Master Thesis

Automatic Retrieval of Topics Using Topic Modeling Techniques from Customer Conversations in the Airline Domain

Konstantina Andronikou

a thesis submitted in partial fulfilment of the requirements for the degree of

MA Linguistics
(Text Mining)

Vrije Universiteit Amsterdam

Computational Lexicology and Terminology Lab
Department of Language and Communication
Faculty of Humanities

Supervised by: Ilia Markov, Gabriele Catanese
2^{nd} reader: Sophie Arnoult

Submitted: July 1, 2022

Abstract

This thesis project focuses on the evaluation of automatic retrieval of topics using topic modeling techniques from customer conversations in the airline domain. The main idea of a topic modeling task is to produce a concise summary highlighting the most common topics from a corpus of thousands of documents. This procedure can detect themes in an unstructured text corpus. Retrieving the information within the conversational data between a customer and an agent can provide in-depth insight into the passenger's satisfaction or issues. This is essential for the airline company as it can help improve customer experience. Three different topic modeling methods were implemented, LDA, NMF and BERTopic. A comparative analysis in terms of predictive performance and topic quality proved the superiority of BERTopic. This project aims to address the gap in related work on this domain, as a great amount of previous research is done on a different type of data. Moreover, as the evaluation of an unsupervised environment can be challenging, an additional in depth evaluation technique is presented based on previous research. This technique aims to analyse the semantic correlation in terms of pairwise cosine similarity for all possible combinations within a topic. Finally, this thesis project contributes to extending the research on automatic topic retrieval on conversational data derived from the airline domain.

Keywords: Text Mining, Topic Modeling, Airline domain, NLP, Unsupervised machine learning, BERTopic, LDA, NMF, Conversational data, Customer experience

Declaration of Authorship

I, Konstantina Andronikou, declare that this thesis, titled *Automatic Retrieval of Topics Using Topic Modeling Techniques from Customer Conversations in the Airline Domain* and the work presented in it are my own. I confirm that:

- This work was done wholly or mainly while in candidature for a degree degree at this University.
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated.
- Where I have consulted the published work of others, this is always clearly attributed.
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work.
- I have acknowledged all main sources of help.
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself.

Date: 28th of June 2022

Signed: Konstantina Andronikou

Acknowledgments

First of all, I would like to thank my supervisors, Dr. Ilia Markov and Gabriele Catanese for the guidance and unlimited support they provided during the entire process of this thesis project. I would also like to thank my dearest friends and classmates Elena Weber, Rorick Terlou, and Mira Reisinger for their unconditional support during this project. I would also like to express my gratitude to the collaborative company, Underlined, for the given opportunity. Moreover, I would like to thank the CLTL staff of the VU Amsterdam for all the valuable knowledge they had to offer during the text mining program. Finally, a major thank you to my family for the love and support they provided me; without them, I would not be able to do what I love.

List of Figures

Contents

Chapter 1

Introduction

The rapid development of technology has contributed to developing new technological tools to improve Customer Experience (CX). As customer experience can be crucial in maintaining a successful business, many employers have adapted to newly developed technological tools. Among the tools used for analyzing and improving customer experience, a significant amount is within the field of Natural Language Processing (NLP). NLP is a sub-field of Artificial Intelligence (AI) that deals with the autonomous manipulation of natural language, whether in speech or text form. Different types of data extract meaningful information for a company's objectives, such as reviews, customer and agent conversations, tweets, etc. As the market for customer experience solutions grows, so will the need to stay current on text analytics and Natural Language Processing capabilities. CX platforms should guarantee that their text analytics systems have capabilities to provide a deep, domain-level understanding. It is no longer acceptable for CX platforms to only have rudimentary tools for analyzing data. To fully understand clients' needs, companies must use a wide range of technological tools to analyze which characteristics impact customer experience. This project aims to implement and evaluate traditional and new developed NLP techniques on user-generated data.

1.1 Problem Definition

This thesis project focuses on implementing NLP techniques to CX in collaboration with the company Underlined [1]. The company's primary goal is to help businesses to optimize customer experience using data-driven approaches and text mining tools. This collaboration project concerns the request of a famous Dutch airline company to create a tool for identifying topics in English customer and agent conversational data. Due to the demand, this project will focus on an automatic retrieval of topics using topic modeling techniques.

In business settings, topic modeling insights can improve a company's strategy and assist in developing marketing platforms (Reisenbichler and Reutterer, 2019). Sometimes, it is the case that companies collect a great amount of data concerning customer experience but have difficulties figuring out precisely the topics within the data. In this case, we are dealing with customer conversations in the airline domain - but what

[1] `https://underlined.eu/`

are the conversations about? Are the customers contacting the company for potential problems with their flight? Are they contacting to discuss a concern about a service the company offers?

