hauser [1], identifying customer needs was considered a qualitative research task primarily by conducting group or one-on-one interviews. With the development of technologies, collecting customer feedback expanded into various formats such as online customer surveys, live chat, and online customer reviews. Among all these methods, online customer surveys have been adopted widely since the survey can be designed with both open-end and close-end questions. Open-end survey questions are believed to capture diversity in responses and provide alternative explanations compared to close-end questions [2]–[4]. However, most data collected from the customer surveys are unstructured, meaning that the data is not structured via predefined data models or schema, and they are primarily text [5].

The processing of such unstructured text survey responses is usually time-consuming and resource-demanding with its traditional way of processing them. Such a conventional approach is labor-intensive and time-consuming since analysts must read and process all the responses manually [6], [7]. A widely used type of analysis that resolves this dilemma is topic modeling. Topic modeling is commonly used in natural language processing (NLP) to topic discovery and semantic mining from unordered documents [8].

1.1 Motivation

Topic modeling is one of the potential automated techniques to extract insights from unstructured survey responses. In the scientific community, different types of topic models have been developed. However, the selection, configuration, and application of these state-of-the-art topic models for unstructured survey responses remain a difficult task for the following reasons:

- The difficulty of applying topic models is complicated by the specific textual characteristics of unstructured survey responses. For example, unstructured survey responses may contain informal, incomplete, and short sentences, while other datasets such as news articles are generally longer and written in a more formal manner.

- Determining the quality of the topic modeling results is not straightforward, especially because ground truth labels are absent. In most works conducted relating to topic modeling, evaluation of topic modelings results are evaluated quantitatively and qualitatively [7], [9]–[11]. Nevertheless, the choices of quantitative evaluation metrics to determine the quality of the topic modeling results are different, there is no conclusive evidence suggesting that a particular metric is the best.

The goal of this project is to explore these challenges by presenting a comprehensive empirical comparison of state-of-the-art topic models in the context of unstructured survey responses. Multiple topic models with varying theoretical backgrounds are considered, namely the Latent Dirichlet Allocation Model [8], Non-negative Matrix Factorization model [12] and SBERT clustering model [13]. They are applied to a real dataset from the financial service company.

1.2 Research questions

The main research question is formulated as follows: **What are the differences between state-of-the-art topic models in the context of unstructured survey responses?**

This goal is split into the following sub-questions:

- What are the differences between the topic modeling results obtained with several state-of-the-art topic models in terms of topic quality?

- What is the sensitivity of the topic modeling results with respect to the language correctness of the input data?

- What are the practical aspects of each topic model in the context of unstructured survey responses?

1.3 Approach

We follow the methodology of CRISP-DM (Cross-Industry Stand Process for Data Mining) in this project as shown in figure 1.1 [14]. There are mainly six phases involved in the process and iterations between different phases are also usually involved in the process. Moreover, different subsets of data and models are applied in the modeling phase and their results are evaluated in the evaluation phase. Applying these methodologies, the project is conducted as follows:

Business understanding

In this stage, the research goals are going to be identified as well as the field of content that should be taken into consideration when executing the project are identified.

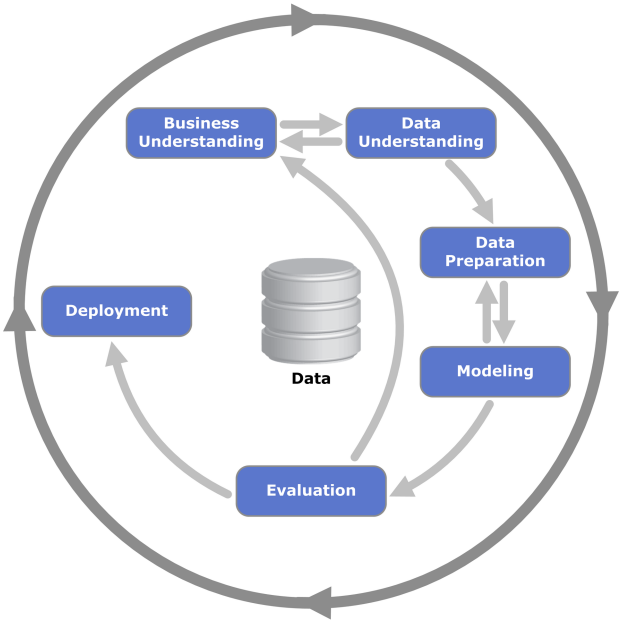


Figure 1.1: The process of CRISP-DM consists of six phases [14]

Data understanding

A real-life dataset containing unstructured survey responses from the financial service company is given for the project. The goal of this phase is to analyze the given dataset to assess its quality. If needed, additional datasets are selected to successfully reach the research goals.

At this phase, we focus on understanding the content of the data, understanding the structure and the format of the data. Moreover, the contents of the data and the quality of the data also present its characteristics and the direction of solving the research questions on the data.

Data preparation

This phase is the follow-up of the data preparation phase. Different pre-processing techniques are applied to the data so that they have a suitable format for the modeling phase. It can be the case that iterations occur to ensure the data is adequate for the model.

1.3.1 Modeling

Based on the defined research questions and business understanding, we first define the evaluation criteria with motivation. Different types of textual datasets are chosen for model results evaluation. We first start by presenting the current tool available for the financial service company and present the corresponding results obtained with it. Three different state-of-the-art topic models conduct topic modeling: a generative probabilistic model, a non-negative matrix factorization model, and an SBERT clustering model.

1.3.2 Evaluation

At the evaluation phase, we evaluate the results obtained from different models and different datasets based on the five evaluation criteria from the context of unstructured survey responses. The evaluations consist of quantitative and qualitative evaluations. Moreover, we also reflect on the work accomplished and see whether all the research goals are achieved.

1.4 Thesis structure

This section briefly presents the structure of the thesis: In Chapter 2, we introduce the preliminary concepts of text mining and common pre-processing techniques, concepts of topic models such as Latent Dirichlet Allocation (LDA), and some related literature studies. In Chapter 3, we describe the project from the business perspective of the financial service company that offers the company data to the project as a case study. We also discuss the current approach of the company and the challenges the company faces. We introduce how the case study is collected and some basic descriptions of the company data. Moreover, additional datasets used in the project are introduced together with the motivations. In Chapter 4, we motivate why specific evaluation criteria need to be defined in the project and specified five evaluation criteria that are used for results evaluation in this project. In Chapter 5, we first present and assess the currently available tool given by the financial service company. Then we present the other three modeling experiment processes and implementation details based on the selected topic modeling methods. In Chapter 6, we present the topic modeling results and evaluate them based on the defined evaluation criteria. In Chapter 7, we discuss the results obtained, the limitations of the conducted work, and conclude the thesis by reflecting both on the business needs and the raised research questions. We also address the possibilities for future work.

Chapter 2

Preliminaries and literature study

In this chapter, we first introduce preliminary concepts such as text mining, natural language processing, and topic modeling. Since they intertwine with each other and serve as the theoretical background of this project. Then, common techniques used to represent documentation are also introduced so that we understand how unstructured textual data can be transformed to formats that computers can understand. After this, other techniques such as word embedding and clustering analysis are introduced as they are also adopted throughout the research project. Moreover, we introduce the topic models that we implemented in the project.

2.1 Text mining

The concept of text mining (also referred to as knowledge discovery from text) means dealing with the machine-supported analysis of the text [15], [16]. Being a variation on data mining, text mining uses techniques from information retrieval, information extraction, natural language processing(NLP) and is further connected to algorithms and methods of knowledge discovery [16], [17]. Together, these techniques are often applied to business and are believed to bring commercial potential values [18].

It has been estimated that 85% of business information lives in the form of text. which contain hidden knowledge for businesses to leverage for a competitive edge. For example, information such as interesting patterns, identify areas of improvement can be used to win competitive edges for companies [16], [19], [20].

2.2 Natural language processing

Natural Language Processing (NLP) is the computerized approach to understand and analyze natural language text or speech that is based on both theories and technologies [21]. The original concept of NLP can be traced back as early as the 1950s when the Georgetown experiment conducted research in automatic translation from Russian into English [22], [23]. However, the advances of NLP in several decades after this experiment were not quite successful due to the variability, ambiguity, and context-dependent interpretation of human languages [22]. This situation got largely improved from the 1980s. When the computational power became more available, then a large number of empirical language data got used for building models and the use of the

statistical approach. All of this is attributed to the advance of NLP [22]. In recent years, NLP has grown rapidly due to increased computational power, advancement for intertwining fields such as deep learning and etc.

The Natural Language Processing field has two main components: language processing and language generation [21]. This project focuses on language processing since it is meant to perform language analysis and obtain meaningful representations from it. It is widely accepted by academia to define the capabilities of an NLP system utilizing language into the following levels [21], [24], [25]. A more capable NLP system should be able to utilize more levels of languages:

Morphology This language level studies the structure of the words. So we know the components of words based on the morphemes (their smallest units of meaning). For example, when the suffix *-ed* is added to a verb(e.g, walk), then this verb describes an action that took place in the past(e.g., walked)

Lexical This language level studies/interprets the meaning of a single word, a part of a word, or a chain of words. For example,

Syntactic This language level studies analyzing the words in a sentence based on the grammatical structure of the language. It can be used to demonstrate the dependency relationships between the words of a parser. Different words have served as different syntax in such a structure in a sentence. For example, two sentences: "Amy called Bram." and "Bram called Amy." are different only in terms of syntax while they convey different meanings.

Semantic This language level studies the meaning in the language, all the levels mentioned above are contributing to how the language is being interpreted.

Discourse This language level studies the meaning of the language by making connections between component sentences. It is different than syntactic and semantic since these two levels look at the language from the sentence perspective. The two most common discourse processing methods are anaphora resolution and discourse/text structure recognition. Anaphora resolution can replace semantically void words like pronouns with appropriate entities that can be referred to.

Pragmatic This level studies the targeted use of languages by considering their context. It often requires incorporating general world knowledge more than the text itself, such as understanding the intended use of the sentence. For example, "I like running in summer as it is healthy for me." and "I like running in summer as it is good weather." are two sentences that need a resolution of the anaphoric term 'it', but the resolution needs pragmatic or general world knowledge.

2.2.1 Text preprocessing techniques

In order to conduct complex NLP tasks such as text clustering. The raw natural language usually needs to be cleaned or prepared since the raw text may affect further the execution and the results of the task [26]. In this subsection, we will introduce some common text preprocessing techniques.

1. Contraction expansion

Input:*text response*
Output:*text response without contractions*

This step is meant to deal with the contractions in the textual data. Contraction is a common habit of many people writing nowadays. It means that we write words or combinations of words that are shortened by dropping letters and replacing them with an apostrophe. The contractions should be processed mainly because computers would not understand that contractions are shortened words. It is easy for humans to understand *we're* and *we are* means the same while computers fail to do so.

This task is performed by employing the python package *"contractions"*. The English contractions and slang can be resolved with it, such a step is presented below. For ambiguous cases, the contractions are resolved to their most common cases. For example, *"he's"* is resolved to the most common case *"he is"*, instead of *"he has"*.

Example:
Input: 'Overall we've made progress.'
Output: 'Overall we have made progress.'

Example:
Input: '['overall','we','have', 'made','progress','.']'
Output: "['overall','we','have', 'made','progress']"

2. Lowercasing

Input:*text response without contractions*
Output:*lower-cased text response without contractions*

This step is used to lowercase all the letters. This step is needed to avoid confusion for computers, as "Canada" and "CANADA" have the same meaning, while computers think they are different. By lowercasing, the same words with different cases are mapped to the same lowercase form. Note that not all textual processing data need this step. For example, 'Apple' representing the company and 'apple' representing the fruit cannot be differentiated if they are both lowercase. However, this situation in which differentiating is needed is not an essential part of our case study. Therefore, we consider this cleaning step a rational choice here.

Example:
Input: 'Overall we have made progress.'
Output: 'overall we have made progress'

3. Tokenization

Input:*lower-cased text response without contractions*
Output:*collection of tokenized lower-cased words without contractions*

This step is used to divide the text responses into a collection of tokens/words. Word tokenization is needed as by combining different sequences of tokens, the context meaning can be better understood by computers [16]. This is executed by employing the *word_tokenize(text)* from the NLTK(Natural Language Toolkit) python package.

Example:
Input: 'overall we have made progress'
Output: '['overall','we','have', 'made','progress','.']'

4. Punctuation removal

Input:*collection of tokenized lower-cased words without contractions*
Output:*collection of tokenized lower-cased words without contractions and punctuation's*

This step is meant to remove punctuation and special symbols of from the data, such as commas, questions marks. This step is needed since usually as the punctuation adds up noise that brings ambiguity while training the model [27].

5. Stop words removal

Input:*collection of tokenized lower-cased words without contractions and punctuation's*
Output:*collection of tokenized lower-cased words without contractions, punctuation's and stop-words*
This step is used to remove words that are commonly used but carry Little useful information. Such words can be 'a', 'the', 'is' and etc. Removal of stop words can be done by utilizing the default *'STOP_WORDS'* list from spaCy. Notably, the list of stop words may need to be extended or exempted depends on the data. For example, 'not' and 'no' should not be treated as stop words for our case study data as they may convey sentiment relating to certain objects.

Example:
Input: '['overall','we','have', 'made', 'progress','.']'
Output: '['overall', 'made', 'progress']'

6. Lemmatization with POS tag

Input:*collection of tokenized lower-cased words without contractions, punctuation's and stop words, en_core_web_sm*
Output:*collection of filtered lemmas*

This step is meant to identify the part-of-speech(POS) tagging for words and find the lemma of certain words based on their POS tag [16]. The POS tagging is a way to categorize words of the sentence into nouns, verbs, etc. We use the trained pipeline *en_core_web_sm* from Spacy and its built-in function *token. tag_* to predicts the most likely applicable tag for each word in the context. Figure 2.1 shows the POS tag of each word from some text responses chosen from the case study data. Usually, nouns are the words that convey useful information while adjectives and adverbs are words that indicate the sentiment orientation. Therefore, we keep the words that are nouns, adverbs or adjectives for further usages.

Example:
Input:'['overall', 'made', 'progress']'
Output: '['progress']'

The last two processing steps are not needed for all the models employed in this project. For certain models utilized pre-trained sentence transformers, retaining stop words and skip lemmatization would not largely impact the performance of the model. After prepare and clean the raw text, we will now discuss some common methods used in transforming the natural language into the formats that machines can utilize.

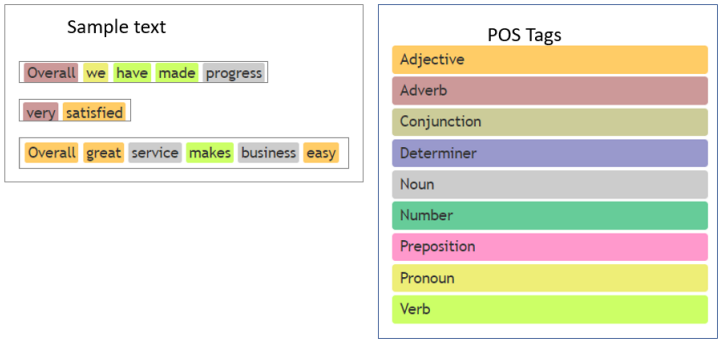


Figure 2.1: An example of visualizing POS tagging NLP task which highlights the POS tag of each word

2.3 Document representation

How to transform the natural language into formats that machines can also understand is the fundamental part of natural language processing. In this project, we adopted three state-of-the-art document representation approaches to transform the textual document into formats machines can utilize. We now introduce them in this section.

2.3.1 Bag-of-words

The bag-of-words (BOW) is a document representation method that represents a sentence or a document as a bag containing sets of words [28]. The grammar and ordering are disregarded while the frequency is preserved. The BOW converts the document to $n \times 1$ vector where n represents the number of tokens/words of the input documents. We now demonstrate the BOW method with an example of the documents: d1: "I am satisfied with the services: d2: "I am unsatisfied with the services". The tokens/words contained in the two documents are:

['I', 'am', 'satisfied', 'unsatisfied', 'with', 'the', 'services']

Therefore, the corresponding BOW representation for these two documents are: d1: [1, 1, 1, 0, 1, 1, 1] and d2: [1, 1, 0, 1, 1, 1, 1]. Since the BOW method preserves the word/token frequency of the document, so it is likely that similar documents may have similar token/words frequencies. However, all the words are given the same weights (importance) with the BOW method which can be too loose in processing the documents. For example, words like 'satisfied' and 'unsatisfied' in documents d1 and d2 conveys more information than words 'am' and 'with'. Thus, other document representation approaches with better weighting strategies such as Term Frequency - Inverse Document Frequency is also often used in represent documents.

2.3.2 Term Frequency-Inverse Document Frequency

The Term Frequency- Inverse Document Frequency (TF-IDF) is a document representation method that represents the document according to the contained words while the weights given are based on the TF-IDF relevance [28]. This metric essentially shows how relevant a word is to a document in a collection of documents. The TF-IDF method consists of two systematical measures: Term Frequency and Inverse Document Frequency.

The Term Frequency TF $tf(t, d)$ is calculates the frequency of each word/term t per document d. It can be formulated as in Equation 2.1.

$$tf(t, d) = \frac{\text{count of } t \text{ in } d}{\text{number of words in } d} \tag{2.1}$$

The Inverse Document Frequency IDF $idf(t)$ measures the informativeness of a term/word t so that the importance of some common words such as 'and' that may be used in all documents of the whole document set can be discounted. In another word, words/terms that appear frequently in the whole document set but are less informative would have low IDF scores. IDF can be formally formulated as in Equation 2.2.

$$idf(t, d, D) = log\frac{|D|}{|d \in D : t \in d| + 1} \tag{2.2}$$

where t is the target term/word, D is the collection of all documents.

The weight of each term in a document can be obtained by multiplying the corresponding TF and IDF obtained, as shown in Equation 2.3.

$$tfidf(t, d, D) = tf(t, d) \times idf(t, d, D) \tag{2.3}$$

A higher term/word tfidf score means a better representation of the term/word is to the document.

The aforementioned two methods of representing documents are mainly based on the word count or its relating ratio while the semantics of the word or the context of the word is not taken into consideration when the documents are represented in machine-understandable formats. In the coming subsection, we will introduce document representation where documents are represented by vectors such as *Word2vec* and *BERT*(Bidirectional Encoder Representations from Transformers).

2.3.3 Embedding

Word Embedding has been proven to be beneficial in various NLP tasks such as sentiment analysis(Glorot, Bordes and Bengio [29]) and parsing(Socher, Lin, Ng *et al.* [30]). With embedding, the words or sentences are mapped into vectors in the vector space. Empowered by the development of the neural network, conducting words or sentence embedding with a neural network has become a popular choice [31]. In the work of Bengio, Ducharme, Vincent *et al.* [32], they first conducted the word embedding with the neural network approach. This work elicited the feedforward neural network in solving the 'curse of dimension' by learning a distributed representation for word vector using the linguistic context in which the word occurs. Based on this work embedding method, more word embedding methods have been proposed such as *Word2vec* conducted by Mikolov, Chen, Corrado *et al.* [33].

Word2Vec

Word2vec mainly consists of two model architectures for computing continuous word vectors and these vector representations were proven to have good performance in word similarity tasks with very large unstructured text data sets [33]. *Word2vec* is based on the distributional hypothesis that words with similar meanings are likely to occur in similar contexts [34]–[36]. Hence, the distance of the generated word vectors represents the extent of semantic similarity between words. The two proposed model architectures are the continuous bag-of-words model(CBOW) and the continuous skip-gram. We will now introduce them briefly below

The CBOW model is meant to learn the target word x_0 from its contextual/surrounding words C within a sliding window. Its objective is to maximize the $P(x_0|C)$ over the training set. The skip-gram model has the reversed model structure as that of the CBOW model, it is meant to learn the word representations that can be used in predicting the contextual/surrounding words in a sentence or a document. Its objective is to maximize the average log probability $logP(x_1, x_2, ..., x_c|x_t$ for the surrounding/contextual words of the target word x_t within the sliding window with size C.

With these new proposed word embedding model architecture, the embeddings can be learned more efficiently with low comparing to the earliest neural network-based embedding proposed [32]. The distributional semantic representation on the word level of the textual data has been largely improved with the advancement of artificial neural networks. Based on the word embedding, approaches such as averaging the word embeddings of all words in a document(Faruqui, Dodge, Jauhar *et al.* [37] and Kenter and De Rijke [38]) or weight the averaged word embedding (De Boom, Van Canneyt, Demeester *et al.* [39]) by incorporating TF-IDF elicit the word-embedding to achieve the goal of sentence-level or document-level text data representation. However, these approaches are corpus-dependent and do not preserve the word order nor differentiate different meanings of the same words under different contexts. Therefore, approached that alleviate these issues such as *BERT*(Devlin, Chang, Lee *et al.* [40]) and *Sentence-BERT*(Reimers and Gurevych [41] have been proposed. Since we use Sentence-VERT in our project, so we will introduce it in the coming section.

Sentence-BERT

Bidirectional Encoder Representations from Transformers (BERT) utilizes bidirectional pre-training to represent language with masked language models [40]. It also makes use of pre-training has been proven to be effective in improving many NLP tasks [42]–[44] such as sentence-level language inference [45]. The bidirectional model structure ensures that the contextual representation of the before and after the context of each token is learned by BERT. And the Transformer structure proposed by Vaswani, Shazeer, Parmar *et al.* [46] transfers the neural network model parameters learned from supervised tasks with large datasets has also to be effective in natural language inference [47]. However, BERT still has some disadvantages such as the massive computational overhead as it requires two input data for sentence-pair tasks like semantic textual similarity and the construction of BERT makes it unsuitable for unsupervised tasks like clustering [41]. The Sentence-BERT (SBERT) proposed by Reimers and Gurevych [41] are meant to solve these issues.

SBERT modifies the BERT model structure using the siamese network to produce meaningful sentence embeddings where semantically similar sentences are close in vector space [41]. The application of siamese network structure ensures that fixed-size vectors can be derived regardless of the input text data length [41]. This enables SBERT in conducting semantic similarity search and clustering with similarity measures such as cosine similarity extremely efficiently. It was shown in [41], the SBERT outperforms other state-of-the-art methods such as BERT and Universal Sentence Encoder [48].

Summary for embedding

There are mainly the following reasons for selecting SBERT out of in the project:

- SBERT is able to differentiate the different meanings of the same word in a different contexts by generating different vectors. Since in *Word2vec*, each word would just have one vector representation regardless of its senses.

- SBERT is better in capturing both the word semantics and the contextual information from the text data with its bidirectional structure. *Word2vec* generates embeddings on word-level while unable to capture the semantics of words as the word order is not taken into account.

- SBERT is better at language inference. While the *Word2vec* can only generate the embeddings based on the training corpus, the SBERT utilized the transfer learning from large pre-trained models and generate the embeddings for all vocabularies. This also saves time needed for training the model on the corpus.

- SBERT is suitable for semantically similarity search task such as clustering. And this is what we need in this project.

2.4 Topic modeling

Topic models have been widely used in text mining and information retrieval recently. A topic model is a model utilize generative probabilistic process in identify abstract "topics" from documents or textual corpora. The "topics" identified by the topic model are cluster of words that considered as describing a certain topic. The first genuine topic model that exploits the probabilistic is the probabilistic latent semantic analysis (PLSA) introduced by Hofmann [49]. After this. a more complete genuine topic model which extends the PLSA was introduced by Blei, Ng and Jordan [8], namely the Latent Dirichlet Allocation (LDA). With the advancement and emerging of technology, more topic models which may have not considered to be genuine yet still shown good topic modeling results have been introduced, such as the Non-negative Matrix Factorization (NMF) [50]. In this project, we conduct topic modeling mainly based on three topic modeling methods: Latent Dirichlet Allocation (LDA) [8], Non-negative Matrix Factorization (NMF) [50] and SBERT topic modeling [13], [41], [51]. In this section, we will introduce these three methods.

2.4.1 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) Blei, Ng and Jordan [8] is a generative probabilistic model. The term "generative" here indicates that such a model is capable of generating a novel corpus through a probabilistic sampling process. How LDA can be used for topic modeling on a given corpus can best be explained by first explaining how the generative process works and subsequently reverse-engineering the process.

The generative process of LDA is based on the specification of multiple classes of probability distributions, each controlling a different aspect of the generated corpus. The process for generating a corpus of N documents and k topics consists of the following steps:

1. k samples, $\phi_i, i = 1, ..., k$, are drawn from a Dirichlet distribution $\text{Dir}(\beta)$ that generates the categorical distribution $\text{Cat}(\phi_i)$ of words occurring in each topic.

2. N samples, $\theta_i, i = 1, ..., N$, are drawn from a Dirichlet distribution $\text{Dir}(\alpha)$ that generates the categorical distribution $\text{Cat}(\theta_i)$ of topics occurring in each document.

3. N samples are drawn from some pre-specified discrete probability distribution that generates the number of words of each document. A possible choice is the Poisson distribution, although a more suitable choice could be made by incorporating prior knowledge from the application context.

4. To each word position in each document, a topic, and a specific word are assigned by first sampling a topic from the document-topic distribution $\text{Cat}(\theta_i)$ and then a word from the topic-word distribution $\text{Cat}(\phi_j)$ corresponding to that topic.

Based on the generative process detailed above, the problem of finding topics in a given corpus can be seen as finding appropriate hyper-parameters α and β for the Dirichlet distributions. This is called *inference* and several techniques can be applied to it, such as Monte Carlo simulation, variational Bayes and likelihood optimization [52].

2.4.2 Non-negative Matrix Factorization

Non-negative Matrix Factorization (NMF) [53] is a linear-algebraic model that decomposes high-dimensional vectors into lower-dimensional representations. It transforms the processed input data of M documents to a Document-Term matrix with dimension $M \times N$ where N is the number of unique words/tokens occurring in the documents.

The NMF model decomposes the input matrix based on the given number of topics k into two matrices: the component matrix and the feature matrix. The component matrix has dimension $k \times N$ so that each row represents a topic with words in it, the feature matrix has dimension $M \times k$ so that each row represents a document and each column the weight of the corresponding topic in the document. This is illustrated as in Figure 2.2. The matrices H and W are determined by minimizing an error function. One common error function used for NMF is the L2 norm: $||A - HW||^2$. The values for H and W are updated iteratively until the error function has converged.