Imagine entering a bookstore to buy a cooking book and being unable to locate the part of the store where the book is located, presuming the bookstore has just placed all types of books together [2]. In this case, the importance of dividing the bookstore into distinct sections based on the type of book becomes apparent. Topic Modeling is a process of detecting themes in a text corpus, similar to splitting a bookshop depending on the content of the books. Despite its growing popularity, there is currently no systematic assessment of the state of research for topic modeling in customer conversations in the airline domain. This project aims to solve this issue by helping Underlined supply a tool to help customers understand their data by giving them a hint of possible topics their data contains. With the implementation of such model, customers can use this tool and retrieve hidden topics within the data.

Over the past two decades, topic modeling has succeeded in various applications in the field of Natural Language Processing and Machine Learning. The main aim of a topic model is to analyze and identify hidden topics within unstructured data. This type of data contains information not organized in a pre-determined way. For example, with text mining tools, words and phrase patterns can be detected within a set of documents (Pascual, 2019). Based on previous research, the most frequent data used for the airline domain for topic modeling are customer tweets and customer reviews. In this scenario, the message is intended directly to the organization for assistance, whereas a tweet is directed to the digital audience. As there is no significant research on conversational data in the airline domain, the needed pre-processing steps to get the data ready for the task are also under investigation. Another aspect under examination for this project is the evaluation of topic modeling. Even though topic modeling approaches have been a subject under investigation since 1980, the evaluation process is a matter under development. The unsupervised nature of topic models makes the model selection problematic; therefore, evaluation is a crucial issue (Wallach et al., 2009). The aim is to investigate and evaluate the topic quality and the predictive performance of different unsupervised machine learning approaches in topic modeling on customer conversation data in the airline domain.

1.2 Research Questions

Based on the data limitations and techniques mentioned the following research questions were formulated:

RQ: How does the performance of different topic modeling techniques differ in terms of predictive performance and topic quality in customer conversations in the airline domain?

Sub-Q1: What additional pre-processing steps can be combined with traditional ones to improve the performance of topic modeling in the airline domain?

[2] This example was inspired by Dutta (2021)

Sub-Q2: Will existent evaluation methods reflect the topic quality in user-generated content in this domain?

1.3 Approach

This project aims to test three topic model approaches regarding topic quality and predictive performance. A comparison study will take place to evaluate the performance of two traditional models and one deep learning model. The most well-known and frequently used model within the field of topic modeling is Latent Dirichlet Allocation (LDA) (Blei et al., 2003). This model is considered state of the art in topic modeling. This project will use this model as one of the traditional approaches. The second model in the traditional category is Non-Negative Matrix Factorization(NMF) (Paatero and Tapper, 1994). The third model representing the deep learning approach is BERTopic (Grootendorst, 2020). All models will be adapted and evaluated on customer conversations in the airline domain. The following Figure 1.1 provides a simplified overview of the procedure of this project.

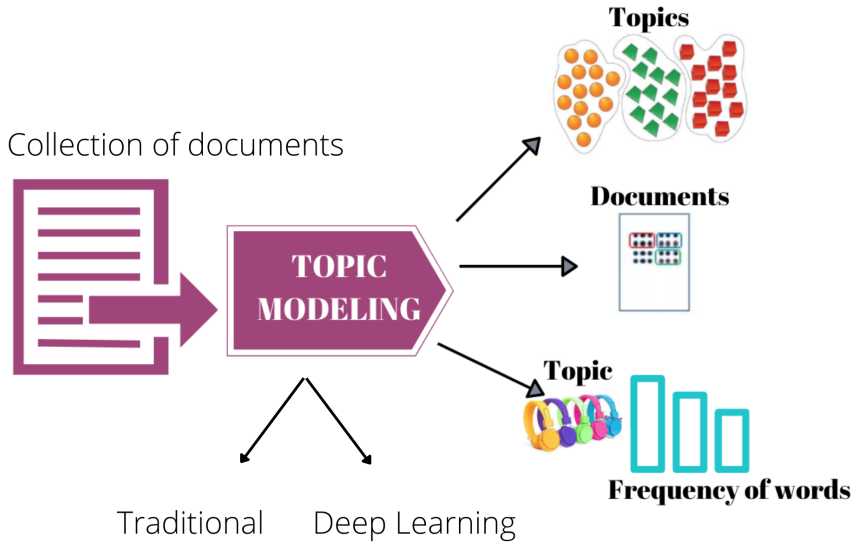

Figure 1.1: Topic Model Procedure (Adaptation from Sinivasan (2020))

1.4 Outline of the Chapters

This report is structured as follows. **Chapter 2** provides a review of previous research done on the matter. **Chapter 3** presents the data, the pre-processing and a deeper explanation of the topic modeling techniques. **Chapter 4** provides a comparative evaluation of the systems and the results generated. Finally, **Chapter 5** presents the main conclusions and proposes potential directions for future research.