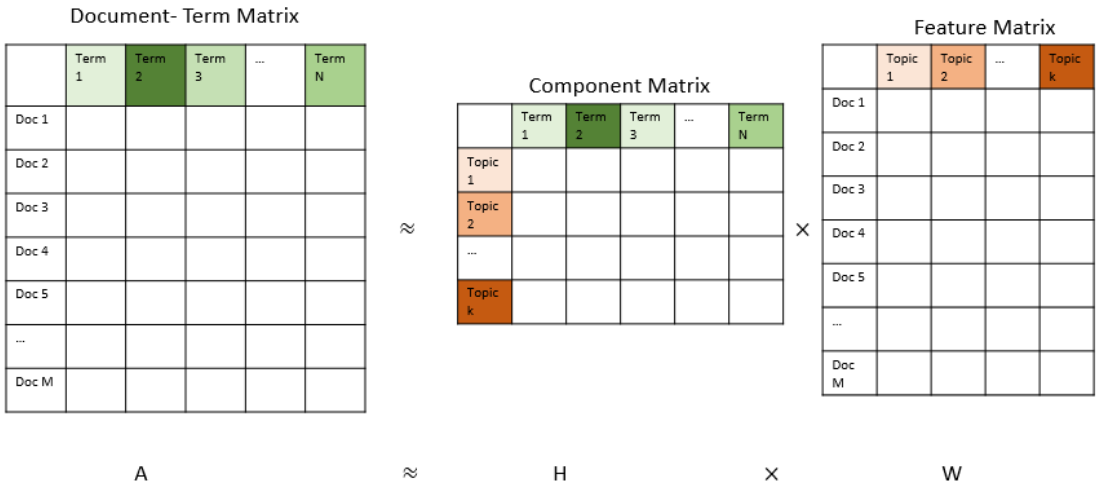


Figure 2.2: Illustration of NMF where the document-term matrix A is decomposed in component matrix H and feature matrix W.

2.4.3 SBERT and clustering topic model

Compared to the two topic models mentioned above, which utilize the count frequency of the text data, this topic modeling process is based on the SBERT model for topic modeling [13], [51]. SBERT aims to better capture word semantics and contextual information. Topic modeling with the SBERT and clustering topic model is done in four phases: 1) document embedding, 2)

dimension reduction, 3) clustering analysis and 4) topic words extraction [13], [51]. Next, these phases will each be described in more detail.

During the first phase, document embedding, the textual data can be encoded with the pre-trained sentence embedding transformer model based on the concept as described in section 2.3.3. In the work of Angelov [13], embeddings are jointly conducted where both the document vectors and word vectors are mapped to the vector space. In our project, we follow the approach from [51], which uses the SBERT model for document embedding. This approach is followed here since the pre-trained models have shown good results for various NLP tasks such as clustering[1]. After encoding the documents, the model will reduce the dimension of the obtained embedding.

During the second phase, dimension reduction, the obtained embeddings of the documents are reduced. This is needed since the obtained embeddings are high dimensional and sparse. If the embedding dimension is too large, then tasks, such as clustering, that rely on distance measurements between data points are computationally intensive. Therefore, dimension reduction is needed in this project. There are multiple dimension reduction algorithms available, such as t-distributed stochastic neighboring embedding (t-sne) [54] and Principle Component Analysis (PCA) [55]. Motivated by the proven performance of using the Uniform Manifold Approximation and Projection (UMAP) algorithm [56] in [13], we adopt UMAP for reducing the embeddinthe g dimension in this project. UMAP was chosen since it preserves the global and local structures in the embedding as well as it has good performance in handling document vectors with high sparsity which usually makes clustering analysis challenging [13], [56], [57]. UMAP reduces the embedding dimension and still preserves the structures of it while the t-sne reduces the dimension with the focus on visualizing the high-dimensional data in 2D dimension.

Clustering analysis

In the previous section, we saw that the SBERT model generates sentence embeddings from textual data that can be used for downstream NLP tasks such as clustering. In this section, we discuss the clustering technique HDBSCAN [58] which is used to identify clusters from the sentence embeddings in our project. This choice is motivated by the work conducted by Angelov [13]. The goal of clustering analysis is to group together similar points in a given dataset, often based on some proximity measure between points [59].

HDBSCAN clustering method is a density-based hierarchical clustering method based on DBSCAN [60] which identifies the densest (significant) clusters [58].

HDBSCAN is a non-parametric method that performs clustering analysis based on the following steps [61]:

1. Definition of a distance metric, the mutual reachability distance, that allows us to properly distinguish regions in space that are dense and sparse. In literature, an estimation of density is typically accommodated by using the standard kth nearest neighbor (denoted by $\text{core}_k(a)$), since this quantity is inexpensive to calculate. Using this quantity, the mutual reachability is defined as follows:

$$d_{\text{reach},k}(a,b) = \max\{\text{core}_k(a), \text{core}_k(b), d(a,b)\}, \tag{2.4}$$

where $d(a,b)$ is the original distance metric in the space. It is easy to verify that $d_{\text{reach},k}$ indeed defines a metric, i.e., it is symmetric, non-negative, and satisfies the triangle inequality. Furthermore, we can understand the effect of transforming the distance metric from d to

[1] SBERT Documentation

$d_{\text{reach},k}$ as spreading out sparse points (as they will have high distance to their kth nearest neighbor).

2. In the next step, a minimum spanning tree is constructed for the points, where the weight of an edge between two points equals the mutual reachability distance between them.

3. Based on the minimum spanning tree, a cluster hierarchy is created by removing edges of the tree in order from highest weight to lowest weight. At each iteration, removing the corresponding edge from the graph increases the number of connected components (clusters) by one. Initially, the minimum spanning tree is fully connected, i.e., there is one connected component. After the final iteration, there are as many connected components as there are data points.

4. In the next phase, the cluster hierarchy is condensed (i.e., simplified) by a parameter denoting the minimum cluster size. If at any point in the hierarchy, clusters are formed containing fewer than the specified minimum size, they are excluded from the cluster hierarchy.

5. Finally, the condensed cluster hierarchy is flattened. In this step, the most significant clusters are selected by considering the area they occupy in the hierarchy. This area is calculated from the cluster lifetime in the hierarchy, i.e., for which mutual reachability threshold does the cluster exist, and the number of points included in the cluster (which also varies over its lifetime as points are dropped out according to the minimum cluster size).

During the last phase, topic words extraction, class-based TF-IDF (c-TF-IDF) [51] is applied to the clusters identified from the last phase to extract topic words. A similar project: TopCat Clifton, Cooley and Rennie [62] was conducted which was meant to identify topics based on data mining techniques. Nevertheless, TopCat first identified key concepts and generated frequent itemsets or words from multi-documents and then conducted clustering which is not quite in line with our goal here. Hence in our project, the c-TF-IDF is a more appropriate choice. With this, all the documents identified as one cluster were concatenated to one long document and then apply TF-IDF to it. This will extract words from the cluster based on their TF-IDF score and use them as the topic representation.

These three topic modeling methods introduced above are used to identify topics from the textual data. In the coming section, we will discuss how can the topic modeling results be quantitatively evaluated.

2.4.4 Evaluation metric

The topic models are mainly used for identifying topics from the textual data in an unsupervised fashion. However, the evaluation of the topic modeling results remains an open research question. In this section, we mainly discuss three literature's which evaluate the topic modeling results from humans perspective, machines perspective, and a collective perspective from both human and machine.

In the work conducted by Chang, Gerrish, Wang *et al.* [63], it questioned the appropriateness of evaluating the topic modeling results with common model fit metrics such as held-out likelihood. By conducting human experiments to evaluate the coherence and relevance of the topic modeling results. The most intruding words were identified from the topic modeling results by human users as the evaluation. They justified their doubt in using model fit metrics for topic modeling evaluation, which means that their experiments presented negative correlations between the model fit metrics and the topic quality measure proposed in their work. However, the method used in this work needs consists of a lot of manual work.

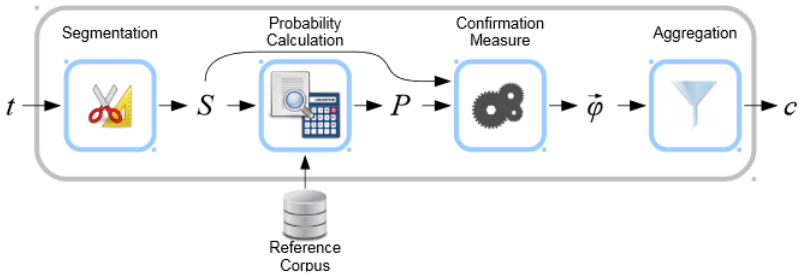


Figure 2.3: The four stages are involved in the unifying framework of obtaining quantified topic coherence scores. The source of the figure is [65]

In the work conducted by Lau, Newman and Baldwin [64], it further contributed proposed two metrics. Firstly, it extended the work of word intrusion conducted [63] by automating it. Secondly, it further experimented with three human-judged observed topic coherence at the model level from one dataset and quantified them, this quantified the topic interpretability as well as emulated human performance. This work was a successful approach in automating and quantifying human judgment on topic coherence of the topic modeling results.

Compared to the two works mentioned above, in the work conducted by Röder, Both and Hinneburg [65], a unifying framework that covers all known coherence measures and is able to combine all main ideas in the context of coherence quantification by the time was introduced. This framework is illustrated as shown in Figure 2.3 where four stages are identified with exchangeable components and can be combined freely to configure the four dimensions for calculating different topic scores.

For the dimension of segmentation, the topic word lists are segmented into smaller pieces and compared with each other, this dimension is noted as $\mathcal{S}$. For the dimension of confirmation measure, the confirmation scores are calculated for the word pieces given by the segmentation. The score obtained represents how strong the word pieces support each other. The different confirmation measures surrogates other topic coherence measures such as *point-wise mutual information* (PMI) [66] which gives the largest correlation with human ratings, UMass coherence [67] which measures asymmetrical confirmation between the segmented pieces. The dimension of confirmation measure is denoted as $\mathcal{M}$. For the dimension of probability calculation, the word probabilities can be used for the confirmation measurement so that different confirmation scores can be calculated. The method to estimate word probabilities is noted as $\mathcal{P}$. For the last dimension, the aggregation means methods of aggregating the scalar values obtained from confirmation scores. The set of aggregation functions is noted as $\sum$. Therefore, the whole framework configures a space of calculating topic coherence through different dimension combinations: $\mathcal{C} = \mathcal{S} \times \mathcal{M} \times \mathcal{P} \times \sum$.

In this work, topic modeling on various datasets such as 20 Newsgroup and Wikipedia subsets were used for calculating topic coherence scores with different configurations based on the four dimensions mentioned above. The computed topic coherence scores were compared to the human rating and the coherence measure C_v was identified as the best measure with its highest correlation. Therefore, the topic coherence measure C_v is adopted in this project.

2.5 Topic modeling for survey responses

In previous sections, we have introduced some preliminaries on NLP and topic modeling. In this section, the impact of conducting topic modeling on textual survey responses is discussed. Moreover, some related works are discussed.

Using surveys to understand customer opinions and hear the voice of the customer has been widely used by companies, there is empirical evidence that shows that customers' satisfaction levels with the company are often affected by how their opinions and feedbacks are resolved [68], [69] . This in return, would have impacts on the business of the company, the reputation of the company, and customer retention [70], [71]. Therefore, not only the voice of the customer should be heard, the feedbacks and opinions collected by surveys should also be processed.

As we earlier mentioned in the introduction, the processing of the unstructured survey responses is not easy. Since the analysis of textual survey responses can be costly and labor-intensive as a lot of human action may be involved [2], [4], [9], [72]. Moreover, the collected responses may vary in their formality and length. The application of topic modeling is not new. There are some works conducted in applying the topic models to the textual survey responses. In the work conducted by Roberts, Stewart, Tingley *et al.* [9], the Structural Topic Model (STM) was proposed. It was argued that STM was helpful in raising the concerns of having a suitable topic model in apply topic modeling to the textual survey responses and the exploration of data with little pre-knowledge. The sentiment analysis was also conducted by human labeling. There was no separate evaluation on the obtained topic words themselves. The results were evaluated based on the priority matrix constructed based on the sentiment analysis and the topic modeling results.

In the work recently conducted by Cammel, De Vos, Soest *et al.* [10], it examined whether the NMF topic model can be applied to patient textual survey responses and its transferability for data from another hospital. In the work conducted Pietsch and Lessmann [7], three topic models: Latent Feature Latent Dirichlet Allocation (LFLDA) [73], Biterm Topic Model(BTM) [11] and Word Network Topic Model(WNTM) [74]. These three models were applied to one real-life survey responses data given from a market research company where nine labels and "other" labels were assigned to the documents. In the end, the topic modeling results were evaluated from both the topic quality and the topic distribution perspectives quantitatively and qualitatively. The topic quality was evaluated by the PMI topic coherence score quantitatively and opinions from human experts in pointing out confusing words qualitatively. The topic distribution was evaluated by comparing the obtained topic distribution to the classification of the data based on the original label quantitatively and how consistent documents with certain labels are identified with the topic qualitatively. It was concluded that the BTM and WNTM topic models achieved promising results.

Chapter 3

Business and data understanding

In this project, survey responses from a customer loyalty survey conducted by a financial service company were used as a case study. The objective of this case study is to better understand the common problems that may occur in processing unstructured survey responses.

Conducting customer loyalty surveys is a popular and useful way to hear the VOC (Voice of Customers) and the level of customers' satisfaction can be improved if corresponding actions are taken according to the collected VOC. In this chapter, we focus on understanding the project from a business perspective. We present typical business processes for the offered service from the financial service company. Moreover, we summarize the survey structure, the types of questions asked, and the current workflow of analyzing the collected responses.

3.1 Customer journey

For companies offering financial services, the processes involved can be quite diverse and complex. In this project, we are mainly dealing with the financing process which can be summarized as "dealer offers financing to their customers and the financial service company offers complete financing advice, solutions and supports to the dealers." Countries may carry out the process differently due to different business cultures and/or different legal obligations.

A general customer journey was identified by the company. There are in total six processes involved in the identified customer journey: Quote, Apply, Enroll, Fund, Manage and Upgrade.

To improve the overall customer satisfaction throughout the customer journey, it is important for companies to hear the Voice of the Customer (VOC) so that they can identify and interpret customer needs and respond to customer requirements. There are multiple ways to capture the VOC, such as interviewing customers and conducting surveys.

In this project, we are using the customer loyalty survey responses collected by a financial service company from 2017 to 2020. The company uses these surveys to improve its services by understanding what aspects customers are satisfied or dissatisfied with.

3.2 Case study

The customer loyalty survey consists of different types of questions. The structure of the survey used by the company can be seen in Appendix A. After 3 questions, the survey used in the case study was split into two branches based on the Net Promoter Score (NPS) scores given from the respondents. NPS is a numeric metric widely used for market research as an indicator of customer loyalty with a scale from 1 to 10. It is presented in the form of a single survey question asking respondents to rate the likelihood of them recommending the company to others.

Based on the given NPS score, respondents will be presented one of the branches based on the NPS score they provided. All the choices in the current survey are exclusive and there are in total 19 possible text inputs that can be collected. However, only two questions (Q20 and Q21) are compulsory text input questions for non-promoters and promoters, respectively. Therefore, the number of responses collected per question varies.

It is therefore not a logical choice to combine all the textual responses collected into one set and conduct topic modeling upon considering such a nested survey structure. We would choose the text responses collected from Q21 as the case study data. In a previous question of the survey, the respondents who gave an NPS score selected a specific area where they think the company should improve upon. In Q21, these respondents explained their choices made in this previous question. This question was chosen since it has the highest number of responses collected (1767) responses and it is also widely accepted that a larger data size provides more potential for recognizing patterns and trends from the data.

The current survey is built, published, distributed, collected, and analyzed on a customer experience management platform called Qualtrics CustomerXM. The responses consist of unstructured data (e.g., text data) and structured data(e.g., NPS score, the choice made for question). Currently, only the collected structured data are utilized for gaining information from customers' feedback. In this way, visualizations for structured data such as distributions over the NPS score or selected choice in multiple-choice questions are available in the form of dashboards. These results help the company to have a general idea of what areas need to be improved and what area earns a high level of satisfaction from the customer.

3.3 Challenges

With the current workflow in handling the collected customer survey as mentioned above, only the collected categorized data based on the exclusive choices are used to bring insights from the customer feedback. The unstructured data is not analyzed, although it could contain valuable insights, too. Hence, the main business challenge addressed in this report is developing tools that the company can use to analyze unstructured data. Specifically, the tools that are explored in this work are topic models.

From the business perspective, one of the potential tools that are relevant to include in this study is the Qualtrics CustomerXM platform. As mentioned above, the financial service company currently relies on this platform for processing structured survey responses. Processing unstructured responses are possible but untested. Hence, conducting topic mining with this tool is included as a part of the project.

Besides the existing Qualtrics CustomerXM platform, a variety of topic models from the NLP literature are considered. A comparison between all topic models is made on the survey dataset to evaluate which of the topic models has the potential to address the stated challenge.

3.4 Case study data description

The financial company conducts business in around 30 countries. Over a time window of 3 years, more than 20,000 customer loyalty survey responses have been recorded. The tool Qualtrics CustomerXM was used to automatically collect and store the survey responses. As we shortly introduced in Chapter 3, the customer loyalty survey is built, distributed, and collected on the Qualtrics CustomerXM platform. Each data entry has attributes defined by the survey designer (e.g., *QID22_TEXT*), as well as metadata (e.g., *User Language*).

The collected survey data consists of structured categorical data and unstructured free-text data. The current workflow in handling the collected survey data conducts good analysis on the structures of categorical data, the focus of this project would be on unstructured free-text data. Although the free-text data contains multiple languages, only English textual data is considered in this project. The motivation for this is that English is a widely spoken language and is often used in the field of text mining.

There are in total 20257 collected responses. By selecting responses that only contain English responses, 4303 responses remain. Due to the current structure of the designed customer loyalty survey, not every question has the same amount of text responses (see Figure 1). After filtering out survey responses that contain no textual responses, there are in total 3035 collected responses containing the text input.

The number of collected text responses per question is shown in Figure 3.1. It can be seen that the differences between the number of collected text responses per question are huge. This is the result of the branch structure of the current survey. This design mainly offers exclusive multiple choices and optional text input for follow-up questions in the branch as shown in the survey structure Figure 1. The respondent enters a given branch depending on the NPS score they gave. Then, the respondent is guided to one of the follow-up questions based on the first choice they made in the branch. Hence, it is possible that only one text response is collected from a respondent who completed the survey. In this project, we select the question with the most responses collected (QID21_TEXT) and use it as the case study data.

In Figure 3.2, the distribution of the number of words in each document can be seen. This distribution is similar to a Poisson distribution. In Figure 3.3, a box plot is used to visualize the case study data with a five-number summary ("minimum", first quartile (Q1), median, third quartile (Q3), and "maximum"). To conclude, the lengths of the documents can be described as follows: there are around 20 empty responses, the maximum document length is around 200 words, the median of the document length is around 25, and 75% of the documents have a length lower than 50. The interquartile range of document lengths is between 12 and 40. The outliers are documents with a document length larger than 65. by reviewing some of the responses, we noticed that the given responses usually cover a lot of information than the specific answer needed for Q21. We also reviewed documents with short lengths such as 1 word, then answers such as ".." or " none" were observed. This happened may be due to the fact the respondent wanted to skip this compulsory question.

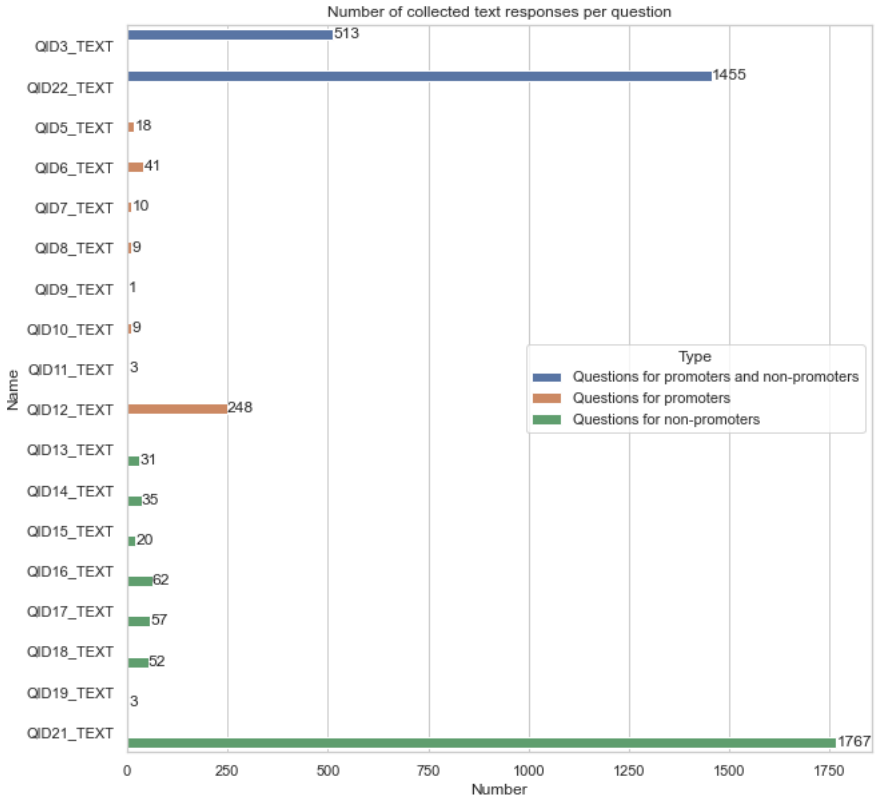


Figure 3.1: Number of collected responses per question

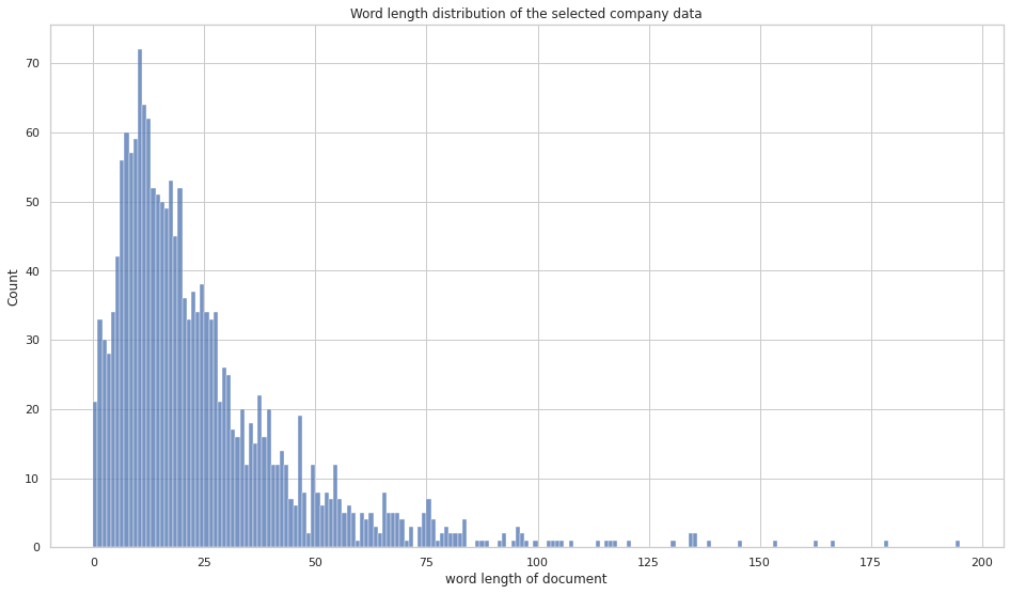


Figure 3.2: Distribution of the document length in words of the company dataset.

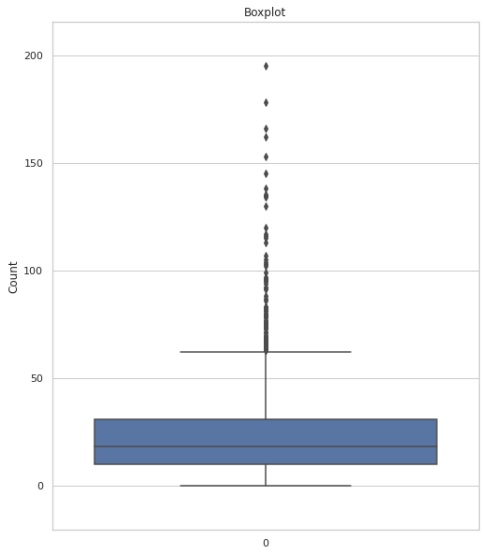


Figure 3.3: Box plot for the word length distribution of the selected company data.

3.5 Dataset selection

Besides the case study data, we selected four additional datasets to successfully reach the research goals:

BBCNews dataset

The BBCNews dataset is commonly used for document classification [75], [76]. It consists of 2225 labeled documents from the BBC website[1]. There are five labels in total: business, entertainment, politics, sport, and tech. A sample document from the business class can be seen as in figure 3.4. The motivation for selecting this dataset for our study is that we expect it to be coherent in the sense that each document is a formal news article and labeled. Although the ideal situation would be to also include a labeled unstructured survey response dataset, such a dataset is not publicly available to the best of our knowledge. Still, we believe the BBCNews dataset provides an interesting contrast to the less formal (and short-text) unstructured survey responses.

Amazon fine food review dataset

The Amazon Fine Food Review dataset[2] consists of reviews of fine foods from Amazon. In this project, we are using subsets based on the work conducted by Shah and de Melo [77]. We took 2000 unique unlabeled documents from the clean text data without any typographical error and the corrupted data with 3.5% and 7.5%character-level typographical error introduced [77] correspondingly. A sample document can be seen in figure 3.5a with its clean version and in figure 3.5b in its corrupted version. These datasets were used to observe the model sensitivity to corrupted data with typographical errors.

Real-life student survey dataset

This dataset consists of text responses collected from the survey dataset collected from a student survey regarding the question: *Which course activities (e.g., lectures, lab sessions, peer review, ..) would better be held online and which on-campus?* There are in total 575 text responses collected where 525 responses are unique. Some duplicated responses can be seen in figure 3.6. The student survey dataset was chosen to show whether the models would also work on other survey data.

[1] BBCNews dataset
[2] Amazon Fine Food Review data on Kaggle

UK economy facing 'major risks'

The UK manufacturing sector will continue to face "serious challenges" over the next two years, the British Chamber of Commerce (BCC) has said.

… …

… …

"Despite some positive news for the export sector, there are worrying signs for manufacturing," the BCC said. "These results reinforce our concern over the sector's persistent inability to sustain recovery." The outlook for the service sector was "uncertain" despite an increase in exports and orders over the quarter, the BCC noted.

The BCC found confidence increased in the quarter across both the manufacturing and service sectors although overall it failed to reach the levels at the start of 2004. The reduced threat of interest rate increases had contributed to improved confidence, it said. The Bank of England raised interest rates five times between November 2003 and August last year. But rates have been kept on hold since then amid signs of falling consumer confidence and a slowdown in output. "The pressure on costs and margins, the relentless increase in regulations, and the threat of higher taxes remain serious problems," BCC director general David Frost said. "While consumer spending is set to decelerate significantly over the next 12-18 months, it is unlikely that investment and exports will rise sufficiently strongly to pick up the slack."

Figure 3.4: A sample document with a business label from the BBCNews dataset, note that not all characters are included

I purchased this item to help my son and grandmother get more fiber in there diet and it has helped them both to get on a more regular track.

(a) A sample document in its clean format

I purchased tgis item to hepl my sen and racdmothur get more fiber in there deet nad ot hadsq hepled tnem both to get on | mor requlars otrack.

(b) A sample document in its corrupted format

Figure 3.5: A sample document for both the clean text data and corrupted text data from the Amazon Fine Food Review dataset

Lectures
Lab sessions on campus lectures online
All on campus
Lab sessions
Peer review
Lab sessions could be online
...

Figure 3.6: Some sample duplicated responses observed from the student survey dataset

Chapter 4

Evaluation criteria

Comparing the results from different topic models is challenging: natural language is complex, and even state-of-the-art topics models are not guaranteed to find topics that would be meaningful for human judgment. Because of this challenge, it is important to define a proper set of evaluation criteria that are suitable in the context of unstructured survey responses. In this section, a selection of quantitative and qualitative evaluation criteria for use in this study are described and motivated.

Before presenting the criteria in the remainder of this section, let us give some general remarks that apply to this particular study:

1. We consider data that has no ground truth labels. Thus, splitting the data into train/test datasets for model performance evaluation is not appropriate here. It is also not our goal to build predictive topic models.

2. The general assumption made for topic modeling is that meaningful and useful topics are found, but the evaluation of that is difficult due to the lack of availability from domain experts to evaluate the meaningfulness and the usefulness of the topic modeling results obtained as well as how the consistency of their evaluation. The interpretation of the word lists of topics found has no fixed rule that we can refer to. Different topic model evaluation methods have been brought out [63]–[65] from the perspectives of how humans and machines interpret the topic modeling results. The choice of an appropriate quantitative evaluation metric that can be applied to models implemented in this project would be helpful.

3. Besides the quantitative evaluation, we also need the qualitative evaluation. Such a qualitative evaluation is vital since we are exploring the potential of different topic models being applied to the survey responses collected by the financial service company. The domain experts can therefore share their opinion on the coherency of the topic modeling results obtained, as well as the meaningfulness of the results in the context of the company business.

Based on the research goal defined in Section 1.2 and the motivation listed above, the evaluation criteria should enable us to conduct a thorough empirical comparison between state-of-the-art topic models in the context of unstructured survey responses. The criteria have been divided into three categories: topic quality, model sensitivity, and practical aspects.

4.1 Topic quality

A requirement for successfully applying topic models as a tool to obtain topics from unstructured survey responses is that the topics are of sufficient quality. The difficulty is that there is no uniquely defined way to measure topic quality. Therefore, it is important to combine business and data understanding to determine the topic quality. The topic quality is evaluated from two distinct perspectives: topic coherence and interpretability.

4.1.1 Topic coherence

There are multiple quantitative metrics available to measure how semantically meaningful the inferred topics are. Chang, Gerrish, Wang *et al.* [63] and Lau, Newman and Baldwin [64] presented how topic model results can be evaluated quantitatively, considering how humans and machines interpret the results correspondingly. Later on, Röder, Both and Hinneburg [65] extended these metrics to a framework that spans the space of all known coherence measures, as well as combining all main ideas in the context of coherence quantification. Knowing this, we choose the best-performed topic coherence measure C_v topic coherence score [65] and use it to quantify the topic coherence of the modeling results.

Having the unified quantitative topic coherence measures for topic models mentioned above for the case study data is a good start, yet we would still like to verify how appropriate the choice of topic coherence is. To do so, we choose the BBC News dataset [1] and apply the topic models to it. Finally, we compare the topic coherence obtained. The BBC News dataset contains 2225 articles, each labeled as one of 5 categories: business, entertainment, politics, sport, or technology. This dataset is chosen since the topic/news categories are known and concrete, while the case study data does not have defined categories. With data having known topics, a higher topic coherence score is expected, and we would like to verify that.

Besides the quantitative topic coherence score, we would like to measure the topic based on human judgment. The domain expert gives their judgment on the coherence of the discovered topic word lists.

4.1.2 Interpretability

According to the Merriam-Webster dictionary, *interpret* is defined as *to explain or tell the meaning of: present in understandable terms*[2] and *explain*[3] is defined as *to show the logical development or relationships of.* While it is common for some researchers to use interpretability and explainability when it comes to model evaluation such as in [78], we use these two terms for evaluation from different perspectives. Therefore, the topic models have compared these components: algorithms explainability and results in interpretability. The models are compared to see how explainable/transparent the algorithm is of the rationale that the model used to make decisions. The results are compared to see how interpretable they are for humans and whether the results and informative enough for domain experts to use as business insights.

[1] BBC News Dataset
[2] Merriam-Webster Dictionary, accessed 2021-08-19
[3] Merriam-Webster Dictionary, accessed 2021-08-19

4.2 Model sensitivity

The evaluation of model sensitivity is essential as real-life datasets commonly contain language errors. This is especially true in the context of unstructured survey responses where informal or incomplete responses can be given. The proposed topic consistency measure can give us a sense of the model sensitivity with respect to typographical errors by choosing for $\mathcal{A}$ and $\mathcal{B}$ the topic modeling results associated with the clean and corrupted dataset, respectively. Many current works in topic model evaluation focus on defining a good quantitative metric that incorporates how both machines and humans interpret topic modeling results (see, e.g., [64], [65], [79]). There is also research such as in [80], [81] that focuses on the short-text topic model performance. While there are researches conducted such as in [82] which investigate the impact of topic cardinality (number of topics) on model sensitivity, model sensitivity is still a less actively researched area.

We choose the *Amazon Fine Food Review* dataset [4] from Shah and de Melo [77] to conduct the sensitivity analysis. The dataset contains 2 variants of the same set of reviews from Amazon, each with a different percentage of character-level typographical errors, namely, 3.75% and 7.5%. and evaluate the corresponding model behaviors in terms of the sensitivity defined in the next section.

Proposed model sensitivity estimator

In this section, the model sensitivity estimator is derived. To our knowledge, the estimator we propose here is novel and has not been used for model sensitivity analysis in topic modeling before. Intuitively, the sensitivity estimator is based on measuring the consistency between two topic modeling results on the same corpus. More specifically, it will be shown that this consistency measure is the following probability: *Given two documents that share a topic in topic modeling result 1, what is the probability that these two documents also share a topic in topic modeling result 2?* We denote this probability as the *topic consistency* associated with the two topic modeling results.

However, this estimator also has its limitations. For example, when the total number of topics in topic modeling result 1 is larger than that in result 2, then this consistency measure would punish the reduce of dimension. Therefore, this topic consistency measure should been seen as an abstraction of model sensitivity rather than a quality measure for a topic model or topic modeling results. To make sure appropriate usage of the topic consistency, we define two use scenarios for it:

- Measure the topic consistency when applying the same data to two different models.
- Measure the topic consistency when using the same model to two data sets where one is original and the another one is corrupted version of the orignal data.

We now derive the mathematical expressions for the sensitivity analysis based on this consistency measure. We start by denoting the pair of topic modeling results we want to compare as $\mathcal{A}$ and $\mathcal{B}$. Each can be expressed as a sequence for which each element gives the topic number of the corresponding document, i.e.:

$$\begin{aligned} \mathcal{A} &= (a_i)_{i=1}^{N}, \quad a_i \in \{0, ..., N_a - 1\} \\ \mathcal{B} &= (b_i)_{i=1}^{N}, \quad b_i \in \{0, ..., N_b - 1\} \end{aligned} \tag{4.1}$$

[4] Typological error dataset

Note that there is no assumption that the number of topics between the topic modeling results (N_a and N_b) is the same. Only the number of documents (N) must be the same. The topic modeling results are combined by defining a third sequence, $\mathcal{T}$, containing the assigned topic in result $\mathcal{A}$ and result $\mathcal{B}$ for each document, i.e.:

$$\mathcal{T} := (a_i, b_i)_{i=1}^{N} \tag{4.2}$$

Then, we define the count matrix $T \in \mathbb{N}^{N_a \times N_b}$ such that the element in the ith row and jth column ($T_{i,j}$) counts how often a document is assigned topic i in $\mathcal{A}$ and topic j in $\mathcal{B}$:

$$T_{i,j} = |\{(a, b) \in \mathcal{T} | a = i, b = j\}| \tag{4.3}$$

Without loss of generality, we assume that each row and each column in T contains at least 1 non-zero element. If a zero row does exist, it indicates that no documents are assigned the corresponding topic in $\mathcal{A}$, and the row can be removed. The same argument holds for zero columns.

We will now show how topic consistency can be calculated from T. First, we count the number of documents associated with each topic in $\mathcal{A}$, and collect these counts in a vector s:

$$s = T\mathbf{1}_{N_b} \in \mathbb{N}^{N_a}, \tag{4.4}$$

where $\mathbf{1}_n \in \mathbb{N}^n$ is a column vector of ones. The count s is used to calculate a relative weight, w_i, for topic i:

$$w_i = \frac{s_i}{\sum_{i=1}^{N_a} s_i} \tag{4.5}$$

T is row-normalized and the result is stored in $\tilde{T}$:

$$\tilde{T}_{i,j} = \frac{T_{i,j}}{s_i} \tag{4.6}$$

Row i in $\tilde{T}$ contains the distribution of topics in $\mathcal{B}$ for topic i in $\mathcal{A}$. It can be used to calculate the probability p_i that 2 documents randomly chosen (with replacement) from topic i in $\mathcal{A}$ are mapped to the same topic in $\mathcal{B}$:

$$p_i = \sum_{j=1}^{N_b} \tilde{T}_{i,j}^2 \tag{4.7}$$

Finally, by weighting the probability p_i by w_i and summing the result over all topics in $\mathcal{A}$, the topic consistency is obtained:

$$\text{topic consistency} = \left(\sum_{i=1}^{N_a} w_i p_i \right) \cdot 100\% \tag{4.8}$$

However, in this way a very high topic consistency could be achieved even if the model is in fact sensitive. Consider, for example, the situation that the model maps nearly all documents in the corrupted dataset to a single topic: the topic consistency will be nearly 100%, although clearly the topic modeling results are sensitive to corruption in the dataset. To remedy this, the topic consistency can be calculated twice, interchanging the order of the topic modeling results $\mathcal{A}$ and $\mathcal{B}$, and subsequently taking the minimum of the two values.

Practical aspects One of the goals of this work is to provide a comprehensive comparison of state-of-the-art topic models in an applied context. With this in mind, this report should contain a discussion on practical aspects that are useful for a reader considering to use the topic models. Specifically, the implementation efforts are considered

We will evaluate the implementation effort needed for each topic model including Qualtrics. The main difference between them is that for some topic models, every preprocessing step is needed, while this is less crucial for other models. Therefore, it might be an impacting factor to take into account to conduct topic modeling with a certain model.Another aspect of the implementation effort that we consider is the amount of training that is needed.

Chapter 5

Methodology

In this chapter, we introduce the overall methodology defined to resolve the research questions raised in Section 1.2. First, we discuss the guideline of the approach taken in this project. Then, we present the details of different models used in this project.

5.1 Design guideline

The main purpose of this research is to investigate how topic modeling can be used in processing unstructured survey text responses. Therefore, the designed method will implement different topic models to the data and obtain topic words.

The method consists of two main phases. The first phase is the data preparation, where the details for techniques used to clean and process data are described. The second phase is the modeling where we show the configuration, algorithms, and measurement for different topic modeling approaches. Including the default Qualtrics text analysis tool, there are in total four models used in the thesis.

1. The first model is part of the text analysis functionality offered in Qualtrics.
2. The second model is a probabilistic topic model: LDA (Latent Dirichlet Allocation) which is one of the most popular topic models.
3. The third model is another popular topic model NMF (Non-negative Matrix Factorization) that is based on a linear-algebraic optimization algorithm.
4. The fourth model is an SBERT and clustering topic model. In this model, we first use the pre-trained sentence embedding model 'paraphrase-mpnet-base-v2'[1] to directly encoding the text to embedding. Then we reduce the dimension of the embedding with UMAP dimension reduction[2] and conduct clustering on the reduced embedding. In the end, we extract words per dense cluster based on the TF-IDF scores and use these words to represent the topics/clusters.

Moreover, iterations between modeling and data preparations were performed based on the observations of the modeling results, this is also in line with the CRISP-DM research mentioned

[1] Sentence Transformers Documentation
[2] UMAP documentation

in Section 1.3. Additionally, we implemented the models so that each model had a corresponding coherence model so that the measurable topic coherence score can be calculated. The results obtained from topic modeling were integrated for visualization. Finally, the topic modeling results are evaluated with the evaluation criterion as specified in Chapter 4.

5.2 Modeling for Qualtrics

As we previously mentioned, the financial service company faces challenges in processing the unstructured textual data, although Qualtrics does offer the functionality of text mining. The purpose of this experiment is to investigate how the text analysis functionality in Qualtrics can be used for topic modeling.

There are currently two ways of identifying "topics" with the text mining function offered within Qualtrics: the built-in automatic recommend topics and a custom way of identifying topics with self defined queries. A concrete example for the chosen data (response collected for Q21) with the topics identified by the built-in automatic recommend recommendation is as shown in Figure 5.1. These results are considered too general to be used. According to the domain expert, the words used for constructing the topic queries are too sparse to extract any information with its current level of granularity. In Qualtrics, 'topics' are identified by queries as shown on the right of the Figure 5.1: three queries for 'Time', 'Market' and 'Request' are shown. It was also observed that all the automatically recommended topics have topic queries constructed only with logic operator '||' i.e., Hence, it was concluded that the built-in automatic recommendation is not suitable for our application.

Now that the built-in automatic topic recommendation is not considered in the sequel. We use the custom way of identifying topics with self defined queries from Qualtrics in this experiment.

Process

We demonstrated above that the built-in automatic topic recommendations offered by Qualtrics is considered too general to be used by the company. Therefore, the process of the experiment quickly emerged to the process illustrated in Figure 5.2 where an alternative approach of identifying topics with self defined queries is taken. This alternative approach within Qualtrics follows the vendor business customer journey introduced in Section 3 and the knowledge from the domain expert as the main guideline. With this alternative approach, the text analysis becomes a human-in-loop modeling procedure.

The experiment is divided into the following sequence of steps:

1. An open-ended question is selected and its corresponding English responses are collected. The data cleaning and pre-processing steps mentioned in section 2.2.1 are not applied since the model in Qualtrics operates on raw textual data. The text analysis tool 'Text iQ' in Qualtrics [83] is used to conduct text analysis for the selected question.

2. We refer to the customer journey and incorporate the knowledge from the domain expert in identifying language patterns that can represent certain phases of the customer journey.

3. We also define queries for language patterns that are not defined in the customer journey but observed frequently.

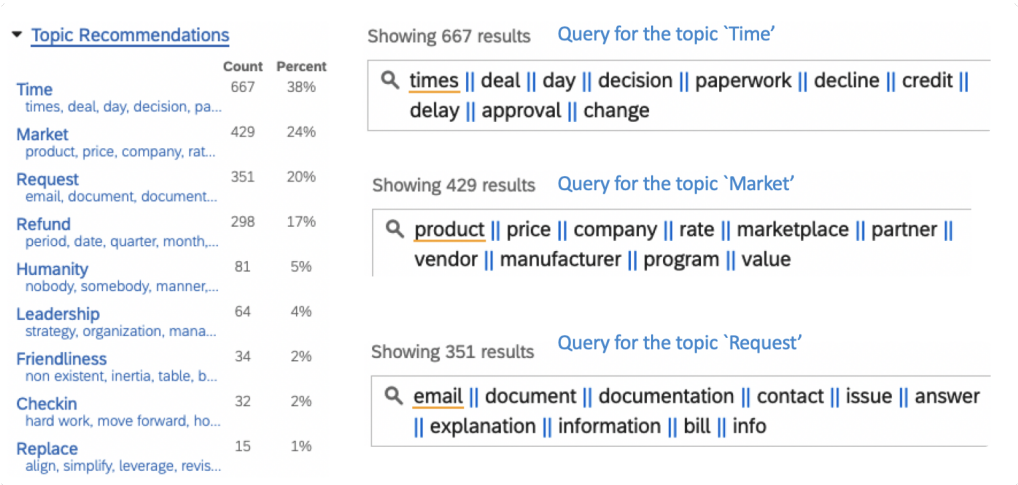


Figure 5.1: For Q21 (Explain the choice made for the selected improvement area): The left side of this figure shows the topics automatically recommended by Qualtrics Text iQ; the right side are the logic topic queries for the first three topics

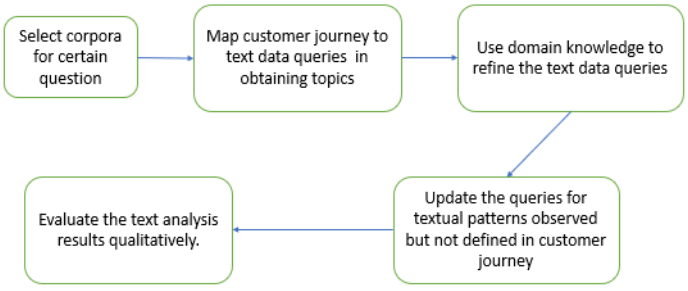


Figure 5.2: The text analysis process for experiment 1

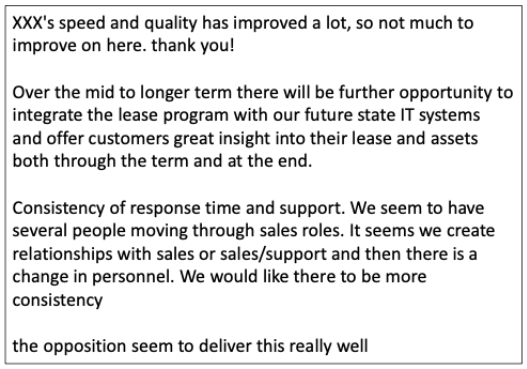


Figure 5.3: Documents randomly sampled for Q21(Explain the choice made for the selected improvement area)

4. We repeat the steps 2 and 3 until responses were not able to be mapped to a meaningful topic according to the domain expert or when the query is too specific to match several documents.

5. We evaluate the results. For the topic modeling, the quality of the topics is evaluated by a domain expert.

The evaluation together with domain experts is needed since there is no objective ground truth for the text analysis results. A quantifiable mathematical expression to represent the text analysis quality is also not available here. Therefore, we conduct the qualitative evaluation performed by the domain experts.

Data

In this project, all the responses collected for certain question is considered as a corpus and each response is considered as a document. Within Qualtrics, the raw text data from the collected English responses is directly used. For example, if we select a question such as Q21 "Explain the choice made for the improvement area", some sample responses collected for this question are as shown in figure 5.3.

5.3 Modeling for state-of-the-art topic models

Beside the Qualtrics tool, different state-of-the-art models are chosen for the following reasons: 1) the selected models are widely used in literature; and 2) the topic models have a diverse theoretical background, making a comparison interesting.

There are various topic models like LDA(Latent Dirchlet Allocation), LSA (Latent Semantic Analysis), NMF(Nonnegative Matrix Factorization) and etc. It is not quite practical to implement and compare all of them considering the limited time. Therefore, we choose LDA and NMF topic models to implement as they are the two most popular topic models. The details of the theories for these two models have been introduced in Section 2.4.1 and Section 2.4.2. Although these two models are based on different theories, the process of performing topic modeling with these two models are similar: Firstly, the data is cleaned and bi-grams are created fom the cleaned data, the cleaned textual data is used to construct corpus and dictionary that are used as input for

Table 5.1: Overview of needed text data cleaning techniques for the models

	LDA	NMF	SBERT and clustering
Contraction expansion	x	x	
Lowercasing	x	x	
Tokenization	x	x	
Punctuation Removal	x	x	x
Stop words removal	x	x	
Lemmatization with POS tag	x	x	

models. Secondly, the topic model is trained based on the input, the topic modeling results are represented in the format of a probability distribution over words. Thirdly, the model is tuned by the number of topics that the model needs to identify with the goal of obtaining high C_v topic coherence score. Lastly, a dominant topic is identified for each document based on the words and their probabilities contained in the document. We integrate these results and visualize them in a dashboard that can be used by the end-users.

The SBERT and clustering model is a combination approach that consists of embedding textual data by a state-of-the-art SBERT model, dimension reduction, clustering analysis and class-based TF-IDF [13], [41], [51]. This experiment setup for topic modeling assumes that most semantically similar documents are close in vector space and likely to share underlying topics [13]. Unlike the previous two topic models that focus more on the probability of the word occurrences and are unable to consider the semantic meaning and context of documents, the setup in this experiment is considered to do so [84]. Using state-of-the-art pre-trained SBERT models are much faster than directly train a neural network based on the data since these models are trained on large corpus(for e.g. WikiAtomicEdits with around 43 million edits across 8 languages[3]) and evaluated by tests such as clustering and paraphrasing[4], therefore more accurate representations of words and sentences shall be obtained with performance proved [41], [85].

The end results obtained with SBERT and clustering model model are also word lists, while its process is a combination of sentence embedding and clustering analysis. Moreover, the number of topics identified by this model depends on the number of dense clusters identified instead of a given parameter. Each dense cluster identified is considered to be a (abstract) topic.

In the coming sections, we illustrate differences between different models from mainly three perspectives:

5.3.1 Pre-processing

The data cleaning techniques used to clean the textual data are as described in section 2.2.1. However, not all data cleaning techniques are needed for all the topic models used in this project. For example, the LDA model need the input data to be in the BOW format while the SBERT and clustering model needs needs the input in the embedding format. Here, we present an overview for the needed data cleaning steps as shown in Table 5.1.

After conducting the data cleaning, we also extract bi-grams from the cleaned text hoping to increase the interpretability of the topic word lists obtained from topic modeling. For the LDA and NMF topic models, the bi-grams were identified from the text data without stopwords using the phrase model constructed. A phrase model automatically detects common phrases from a stream of sentences. For the SBERT and clustering model, bi-grams were constructed per identified cluster

[3]WikiAtomicEdits
[4]Tuned Sentence Embedding Models

by extracting bi-grams from the document matrices from the cluster. N-grams are essentially N words occurring frequently together in the documents. The n-grams are used since by treating n words that frequently occur together as single term, it helps the model to handle implicit semantic structure better. For example, *"New York"* becomes *"new_york"* by constructing bi-gram model. Such processing preserves the implicit semantic structure from the tokenized textual data so that *"new"* and *"york"* are not considered separately and assigned to different topics in which case will lost their meaning and its context [86]. Some specific bi-grams obtained in the case study are words such as *"long_term"*.

5.3.2 Topic words and dominant topic

The results of topic modeling are usually represented by lists of words. Although LDA topic model is a generative probabilistic model and NMF topic model is a matrix factorization model, they both agree on that: each (abstract) topic identified is a probability distribution of words, each document is considered as containing different topics. Using these characteristics, each document would be able to be identified with a dominant topic based on the probability distribution of the words contained in it.

For SBERT and clustering model, it discovers the topics from the documents based on the semantic meanings and word orders of the documents instead of the word co-occurrences. This textual data is first encoded to embedding with a pre-trained sentence transformer. The obtained embedding are the results of mapping sentences to vectors of real numbers. The embedding obtained is very sparse considering the words length distribution of our text data. Therefore, dimension reduction is performed on the embedding such that the clusters can be more easily discovered [13], [56].

The dimension reducer UMAP has several hyperparameters[5] that determines how the dimension reduction can be performed. In our project, we adopted the findings discovered by Angelov [13], it was found that the *number of nearest neighbors* delivers the best performance when its value set to 15 which means that the reducer gives more emphasis on local structure than the global structure. While this is indeed needed for the task of clustering which meant to find dense areas of documents that are close to each other in the high dimensional space. Therefore, we adopt the same value 15 for the *number of nearest neighbours*. In another word, each identified cluster/topic consists of at least 15 documents. Each identified cluster consists of encoded documents that are considered having similar meanings and therefore sharing a similar underlying topic. We also adopt the *cosine similarity* as the distance metric used by the UMAP reducer as it is a common and popular distance metric used when it comes to vector space due to its characteristic of measuring documents' similarity regardless of their size [33], [87].

The topic word list is extracted using TF-IDF for each cluster that is identified by the clustering analysis. The topic coherence score per model was obtained by taking the average of the C_V topic coherence score calculated based on the corresponding coherence model. The details of embedding and clustering analysis used in this project are described in Section 2.3.3.

5.3.3 Tuning

Hyper-parameters are tuned with the goal of maximizing the C_v topic coherence score. For the LDA and NMF topic models, the number of topics that the topic model needs to discover was set as the tuning parameter. While the SBERT and clustering model, the number of the topics identified by the model depends on the clustering analysis results instead of a given model

[5] UMAP documentation

parameter. Therefore, we use *n_components* as a tuning parameter in our project. This parameter decides the embedding dimension after reducing, considering the diversity of the datasets used in our project as well as the goal of obtaining high topic coherence scores. Since the datasets used [13] did not contain survey response types of data that are relatively short and they were all unlabelled data. Hence, we would like to find out which embedding dimension used would give the best topic coherence score for the corresponding coherence model implemented. Therefore, the number of dimensions for the reduced embedding is set to be the tuning parameter in this project hoping to obtain the best possible C_v topic coherence score.

5.3.4 Evaluation

The results obtained by the state-of-the-art models are evaluated based on the evaluation criteria defined in Chapter 4. Specifically, to evaluate the topic coherence quantitatively for the topic modeling results, corresponding coherence models were constructed so that the topic coherence scores can be calculated. For model sensitivity experiments defined in the evaluation criteria, the topic consistencies based on the proposed model sensitivity estimator were also calculated.

Overall, the processes of conducting topic modeling with different topic models can be concluded as shown in Figure 5.4 and Figure 5.5. The process chart use different colors to indicate different steps: The blue color is used for steps needed for pre-processing, the orange color is used for steps need for obtaining topic words and dominant topic, the green color is used for steps that needed for tuning the model and the purple color is used for steps that integrating the results and the evaluation.