Chapter 2

Literature Review

This chapter gives a summary of previous topic modeling research as well as the research gaps that this project aims to fill. The sub-sections that follow present topic modeling in greater detail. The first sub-section shows the task's roots and the first model implemented, while the second sub-section discuses previous research on the subject.

2.1 Background

User-generated content, such as online reviews and tweets, has been demonstrated in numerous studies to improve customer happiness, service quality, brand reputation, (Goyal, 2021). A topic modeling task is implemented to identify potential indicators for the customer's satisfaction or the company's quality. This task is considered an unsupervised machine learning technique capable of scanning a series of documents to detect and extract hidden semantics. It is one of the most commonly used approaches for processing unstructured textual data. Topic modeling emerged in the 1980s from the 'generative probabilistic modeling' field. These models are used to solve tasks such as likelihood estimates, data modeling, and class distinction using probabilities. The first method used to carry out a topic modeling task was the TF-IDF (term frequency-inverse document frequency) reduction scheme developed in 1983 by Salton and McGill (1986). This method can categorize each document within a corpus that is made up of a vocabulary based on one factor the frequency of occurrence of each word over the whole corpus (Vayansky and Kumar, 2020). The main idea behind this task is to produce a concise summary highlighting the most common topics from a corpus of thousands of documents. This model takes a set of documents as input and generates a set of topics that accurately and coherently describe the content of the documents.

Topic models are continually evolving in sync with technological developments. As a result, new topic modeling techniques have been developed since the 1980s. The Figure below illustrates some topic modeling techniques developed between 1983 and 2003.

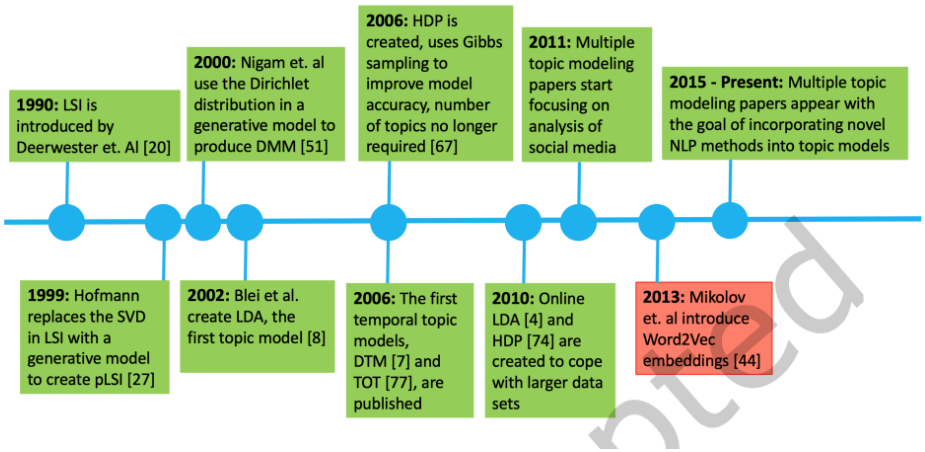

Figure 2.1: Topic Models Timeline 1990-2013
(Churchill and Singh, 2021)

As topic modeling increases, it can be seen that not all topic modeling techniques can be used and be suitable for all types of data (Churchill and Singh, 2021). For example, the algorithm used to retrieve hidden topics on social media data might not perform well for scientific articles due to the different patterns of words. Each data and domain characteristic, such as document length, and sparsity, must be considered before implementing a topic modeling algorithm (Churchill and Singh, 2021). Due to this variation of different types of data, new topic modeling approaches are developed. Algorithm development, of course did not stop in 2013; since then, more models have been created. The following Figure shows an overview of some models developed between 2013 and 2022.