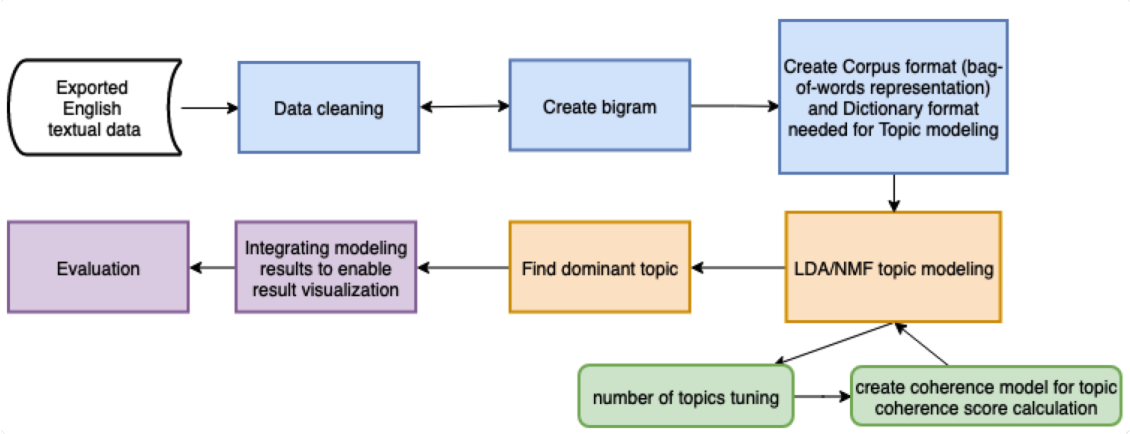


Figure 5.4: The topic modeling process for LDA and NMF topic models. Blocks with different colors indicate the corresponding perspectives as what we mentioned above

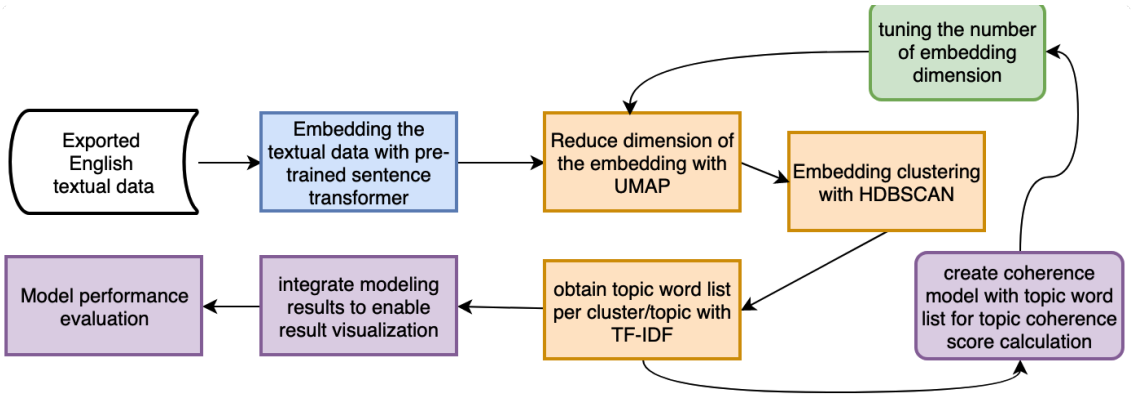


Figure 5.5: The topic modeling process for SBERT and clustering model. Blocks with different colors indicate the corresponding perspectives as what we mentioned above. The identification of the dominant topic is not needed for this model.

Chapter 6

Results

In this chapter, we first evaluate the results obtained using Qualtrics as a tool to perform text analysis for the case study. Then, we assess the results of the state-of-the-art topic models based on the criteria as identified in Section 4.

6.1 Qualtrics text analysis for the case study

The approach of self-constructing topic queries based on the customer journey was used as described in Section 5.2. There are 41 topics defined: 21 topics are related to the illustrated customer journey, while the other 20 are constructed based on observations. The overview is as shown in figure 6.1. The total number of collected English responses for Q21 is 1767. Of these responses, 1190 (67%) satisfied one or more self-defined queries.

A query represents each topic in Qualtrics. Each query operates as a pattern matcher on survey responses. In this way, each survey response can be matched to no, one, or multiple queries. The pattern matching works as follows: each query is composed of a sequence of patterns that are connected via standard boolean operators (e.g., "OR", "AND", etc.). Some of the manually constructed topics are shown in figure 6.2. All the topic queries were constructed by combining the keywords from the business process identified by domain experts as well as manual pattern observations. For example, the topic query for 'Digital quoting tool' has ten components that are connected with the logic operator | (i.e.'OR') as shown in Figure 6.1, each containing a word group and an integer value (4) that specifies how many words can separate the words in the word group. This topic query would find all the responses that fulfill the condition. Some randomly selected responses based on this query can be seen in figure 6.3.

There are also 582 responses in this case not identified by any topic. Together with the domain expert, it was concluded that these responses are considered too general, hard to understand without context, or too specific to be mapped into any specific business process or topic. Some responses identified by this query can be seen as in figure 6.4.

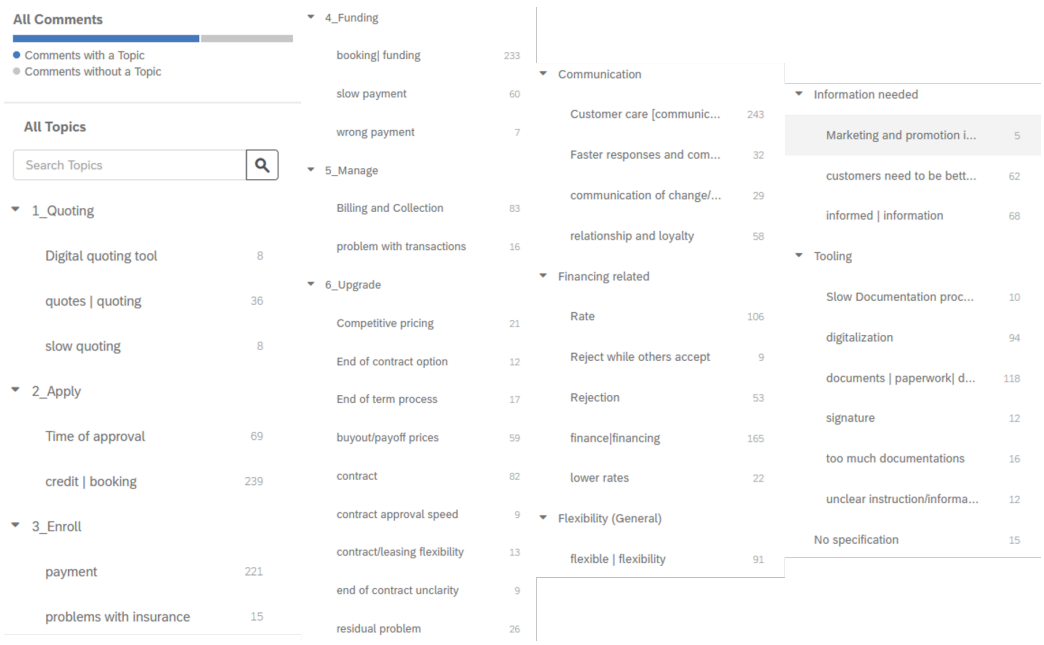


Figure 6.1: The overview of all the self-defined Qulatrics topics for the case study data

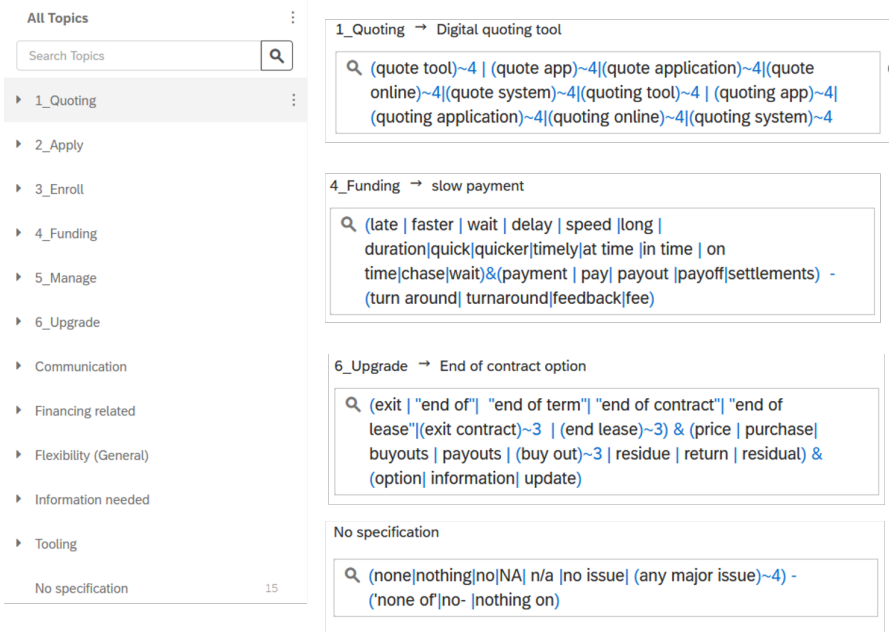


Figure 6.2: For Q21 (Explain the choice made for the selected improvement area): the left side with topics manually constructed based on the customer journey; the right side with four topic query

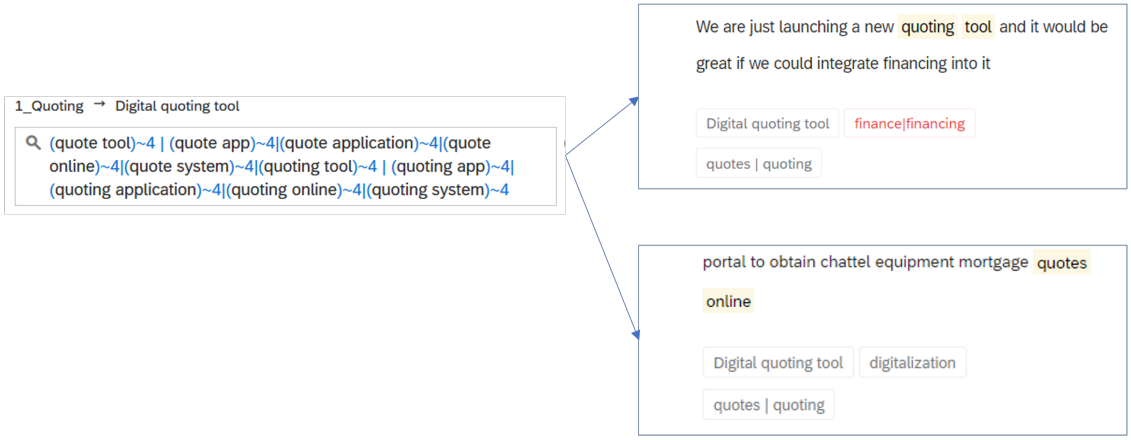


Figure 6.3: Some responses identified by the query defined for 'Digital quoting tool' from the case study data

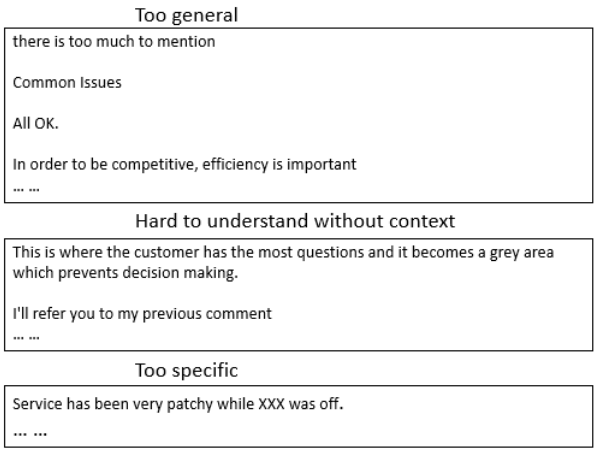


Figure 6.4: Some responses that cannot be identified by the constructed topics for the case study data

6.2 Topic quality

We evaluate the topic quality of the topic word lists obtained from topic modeling with different models. This evaluation was mainly conducted from two perspectives: topic coherence and interpretability.

6.2.1 Topic coherence

This section evaluates the topic coherence obtained from different models for each dataset, both quantitatively and qualitatively. Note that the results obtained from the Qualtrics analysis cannot be evaluated as the tool itself does not offer the possibility of conducting such a measure.

Quantitative observations on topic coherence score:

Let us first compare the topic coherence scores obtained when each model was tuned for a maximal score. An aggregate overview of these scores for each combination of model and dataset can be seen in Table 6.2. From this table, we observe several points: First, the topic coherence scores obtained by the SBERT and clustering topic model are the highest for each dataset. Second, the topic coherence scores obtained by the NMF topic model are the lowest for all the datasets. Third, the SBERT and clustering topic model applied to the BBCNews dataset obtained the highest topic coherence score (0.7528) among all the combinations of models and datasets.

Some remarks are needed to properly interpret the topic coherence scores for the Amazon clean and Amazon corrupted datasets. In the work of Röder, Both and Hinneburg [65], it was stated that the topic coherence measure Cv has the strongest correlation with human ratings in judging the topic coherence, as we earlier discussed in Section 2.4.4. A higher topic coherence score should indicate that the topic words have better support for each other within one topic. With this in mind, it is noted that the topic coherence scores obtained for the corrupted Amazon Fine Food Review dataset with 3.5% character-level typographical errors are all higher than that of the ones obtained with the clean data. This is unexpected since the introduction of typographical errors decreases human interpretability of the topics, and therefore should logically lead to a decrease in topic coherence score.

Now, let us point out another salient detail about the table. Specifically, compare the results for the BBCNews dataset with the Amazon corrupted dataset. On the one hand, we know that the BBCNews dataset is a coherent dataset as it is a labeled collection of formal news articles. On

Table 6.1: Number of total documents and tokens before and after preprocessing for all the datasets used in the project

	Total number of document	Total number of tokens
original $_{\text{case study}}$	1767	42457
after preprocessing $_{\text{case study}}$	1746	12116
original $_{\text{BBCNews}}$	2225	859740
after preprocessing $_{\text{BBCNews}}$	2225	223645
original $_{\text{Amazon}}$	2000	111287
after preprocessing $_{\text{Amazon\_clean}}$	2000	29206
after preprocessing $_{\text{Amazon\_corrupted\_7.5\%}}$	2000	26523
original $_{\text{student}}$	575	8043
after preprocessing $_{\text{student}}$	549	2636

the other hand, it is reasonable to assume that the Amazon corrupted dataset is a *less* coherent: it is composed of informal internet reviews corrupted with typographical errors. Thus, one would expect each model to achieve a higher coherence score for the BBCNew dataset than the Amazon corrupted dataset. However, the topics found by the LDA and NMF models for the latter are awarded a *higher* coherence score, which is a similar observation as in the previous paragraph and further supports the argument that the topic coherence score may fail to be robust when anomalies, such as typographical errors, are present in the data.

It can also be seen that the SBERT model classifies a substantial portion of the documents as outliers. We consider this to be double-sided: on the one hand, the disadvantage is underutilization of the dataset, while the advantage is that the model can cherry-pick data entries to ensure higher semantic similarity within clusters, possibly leading to more coherent topics.

Table 6.2: Highest topic coherence scores and the corresponding number of topics obtained from models each dataset

	company	BBC	Amazon clean	Amazon corrupted $_{7.5\%}$	student survey
LDA	0.5587	0.5045	0.5226	0.5744	0.5219
NMF	0.403	0.5023	0.3904	0.4725	0.4387
SBERT and clustering	0.5872	0.7528	0.6061	0.657	0.5223
number of topics_LDA	32	3	30	24	12
number of topics_NMF	38	10	48	44	46
number of topics_SBERT and clustering	19	5	24	18	10

Looking at Table 6.2, we will now provide a possible explanation for some of the differences in topic coherence scores between models.

First, we point out that the clustering model achieves the highest topic coherence score for all datasets. The clustering model identifies topics based on the semantic meaning of the documents, while the LDA and NMF models perform topic modeling based on generative probabilistic models and non-negative matrix factorization, respectively. Thus, documents within clusters from the cluster model are more likely to be coherent in terms of semantic meaning. In contrast, the topic word lists identified by the LDA and NMF models represent how a topic can be expressed as a (weighted) group of words without considering the semantic meaning of documents. Thus, the final topic coherence score calculated as the average of the topic coherence score per topic is higher than that of the LDA and NMF models.

Second, the NMF model produced overall the lowest coherence score. This observation came out as unexpected: according to work conducted by O'Callaghan, Greene, Conway *et al.* [88] Chen, Zhang, Liu *et al.* [12] and Albalawi, Yeap and Benyoucef [81], NMF models are considered to be able to deliver better topics than LDA models for short-text topic modeling. Nevertheless, our quantitative results obtained show otherwise for all datasets with various average bi-gram token lengths used for models, as shown in figure 6.5. Among all the datasets, none of them has achieved the highest topic coherence score with the NMF topic model. A possible explanation for this is that the previously mentioned works employed an evaluation metric different from the topic coherence score. The previous three works use measurement metrics such as document similarity and point-wise mutual information (which focus more on word co-occurrence) and machine learning performance metrics such as F1-score and recall. Instead, we use the evaluation metric C_v topic coherence score, which aims to simulate human judgment into its score. For this reason, it is

considered to be a more thorough and complete topic modeling evaluation metric according to [63]–[65]. However, as we observed earlier in this section when inspecting the results for the Amazon corrupted dataset, the C_v topic coherence score may not be a suitable metric for our usage scenario.

Considering the above observations which hinted at practical limits of the C_v topic coherence score, we additionally evaluated the topic coherence by looking at the correlation between the topic coherence scores obtained for a selection of topics per model and their domain expert rating given by domain experts for the company data as shown in Appendix B. The three-domain experts rate the coherence of the topic word lists as 'yes', 'no' , or 'somehow' , and we encoded them numerically to 1, 0, and 0.5, respectively. By fitting a linear correlation model to the resulting data points as shown in Figure 6.6, we see that the encoded average expert rating on the topic coherence slightly decreases as the topic coherence score increases. This trend is the same for all models. These results again indicate that the C_v topic coherence score may not be a good evaluation metric for the unstructured survey text responses. However, it should be mentioned that the results are based on a relatively small sample, namely, 15 topic word lists with 3 ratings for each word list. Furthermore, the negative correlation between topic coherence score and human judgment is reminiscent of a widely recognized phenomenon in the machine learning community, namely that the degree of explainability of a machine learning model decreases as its prediction accuracy increases [89].

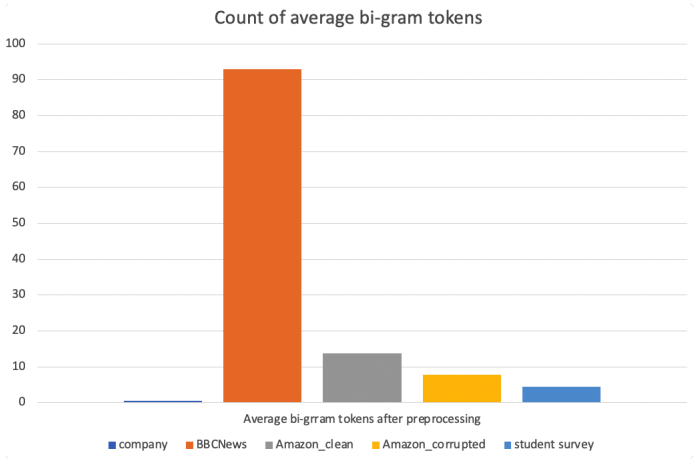


Figure 6.5: The count of average bi-gram tokens used by models per dataset

Besides the topic coherence scores, we can also see that the number of topics identified per model varied. The number of topics identified by the SBERT and clustering topic model is the lowest for almost all datasets apart from the BBCNews dataset. This may be due to their different theoretical base, although they are all considered unsupervised learning models. With the clustering model, for documents that do not consider semantically similar enough to be included in any cluster, they are considered outliers. The LDA topic model has the assumption that each document is a mixture of topics and each topic is a mixture of tokens/words. With the NMF topic model, each document is represented by a weighted sum of topics. For the LDA and NMF model, only empty documents after pre-processing would not be assigned with the dominant topic based on weights. Hence, this offers more potential to discover (more) latent topics from all the documents. It has also been observed that with the same dataset, the magnitude of the topic coherence score obtained from different models decreases when the optimized number of topics per model increases, although the changes of the topic scores within the same model per data do not follow such a trend. This may imply the trade-off in choosing a certain model based on the topic coherence. If one wants to discover as many topics as possible with a decent topic coherence, then the LDA topic model or NMF topic model is a better choice than the SBERT and clustering

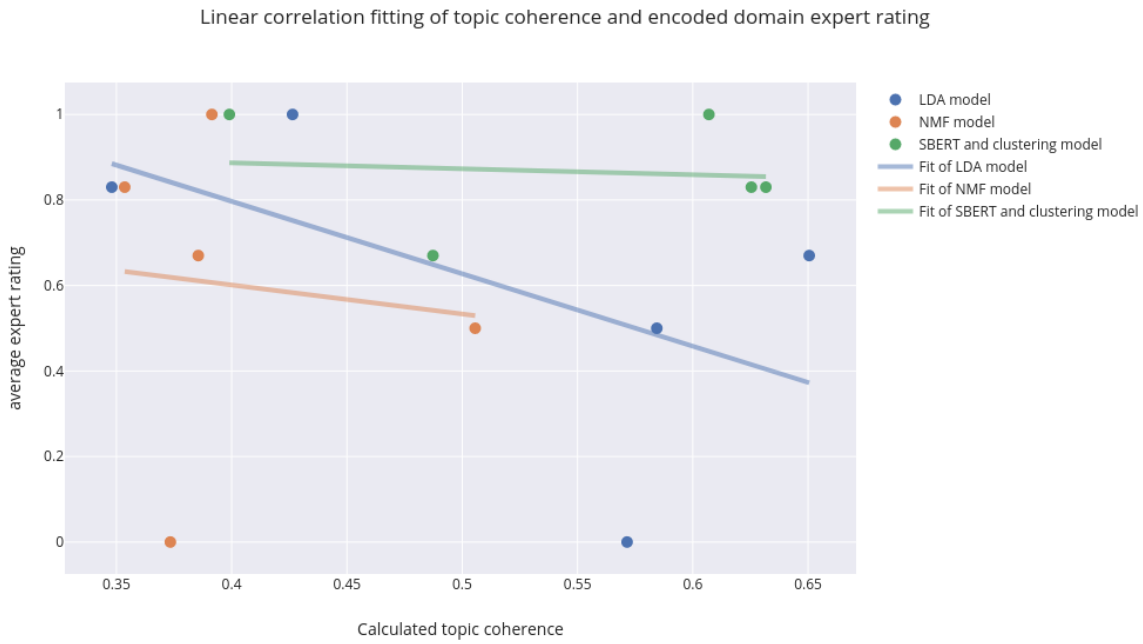


Figure 6.6: Correlation fitting between the calculated topic coherence scores and the encoded expert ratings

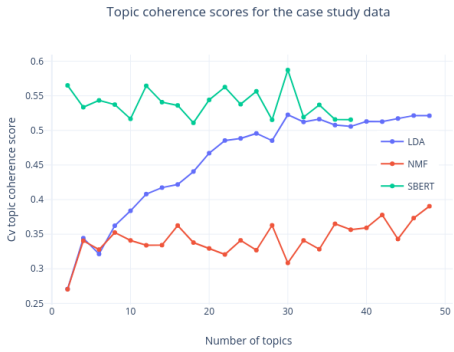
topic model and vice versa.

After comparing the absolute topic coherence scores and the number of topics between models and datasets, we now look at the trend for topic coherence scores. It was observed that the changes of topic coherence scores do not show obvious fluctuation but the small noise-like variation with the SBERT and clustering topic model for all the datasets, as shown in Figure 6.7. While the obtained results from other models per dataset show trends of decreasing. This may also be due to the characteristics of the clustering model that each document is only included in one cluster.

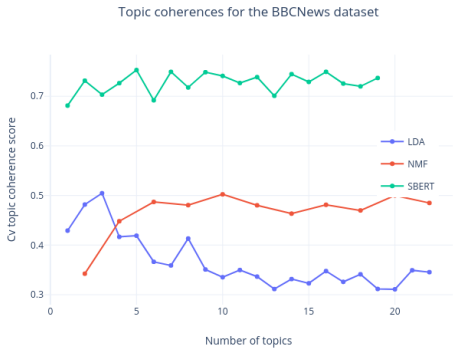
Qualitative observations of the extracted topics:

Among all the datasets used in the project, only the BBCNew dataset had labels assigned to each document, where each label indicates which of 5 topics best describes the document. Based on the fact that each document can be classified into these 5 topics, it is reasonable to expect that topic models find around 5 topics. However, the corresponding optimal number of topics obtained was 31 instead of 5 and there were also around 700 documents considered as outliers by the SBERT and clustering topic model. Moreover, we can see the inconsistent document numbers per topic/label if we compare the obtained topic distribution over the document set by manually setting the number of topics to 5 for the LDA model, as shown in figure 6.8b to the original topic distribution shown in6.8a.

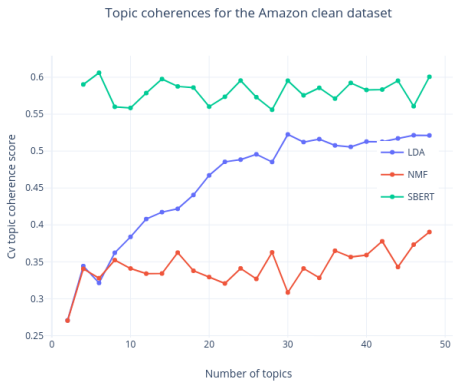
When combining this observation with the original textual document, we think such incoherence may result from the granularity of the topics captured by the topic models. For example, a document such as the one shown in Figure 6.9 which is news about Davos annual World Economic Forum (WEF) has a label "Business" which. While with interpretation from humans, "politics" is also a logical label that can be assigned to this document. In other words, the original document



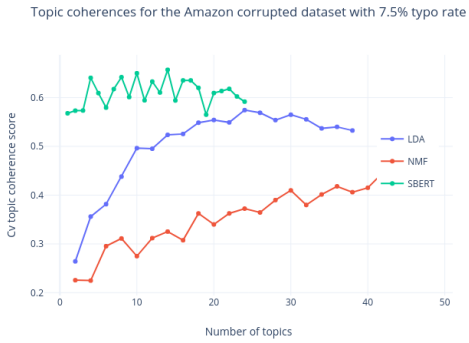
(a) Changes of topic coherence scores for the case study data



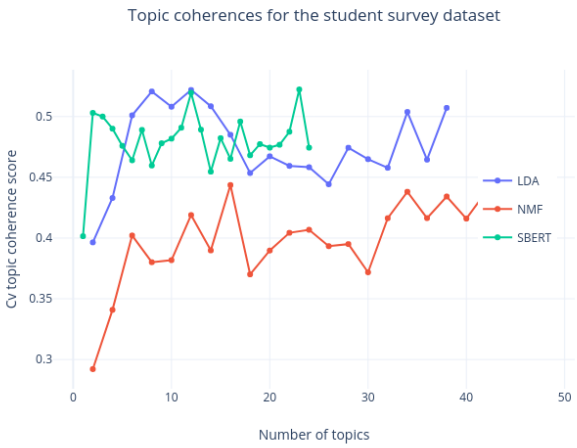
(b) Changes of topic coherence scores for the BBCNews dataset



(c) Changes of topic coherence scores for the clean Amazon Fine Food Review dataset

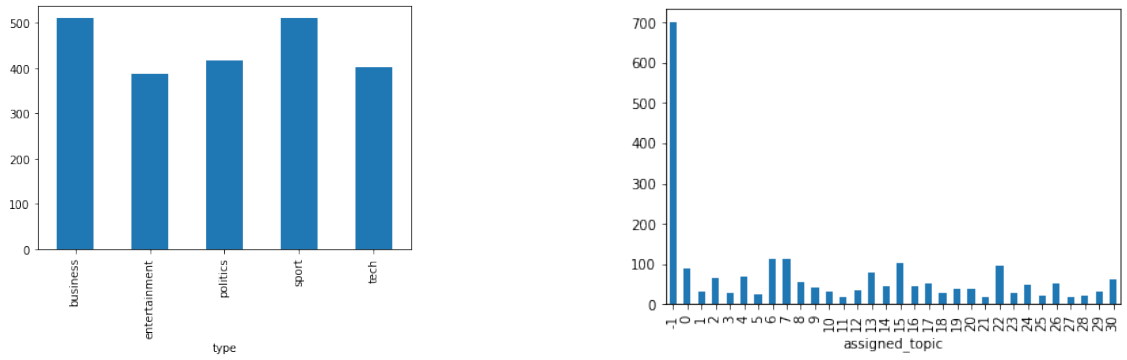


(d) Changes of topic coherence scores for the corrupted Amazon Fine Food Review dataset with 7.5% typographical error rate



(e) Changes of topic coherence scores for the student survey dataset

Figure 6.7: The change of C_v topic coherence scores corresponding to the number of topics identified for each datasets with different models



(a) Distribution of 5 original classes of the BBCNews dataset

(b) Results based on the SBERT and clustering model obtained for the BBCNews dataset

Figure 6.8: Topic distributions based on the given labels and the SBERT and clustering model for the BBCNews data

World leaders gather to face uncertainty More than 2,000 business and political leaders from around the globe are arriving in the Swiss mountain resort Davos for the annual World Economic Forum (WEF). For five days, they will discuss issues ranging from China's economic power to Iraq's future after this Sunday's elections... ...against globalisation, for fair trade, and many other causes, have promised to set an alternative agenda to that of the Swiss summit. .....Ultimately, the forum will be dominated by business issues - from outsourcing to corporate leadership - with bosses of more than a fifth of the world's 500 largest companies scheduled to attend. But much of the media focus will be on the political leaders coming to Davos, ... Microsoft founder Bill Gates, the world's richest man and a regular at Davos, will focus on campaigning for good causes, though business interests will not be ....