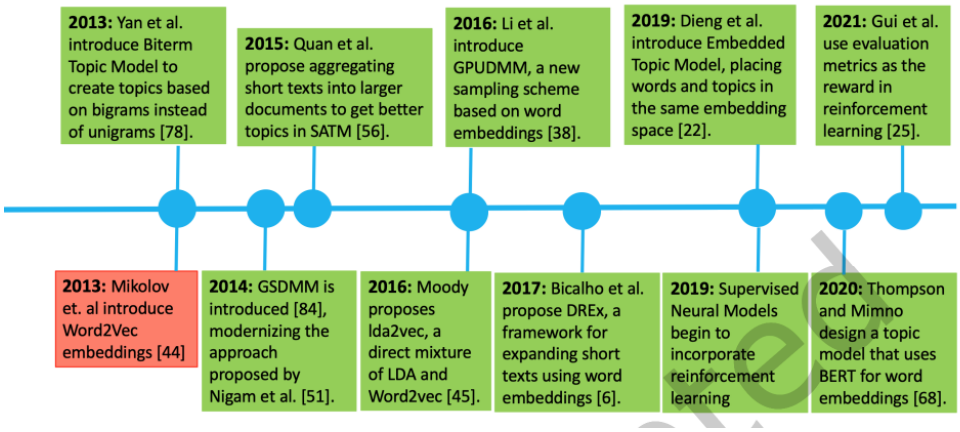

Figure 2.2: Topic Models Timeline 2013-2022
(Churchill and Singh, 2021)

Topic modeling as a subject has been under investigation for almost 40 years and has been used across many domains such as airlines, tourism, and more. Customers' voices are crucial in understanding their expectations and improving the service quality. A variety of previous research in the airline domain was evaluated and is presented in the following section.

2.2 Airline Domain

Kwon et al. (2021) researched topic modeling and sentiment analysis on customer reviews posted on Skytrax. Online reviews from 27 airlines were gathered, totalling over 14,000 reviews. With the acquired data, topic modeling and sentiment analysis were utilized to determine what essential words appear in the online reviews. This study used LDA for generating potential topics in the data. By conducting a topic modeling study through structurally transformed papers, statistical text processing techniques were used to assess the probability of the emergence of topics. With this topic modeling technique, five topics were generated. Based on these topics and the sentiment analysis, it was recorded that through frequency analysis, 'seat', 'service', and 'meal' were identified as important issues. Moreover, the results demonstrated that the main problem that can cause customer dissatisfaction was a delay, whereas 'staff services' based on the sentiment analysis is a crucial factor for the customer's satisfaction.

Liau and Tan (2014) executed a topic modeling task to gain customer knowledge in low-cost airlines. This study implements a different topic model technique than the ones explored in this report. For this project data was collected in the time frame of two and a half months. The collected data consisted of 10,895 tweets. Clustering was used to extract all relevant topics in the dataset, as it can automatically organize, discover, and summarize latent information from unstructured text documents. Two algorithms for topic detection and two for sentiment analysis were tested to investigate customers' views about the airline. Before implementing any algorithms, the corpus was converted to a term-document matrix. In this mathematical matrix, its rows correspond to the tweets in the corpus, and columns correspond to the appearance frequency of the terms in the corpus. The two clustering algorithms used were K-Means and SK-Means. K-means (MacQueen, 1967) is one of the most basic clustering algorithms for detecting common patterns in disorganized data sets. The algorithm divides all data points into K clusters by selecting close data points based on their similarities. While 'SK-means clustering is a modification of K-Means clustering with cosine similarity that works on the vectors that lie on the unit sphere' (Liau and Tan, 2014, p.1350). Based on the results, both algorithms yield very similar clusters, the only difference being the order of the words presented. Topics such as 'flight delays', 'customer service', 'ticket promotions' were the the main topics discussed by the customers.

Srinivas and Ramachandiran (2020) implemented topic modeling techniques to investigate the service quality of airline companies based on online customer reviews. This research aimed to use an unsupervised text analytics approach to extract company and competitor-specific intelligence from 99,147 airline reviews of a US-based target carrier. Three topic models were used to extract the essential topics: probabilistic Latent Semantic Analysis (pLSA) (Hofmann, 1999) and two types of Latent Dirichlet Allocation (LDA-VI and LDA-GS). Based on the results generated, it was evident that 18 topics were recorded for pLSA and 23 topics for the other two topic models. The topics that contained similar keywords were grouped, and less coherent topics were discarded. Based on this action, 11 meaningful topics were identified from each of the three topic models. Based on the topic, it was recorded that the most discussed aspects from the online reviews were seating and baggage services, ticketing services, and reward programs. In addition to the topics, the researchers decided to implement sentiment

analysis to see if the topics were positively or negatively discussed. The proportion of positive and negative attitudes among the review sentences is balanced with 38% for negative and positive sentiment and 24% neutral.