Figure 6.9: A document with a given label "Business" from the BBCNews dataset, while the topic word list is "people, user, service, firm, company, new, net, technology, software, system" based on its LDA topic modeling results.

may be identified with multiple topics besides the given one. We demonstrate this by visualizing how the topics that were identified by the LDA models are related to these labels. The visualization is a heatmap, shown in Figure 6.10, based on the count matrix T that was previously introduced for sensitivity analysis in Section 4.2. The vertical axis is the optimized LDA topic model where three topics were identified, the Some interesting observations can be made from this heatmap. First, we see that topics 1 and 2 are mainly associated with the labels "Sport" and "Entertainment", respectively, because the distribution of documents for these topics is mainly concentrated at those labels. Second, the LDA model has more difficulty in differentiating documents from the "Business", "Politics" and "Tech" labels. Although it would be more meaningful to select 5 topics with the LDA topic modeling for the heatmap. However, we did not make the heatmap for it due to limited time. But based on these observations, we think that a trade-off may exist between obtaining the most coherent topic words and discovering as many hidden topics as possible when choosing the most suitable topic model.

When observing the results obtained from the another student survey responses, we noticed the results were slightly different than that of the case study. The distribution of topics discovered by the LDA model shows that 96.9% of the documents were decided with the same dominant topic as shown in figure 6.11b. The corresponding topic words for it: "campus, session, lab, lecture, online, peer_review, review, question, live, use" can already offer the insight that the majority of students would like to have lab sessions at the campus, lecture online and peer review on campus.

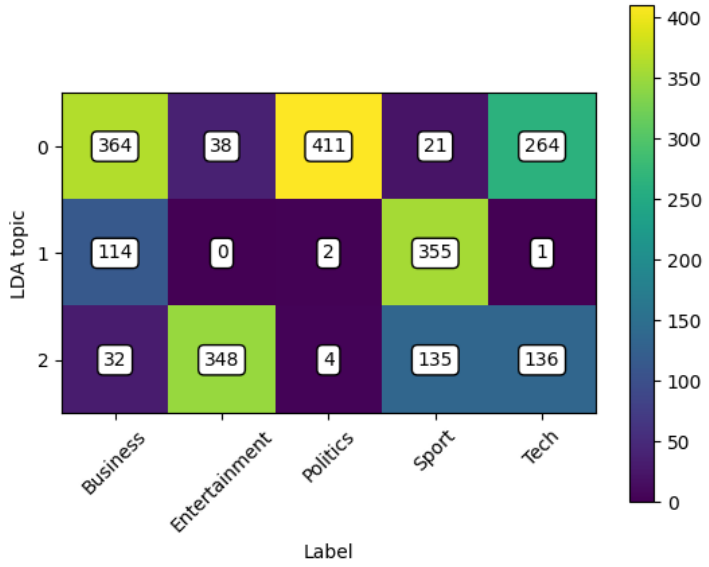


Figure 6.10: A heatmap showing the number of documents in the BBCNews dataset with their original labels (on the x-axis) and classified topics by the LDA model (on the y-axis)

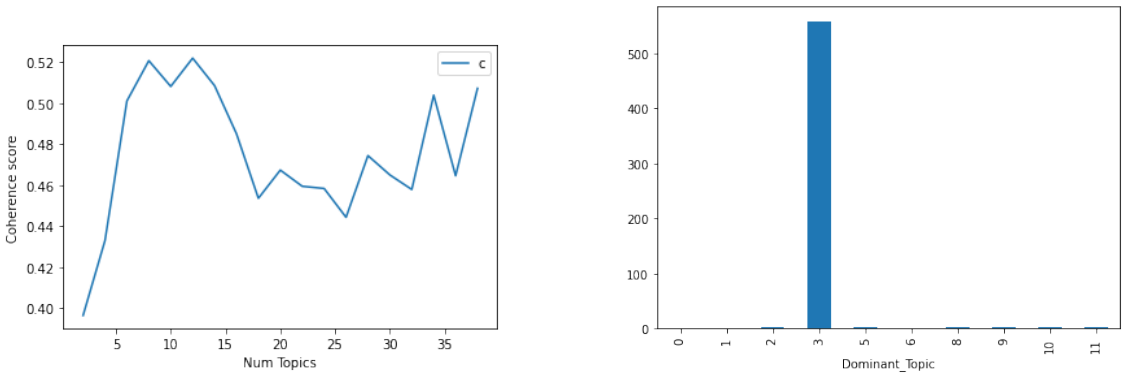
This interpretation is confirmed by us by looking at the original textual responses. Nevertheless, the same behavior was not observed for the company dataset which is also unstructured survey text responses. It may be resulted from how the survey question was formed. The student survey question was formed in a way that concise answers are meant to be answered while the company survey question was formed that widely spanned answers can be answered. Hence, the overall granularity of textual datasets collected was different.

6.2.2 Interpretability

The models are compared to see how explainable/transparent the algorithm is of the model's rationale for making decisions. The state-of-the-art models used in this project are based on different theoretical grounds. The LDA topic model is a generative probabilistic model, it is transparent in the sense that we know the process is based on the Dirichlet process. The NMF topic model conducts matrix factorization which reduces the dimensions of the Document-Term matrix once the number of total topics k is defined. The algorithms used in two models are explainable than that for SBERT and clustering model. Since in SBERT and clustering model, the clustering analysis is mainly based on the distance between encoded data points, there is no concisely defined mathematical model. Therefore, it is considered to be less transparent.

The topic word lists obtained are not always interpretable. Here, interpretability means that the user can understand certain behaviors under given situations [90]. The majority of topic modeling literature mainly focuses on evaluating the topic word lists obtained [12].

The feedbacks collected from the domain experts during the topic modeling results evaluation as shown in table 1 also implies that the topic coherence score calculated does not necessarily correlate with human judgment on topic coherence. There are several situations where it becomes problematic to interpret the results. First, when the topic word list obtained is too sparse, such as the ones for LDA_3 (speed, market, role, key, price, grey, offer, talk, environment, traditional) and



(a) Changes of topic coherence scores based on the number of topics

(b) Distribution of all topics identified over the document set

Figure 6.11: Results based on the optimized LDA topic model obtained for the student survey data

NMF_2 (dealer, treat, work, area, lender, frustrating, financing, end_user, figure, bad) as shown in Table 1. In this scenario, it is not even possible to identify a single theme. An example document with LDA_3 identified as its dominant topic would be: *"Speed is of the essence once agreed to start equipment financing in a new market."*. It can be observed that the first few words in this document have a good fit with the topic word lists of LDA_3, while some topic words of it do not seem to have any relation to this document.

Second, when the topic word list obtained includes multiple themes, such as pretrained_3 (online, user-friendly, friendly, navigate, requirements, user, documentation, docs, difficult, having): this example was considered by the domain expert to be ambiguous: it could be a topic indicating that the online portal needs to be designed more user-friendly, or that the documentation is difficult to understand. For example, for two documents both have this topic assigned as their dominant topics. Document A: *"Website is not user friendly..."* talks about the website while Document B: *"I like using the online submission but ... user friendly."* talks about the submission portal.

There are two more interpretability issues that were mainly observed when employing the SBERT and clustering topic model. The first issue observed specifically for the SBERT and clustering topic model is that repetition of words having similar meanings were observed, such as 'friend' and 'friendly', 'approval' and 'approvals'. This is related to the fact that the SBERT and clustering topic model does not require extensive pre-processing steps such as lemmatization and tokenization that are needed by the other two models. The second issue is that the clustering model identifies many more outlier documents compared to the other models. Outlier documents generally decrease the portion of the dataset that is given an interpretable result. Note that in contrast, with LDA and NMF models, a document is only identified as an outlier document when the length of the document is 0 after data cleaning and preparation. But with clustering, all the documents that are not considered semantically similar to any dense cluster based on distance metric measurement will be considered outliers. This result may be altered by tuning the hyperparameters such as the minimum size of clusters and minimum distance, but the process can be quite time and resource-demanding.

Nevertheless, our methodology used in the project enables a better interpretation of the results. We train the model based on all the identifiable bi-gram tokens and offer to connect all the documents to the identifiable topics for all the models. These improvements were also confirmed to be helpful by the domain experts in resolving the mentioned problematic occasions.

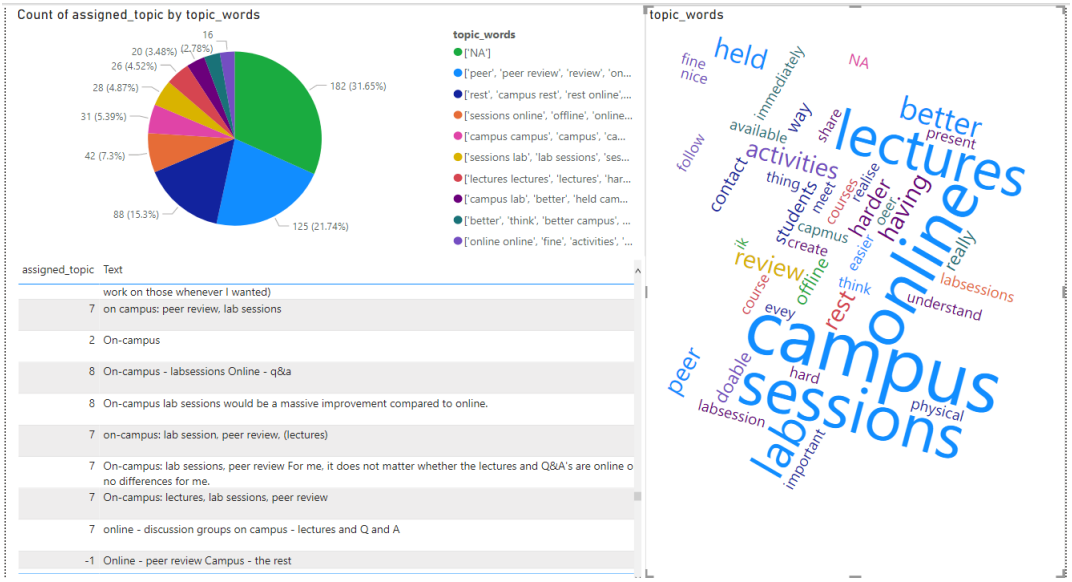


Figure 6.12: The overall impression of the visualization dashboard to present topic modeling results.

By utilizing the PowerBI tool, we created a dashboard for visualizing the topic modeling results. Figure 6.12 is an impression of the interactive visualization dashboard. The bar chart shows the composition of the text data based on the topics identified, the word cloud represents the topic words based on how often they are shown in the document, and the original text can also be shown. The dashboard is interactive, so the end-user can also change the contents based on their own selection, as shown in Figure 6.13 Such visualization is helpful since it allows the user to connect the abstract topic modeling results with the concrete natural language without losing a high-level overview of the whole documents. Additionally, it gives the end-user the possibility of gaining a better understanding of the original text. This was also recognized by the domain experts.

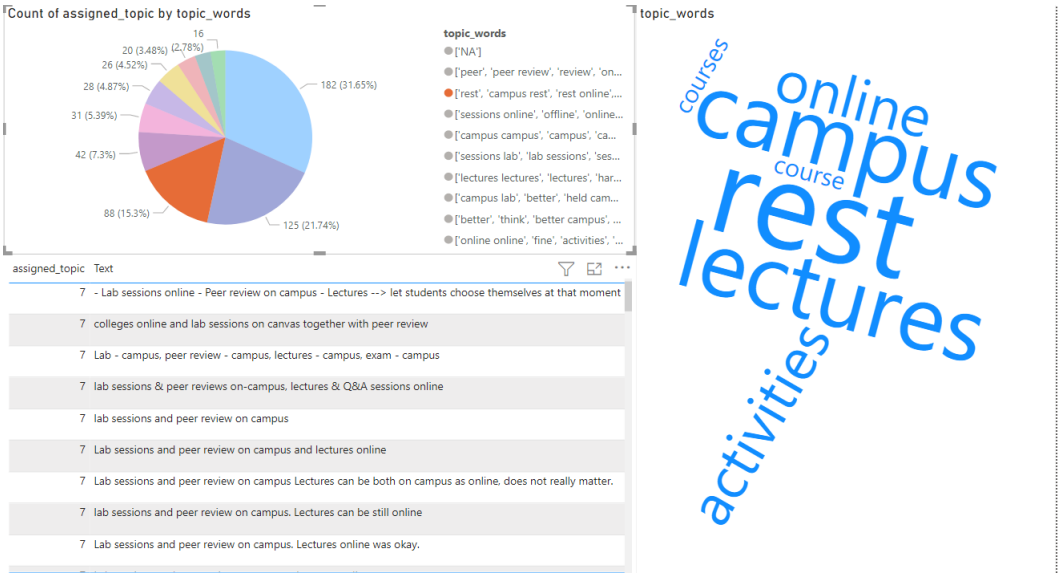


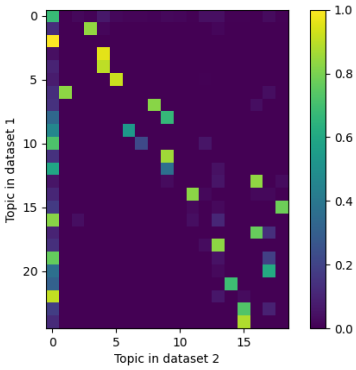
Figure 6.13: The visualization dashboard can present topic modeling results interactively based on the end-user choice.

6.3 Model sensitivity

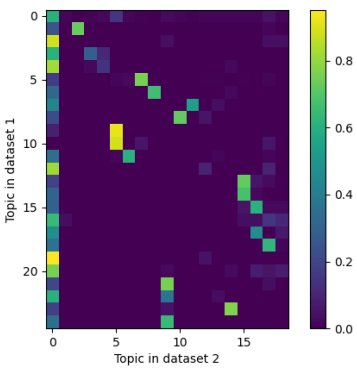
Model consistency	0% → 3.75%	3.75% → 0%	0% → 7.5%	7.5% → 0%
LDA	6%	33%	6%	32%
NMF	18%	19%	13%	14%
SBERT	60%	55%	48%	45%

Table 6.3: Comparison of the topic consistency between corrupted and clean Amazon Fine Food Review dataset. The arrows indicate the direction the topic consistency is measured between different percentages of character-level typographical errors.

The computed topic consistency for all models can be seen in table 6.3. As described in Section 4.2, the topic consistency is a percentage expressing the probability that two documents from the same topic in one topic modeling result are also in the same topic in the other topic modeling result. As can be seen from the table, the topic consistency monotonically decreases with an increase in the character-level typographical error rate. Large differences exist between the models. The topic consistency of the SBERT and clustering analysis model is the highest for all scenarios. The difference is largest between LDA and SBERT when considering the topic consistency from 0% to 3.75% character-level error: LDA achieves a consistency of only 6%, SBERT 10 times that 60%. To further illustrate the concept of topic consistency, Figure 6.14 contains two heatmaps for the SBERT model. Each heatmap represents the row-normalized count matrix T, i.e., $\tilde{T}$, (see Section 4.2) from which the topic consistency scores in the table can be derived. Note that the outlier topic is label 0, dataset 1 is the clean dataset and dataset 2 is the corrupted dataset. For example, in Figure 6.14a where dataset 2 is the corrupted dataset with 3.5% typographical error rate, the documents identified as topic 2 in dataset 1 (clean data) have a probability of 1.0 to be identified as topic 0 in dataset 2 (corrupted data). Moreover, we can see highlighted cells located diagonally in Figure 6.14a become more sparse in Figure 6.14b, and the column corresponding to the outliers (topic 0) has more highlighted cells. This implies that the topic consistency decreases when the typographical error rate in corrupted data increases from 3.5% to 7.5%.



(a) Topic consistency heatmap between dataset 1 (clean data) and dataset 2 (corrupted dataset with 3.5% character-level typographical error)



(b) Topic consistency heatmap between dataset 1 (clean data) and dataset 2 (corrupted dataset with 7.5% character-level typographical error)

Figure 6.14: Results based on the optimized LDA topic model obtained for the Amazon Fine Food Review datasets. The dataset 2 in both figures are the same clean Amazon Fine Food Review data.

Besides the topic consistency, the model sensitivity is explored in terms of topic coherence score and topic distribution. In this aspect, as we mentioned earlier in section 6.2.1, it was unexpected that the topic coherence scores obtained from the corrupted data were higher than the ones obtained from the clean data for all models. But different behaviors were observed per model type. When comparing the results of these two settings for the LDA model, both the topic coherence score as a function of the number of topics and the topic distribution showed large differences. The changes of topic coherence scores transitioned from increasing to stabilizing with the clean data, while a mild peak was observed with the corrupted data. For the clean data, the topic distribution shows a big concentration on topic 5 which contains nearly all the documents, while the topic distribution for the corrupted data is much more evenly distributed. In contrast, the NMF and SBERT and clustering topic models appear to be less sensitive to the data with different character-level typo rates. Both models show a similar trend for the changes of the topic coherence scores, although the absolute values may differ.

6.4 Practical aspects

In this section, we discuss the practical aspects as defined in Section 4.

First, the topic models have different requirements in terms of data pre-processing required. This is depicted in Table 5.1. It indicates that for LDA and NMF models, more data pre-processing steps are needed than for the SBERT and clustering model. Besides the pre-processing, an important practical aspect to keep in mind when applying the SBERT and clustering model is the potential under-utilization of documents in the dataset if they are classified as outliers. Thus, when documents in the dataset are composed of more than 1 topic, SBERT fails to model such a mixture of topics, whereas the LDA and NMF models inherently have the ability to assign multiple topics to documents.

Finally, we investigated in more detail what the advantages and disadvantages of the Qualtrics platform are in the context of topic modeling for unstructured survey responses. The major

advantages are as follows. First, Qualtrics is a commercial customer experience platform that offers convenient access to the original text data. Second, it includes powerful topic modeling results visualization functionalities. On the other hand, some disadvantages were observed. First, the recommended topics from the built-in text mining functionality only support uni-gram-based topic words. If N-gram-based topic words are desired, the user needs to put in additional effort to self-define queries. Second, the method powering the built-in text mining functionality is unknown. Third, if self-defined topics are required (for example because the built-in functionality does not provide satisfactory results), this requires considerable manual analysis and domain knowledge. Fourth, flexibility to self-define topic words is limited. For example, it is not possible to define a query to exactly fit a given pattern (e.g., matching "None" without matching "None of"). Fifth, it is currently not possible to integrate external topic models with the platform.

Chapter 7

Conclusion

In this chapter, the conclusions of this project are presented. First, the research goals are revisited and discussed whether and how they have been addressed. Then we illustrate the main contributions of the thesis. Finally, limitations and possible future work are presented.

7.1 Research questions

Recollect the main research goal defined in section 1.2: ***What are the differences between state-of-the-art topic models in the context of unstructured survey responses?***. This question was addressed by subdividing it into the following two sub research questions:

What are the differences between the topic modeling results obtained with several state-of-the-art topic models in terms of topic quality?

This sub research question was addressed in the following way: a potential evaluation metric for topic modeling, the C_v topic coherence score, was chosen. In the presented work from Lau, Newman and Baldwin [64], Chang, Gerrish, Wang *et al.* [63] and Röder, Both and Hinneburg [65], it was observed that, among all considered alternatives, the C_v topic coherence score has the highest correlation with human judgment. Although the primary target application of this study is in the context of unstructured survey responses, four additional datasets were included in the study to enable us to consider the results from a wider perspective as described in Section 3.5. Next, we calculate the topic coherence score for all combinations of topic models and datasets. The highest topic coherence score was determined by tuning the number of topics for the LDA and NMF models and the number of embedding dimensions for the pre-trained sentence embedding and clustering model.

We also included judgment on the topic quality from human domain experts for two reasons: to compare the ratings with the calculated topic coherence scores and to evaluate the difference in topic interpretability between the topic models. Unexpectedly, no positive linear correlation was observed between topic coherence score and human judgment for all models used on the company dataset, as shown in figure 6.6. On the other hand, 11 out of 15 selected topic word lists obtained from the company data were still identified as being somewhat coherent to coherent.

When comparing the interpretability of the topic modeling results from the SBERT and clustering model on the one hand and the LDA and NMF models, on the other hand, two issues negatively impacting the topic word list interpretability were observed for the SBERT and clus-

tering model: first, repetitions of words, such as 'approval' and 'approvals,' were frequently present in the topic word lists; and second, many documents were classified as outliers. The repetition was a result of the lack of data cleaning. The presence of outliers is due to the nature of clustering analysis based on similarity measurement.

Regarding the topic quality results for the unstructured survey responses from the company, the processing of survey responses by topic models presented in this work was recognized by the domain experts as a good start. By visualizing these results, it created convenience for the end-user to interpret the topic modeling results better. Specifically, the ability to categorize a document by its identified dominant topic and subsequently showing the user the initial responses enable a better interpretation of the obtained topics. Moreover, we demonstrate how topic modeling can be performed in Qualtrics by self-defining queries, the advantages and limitations of using Qualtrics for topic modeling were discussed.

What is the sensitivity of the topic modeling results concerning the language correctness of the input data?

Second, we conducted an evaluation of model sensitivity to data with typographical errors. Unstructured survey text responses are usually informally written with possible typographical errors. However, the impact of data with typographical error on the topic modeling results has not been studied before, according to our knowledge. One work related to sensitivity analysis studied the impact of topic cardinality on the topic coherence evaluation [82]. Because such a gap in the literature exists, we selected a dataset containing varying rates of character-level typographical errors to investigate topic model sensitivity. Specifically, we used the modified datasets based on the Amazon Fine Food Reviews dataset where Shah and de Melo [77] prepared clean text data and corrupted text data with varying character-level typographical error rates.

The sensitivity analysis consisted of two aspects. First, we applied our proposed model sensitivity estimator based on topic consistency between two topic modeling results, as shown in table 6.3. It was concluded that the SBERT and clustering topic model is significantly less sensitive to typographical errors than the LDA and NMF models.

For the second aspect of our sensitivity analysis, we observed the change in topic coherence score and topic distribution. These results further supported the observation obtained in figure 6.6 that the C_v topic coherence score is not well aligned with human judgment: the topic coherence scores obtained from the corrupted dataset were higher than that obtained from the clean dataset, which is not logical. Yet, the changes for topic coherence scores and the topic distribution over the document set obtained from different models demonstrated that the LDA topic model is the most sensitive to the corrupted data. In contrast, the other two models reacted little to the typographical error corruption.

What are the practical aspects of each topic model in the context of unstructured survey responses?

Regarding the practical aspects of the topic models, several conclusions can be drawn. First, there is a difference between the topic models regarding the work needed to pre-process the data. Second, the SBERT and clustering topic model showed a tendency to underutilize data by classifying many documents as outliers. In contrast, the topic coherence score indicated that the identified clusters are semantically similar. Third, the advantages and limitations of the Qualtrics platform were evaluated from a practical standpoint. Broadly speaking, the platform can produce topic modeling results with higher levels of granularity, with the trade-off that it requires higher levels of manual labor.

Based on the above, the following conclusions to the main research goal can be stated. First, the topic models considered in this empirical comparison have value as a starting point in the

context of unstructured survey responses: this was confirmed by domain experts on the company dataset. However, the granularity of results is limited because some topic word lists can be pretty abstract and require additional human effort to interpret and generate concrete insights. No one model with the highest topic quality can be identified among the topic models used in this project. However, some general observations were noticed that could be helpful pointers for future research. Namely, the NMF model tends to find more topics than the SBERT and clustering topic model. Moreover, the SBERT and clustering topic model usually finds topics with a higher C_v topic coherence score than the NMF and LDA models. On the other hand, several possible limitations of the selected C_v topic coherence score were demonstrated by comparing different combinations of models and datasets. Third, a sensitivity analysis based on a proposed consistency measure illustrated that the SBERT and clustering topic model is less sensitive to typographical errors than the LDA and NMF topic models.

7.2 Contributions

The contributions of this project are as follows:

- An empirical comparison of multiple state-of-the-art topic models in the context of unstructured survey responses. This comparison was supported with a wide variety of datasets. Other studies did not observe such diverse datasets (including unstructured survey responses, formal news articles, and product reviews with and without typographical errors). As a result, the results of this study offer more dimensions in empirical comparison.

- An extensive investigation of limitations of the C_v topic coherence score when applied to different types of datasets. Our investigation pointed out that the C_v topic coherence score has some flaws. For example, when adding typographical errors to a dataset, the C_v topic coherence score of the topic modeling results increased. Furthermore, the C_v topic coherence score was negatively correlated with domain expert judgment. In conclusion, we hope that our work will inspire additional research into robust topic coherence measurement.

- A sensitivity analysis based on a novel topic consistency measure was proposed. The topic consistency can be used for determining how sensitive a topic model is to perturbations in the input data, without requiring any ground truth labels. The basic principle is that consistency can be defined as the probability that if two documents were assigned the same topic at first, they were also assigned the same topic when the perturbation is applied.