Another study was conducted by Korfiatis et al. (2019) to investigate and measure the service quality in the airline domain. The data used for this project was 557,203 online customer reviews sourced from TripAdvisor. The reviews included information such as the flight date, name of the airline company, inbound and outbound destination, as well as the cabin class of the passenger. Structural Topic Model (STM) (Roberts et al., 2016) was chosen for this study, where topic coverage and word distribution are estimated by Bayesian inference. In this case, the number of topics was selected based on three factors, heldout likelihood, semantic coherence of the words to each topic and exclusivity of topic words. Service quality indicators were captured based on the topics generated from the topic model. Passengers' loyalty and satisfaction depend mainly on cost and comfort.

Finally, a similar study to the aim of this project was conducted by Egger and Yu (2022) to gain insights into human behaviours and travel experiences during the pandemic. The data used for this project was 50,000 tweets in English that included the terms 'covidtravel' and the combination of 'covid' and 'travel'. The models used to retrieve the hidden topics from the tweets were: LDA, NMF, BERTopic, and Top2Vec. The number of topics used for this project was: 14 topics for LDA, 10 for NMF, 6 for BERTopic, and 5 for Top2Vec. To identify the number of topics for each model different methods were used. In the case of LDA, a grid search was performed, while for NMF, the coherent score was measured, and the highest corresponding number of topics was the optimal one. For the other two approaches, the number of topics depended on the number of documents that could be grouped. The topics include discussions and information about travelling during COVID-19, such as 'covid tests', 'travel ban', and 'travel pass'. Based on human interpretation, it was supported that BERTopic and NMF performed better in evaluating Twitter data, followed by Top2Vec and LDA.

Even though topic modeling models have been a subject under investigation since 1980, the evaluation process is a matter under development. The following section presents the challenges in the evaluation processes and the limitations that can occur.

2.3 Evaluation

The unsupervised nature of topic models makes the model selection problematic; therefore, evaluation is an important issue (Wallach et al., 2009). In most evaluations, the use of gold standards makes the process easier. In this case, no gold labels are available for comparing the results (Kapadia, 2019). However, it is vital to determine whether a trained model performs well. As this project is using customer-oriented data, research was done on evaluation methods for this specific domain (Reisenbichler and Reutterer, 2019). Appendix A presents a variety of different evaluation methods on the customer/marketing domain proposed by Reisenbichler and Reutterer (2019). This section focuses on the most frequently used methods.

The most frequently used approaches for evaluating a topic modeling task are: Hu-

man Judgment, Intrinsic and Extrinsic Evaluation Metrics (Kapadia, 2019). Intrinsic evaluation can include tasks such as capturing model semantics and topics interpretability such as coherence, while the extrinsic checks if the model can perform well on predefined tasks. In the case of topic modeling, the choice of the evaluation procedure depends not only on the machine algorithm used to execute the task but also on the data. Furthermore, topic modeling is highly dependent on arbitrary aspects such as setting hyperparameters such as the number of topics (Reisenbichler and Reutterer, 2019).

2.3.1 Intrinsic Evaluation

The most frequently used intrinsic evaluation methods are *coherence score* and *cosine similarity*. In the NLP community, coherence measures have been developed to evaluate topics created by a topic model. A coherence score calculates the average or median pairwise word similarities created by a topic's top terms. It is a metric that evaluates how semantically related high-scoring terms are within a single topic (Kapadia, 2019). In the case of cosine similarity, the angel of two words is identified and the similarity between two documents is calculated. The metric generated represents a written document as a term vector and calculates the similarity between two documents by computing the cosine value between their term vectors (Rahutomo et al., 2012).

2.3.2 Human Judgment

Based on (Chang et al., 2009) there are two ways of topics being evaluated by a human, observation-based and interpretation-based. Observation-based is an way to assess a topic model by looking at the most probable words in the topic. However, the interpretation-based approach is more complex; with this method, humans can evaluate two components of the latent space of a topic model. First, a task is generated to determine whether a topic is semantically coherent in a way humans can identify. This process is referred to as *word incursion*, subjects must recognize a fictitious word that has been introduced into a topic (Chang et al., 2009). The second task determines whether a document's and topic association is logical. This task is known as *topic incursion* and an individual is required to identify a topic that the model did not correlate with the document (Chang et al., 2009). The following Figure presents the process of these evaluation methods.