7.3 Recommendation for company

The topic model choice is driven by the aspects that are important for the analysis. If the company is interested in finding many topics, it is recommended to use NMF modeling. However, it is not guaranteed that all identified topics would be meaningful in the company business context. On the other hand, if the company is interested in finding the most discussed and coherent topics, the SBERT model is recommended. However, the SBERT and clustering model does has the issue of under-utilizing data. This was observed for both survey response datasets as well as datasets consists of news or online reviews. If topics with good granularity in the company business context are required, then Qualtrics is recommended since the topics are defined in the context of the company business.

7.4 Discussion

The following limitations and suggestions for future work are identified:

- **additional topic models** In this project, we have implemented a variety of topic models for datasets with different characteristics. More advanced topic models were not considered as they could not be managed within the designated time frame.
- **threshold of forming bi-gram tokens** We chose to identify possible bi-gram tokens if two tokens were observed appearing together more than two times. This can be changed based on the need of the user. Such a change may affect the topic coherence and the optimal number of topics.
- **employing other topic coherence scores** Our results and evaluations of topic coherence obtained in Chapter 6 indicated that the choice of C_v topic coherence score to choose an evaluation metric that is close to how human evaluates was not successful. However, besides the C_v topic coherence score, many other measures exist in the literature [65]. Thus, it is worthwhile to investigate whether measures can be found that are more suitable in the context of unstructured survey responses.
- **ground truth labels** The topic modeling in this project is unsupervised. The motivation for including the labeled BBCNews dataset was to investigate whether topic models produce topics that are consistent with the given labels. It was not the case since it was noticed that some documents might appear to be eligible for several labels, as shown in Figure 6.10. This may imply that the identified topics are appropriate but not necessarily at the best level of granularity (which also explains why, in this particular example, only three topics instead of 5 were identified by the LDA model). To get a precise and sound conclusion on this, further work such as more detailed labeling making use of ontology or/and more domain experts involved in assessing the tropic modeling results would be helpful. an

Bibliography

[1] A. Griffin and J. R. Hauser, 'The voice of the customer,' *Marketing science*, vol. 12, no. 1, pp. 1–27, 1993 (cited on p. 1).

[2] K. M. Jackson and W. M. Trochim, 'Concept mapping as an alternative approach for the analysis of open-ended survey responses,' *Organizational research methods*, vol. 5, no. 4, pp. 307–336, 2002 (cited on pp. 1, 17).

[3] A.-M. Pothas, A. G. De Wet and J. M. De Wet, 'Customer satisfaction: Keeping tabs on the issues that matter,' *Total Quality Management*, vol. 12, no. 1, pp. 83–94, 2001 (cited on p. 1).

[4] M. B. Miles and A. M. Huberman, *Qualitative data analysis: An expanded sourcebook*. sage, 1994 (cited on pp. 1, 17).

[5] C. Taylor, *Structured vs unstructured data 101: Top guide*, Link, Aug. 2021 (cited on p. 1).

[6] W. Penfield and L. Roberts, *Speech and brain mechanisms*. Princeton University Press, 2014 (cited on p. 1).

[7] A.-S. Pietsch and S. Lessmann, 'Topic modeling for analyzing open-ended survey responses,' *Journal of Business Analytics*, vol. 1, no. 2, pp. 93–116, 2018 (cited on pp. 1, 2, 17).

[8] D. M. Blei, A. Y. Ng and M. I. Jordan, 'Latent dirichlet allocation,' *the Journal of machine Learning research*, vol. 3, pp. 993–1022, 2003 (cited on pp. 1, 2, 12).

[9] M. E. Roberts, B. M. Stewart, D. Tingley, C. Lucas, J. Leder-Luis, S. K. Gadarian, B. Albertson and D. G. Rand, 'Structural topic models for open-ended survey responses,' *American Journal of Political Science*, vol. 58, no. 4, pp. 1064–1082, 2014 (cited on pp. 2, 17).

[10] S. A. Cammel, M. S. De Vos, D. van Soest, K. M. Hettne, F. Boer, E. W. Steyerberg and H. Boosman, 'How to automatically turn patient experience free-text responses into actionable insights: A natural language programming (nlp) approach,' *BMC medical informatics and decision making*, vol. 20, pp. 1–10, 2020 (cited on pp. 2, 17).

[11] X. Yan, J. Guo, Y. Lan and X. Cheng, 'A biterm topic model for short texts,' in *Proceedings of the 22nd international conference on World Wide Web*, 2013, pp. 1445–1456 (cited on pp. 2, 17).

[12] Y. Chen, H. Zhang, R. Liu, Z. Ye and J. Lin, 'Experimental explorations on short text topic mining between lda and nmf based schemes,' *Knowledge-Based Systems*, vol. 163, pp. 1–13, 2019 (cited on pp. 2, 43, 48).

[13] D. Angelov, 'Top2vec: Distributed representations of topics,' 2020. arXiv: 2008.09470 [cs.CL] (cited on pp. 2, 12–14, 35–37).

[14] C. Shearer, 'The crisp-dm model: The new blueprint for data mining,' *Journal of data warehousing*, vol. 5, no. 4, pp. 13–22, 2000 (cited on pp. 2, 3).

[15] R. Feldman and I. Dagan, 'Knowledge discovery in textual databases (kdt).,' in *KDD*, vol. 95, 1995, pp. 112–117 (cited on p. 5).

[16] A. Hotho, A. Nürnberger and G. Paaß, 'A brief survey of text mining.,' in *Ldv Forum*, Citeseer, vol. 20, 2005, pp. 19–62 (cited on pp. 5, 7, 8).

[17] R. Feldman, J. Sanger *et al.*, *The text mining handbook: advanced approaches in analyzing unstructured data*. Cambridge university press, 2007 (cited on p. 5).

[18] V. Gupta, G. S. Lehal *et al.*, 'A survey of text mining techniques and applications,' *Journal of emerging technologies in web intelligence*, vol. 1, no. 1, pp. 60–76, 2009 (cited on p. 5).

[19] L. Kumar and P. K. Bhatia, 'Text mining: Concepts, process and applications,' *Journal of Global Research in Computer Science*, vol. 4, no. 3, pp. 36–39, 2013 (cited on p. 5).

[20] L. Dey, S. M. Haque, A. Khurdiya and G. Shroff, 'Acquiring competitive intelligence from social media,' in *Proceedings of the 2011 joint workshop on multilingual OCR and analytics for noisy unstructured text data*, 2011, pp. 1–9 (cited on p. 5).

[21] E. D. Liddy, 'Natural language processing,' 2001 (cited on pp. 5, 6).

[22] J. Hirschberg and C. D. Manning, 'Advances in natural language processing,' *Science*, vol. 349, no. 6245, pp. 261–266, 2015 (cited on pp. 5, 6).

[23] *Georgetown–ibm experiment*, Link, Dec. 2020 (cited on p. 5).

[24] A. Kao and S. R. Poteet, *Natural language processing and text mining*. Springer Science & Business Media, 2007 (cited on p. 6).

[25] J. Allen, *Natural language understanding*. Benjamin-Cummings Publishing Co., Inc., 1988 (cited on p. 6).

[26] A. K. Uysal and S. Gunal, 'The impact of preprocessing on text classification,' *Information processing & management*, vol. 50, no. 1, pp. 104–112, 2014 (cited on p. 6).

[27] B. E. M. Jones, 'Exploring the role of punctuation in parsing natural text,' in *Proceedings of the 15th Conference on Computational Linguistics - Volume 1*, ser. COLING '94, Link, Kyoto, Japan: Association for Computational Linguistics, 1994, pp. 421–425. DOI: 10.3115/991886.991960 (cited on p. 8).

[28] G. Salton and M. J. McGill, *Introduction to modern information retrieval*. mcgraw-hill, 1983 (cited on p. 9).

[29] X. Glorot, A. Bordes and Y. Bengio, 'Domain adaptation for large-scale sentiment classification: A deep learning approach,' in *ICML*, 2011 (cited on p. 10).

[30] R. Socher, C. C.-Y. Lin, A. Y. Ng and C. D. Manning, 'Parsing natural scenes and natural language with recursive neural networks,' in *ICML*, 2011 (cited on p. 10).

[31] O. Levy and Y. Goldberg, 'Neural word embedding as implicit matrix factorization,' *Advances in neural information processing systems*, vol. 27, pp. 2177–2185, 2014 (cited on p. 10).

[32] Y. Bengio, R. Ducharme, P. Vincent and C. Janvin, 'A neural probabilistic language model,' *The journal of machine learning research*, vol. 3, pp. 1137–1155, 2003 (cited on pp. 10, 11).

[33] T. Mikolov, K. Chen, G. Corrado and J. Dean, 'Efficient estimation of word representations in vector space,' *arXiv preprint arXiv:1301.3781*, 2013 (cited on pp. 10, 36).

[34] S. McDonald and M. Ramscar, 'Testing the distributioanl hypothesis: The influence of context on judgements of semantic similarity,' in *Proceedings of the Annual Meeting of the Cognitive Science Society*, vol. 23, 2001 (cited on p. 10).

[35] M. Sahlgren, 'The distributional hypothesis,' *Italian Journal of Disability Studies*, vol. 20, pp. 33–53, 2008 (cited on p. 10).

[36] Z. S. Harris, 'Distributional structure,' *Word*, vol. 10, no. 2-3, pp. 146–162, 1954 (cited on p. 10).

[37] M. Faruqui, J. Dodge, S. K. Jauhar, C. Dyer, E. Hovy and N. A. Smith, 'Retrofitting word vectors to semantic lexicons,' *arXiv preprint arXiv:1411.4166*, 2014 (cited on p. 11).

[38] T. Kenter and M. De Rijke, 'Short text similarity with word embeddings,' in *Proceedings of the 24th ACM international on conference on information and knowledge management*, 2015, pp. 1411–1420 (cited on p. 11).

[39] C. De Boom, S. Van Canneyt, T. Demeester and B. Dhoedt, 'Representation learning for very short texts using weighted word embedding aggregation,' *Pattern Recognition Letters*, vol. 80, pp. 150–156, 2016 (cited on p. 11).

[40] J. Devlin, M.-W. Chang, K. Lee and K. Toutanova, 'Bert: Pre-training of deep bidirectional transformers for language understanding,' *arXiv preprint arXiv:1810.04805*, 2018 (cited on p. 11).

[41] N. Reimers and I. Gurevych, 'Sentence-bert: Sentence embeddings using siamese bert-networks,' in *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, https://arxiv.org/abs/1908.10084Link, Association for Computational Linguistics, Nov. 2019 (cited on pp. 11, 12, 35, 87).

[42] A. W. Yu, D. Dohan, M.-T. Luong, R. Zhao, K. Chen, M. Norouzi and Q. V. Le, 'Qanet: Combining local convolution with global self-attention for reading comprehension,' *arXiv preprint arXiv:1804.09541*, 2018 (cited on p. 11).

[43] M. E. Peters, M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee and L. Zettlemoyer, 'Deep contextualized word representations,' *arXiv preprint arXiv:1802.05365*, 2018 (cited on p. 11).

[44] J. Howard and S. Ruder, 'Universal language model fine-tuning for text classification,' *arXiv preprint arXiv:1801.06146*, 2018 (cited on p. 11).

[45] S. R. Bowman, G. Angeli, C. Potts and C. D. Manning, 'A large annotated corpus for learning natural language inference,' *arXiv preprint arXiv:1508.05326*, 2015 (cited on p. 11).

[46] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser and I. Polosukhin, 'Attention is all you need,' in *Advances in neural information processing systems*, 2017, pp. 5998–6008 (cited on p. 11).

[47] A. Conneau, D. Kiela, H. Schwenk, L. Barrault and A. Bordes, 'Supervised learning of universal sentence representations from natural language inference data,' *arXiv preprint arXiv:1705.02364*, 2017 (cited on p. 11).

[48] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Céspedes, S. Yuan, C. Tar *et al.*, 'Universal sentence encoder,' *arXiv preprint arXiv:1803.11175*, 2018 (cited on p. 11).

[49] T. Hofmann, 'Probabilistic latent semantic analysis,' *arXiv preprint arXiv:1301.6705*, 2013 (cited on p. 12).

[50] W. Xu, X. Liu and Y. Gong, 'Document clustering based on non-negative matrix factorization,' in *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, 2003, pp. 267–273 (cited on p. 12).

[51] M. Grootendorst, *Bertopic: Leveraging bert and c-tf-idf to create easily interpretable topics.* Version v0.7.0, Link, 2020. DOI: 10.5281/zenodo.4381785 (cited on pp. 12–15, 35).

[52] K. Canini, L. Shi and T. Griffiths, 'Online inference of topics with latent dirichlet allocation,' in *Artificial Intelligence and Statistics*, PMLR, 2009, pp. 65–72 (cited on p. 13).

[53] S. Tsuge, M. Shishibori, S. Kuroiwa and K. Kita, 'Dimensionality reduction using non-negative matrix factorization for information retrieval,' in *2001 IEEE International Conference on Systems, Man and Cybernetics. e-Systems and e-Man for Cybernetics in Cyberspace (Cat. No. 01CH37236)*, IEEE, vol. 2, 2001, pp. 960–965 (cited on p. 13).

[54] L. Van der Maaten and G. Hinton, 'Visualizing data using t-sne.,' *Journal of machine learning research*, vol. 9, no. 11, 2008 (cited on p. 14).

[55] I. Jolliffe, 'Principal component analysis,' *Encyclopedia of statistics in behavioral science*, 2005 (cited on p. 14).

[56] L. McInnes, J. Healy and J. Melville, *Umap: Uniform manifold approximation and projection for dimension reduction*, 2020. arXiv: 1802.03426 [stat.ML] (cited on pp. 14, 36).

[57] R. B. Marimont and M. B. Shapiro, 'Nearest neighbour searches and the curse of dimensionality,' *IMA Journal of Applied Mathematics*, vol. 24, no. 1, pp. 59–70, 1979 (cited on p. 14).

[58] R. J. Campello, D. Moulavi and J. Sander, 'Density-based clustering based on hierarchical density estimates,' in *Pacific-Asia conference on knowledge discovery and data mining*, Springer, 2013, pp. 160–172 (cited on p. 14).

[59] L. Kaufman and P. J. Rousseeuw, *Finding groups in data: an introduction to cluster analysis*. John Wiley & Sons, 2009, vol. 344 (cited on p. 14).

[60] M. Ester, H.-P. Kriegel, J. Sander, X. Xu *et al.*, 'A density-based algorithm for discovering clusters in large spatial databases with noise.,' in *kdd*, vol. 96, 1996, pp. 226–231 (cited on p. 14).

[61] L. McInnes, J. Healy and S. Astels, *How hdbscan works*, Link, 2016 (cited on p. 14).

[62] C. Clifton, R. Cooley and J. Rennie, 'Topcat: Data mining for topic identification in a text corpus,' *IEEE transactions on knowledge and data engineering*, vol. 16, no. 8, pp. 949–964, 2004 (cited on p. 15).

[63] J. Chang, S. Gerrish, C. Wang, J. L. Boyd-Graber and D. M. Blei, 'Reading tea leaves: How humans interpret topic models,' in *Advances in neural information processing systems*, 2009, pp. 288–296 (cited on pp. 15, 16, 26, 27, 44, 55).

[64] J. H. Lau, D. Newman and T. Baldwin, 'Machine reading tea leaves: Automatically evaluating topic coherence and topic model quality,' in *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics*, 2014, pp. 530–539 (cited on pp. 16, 26–28, 44, 55).

[65] M. Röder, A. Both and A. Hinneburg, 'Exploring the space of topic coherence measures,' in *Proceedings of the eighth ACM international conference on Web search and data mining*, 2015, pp. 399–408 (cited on pp. 16, 26–28, 42, 44, 55, 58).

[66] D. Newman, J. H. Lau, K. Grieser and T. Baldwin, 'Automatic evaluation of topic coherence,' in *Human language technologies: The 2010 annual conference of the North American chapter of the association for computational linguistics*, 2010, pp. 100–108 (cited on p. 16).

[67] D. Mimno, H. Wallach, E. Talley, M. Leenders and A. McCallum, 'Optimizing semantic coherence in topic models,' in *Proceedings of the 2011 conference on empirical methods in natural language processing*, 2011, pp. 262–272 (cited on p. 16).

[68] R. A. Spreng, G. D. Harrell and R. D. Mackoy, 'Service recovery: Impact on satisfaction and intentions,' *Journal of Services marketing*, 1995 (cited on p. 17).

[69] A. R. Andreasen and A. Best, 'Consumers complain-does business respond,' *Harvard Business Review*, vol. 55, no. 4, pp. 93–101, 1977 (cited on p. 17).

[70] F. F. Reichheld and W. E. Sasser, 'Zero defeofions: Quoliiy comes to services,' *Harvard business review*, vol. 68, no. 5, pp. 105–111, 1990 (cited on p. 17).

[71] C. W. Hart, J. L. Heskett and W. E. Sasser Jr, 'The profitable art of service recovery.,' *Harvard business review*, vol. 68, no. 4, pp. 148–156, 1990 (cited on p. 17).

[72] H. Schuman and S. Presser, *Questions and answers in attitude surveys: Experiments on question form, wording, and context*. Sage, 1996 (cited on p. 17).

[73] D. Q. Nguyen, R. Billingsley, L. Du and M. Johnson, 'Improving topic models with latent feature word representations,' *Transactions of the Association for Computational Linguistics*, vol. 3, pp. 299–313, 2015 (cited on p. 17).

[74] Y. Zuo, J. Zhao and K. Xu, 'Word network topic model: A simple but general solution for short and imbalanced texts,' *Knowledge and Information Systems*, vol. 48, no. 2, pp. 379–398, 2016 (cited on p. 17).

[75] S. M. H. Dadgar, M. S. Araghi and M. M. Farahani, 'A novel text mining approach based on tf-idf and support vector machine for news classification,' in *2016 IEEE International Conference on Engineering and Technology (ICETECH)*, IEEE, 2016, pp. 112–116 (cited on p. 24).

[76] L. Q. Trieu, H. Q. Tran and M.-T. Tran, 'News classification from social media using twitter-based doc2vec model and automatic query expansion,' in *Proceedings of the Eighth International Symposium on Information and Communication Technology*, 2017, pp. 460–467 (cited on p. 24).

[77] K. Shah and G. de Melo, 'Correcting the autocorrect: Context-aware typographical error correction via training data augmentation,' in *Proceedings of the 12th Language Resources and Evaluation Conference (LREC 2020)*, Marseille, France, 2020 (cited on pp. 24, 28, 56).

[78] D. V. Carvalho, E. M. Pereira and J. S. Cardoso, 'Machine learning interpretability: A survey on methods and metrics,' *Electronics*, vol. 8, no. 8, p. 832, 2019 (cited on p. 27).

[79] H. M. Wallach, I. Murray, R. Salakhutdinov and D. Mimno, 'Evaluation methods for topic models,' in *Proceedings of the 26th annual international conference on machine learning*, 2009, pp. 1105–1112 (cited on p. 28).

[80] Y. Chen, H. Zhang, R. Liu, Z. Ye and J. Lin, 'Experimental explorations on short text topic mining between lda and nmf based schemes,' *Knowledge-Based Systems*, vol. 163, pp. 1–13, 2019 (cited on p. 28).

[81] R. Albalawi, T. H. Yeap and M. Benyoucef, 'Using topic modeling methods for short-text data: A comparative analysis,' *Frontiers in Artificial Intelligence*, vol. 3, p. 42, 2020 (cited on pp. 28, 43).

[82] J. H. Lau and T. Baldwin, 'The sensitivity of topic coherence evaluation to topic cardinality,' in *Proceedings of the 2016 conference of the North American chapter of the Association for Computational Linguistics: Human language technologies*, 2016, pp. 483–487 (cited on pp. 28, 56).

[83] *Documentation website for qualtrics text analysis: Text iq*, Link, Apr. 2021 (cited on p. 32).

[84] L. Logeswaran and H. Lee, 'An efficient framework for learning sentence representations,' in *International Conference on Learning Representations*, Link, 2018 (cited on p. 35).

[85] S. R. Bowman, G. Angeli, C. Potts and C. D. Manning, 'A large annotated corpus for learning natural language inference,' in *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, https://aclanthology.org/D15-1075 Link, Lisbon, Portugal: Association for Computational Linguistics, Sep. 2015, pp. 632–642. DOI: 10.18653/v1/D15-1075 (cited on p. 35).

[86] P. Kherwa and P. Bansal, 'Semantic n-gram topic modeling,' *EAI Endorsed Transactions on Scalable Information Systems*, vol. 7, no. 26, 2020 (cited on p. 36).

[87] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado and J. Dean, 'Distributed representations of words and phrases and their compositionality,' in *Advances in neural information processing systems*, 2013, pp. 3111–3119 (cited on p. 36).

[88] D. O'Callaghan, D. Greene, M. Conway, J. Carthy and P. Cunningham, 'Down the (white) rabbit hole: The extreme right and online recommender systems,' *Social Science Computer Review*, vol. 33, no. 4, pp. 459–478, 2015 (cited on p. 43).

[89] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao and J. Zhu, 'Explainable ai: A brief survey on history, research areas, approaches and challenges,' in *CCF international conference on natural language processing and Chinese computing*, Springer, 2019, pp. 563–574 (cited on p. 44).

[90] F. Sperrle, M. El-Assady, G. Guo, R. Borgo, D. H. Chau, A. Endert and D. Keim, 'A survey of human-centered evaluations in human-centered machine learning,' *Computer Graphics Forum*, 2021 (cited on p. 48).

[91] M. Honnibal, I. Montani, S. Van Landeghem and A. Boyd, *spaCy: Industrial-strength Natural Language Processing in Python*, Link, 2020. DOI: 10.5281/zenodo.1212303 (cited on p. 87).

[92] R. Rehurek and P. Sojka, 'Gensim–python framework for vector space modelling,' *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, vol. 3, no. 2, 2011 (cited on p. 87).

[93] L. McInnes, J. Healy, N. Saul and L. Grossberger, 'Umap: Uniform manifold approximation and projection,' *The Journal of Open Source Software*, vol. 3, no. 29, p. 861, 2018 (cited on p. 87).

[94] L. McInnes, J. Healy and S. Astels, 'Hdbscan: Hierarchical density based clustering,' *The Journal of Open Source Software*, vol. 2, no. 11, Mar. 2017, Link. DOI: 10.21105/joss.00205 (cited on p. 87).

Appendices

A Case study survey structure

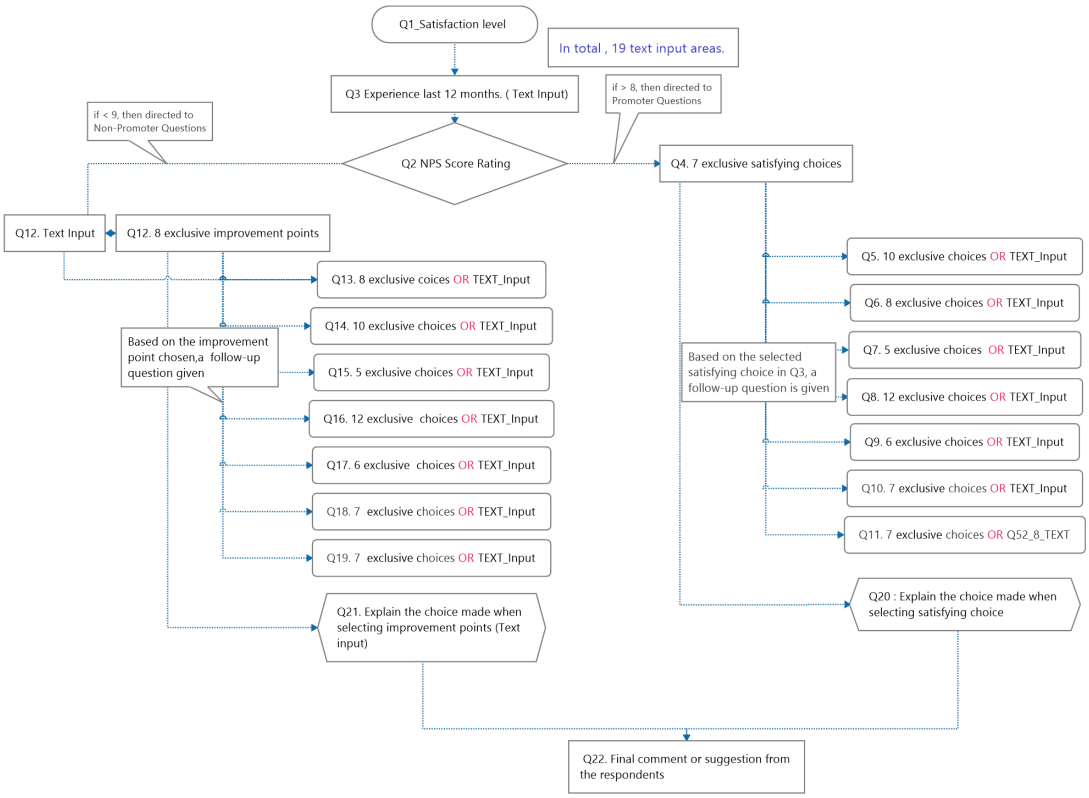


Figure 1: An example of the survey structure

B Topic coherence scores given from domain experts

Table 1: The calculated topic coherence scores and encoded domain experts rating for a selection of 5 topics per model

	Topic word lists	Calculated topic coherence score	Expert 1 rating	Expert 2 rating	Expert 3 rating	Average Expert rating
LDA_1	equipment, easy, option, financing, machine, vendor, money, purchase, flexible, end	0.4264	1	1	1	1
LDA_2	consistency, credit_approval, great, personnel, portal, cost, lender, bill, signer, detail	0.5844	1	0	0.5	0.5
LDA_3	speed, market, role, key, price, grey, offer, talk, environment, traditional	0.5715	0	0	0	0
LDA_4	partner, financial, loyalty, genuine, tax, comment, efficiency, strategy, sensitive, rest	0.6505	0	1	1	0.67
LDA_5	lease, new, end_lease, insight, integrate, mid, fact, moment, old, letter	0.3480	1	1	0.5	0.83
NMF_1	option, finance, lease, able, tool, end, end_lease, retail, inventory, use	0.3914	1	1	0.5	0.83
NMF_2	customer, documentation, month, condition, solution, wrong, phone, paperwork, email, experience	0.5056	1	0.5	0.5	0.67
NMF_3	rate, important, high, competitive, competition, lender, financing, current, sheet, increase	0.3734	1	0.5	1	0.83