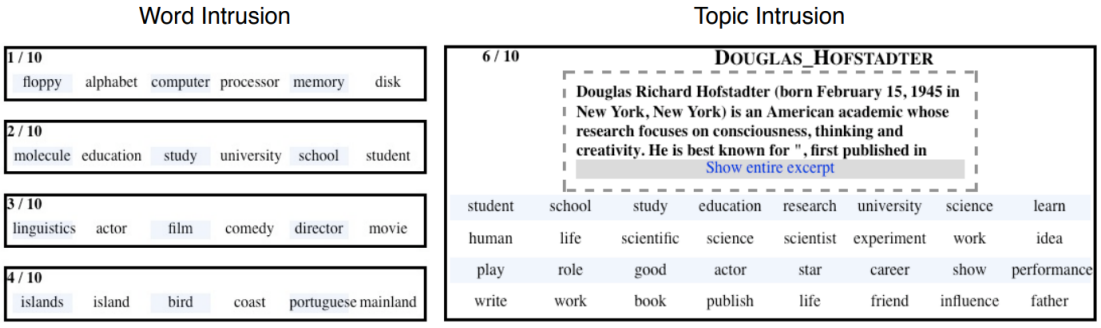

Figure 2.3: Evaluation Method by (Chang et al., 2009)

The following sub-sections present in greater detail the individual elements generated for the human based evaluation methods.

2.3.3 Word Incursion

The subject is given six randomly arranged words to complete the word incursion challenge. The user's objective is to locate the invader, a word that does not belong to the others. This task's primary goal is to evaluate how closely the inferred topics correspond to human understanding. First, a random topic is selected, and the five most reasonable terms from the topic are chosen. To reduce the possibility that the invader comes from the same semantic group, a word is randomly selected from a pool of terms with a low likelihood. However, the intruder word needs to have a high probability in another topic to avoid being rejected due to low frequency of occurrence. The subject is then given the set of all words, for example {'cat','dog', 'horse', **'banana'**,'cow', 'pig'} [1].

2.3.4 Topic Incursion

Examining if a mixture of a topic model's decomposition of documents is in agreement with human judgment is called topic incursion. Subjects are given four topics, and the eight highest-probability terms represent each topic. Three of those topics were allocated to the document with the highest probability. In combination, low-probability topics are randomly chosen as the remaining intruder topic. The subject must choose a topic that is unrelated to the document. This task's main aim is to easily assess the quality of document-subject assignments discovered by topic models.

2.4 Evaluation Limitations

The difficulty[2] with the evaluation metrics in topic modeling is that they attempt to measure something objective, like topic coherence. Even if the coherence score is high, it does not mean that the topic quality is accurate. In those cases, it would help to combine with human judgment. However, there is also a risk that a human might be able to tell whether a topic representation is coherent but not necessarily accurate. Topic accuracy and quality are subjective depending on the person evaluating the results; this can create bias in the results. A combination of various evaluation methods can help minimize any subjectivity and bias in the results.

This report seeks to address these cases and limitations in the literature of customer conversations in the airline domain. It can be seen that the most frequently used data type for topic modeling in this domain is tweets and online reviews. This project investigates and implements various evaluation approaches such as topic coherence, cosine similarity, and observation-based human judgment. These methods analyze the performance of three topic models in terms of topic quality and predictive performance by using customer conversations as input. The following chapter presents and describe in greater detail the data provided for this project, the steps implemented to prepare the data, and the topic models.

[1] This example was inspired by Chang et al. (2009)

[2] This section was inspired by a discussion with Grootendorst (2020), creator of BERTopic.

Chapter 3

Methodology

The following sections presents the experimental setup for this project. The data and the required pre-processing steps, as well as the model's architecture, implementation, and limitations are presented.

3.1 Data

The data for this project is provided by Underlined's client VANAD Engage [1]. The information is customer conversations executed on WhatsApp. It consists of about 100,000 customer conversations in English from a famous Dutch airline stored in a JSON format. For privacy reasons, personal and sensitive information regarding the customer and the agent in the data was anonymized with a PII (Personally Identifiable Information) masking tool. Personal information such as names, phone numbers, addresses, etc., are masked to ensure the privacy of both the agent and the customer. As it can be seen from the demo conversation 3.1, the name of the client gets replaced with a masking label ([PII_NAME]). The data shall not be shared with third parties due to privacy reasons. Additional information such as the demographic, age, gender, and socioeconomic status of the speakers is unknown as the data is anonymous. Due to the data privacy restrictions, the following Figure 3.1 is an artificial example.