NMF_4	dealer, treat, work, area, lender, frustrating, financing, end_user, figure, bad	0.3855	0	0	0	0
NMF_5	approval, condition, easy, paperwork, process, address, ready, client, sale, quick	0.3536	1	0.5	0.5	0.67
pretrained_1	communication, transition, thinking, little communication, better communication, outside, information, better, poor, high	0.3990	1	1	1	1
pretrained_2	credit, approval, approved, declined, credit box, box, need, denied, rate, deal	0.6254	1	0.5	1	0.83
pretrained_3	online, user friendly, friendly, navigate, requirements, user, documentation, docs, difficult, having	0.6317	0.5	1	1	0.83
pretrained_4	quickly, speed, deal, 30 faster, paid, quick, close, paid quickly, sales	0.6070	1	1	1	1
pretrained_5	approval, approvals, approval follow, getting, follow, structured, faster, process, approval time, udc	0.4873	0.5	0.5	1	0.67

C Complete topic modeling results for the BBCNews data

C.1 LDA topic model

The complete topic word lists obtained from LDA topic modeling:

Table 2: The complete topic word lists for the optimal LDA topic model obtained for the BBCNews dataset w

Topic number	Topic word list	Number of documents

0	people, 0.017446937 user, 0.010964086 service, 0.008972321 firm, 0.008774997 company, 0.008286212 new, 0.00792276 net, 0.0077514704 technology, 0.0075995373 software, 0.0074723586 system, 0.007089208	1098
1	game, 0.011190611 year, 0.009659261 match, 0.008944348 second, 0.008788273 minute, 0.008349499 time, 0.008265683 final, 0.008157449 side, 0.0080122575 goal, 0.0074958056 coach, 0.0072581833	472
2	game, 0.025611436 player, 0.014059047 mobile, 0.011984815 tv, 0.010701589 technology, 0.009800756 music, 0.009659491 new, 0.009532984 year, 0.009233816 time, 0.00887134 good, 0.008432053	655

C.2 NMF topic model

Table 3: Complete topic word lists obtained from NMF topic modeling results for the BBCNews dataset

Topic Number	Topic word list	Number of documents
0	0.033*people + 0.020*technology + 0.016*system + 0.015*net + 0.014*network + 0.014*firm + 0.014*site + 0.013*music + 0.011*file + 0.011*user,	151
1	0.018*sale + 0.017*law + 0.017*government + 0.014*case + 0.013*court + 0.009*right + 0.009*company + 0.008*deal + 0.008*decision + 0.008*russian,	508
2	0.031*market + 0.021*month + 0.020*dollar + 0.020*economy + 0.019*price + 0.016*analyst + 0.016*growth + 0.015*sale + 0.010*high + 0.009*profit,	265
3	0.043*firm + 0.039*company + 0.038*business + 0.023*minimum_wage + 0.015*wage + 0.014*job + 0.013*increase + 0.009*worker + 0.009*work + 0.008*employer,	0

4	0.036*people + 0.030*new + 0.024*party + 0.010*election + 0.009*time + 0.009*plan + 0.008*issue + 0.008*right + 0.007*way + 0.006*leader,	338
5	0.069*government + 0.016*tax + 0.011*year + 0.011*plan + 0.010*threat + 0.010*terrorist + 0.009*taxis + 0.008*cut + 0.008*budget + 0.008*election,	0
6	0.048*game + 0.026*good + 0.021*time + 0.014*player + 0.010*film + 0.008*title + 0.007*team + 0.007*child + 0.006*win + 0.005*point,	441
7	0.041*country + 0.027*number + 0.020*week + 0.020*job + 0.017*month + 0.014*world + 0.014*figure + 0.014*economy + 0.012*big + 0.010*woman,	116
8	0.094*year + 0.025*good + 0.022*song + 0.020*music + 0.018*award + 0.017*good_song + 0.011*british + 0.010*film + 0.009*sale + 0.009*band,	312
9	0.027*service + 0.026*tv + 0.022*mobile + 0.018*company + 0.014*people + 0.012*technology + 0.012*user + 0.012*phone + 0.009*device + 0.008*content]	94

C.3 SBERT and clustering model

Table 4: Complete topic lists obtained SBERT and clustering model BBCNews dataset

Topic Number	Topic word list	Number of documents
0	labour, blair, mr blair, brown, prime minister, prime, election, mr brown, campaign, minister	87
1	roddick, seed, open, australian, nadal, australian open, tennis, federer, henman, hewitt	31
2	kenteris, iaaf, doping, thanou, greek, conte, drugs, athletes, balco, olympics	66
3	olympic, indoor, race, holmes, championships, champion, athens, marathon, 60m, radcliffe	29
4	search, google, yahoo, blogs, web, jeeves, ask jeeves, desktop, ask, blog	67
5	games, gaming, nintendo, gamers, game, sony, xbox, ds, console, titles	25
6	box office, box, office, 8m, comedy, takings, 7m, fockers, starring, oscar	111
7	virus, security, spam, mail, software, users, site, spyware, windows, microsoft	113
8	album, band, song, music, rock, singer, urban, pop, chart, single	54
9	broadband, bt, tv, net, programmes, digital, service, content, internet, access	40
10	gadgets, gadget, technologies, devices, electronics, digital, portable, ces, technology, robot	32
11	phones, mobile, phone, mobiles, camera, mobile phone, multimedia, cameras, handsets, 3g	16
12	glazer, mr glazer, club, united, manchester united, board, manchester, wembley, proposal, bid	34

13	ireland, gara, italy, penalty, try, half, driscoll, england, irish, victory	78
14	rugby, wales, england, robinson, new zealand, zealand, nations, squad, lions, france	45
15	sri, countries, aid, lanka, sri lanka, aids, indonesia, poverty, hiv, tsunami	103
16	economy, growth, dollar, economic, rates, rate, exports, euro, rise, bank	45
17	kennedy, lib, mr kennedy, vote, dems, lib dems, party, elections, election, parties	51
18	yukos, oil, russian, gazprom, russia, rosneft, yugansk, gas, khodorkovsky, auction	29
19	liverpool, gerrard, parry, benitez, steven, anfield, club, league, madrid, champions league	37
20	chelsea, mourinho, arsenal, fa, rooney, football, united, mutu, referee, ferguson	38
21	sec, marsh, parmalat, insurance, financial, fannie, executives, spitzer, firms, fannie mae	18
22	ballet, musical, dancers, broadway, theatre, dame, spend spend, west end, premiere, spend	94
23	film, actress, actor, festival, oscar, awards, aviator, director, films, oscars	29
24	lse, boerse, deutsche boerse, deutsche, euronext, bid, takeover, exchange, offer, shareholders	48
25	gm, fiat, car, sales, bmw, profits, rover, cars, euros, nissan	22
26	air, airline, boeing, airbus, airlines, aircraft, airways, planes, ba, fuel	52
27	police, terror, human rights, lord, rights, suspects, human, lords, law, trial	17
28	id, id cards, cards, asylum, immigration, clarke, howard, tb, plans, tories	20
29	brown, mr brown, chancellor, budget, tax, election, stability, gordon brown, treasury, balls	30
30	pension, workers, councils, pensions, local, local government, strike, age, council tax, council	63

D Complete topics obtained for the clean Amazon fine food review dataset

D.1 LDA topic model

Table 5: Complete topics obtained with the LDA topic model for the clean Amazon Fine Food Review dataset

Topic number	Topic word list	Number of documents
0	'0.106*small + 0.105*much + 0.077*sure + 0.056*real + 0.053*ounce + 0.051*vanilla + 0.039*mustard + 0.024*side + 0.024*extra + 0.021*seed'),	12

1	'0.133*stuff + 0.054*candy + 0.042*old + 0.041*type + 0.033*color + 0.033*peanut + 0.030*lemon + 0.026*joint + 0.025*lime + 0.023*hip'),	29
2	'0.056*local + 0.049*gum + 0.046*oil + 0.032*steak + 0.028*change + 0.027*leave + 0.022*close + 0.021*addition + 0.018*pepper + 0.017*seasoning'),	32
3	'0.169*dog + 0.126*treat + 0.098*use + 0.046*easy + 0.028*bulk + 0.026*wrong + 0.025*honey + 0.020*warm + 0.019*thank + 0.018*peanut_butter'),	68
4	'0.090*pack + 0.072*gift + 0.044*new + 0.042*top + 0.035*mouth + 0.034*hour + 0.027*worry + 0.023*seaweed + 0.022*everyday + 0.021*clock'),	39
5	'0.139*good + 0.108*great + 0.101*flavor + 0.077*taste + 0.062*tea + 0.060*love + 0.045*time + 0.034*bag + 0.028*well + 0.027*nice'),	1079
6	'0.049*bitter + 0.038*star + 0.033*delivery + 0.030*reviewer + 0.029*little_bit + 0.026*reason + 0.024*cold + 0.024*idea + 0.022*sharp + 0.021*vitamin'),	21
7	'0.058*second + 0.052*amazing + 0.048*disappointed + 0.041*line + 0.036*pound + 0.033*bag + 0.026*stomach + 0.025*wheat + 0.023*soft + 0.022*trash'),	21
8	'0.152*box + 0.099*tasty + 0.092*fresh + 0.088*cookie + 0.048*salty + 0.046*whole + 0.021*home + 0.021*course + 0.015*rock + 0.014*bland'),	31
9	'0.168*day + 0.117*thing + 0.079*bottle + 0.059*blend + 0.057*kid + 0.052*happy + 0.051*hard + 0.035*cost + 0.028*half + 0.021*pill'),	30
10	'0.103*review + 0.070*several + 0.044*flavorful + 0.041*special + 0.037*daily + 0.029*wife + 0.027*area + 0.027*part + 0.026*pleased + 0.021*put'),	22
11	'0.412*product + 0.096*cat + 0.060*problem + 0.035*quality + 0.029*fruit + 0.023*opinion + 0.019*ginger + 0.016*personal + 0.014*senior + 0.012*recommend'),	45
12	'0.079*available + 0.045*dog_food + 0.039*noodle + 0.039*vegetable + 0.038*run + 0.030*heavy + 0.029*supplement + 0.028*cook + 0.025*condition + 0.023*alive'),	26
13	'0.123*strong + 0.103*mix + 0.056*organic + 0.041*first + 0.034*smell + 0.031*grocery_store + 0.028*convenient + 0.025*taste + 0.020*open + 0.019*hazelnut'),	25
14	'0.055*fine + 0.047*difference + 0.041*thought + 0.024*flower + 0.022*consistency + 0.022*ground + 0.022*turn + 0.021*simple + 0.020*clean + 0.019*quantity'),	29
15	'0.095*snack + 0.090*chocolate + 0.062*package + 0.053*free + 0.052*piece + 0.046*salt + 0.032*beef + 0.031*fat + 0.027*rich + 0.019*first_time'),	45
16	'0.085*milk + 0.060*smooth + 0.050*money + 0.037*stick + 0.033*baby + 0.029*red + 0.027*meat + 0.022*similar + 0.022*allergy + 0.022*swiss'),	32

17	'0.364*coffee + 0.063*different + 0.031*variety + 0.022*bold + 0.022*pot + 0.020*stash + 0.016*energy + 0.015*weak + 0.014*cappuccino + 0.014*latte'),	54
18	'0.214*food + 0.163*price + 0.056*cheap + 0.043*buy + 0.040*expensive + 0.024*try + 0.018*issue + 0.016*fast + 0.016*service + 0.013*female'),	46
19	'0.072*long + 0.061*hand + 0.055*right + 0.048*chicken + 0.047*amazon + 0.028*dark + 0.028*shipping + 0.024*result + 0.023*fish + 0.023*surprised'),	25
20	'0.193*little + 0.135*sweet + 0.095*excellent + 0.087*worth + 0.043*protein + 0.042*high + 0.023*glad + 0.020*mine + 0.020*pop + 0.020*sodium'),	18
21	'0.169*favorite + 0.052*can + 0.048*breakfast + 0.041*pancake + 0.041*buying + 0.040*butter + 0.024*grocery + 0.021*round + 0.021*shop + 0.019*dry_food'),	21
22	'0.129*chip + 0.081*people + 0.064*wonderful + 0.064*family + 0.048*light + 0.032*date + 0.025*health + 0.023*hungry + 0.021*white_tea + 0.018*homemade'),	26
23	'0.094*month + 0.057*wine + 0.046*container + 0.038*sauce + 0.033*market + 0.032*dollar + 0.026*aftertaste + 0.024*recipe + 0.024*replacement + 0.022*test'),	40
24	'0.084*delicious + 0.064*ingredient + 0.057*natural + 0.047*husband + 0.045*calorie + 0.043*cup + 0.036*hot + 0.031*green_tea + 0.029*juice + 0.025*machine'),	25
25	'0.113*many + 0.077*bar + 0.059*texture + 0.055*case + 0.050*kind + 0.050*packaging + 0.048*jar + 0.046*one + 0.017*tough + 0.017*version'),	33
26	'0.120*item + 0.074*quick + 0.050*diet + 0.036*dish + 0.032*son + 0.030*need + 0.030*fiber + 0.027*smoothie + 0.026*magnesium + 0.022*help'),	33
27	'0.058*purchase + 0.058*meal + 0.051*weight + 0.039*spicy + 0.039*full + 0.034*work + 0.029*less + 0.024*lunch + 0.024*mind + 0.022*eat'),	31
28	'0.105*order + 0.080*lot + 0.069*bad + 0.068*way + 0.042*drink + 0.042*week + 0.042*size + 0.035*last + 0.035*friend + 0.034*large'),	32
29	'0.099*sugar + 0.092*regular + 0.078*pod + 0.070*enough + 0.051*low + 0.044*cereal + 0.030*cheese + 0.026*thin + 0.024*content + 0.019*bear')]	30

D.2 NMF topic model

Table 6: Complete topics obtained from the NMF model for the clean Amazon Fine Food Review dataset

Topic number	Topic word list	Number of documents
0	0.180*bag + 0.036*fine + 0.021*open + 0.019*bad + 0.015*packaging + 0.015*syrup + 0.014*walnut + 0.014*black + 0.013*fruit + 0.013*chamomile,	8

1	0.078*bottle + 0.068*hair + 0.042*conditioner + 0.031*good + 0.030*thing + 0.024*heavy + 0.021*dry + 0.018*product + 0.017*smell + 0.013*squeeze,	26
2	0.212*coffee + 0.030*strong + 0.025*vanilla + 0.021*morning + 0.018*smooth + 0.017*drink + 0.016*favorite + 0.014*bold + 0.013*starbuck + 0.013*wonderful,	112
3	0.238*flavor + 0.043*different + 0.035*brand + 0.026*texture + 0.022*well + 0.016*favorite + 0.010*beef + 0.010*kid + 0.010*oil + 0.008*spicy,	11
4	0.043*work + 0.034*perfect + 0.022*item + 0.020*milk + 0.016*time + 0.015*worth + 0.013*room + 0.013*minute + 0.012*hand + 0.012*noodle,	69
5	0.090*one + 0.036*fresh + 0.025*t + 0.019*bad + 0.019*box + 0.018*dental + 0.017*lemon + 0.017*well + 0.017*trap + 0.015*soft,	28
6	0.264*coffee + 0.025*taste + 0.018*organic + 0.018*fresh + 0.017*pot + 0.017*regular + 0.015*bean + 0.014*coffee_maker + 0.014*packet + 0.014*weak,	6
7	0.054*plant + 0.048*gum + 0.020*half + 0.020*kind + 0.017*cheese + 0.015*piece + 0.015*leave + 0.015*review + 0.013*yellow + 0.012*money,	22
8	0.166*tea + 0.043*flavor + 0.020*light + 0.018*sugar + 0.016*wonderful + 0.015*excellent + 0.015*milk + 0.015*drink + 0.014*lady + 0.013*breakfast,	84
9	0.040*easy + 0.032*small + 0.025*extra + 0.023*package + 0.019*fruit + 0.018*worth + 0.017*high + 0.013*office + 0.013*eat + 0.013*meal,	64
10	0.583*good + 0.007*problem + 0.007*vegetable + 0.006*free + 0.005*friend + 0.005*salty + 0.005*warm + 0.004*weight + 0.004*hazelnut + 0.004*amazing,	1
11	0.244*tea + 0.057*love + 0.015*cup + 0.010*gift + 0.009*hot + 0.009*aroma + 0.008*loose_tea + 0.008*set + 0.008*different + 0.007*teapot,	5
12	0.140*cookie + 0.019*company + 0.016*fact + 0.015*different + 0.014*favorite + 0.014*gram + 0.014*special + 0.013*health + 0.013*sugar_free + 0.012*cost,	40
13	0.234*product + 0.018*sugar + 0.017*milk + 0.014*potassium + 0.014*grass + 0.013*mind + 0.013*quick + 0.012*energy + 0.011*container + 0.011*sodium,	75
14	0.261*product + 0.031*order + 0.011*reason + 0.010*packaging + 0.009*package + 0.009*jar + 0.009*fact + 0.008*amazon + 0.008*month + 0.007*sleep,	30
15	0.092*sweet + 0.035*sure + 0.032*bit + 0.032*t + 0.028*brand + 0.026*waffle + 0.021*store + 0.016*people + 0.016*sleep + 0.016*year,	55
16	0.135*food + 0.093*store + 0.018*love + 0.016*flavor + 0.013*sauce + 0.013*order + 0.012*cost + 0.011*expensive + 0.011*big + 0.010*variety,	67
17	0.136*food + 0.132*dog + 0.030*dog_food + 0.017*brand + 0.013*problem + 0.012*corn + 0.011*grain + 0.011*eat + 0.011*raw + 0.010*dry_food,	31

18	0.110*treat + 0.105*small + 0.096*dog + 0.027*chew + 0.022*size + 0.017*hard + 0.016*pound + 0.012*hand + 0.010*training + 0.009*expensive,	35
19	0.123*bar + 0.060*t + 0.020*meal + 0.019*kind + 0.019*protein + 0.014*low + 0.012*dark_chocolate + 0.012*nut + 0.011*flavor + 0.011*g_protein,	31
20	0.191*chocolate + 0.019*pop + 0.016*people + 0.013*s + 0.013*smell + 0.011*dark_chocolate + 0.011*tough + 0.011*experience + 0.010*product + 0.010*gift,	29
21	0.122*treat + 0.090*drink + 0.017*juice + 0.012*piece + 0.012*allergy + 0.010*beverage + 0.010*market + 0.010*milk + 0.009*vitamin + 0.009*disc,	38
22	0.069*cereal + 0.042*favorite + 0.029*pod + 0.028*review + 0.027*sugar + 0.025*bit + 0.024*taste + 0.017*milk + 0.017*cold + 0.016*hot,	42
23	0.102*water + 0.020*sweet + 0.017*pod + 0.013*coconut + 0.012*high_fructose + 0.011*corn_syrup + 0.011*thirsty + 0.010*body + 0.010*honey + 0.009*mio,	40
24	0.128*cat + 0.030*cereal + 0.029*love + 0.025*food + 0.021*way + 0.018*cat_food + 0.015*old + 0.015*lot + 0.013*sweet + 0.013*pesto,	34
25	0.276*love + 0.029*treat + 0.021*ingredient + 0.018*brand + 0.017*dog + 0.010*shampoo + 0.009*meal + 0.009*pop + 0.008*excited + 0.008*kid,	14
26	0.155*box + 0.063*cookie + 0.020*trap + 0.016*day + 0.014*order + 0.012*packaging + 0.011*pink + 0.009*husband + 0.009*store + 0.009*refund,	33
27	0.109*nice + 0.052*use + 0.030*way + 0.026*good + 0.018*butter + 0.016*natural + 0.015*stuff + 0.011*twining + 0.011*surprise + 0.011*tea,	45
28	0.038*honey + 0.029*well + 0.024*week + 0.022*personal + 0.021*antibiotic + 0.019*extra + 0.018*strong + 0.017*raw + 0.016*jar + 0.016*problem,	21
29	0.067*lot + 0.043*good + 0.042*milk + 0.035*butter + 0.018*well + 0.017*expensive + 0.015*oil + 0.014*hard + 0.012*easy + 0.011*different,	129
30	0.107*snack + 0.043*delicious + 0.037*seed + 0.029*little + 0.022*calorie + 0.022*well + 0.021*salt + 0.021*people + 0.020*kernel + 0.019*easy,	34
31	0.182*day + 0.020*great + 0.013*way + 0.012*mix + 0.012*chew + 0.012*new + 0.011*plain + 0.010*t + 0.010*nose + 0.009*rawhide,	36
32	0.065*use + 0.031*time + 0.026*salt + 0.019*pasta + 0.017*hot + 0.016*sauce + 0.014*recipe + 0.013*bit + 0.013*water + 0.013*cooking,	76
33	0.426*great + 0.008*sauce + 0.008*breakfast + 0.007*low + 0.007*date + 0.005*happy + 0.005*perfect + 0.005*way + 0.005*alternative + 0.005*hot_sauce,	17
34	0.178*flavor + 0.079*strong + 0.021*packet + 0.021*beverage + 0.020*good + 0.018*juice + 0.017*little + 0.011*black_cherry + 0.010*blend + 0.008*ounce,	65

35	0.085*ingredient + 0.064*natural + 0.033*snack + 0.031*organic + 0.027*fruit + 0.024*sugar + 0.022*allergy + 0.019*free + 0.019*healthy + 0.019*bar,	36
36	0.082*s + 0.047*tasty + 0.022*packaging + 0.020*quality + 0.016*product + 0.014*wonderful + 0.014*pouch + 0.010*incredible + 0.009*soup + 0.008*cheese,	56
37	0.199*price + 0.027*good + 0.025*year + 0.020*item + 0.016*trap + 0.013*shipping + 0.012*order + 0.012*case + 0.010*type + 0.009*low,	67
38	0.212*time + 0.065*year + 0.019*couple + 0.013*bread + 0.011*product + 0.011*eat + 0.011*oil + 0.009*kid + 0.007*mix + 0.007*month,	32
39	0.057*mix + 0.047*calorie + 0.034*gum + 0.031*level + 0.026*blood_sugar + 0.023*fat + 0.016*half + 0.015*insulin + 0.015*ginger + 0.015*drink,	34
40	0.162*tea + 0.109*green_tea + 0.074*twining + 0.035*variety + 0.025*caffeine + 0.021*black_tea + 0.019*color + 0.019*taste + 0.018*strong + 0.017*box,	7
41	0.060*large + 0.031*dog + 0.022*size + 0.021*lot + 0.019*chew + 0.016*snack + 0.015*calorie + 0.015*bone + 0.014*seed + 0.014*t,	43
42	0.302*taste + 0.019*vanilla + 0.016*drink + 0.014*sour + 0.012*bit + 0.011*artificial + 0.010*doesn + 0.010*sure + 0.009*pure + 0.009*salt,	1
43	0.039*soy + 0.035*chicken + 0.033*piece + 0.032*bit + 0.026*soup + 0.016*s + 0.016*possible + 0.014*chemical + 0.013*company + 0.013*box,	33
44	0.173*chip + 0.046*healthy + 0.021*pop + 0.019*salt + 0.017*corn + 0.016*life + 0.015*fat + 0.013*jalapeno_chip + 0.012*satisfy + 0.012*kettle,	29
45	0.168*bag + 0.025*easy + 0.024*sure + 0.015*wonderful + 0.013*way + 0.012*water + 0.011*chip + 0.010*minute + 0.009*trash + 0.009*portion,	36
46	0.098*little + 0.041*stuff + 0.034*brand + 0.029*well + 0.028*great + 0.022*thing + 0.018*ball + 0.016*pack + 0.016*grocery_store + 0.015*item,	68
47	0.127*taste + 0.056*bad + 0.029*thing + 0.022*big + 0.017*bite + 0.016*mix + 0.014*ginger + 0.012*bud + 0.011*lemon + 0.009*tasty]	104

D.3 SBERT and clustering model

Table 7: Complete topics obtained with SBERT and clustering model for the clean Amazon Fine Food Review dataset

Topic Number	Topic word list	Number of documents
0	cats, cat, food, cat food, eat, scratching, weight, pet, like, foods,	56
1	salt, popcorn, butter, table salt, popper, table, oil, sea, pops, kernels,	26

2	treats, dogs, dog, loves, treat, chews, chew, bone, small, size,	74
3	food, dog, dog food, dogs, dry, dry food, loves food, food great, eat, dog loves,	50
4	tea, teas, green, green tea, flavor, taste, drink, chai, drinking, box,	173
5	chips, chip, bag, jalapeno, potato, bags, pop chips, kettle, fried, love,	53
6	cereal, cereals, oatmeal, oats, quaker, breakfast, flakes, clusters, fiber, milk,	28
7	butter, peanut, peanut butter, pb2, mixer, ghee, almond butter, stirring, almond, lot easier,	18
8	coffee, roast, flavored, vanilla, strong, dark, coffees, bold, timothy, like,	24
9	coffee, brew, roast, cup, pod, strong, cup coffee, brewed, ounces, coffee know,	18
10	drink, water, energy, coconut, coconut water, juice, drinks, soda, flavor, beverage,	71
11	sugar, sweetener, blood sugar, use, blood, levels, product, insulin, baking, use product,	22
12	bars, bar, snack, kind bars, kind, protein, granola, nuts, meal, chocolate,	36
13	chocolate, dark chocolate, bar, dark, chocolates, candy, melt, like, gift, better,	46
14	cookies, cookie, free, gluten free, gluten, pretzels, delicious, chewy, sugar, sugar free,	41
15	candy, candies, loved, day, bought, party, brings, giant, girlfriend, gummy,	28
16	almonds, nuts, blue diamond, diamond, blue, delicious, pistachios, snack, bulk, cashews,	30
17	product, price, box, shipping, order, service, dented, good, received, ordered,	88
18	baby, old, food, month old, baby food, organic, veggies, son, month, vegetables,	21
19	snack, kids, snacks, loves, husband, kids love, love, great, sweet, grits,	41
20	sauce, chili, hot, seasoning, sauces, hot sauce, red, salsa,0 mix,	29
21	meat, seasoning, steak, teriyaki, salad, seaweed, dressing, chicken, blackened, salmon,	33
22	noodles, pasta, pastas, cooked, ramen, noodle, cook, rinse, quick, regular pasta,	22
23	soup, chicken, add, wolfgang, soy, soups, broth, extra, easy make, filling	17

E Complete topics obtained for the corruptedAmazon fine food review dataset with 3.5 character-level typographical error rate

E.1 LDA topic model

Table 8: Complete topic lists obtained for the corrupted Amazon Fine Food Review dataset with LDA topic model

Topic Number	Topic word list	Number of documents
0	0.039*mustard + 0.036*gum + 0.036*first + 0.025*warm + 0.023*aftertaste + 0.020*fast + 0.020*care + 0.018*hair + 0.018*daughter + 0.017*plain,	41
1	0.104*year + 0.049*cookie + 0.047*buy + 0.046*purchase + 0.043*able + 0.038*butter + 0.019*early + 0.016*first_time + 0.015*bite + 0.013*dust,	43
2	0.129*many + 0.090*bar + 0.034*choice + 0.027*bold + 0.019*dark_chocolate + 0.017*fur + 0.017*age + 0.014*supplier + 0.011*crisp + 0.011*oatmeal,	58
3	0.081*use + 0.073*perfect + 0.055*pod + 0.050*excellent + 0.044*healthy + 0.044*cup + 0.028*year_old + 0.018*maker + 0.017*cffee + 0.014*worry,	46
4	0.212*flavor + 0.110*good + 0.070*much + 0.069*day + 0.062*nice + 0.046*stuff + 0.024*real + 0.022*ounce + 0.019*happy + 0.018*blend,	108
5	0.215*tea + 0.133*time + 0.062*brand + 0.048*tasty + 0.031*pack + 0.020*fan + 0.016*green_tea + 0.016*husband + 0.014*huge + 0.010*least,	71
6	0.105*price + 0.089*small + 0.057*wine + 0.048*piece + 0.047*kid + 0.036*hand + 0.024*shape + 0.024*chicken + 0.016*reasonable + 0.012*short,	42
7	0.056*review + 0.047*case + 0.031*special + 0.027*meat + 0.025*area + 0.019*description + 0.017*gain + 0.016*help + 0.016*plenty + 0.015*chewy,	59
8	0.083*regular + 0.075*item + 0.068*delicious + 0.067*cheap + 0.033*dish + 0.027*diet + 0.024*seed + 0.023*mind + 0.023*single + 0.022*fiber,	54
9	0.068*natural + 0.024*less + 0.023*buying + 0.022*latte + 0.018*hungry + 0.016*bck + 0.016*half + 0.015*pill + 0.014*popcorn + 0.013*sleep,	70
10	0.288*taste + 0.094*order + 0.064*fresh + 0.037*smooth + 0.031*oil + 0.022*amazing + 0.018*replacement + 0.010*worth_price + 0.010*availble + 0.009*degree,	48
11	0.296*love + 0.092*favorite + 0.019*pound + 0.018*joint + 0.015*tha + 0.015*spot + 0.012*fod + 0.011*shelf + 0.011*hot_cold + 0.011*espresso,	58
12	0.285*product + 0.070*water + 0.041*several + 0.028*salt + 0.028*long + 0.024*bitter + 0.023*fruit + 0.013*sodium + 0.009*cold + 0.008*place,	69

13	0.185*food + 0.091*cat + 0.035*jar + 0.031*quick + 0.023*smell + 0.023*tough + 0.015*sour + 0.013*cat_food + 0.012*senior + 0.011*glass,	59
14	0.116*chip + 0.063*way + 0.051*wonderful + 0.034*friend + 0.033*light + 0.025*steak + 0.022*date + 0.019*stick + 0.014*corn + 0.013*individual,	59
15	0.133*dog + 0.101*treat + 0.076*mix + 0.033*thin + 0.015*theme + 0.013*thought + 0.012*bake + 0.012*cream + 0.011*greenie + 0.011*senseo,	63
16	0.055*packaging + 0.035*available + 0.034*flavorful + 0.024*find + 0.023*waste + 0.020*mushroom + 0.019*supplement + 0.017*stock + 0.015*simple + 0.012*terrific,	49
17	0.064*problem + 0.051*bottle + 0.050*people + 0.042*work + 0.014*course + 0.013*lunch + 0.012*sleepy + 0.011*last_year + 0.011*potassium + 0.008*poupon,	54
18	0.135*store + 0.095*bit + 0.039*machine + 0.034*last + 0.029*red + 0.021*black + 0.020*minute + 0.013*white + 0.012*need + 0.012*carry,	47
19	0.126*sweet + 0.075*package + 0.045*dry + 0.044*right + 0.043*fat + 0.037*breakfast + 0.028*baby + 0.016*recipe + 0.015*result + 0.014*surprised,	46
20	0.354*good + 0.086*bad + 0.033*high + 0.024*mouth + 0.017*ginger + 0.015*pay + 0.015*stash + 0.013*good_price + 0.012*bulk + 0.012*funny,	60
21	0.102*lot + 0.075*snack + 0.043*full + 0.037*money + 0.037*salty + 0.034*beef + 0.033*spicy + 0.030*pot + 0.021*color + 0.019*content,	44
22	0.030*fine + 0.027*top + 0.026*delivery + 0.025*wrong + 0.025*can + 0.021*addition + 0.018*blood + 0.016*close + 0.014*cook + 0.014*blue_diamond,	66
23	0.292*coffee + 0.060*ad + 0.036*drink + 0.036*week + 0.027*bean + 0.019*enjoy + 0.018*life + 0.017*tree + 0.016*ship + 0.012*plant,	65
24	0.064*worth + 0.052*meal + 0.039*big + 0.034*container + 0.032*difference + 0.028*sauce + 0.017*fish + 0.014*moist + 0.014*great_snack + 0.013*syrup,	53
25	0.085*box + 0.053*ingredient + 0.041*size + 0.038*large + 0.031*calorie + 0.026*stomach + 0.025*juice + 0.022*end + 0.019*daily + 0.019*honey,	42
26	0.058*chocolate + 0.035*milk + 0.031*organic + 0.028*soup + 0.026*fact + 0.021*vanilla + 0.021*star + 0.020*pancake + 0.018*market + 0.017*try,	65
27	0.049*cost + 0.043*quality + 0.035*opinion + 0.024*excited + 0.019*pasta + 0.018*tese + 0.017*world + 0.014*bear + 0.014*shipment + 0.014*think,	49
28	0.062*free + 0.046*cereal + 0.041*whole + 0.034*other + 0.022*seller + 0.020*folk + 0.013*drinking + 0.013*trash + 0.013*yogurt + 0.013*average,	52
29	0.121*little + 0.111*bag + 0.083*well + 0.064*sugar + 0.062*strong + 0.046*sure + 0.041*low + 0.032*one + 0.027*expensive + 0.015*weak,	56

30	0.088*month + 0.058*amount + 0.036*variety + 0.023*dollar + 0.021*next + 0.018*health + 0.016*change + 0.014*teriyaki + 0.014*vinegar + 0.013*teaspoon,	53
31	0.106*thing + 0.048*hard + 0.031*line + 0.027*cheese + 0.023*seaweed + 0.021*personal + 0.021*little_bit + 0.017*hot_spicy + 0.017*flower + 0.016*guy,	39
32	0.065*kind + 0.036*protein + 0.034*hour + 0.030*part + 0.027*new + 0.024*toy + 0.021*ball + 0.019*crunchy + 0.018*put + 0.013*alive,	50
33	0.333*great + 0.047*texture + 0.046*easy + 0.036*candy + 0.033*old + 0.017*extra + 0.015*outstanding + 0.014*similar + 0.014*low + 0.013*chew,	61
34	0.043*weight + 0.024*body + 0.023*issue + 0.022*open + 0.022*flvor + 0.020*cool + 0.015*customer + 0.014*emal + 0.013*perfect_size + 0.012*ground,	52
35	0.067*different + 0.054*family + 0.052*local + 0.050*hot + 0.049*gift + 0.046*enough + 0.029*type + 0.019*rich + 0.013*ths + 0.011*gas]	49

E.2 NMF topic model

Table 9: Complete topic word lists for the corrupted Amazon Fine Food Review dataset with NMF topic model

Topic number	Topic word list	Number of documents
0	0.075*cookie + 0.060*box + 0.017*time + 0.017*free + 0.014*year + 0.014*use + 0.012*company + 0.011*fact + 0.007*sugar + 0.007*broken,	32
1	0.032*seed + 0.030*light + 0.025*time + 0.024*lot + 0.018*eat + 0.017*large + 0.017*delicious + 0.017*kernel + 0.015*people + 0.012*type,	30
2	0.111*year + 0.029*able + 0.019*well + 0.012*problem + 0.012*ingredient + 0.009*work + 0.008*hair + 0.008*milk + 0.008*week + 0.008*minute,	53
3	0.069*item + 0.024*pod + 0.023*way + 0.019*coffee + 0.014*buy + 0.013*favorite + 0.013*nice + 0.012*light + 0.010*perfect + 0.009*great,	57
4	0.103*bag + 0.026*pack + 0.016*way + 0.015*open + 0.014*regular + 0.012*cookie + 0.011*problem + 0.010*half + 0.009*package + 0.007*broken,	35
5	0.110*strong + 0.033*sugar + 0.026*sweet + 0.019*favorite + 0.015*hair + 0.010*shampoo + 0.009*brand + 0.007*insulin + 0.007*blood_sugar + 0.006*year,	40
6	0.238*coffee + 0.017*packet + 0.017*morning + 0.012*cffee + 0.012*use + 0.011*love + 0.011*coffee_drinker + 0.011*recommend + 0.010*weak + 0.008*organic,	12
7	0.086*price + 0.058*store + 0.011*product + 0.009*hair + 0.009*trap + 0.009*cheap + 0.007*dry + 0.007*cost + 0.006*ad + 0.006*grocery_store,	88

8	0.082*day + 0.022*price + 0.018*way + 0.009*hour + 0.008*morning + 0.008*energy + 0.008*night + 0.007*case + 0.007*calorie + 0.007*tooth,	70
9	0.102*snack + 0.034*salt + 0.017*little + 0.015*sure + 0.014*healthy + 0.014*natural + 0.013*tasty + 0.013*kid + 0.012*thing + 0.011*hot,	43
10	0.182*love + 0.062*order + 0.026*chip + 0.013*brand + 0.012*perfect + 0.011*bean + 0.010*bag + 0.008*cost + 0.006*shampoo + 0.006*size,	45
11	0.139*little + 0.010*idea + 0.008*great + 0.008*particular + 0.007*tough + 0.007*roast + 0.007*coffee + 0.006*half + 0.006*bed + 0.005*instant,	48
12	0.081*bag + 0.035*water + 0.016*good + 0.015*enjoy + 0.013*small + 0.013*rinse + 0.013*pasta + 0.013*drain + 0.012*liquid + 0.011*noodle,	19
13	0.141*fruit + 0.037*sugar + 0.032*taste + 0.028*texture + 0.026*organic + 0.023*strawberry + 0.020*product + 0.017*ingredient + 0.016*s + 0.015*wonderful,	9
14	0.087*stuff + 0.033*well + 0.022*butter + 0.013*peanut + 0.012*package + 0.011*review + 0.010*store + 0.010*place + 0.010*tasty + 0.009*way,	50
15	0.091*great + 0.033*hot + 0.030*different + 0.020*brand + 0.011*chocolate + 0.010*big + 0.010*sauce + 0.009*package + 0.009*store + 0.008*bread,	132
16	0.206*taste + 0.011*sure + 0.009*favorite + 0.009*salt + 0.007*wine + 0.007*rich + 0.006*sour + 0.006*table_salt + 0.006*great + 0.006*bottle,	34
17	0.084*water + 0.018*product + 0.014*butter + 0.011*calorie + 0.011*bit + 0.011*tree + 0.010*lot + 0.008*drink + 0.007*people + 0.007*sure,	49
18	0.057*bag + 0.030*people + 0.020*honey + 0.019*juice + 0.019*calorie + 0.018*texture + 0.014*bit + 0.013*little + 0.010*d + 0.009*husband,	40
19	0.072*tea + 0.050*cat + 0.033*great + 0.032*quality + 0.020*high + 0.016*local + 0.013*cost + 0.013*sure + 0.009*happy + 0.008*smooth,	77
20	0.053*food + 0.043*family + 0.027*hot + 0.022*allergy + 0.016*dairy + 0.014*water + 0.011*cocoa + 0.008*delicious + 0.008*available + 0.008*seed,	73
21	0.056*ad + 0.025*salt + 0.017*honey + 0.017*well + 0.016*bitter + 0.014*symptom + 0.013*fine + 0.013*extra + 0.012*round + 0.011*jar,	49
22	0.155*dog + 0.051*small + 0.042*treat + 0.012*hard + 0.011*chew + 0.010*big + 0.010*hand + 0.009*piece + 0.009*size + 0.008*dry,	42
23	0.074*box + 0.067*s + 0.017*big + 0.017*valerian + 0.012*chip + 0.011*trap + 0.010*great + 0.009*potato + 0.009*good + 0.007*month,	51
24	0.112*treat + 0.044*one + 0.025*store + 0.016*fresh + 0.014*mini + 0.013*dental + 0.013*greenie + 0.010*time + 0.009*ingredient + 0.009*tese,	38

25	0.238*flavor + 0.013*different + 0.012*bit + 0.006*spicy + 0.006*brand + 0.005*well + 0.005*texture + 0.005*black_cherry + 0.004*recommend + 0.004*convenient,	15
26	0.046*food + 0.037*soup + 0.034*fresh + 0.027*soy + 0.022*chicken + 0.018*thought + 0.016*regular + 0.015*extra + 0.014*oil + 0.012*able,	26
27	0.131*tea + 0.089*green_tea + 0.039*twining + 0.024*variety + 0.022*black_tea + 0.021*color + 0.017*nice + 0.016*taste + 0.014*t + 0.013*light,	7
28	0.265*great + 0.035*love + 0.012*easy + 0.006*date + 0.006*pocket + 0.006*meal + 0.005*cream + 0.005*natural + 0.005*daughter + 0.005*piece,	1
29	0.213*coffee + 0.024*bean + 0.013*bitter + 0.013*big + 0.011*taste + 0.011*blend + 0.011*half + 0.009*bold + 0.008*ounce + 0.008*strong,	53
30	0.244*product + 0.006*quality + 0.006*packaging + 0.005*use + 0.005*happy + 0.004*quick + 0.004*grass + 0.004*mind + 0.004*return + 0.004*body,	63
31	0.083*cereal + 0.060*milk + 0.025*healthy + 0.020*good + 0.020*fresh + 0.018*cold + 0.015*lot + 0.015*taste + 0.014*old + 0.014*way,	25
32	0.196*good + 0.020*use + 0.019*lot + 0.019*well + 0.018*thing + 0.015*real + 0.013*pod + 0.010*fresh + 0.010*machine + 0.008*cup,	94
33	0.169*food + 0.021*healthy + 0.018*cat + 0.012*lot + 0.011*well + 0.011*year_old + 0.010*week + 0.009*dog_food + 0.009*choice + 0.008*cheap,	11
34	0.216*tea + 0.019*gift + 0.013*ice + 0.011*little + 0.010*small + 0.010*excellent + 0.009*black + 0.009*order + 0.008*teaspoon + 0.008*different,	3
35	0.399*good + 0.007*nice + 0.006*rice + 0.005*diet + 0.005*size + 0.004*green + 0.004*calorie + 0.004*sauce + 0.004*little_bit + 0.004*beef,	1
36	0.076*t + 0.069*sweet + 0.023*mix + 0.017*lot + 0.017*treat + 0.016*s + 0.013*cheese + 0.011*wine + 0.009*butter + 0.009*hour,	48
37	0.122*cat + 0.032*t + 0.019*bit + 0.016*formula + 0.016*old + 0.015*smell + 0.011*pavement + 0.011*love + 0.011*day + 0.010*energetic,	21
38	0.131*time + 0.016*jar + 0.011*honey + 0.009*couple + 0.008*good + 0.007*day + 0.006*perfect + 0.006*week + 0.006*symptom + 0.006*request,	46
39	0.177*tea + 0.037*bag + 0.024*bad + 0.024*nice + 0.023*box + 0.014*packaging + 0.014*sugar + 0.013*taste + 0.013*loose_tea + 0.012*brand,	12
40	0.155*bar + 0.034*meal + 0.027*bad + 0.022*fiber + 0.017*kind + 0.016*protein + 0.013*milk + 0.013*g + 0.012*low + 0.011*worth,	17
41	0.065*month + 0.020*small + 0.019*brand + 0.017*bar + 0.017*fat + 0.017*insulin + 0.016*blood_sugar + 0.013*use + 0.013*high + 0.012*level,	51

42	0.057*mix + 0.048*taste + 0.028*bit + 0.019*drink + 0.018*beverage + 0.017*light + 0.016*ginger + 0.013*strong + 0.010*bad + 0.010*flavorful,	54
43	0.082*easy + 0.016*lot + 0.013*small + 0.011*butter + 0.010*packet + 0.009*work + 0.008*couple + 0.008*hard + 0.008*eat + 0.007*extra,	51
44	0.106*chip + 0.043*gum + 0.038*flavor + 0.023*packet + 0.022*salt + 0.017*good + 0.013*regular + 0.013*pop + 0.009*piece + 0.009*oil,	33
45	0.083*drink + 0.071*flavor + 0.018*favorite + 0.016*fruit + 0.014*peach + 0.013*beverage + 0.009*sweetener + 0.009*juice + 0.008*nice + 0.007*line,	63
46	0.122*chocolate + 0.019*bar + 0.017*well + 0.016*wonderful + 0.013*thing + 0.011*tat + 0.008*dairy + 0.008*awful + 0.008*allergy + 0.008*blow,	34
47	0.032*time + 0.022*sour + 0.019*container + 0.015*warm + 0.011*sweet + 0.011*enjoy + 0.011*liquid + 0.010*rinse + 0.010*week + 0.010*drain]	49

E.3 SBERT and clustering model

Table 10: Complete topics obtained with SBERT and clustering model for the corrupted Amazon Fine Food Review dataset

Topic number	Topic word list	Number of documents
0	chips, chip, jalapeno,12 fried, bag, love, bags, potato, kale,	46
1	gum, gums, bears, gummy, cardamom, gummy bears, flavor, pieces, long, piece,	20
2	cats, cat, food, cat food, eat, bag, treats, like, scratching, dry,	56
3	dog, dogs, treats, food, treat, loves, dog loves, eat, dog food, size,	180
4	tea, teas, green, green tea, flavor, taste, drink, bags, really, nice,	178
5	coffee, flavored, bold, cup coffee, love coffee, cup, vanilla, favorite, strong, like bold,	27
6	coffee, packets, cup, make, bag, weak, cffee, box, aftertaste, weak strong,	18
7	cereal, cereals, oatmeal, oats, quaker, fiber, yogurt, cheerios, flakes, crunchy,	31
8	drink, juice, coconut, water, energy, butter, drinks, soda, beverage, product,	104
9	flour, pancake, waffle, gluten, pancakes, mix, mixes, gf, make, waffles,	16
10	chocolate, dark, chocolates, dark chocolate, candy, bar, like, melt, gift, hershey,	44
11	product, taste, tasted, used, like, quality, ve, box, brand, bought,	44
12	product, price, order, amazon, store, happy, good, buying, purchase, bought,	125

13	sauce, hot, chili, hot sauce, pepper, seasoning, salsa, sauces, heat, red,	24
14	soup, pasta, noodles, noodle, chicken, cook, beef, pastas, soups, rinse,	37
15	bars, almonds, nuts, bar, snack, kind bars, kind, protein, good, blue,	58
16	snacks, snack, sweet, great, son, just, loves, delicious, eat, love,	66
17	cookies, free, cookie, gluten, gluten free, chewy, chocolate, allergies, sugar, murray	38

F Complete topics obtained for the student survey data

F.1 LDA topic model

Table 11: Complete topic word lists with LDA topic model for the student survey data

Topic Number	Topic words list	Total responses
0	0.251*easy + 0.081*video + 0.043*time + 0.037*computer + 0.026*university + 0.023*sharing_screen + 0.021*topic + 0.020*quality + 0.016*comunicate + 0.016*rewatched	1
1	0.283*offline + 0.061*effective + 0.048*education + 0.048*environment + 0.041*good + 0.028*previous + 0.014*generation + 0.014*correction + 0.014*guess + 0.014*professor	1
2	0.286*course + 0.133*work + 0.059*main + 0.048*helpful + 0.029*format + 0.028*able + 0.025*study + 0.017*youtube + 0.013*canvas + 0.010*excercize	3
3	0.207*campus + 0.199*session + 0.185*lab + 0.182*lecture + 0.091*online + 0.061*peer_review + 0.038*review + 0.008*question + 0.003*live + 0.002*use	557
4	0.198*fine + 0.122*opinion + 0.109*think + 0.069*help + 0.038*laptop +0.033*place + 0.027*none + 0.027*general + 0.020*weekly + 0.016*canva	0
5	0.188*well + 0.163*rest + 0.085*exam + 0.079*people + 0.069*nice + 0.049*possible + 0.046*thing + 0.037*sure + 0.035*home + 0.019*problem	1
6	0.127*tutor + 0.094*hybrid + 0.083*much + 0.067*fellow + 0.062*trial + 0.036*team + 0.025*hour + 0.018*feedback + 0.018*suit + 0.018*repeat	2
7	0.139*peer + 0.069*necessary + 0.060*top + 0.054*lecturer + 0.051*motivate + 0.037*test + 0.033*lectur + 0.033*improvement + 0.033*massive + 0.027*optional	0
8	0.228*labsession + 0.199*student + 0.118*group + 0.022*element + 0.022*less + 0.022*unclear + 0.022*distraction + 0.022*busy + 0.022*mess + 0.017*member	2

9	0.121*way + 0.072*other + 0.047*preference + 0.034*poster + 0.031*information + 0.029*satisfied + 0.029*face + 0.027*presentation + 0.021*setting + 0.020*appropriate	2
10	0.227*programming + 0.145*activity + 0.110*theory + 0.089*exercise + 0.057*person + 0.032*part + 0.029*available + 0.019*recording + 0.017*discussion + 0.014*efficient	3
11	0.107*doable + 0.105*real_life + 0.059*personal + 0.055*little + 0.053*bond + 0.040*clear + 0.039*useful + 0.033*lot + 0.031*interaction + 0.022*individual	2

F.2 NMF topic model

Table 12: Complete topics obtained with NMF topic model for the student survey data

Topic Number	Topic word list	Total responses
0	0.157*well + 0.140*review + 0.087*session + 0.086*theory + 0.071*lab + 0.061*work + 0.050*campus + 0.027*fine + 0.025*opinion + 0.022*real_life,	46
1	0.090*student + 0.074*home + 0.073*online + 0.059*opinion + 0.053*nice + 0.046*fine + 0.046*programming + 0.019*way + 0.018*poster + 0.018*physical,	22
2	0.349*campus + 0.151*session + 0.065*lab + 0.038*peer_review + 0.025*easy + 0.025*problem + 0.016*screen + 0.014*laptop + 0.014*helpful + 0.013*programming,	27
3	0.470*online + 0.084*activity + 0.058*course + 0.024*offline + 0.021*bit + 0.016*people + 0.015*change + 0.014*education + 0.013*previous + 0.011*test,	10
4	0.301*peer_review + 0.150*session + 0.129*online + 0.124*lab + 0.106*lecture + 0.022*offline + 0.012*effective + 0.010*test + 0.010*team + 0.008*person,	55
5	0.242*well + 0.211*lecture + 0.052*people + 0.046*way + 0.029*great + 0.027*student + 0.026*bit + 0.022*opinion + 0.018*study + 0.018*difficult,	24
6	0.493*lecture + 0.234*campus + 0.033*review + 0.033*rest + 0.025*lab + 0.022*offline + 0.010*watch + 0.009*thing + 0.009*nice + 0.006*session,	29
7	0.403*lab + 0.397*session + 0.096*review + 0.015*fine + 0.011*fun + 0.007*sharing + 0.007*hard + 0.005*peer + 0.004*way + 0.004*small,	7
8	0.120*offline + 0.066*student + 0.048*review + 0.047*lab + 0.046*question + 0.042*nice + 0.041*session + 0.032*follow + 0.030*work + 0.030*people,	29
9	0.321*campus + 0.157*lab + 0.148*online + 0.119*session + 0.039*nice + 0.020*fine + 0.016*possible + 0.011*group + 0.010*work + 0.010*lecture,	25
10	0.180*easy + 0.121*lecture + 0.070*session + 0.047*lab + 0.025*video + 0.024*time + 0.022*review + 0.018*hybrid + 0.016*good + 0.015*information,	55

11	0.252*course + 0.044*activity + 0.041*possible + 0.031*fine + 0.028*helpful + 0.025*internet + 0.024*case + 0.024*canva + 0.019*study + 0.017*self,	22
12	0.064*clear + 0.063*session + 0.062*theory + 0.053*useful + 0.052*lab + 0.051*activity + 0.033*course + 0.032*group + 0.030*real_life + 0.029*rest,	79
13	0.074*education + 0.057*campus + 0.052*previous + 0.042*student + 0.034*exam + 0.031*time + 0.031*able + 0.027*computer + 0.027*quality + 0.027*discussion,	6
14	0.158*peer + 0.104*online + 0.095*rest + 0.093*campus + 0.035*format + 0.020*interaction + 0.019*easy + 0.019*time + 0.019*place + 0.019*ask,	37
15	0.184*campus + 0.114*labsession + 0.048*way + 0.043*tutor + 0.041*programming + 0.040*group + 0.040*peer_review + 0.035*theory + 0.030*video + 0.028*question]	76

F.3 SBERT and clustering model

Table 13: Complete topics obtained with SBERT and clustering model for the student survey dataset

Topic number	Topic word list	Total responses
0	campus rest, rest, rest online, lectures campus, online lectures, activities, courses, course activities, option, rest campus	17
1	lectures lectures, lectures, youtube videos, harder create, having travel, having theory, having share, having lectures, having labsession, having	26
2	sessions lab, lab sessions, sessions lectures, sessions, lab, lectures lab, available immediately, doubts, oeer review, sessions capmus	29
3	campus campus, campus, campus say, campus evey, campus physical, students way, thing campus, realise, contact students, doesn realise	26
4	online online, fine, activities, doable, online nice, follow online, activities held, online, doable online, held online	16
5	sessions campus, campus lab, labsessions campus, labsessions, session campus, campus labsessions, lab, lab sessions, working groups, help greatly	26
6	better, held campus, better campus, lectures better, held, sessions better, better held, campus work, better lectures, campus lab	16
7	sessions online, offline, online online, ik, really lab, lab sessions, lab, sessions offline, online lab, sessions	21
8	lectures campus, theory, programming online, programming, campus lab, online lectures, sessions online, online theory, online programming, campus lectures	87

9	peer, peer review, review, online lab, review campus, campus lectures, review online, sessions peer, campus peer, lectures online	111
10	better, think, better campus, easier, held, peer, fine, lectures, peer review, review	19
0	campus rest, rest, rest online, lectures campus, online lectures, activities, courses, course activities, option, rest campus	17

G Tools

In this project, we use Python 3 to implement the models and the experiments. We also utilized the following tools Python libraries.

1. Spacy [91]. An open-source library that can used to pre-process text for natural language processing tasks, build information extraction system or natural language understanding system. We used this library for pre-processing our text data such as lemmatization.
2. SBERT [41]. A python framework for state-of-the-art sentence, text and image embeddings. We used this library to encode our text data with the pre-trained sentence embedding model offered.
3. gensim [92]. An open-source python framework for vector space modeling and topic modeling. We used this library to train the LDA topic model and the NMF model needed for the experiments.
4. UMAP [93]. An open-source python library based on the Uniform Manifold Approximation and Projection (UMAP algorithm for dimension reduction. We used this library in our project to reduce the dimension of the embedding obtained from the text data.
5. HDBSCAN [94]. An open-source python library that can be used in unsupervised learning to identify clusters or dense areas of the data. We used this tool in our project to conduct clustering analysis.

.