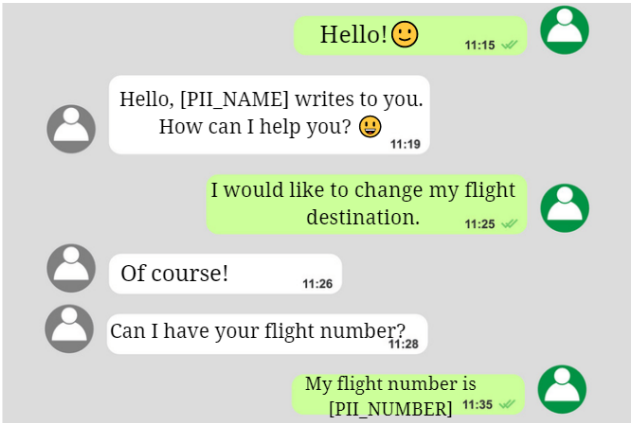

Figure 3.1: Demo Conversation

[1] https://www.vanadengage.com/

3.1.1 Data Analysis

After further analysis of the data, many elements must be considered for preparing the data for the task. As the data provided for this project is conversations between agents and customers through the online platform, WhatsApp's many linguistic features need to be considered. Some of the components of language that can be investigated with sociolinguistic data such as conversational data include speech act, linguistic variety from a diaphasic, diastratic, or diatopic point of view, use of orthographic elements, spelling errors, and the use of emojis in communication (Dorantes et al., 2018). Furthermore, code-switching is a phenomenon that must be considered in virtual interactions across several platforms. Apart from the linguistic features, other conversational elements such as emojis and punctuation also need to be considered. These factors need to be analyzed in greater detail to prepare the data for the topic modeling task.

3.2 Pre-Processing

Taking into consideration all the possible challenges within the data, some pre-processing steps were implemented. This procedure contains two main steps: filtering and applying text mining tools.

3.2.1 Filtering

The first step in the pre-processing procedure was to use a **language detection library**. To deal with the different languages and code-switching within the data, the external library Pycld2 [2] was implemented. This language detection library can detect mono- and multi-lingual sentences. Two parameters were set within the language detection, language code and the number of languages detected. If the language detected from the library is under the language code 'en' (English), and the number of languages detected is one, then the conversation is stored. This way, any code-switching sentences or other languages except English are excluded. After all the English conversations were stored, a **encoding-decoding** function via 'ascii' was implemented to exclude any non-recognizable characters.

As the conversations are executed on WhatsApp, emojis are a frequently used element. The **emojis** detected in the data were filtered out with the help of the external library clean-text [3]. This library was able to clean the text from any emojis as well as **lower case** it. Additionally, to reduce the noise of the data even more, **automatic responses** generated from the company were excluded. In this case, a template was created for the specific company with the most frequent generated responses. This decision was made as these responses did not carry any potential topics, as they were informing customers for business hours or asking the customer to evaluate their experience. After these steps were implemented, the data was reduced to 77,815 customer conversations.

The final stage of filtering was to investigate and analyze which words do not carry any semantic meaning and will not have any influence on the potential topics generated.

[2] `https://pypi.org/project/pycld2/`
[3] `https://pypi.org/project/clean-text/`

The first type of words excluded was **stop words**. Stop words are frequently used in a language; in the case of English words such as conjunctions, pronouns, determiners, etc., are considered stop words. An automatic generated English list was imported from the external library NLTK [4] and was used for this step. This list excluded in total 179 stop-words. After excluding these words, a frequency check was implemented to find and analyze the top frequent words within this domain. Words such as 'Hello', 'Welcome', 'Thanks', etc., were recorded to have the highest frequency. Based on the domain-specific frequency check, an **external list** was manually created. This list contained a total number of 552 words that did not have any contribution to potential generated topics and therefore was excluded from the data.

Finally, punctuation tends to be a frequent element in text data but has no semantic meaning and creates a lot of noise in the data. Therefore it was excluded. After these steps, the data was slightly reduced to the total number of 77,799 conversations. The output of this filtering procedure was used for implementing the next pre-processing step, the text mining tools.

3.2.2 Text Mining Tools

The first text mining tools implemented were **tokenization** and **lemmatization**. A token is 'the word or the punctuation mark as it appears in the sentence' (Abu-Jbara and Radev, 2012) while a lemma is the root form of a token (Abu-Jbara and Radev, 2012); for instance, the word 'undivided' within a sentence is a token and 'divide' would be the corresponding lemma. Lemmatizing redundant topics such as 'flight' and 'flights' will not occur. Both tokens and lemmas are beneficial as they divide the text data into pieces and thus make it easier for a language model to distinguish. In addition to lemmatization, **part-of-speech tags (POS)** were extracted to lemmatize the verbs and avoid topics such as 'cancel' and 'cancels'. POS aims to connect a token in text data to its grammatical definition. These implementations were applied through the external library, NLTK. Finally, with all these text mining tools implemented, the data was ready for the topic modeling task.

3.3 Approaches

The following sections present in greater detail the architecture and training procedure for all three topic modeling approaches.

3.4 Latent Dirichlet Allocation (LDA)

Latent Dirichlet Allocation (LDA)(Blei, Ng, and Jordan, 2003) is a three-level hierarchical generative model. This model is a powerful textual analysis technique based on computational linguistics research that uses statistical correlations between words in a large number of documents to find and quantify the underlying subjects (Jelodar et al., 2019).

'Latent' represents the process of the model to discover the hidden topics within the documents. The word *'Dirichlet'* indicates that the distribution of subjects in a

[4] `https://www.nltk.org/`

document and the distribution of words within topics are both assumed to be Dirichlet distributions. Finally, *'Allocation'* represents the distribution of topics in the document (Ganegedara, 2019). The model implies that textual documents are made up of topics, which are made up of words from a lexicon. The hidden topics are 'a recurring pattern of co-occurring words' (Blei, 2012). A probability distribution can be used to represent each document across latent themes, with a common Dirichlet prior across all documents. However, this generative model does not consider the order of words or the semantic relation while generating topics. Hence it relies only on the bag-of-words (BOW) approach. This approach is responsible for representing information as numerical values disregarding grammar and even word order (Huilgol, 2020). The following Figure 3.2 presents the intuition behind this topic modeling technique.

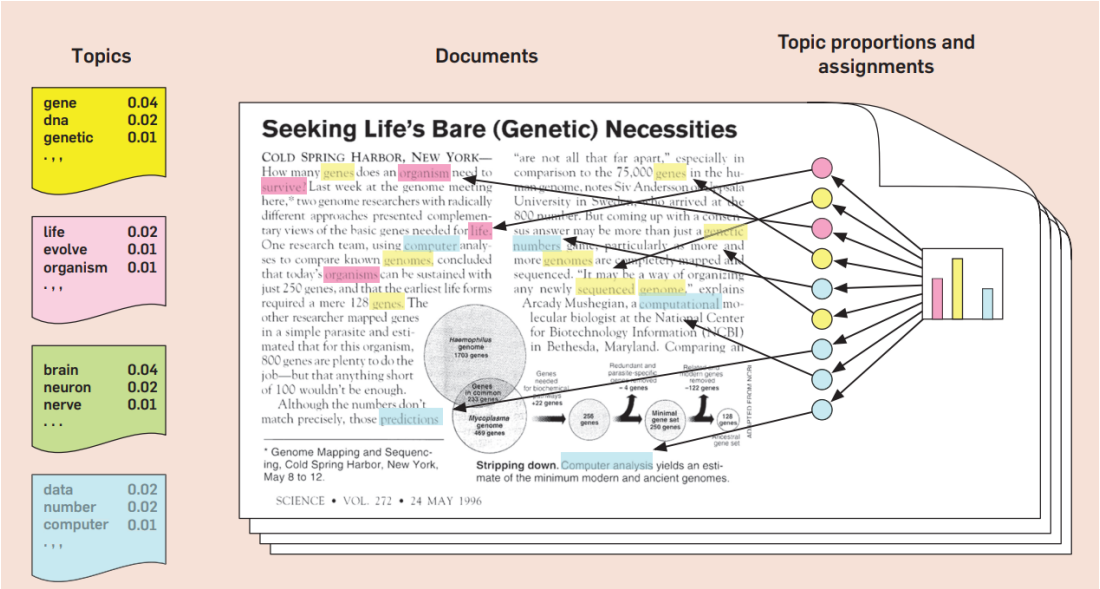

Figure 3.2: LDA Representation (Blei, 2012)

In the case of Figure 3.2, 'We assume that some number of 'topics' which are distributions over words, exist for the whole collection (far left). Each document is assumed to be generated as follows. First, choose a distribution over the topics (the histogram at right); then, for each word, choose a topic assignment (the colored coins) and choose the word from the corresponding topic' (Blei, 2012, p.78). This topic modeling algorithm categorizes the words within a document based on two assumptions: documents are a mixture of topics and topics are a mixture of words. In other words, 'the documents are known as the probability density (or distribution) of topics, and the topics are the probability density (or distribution) of words' (Seth, 2021).

Every corpus that contains a collection of documents can be converted to a document-word/document term matrix (DTM). LDA converts the documents into DTM, a statistical representation describing the frequency of terms that occur within a collection of documents. The DTM gets separated into two sub-matrices: the document-term matrix: which contains the possible topics, and the topic-word matrix: which includes the words that the potential topics contain (Seth, 2021).

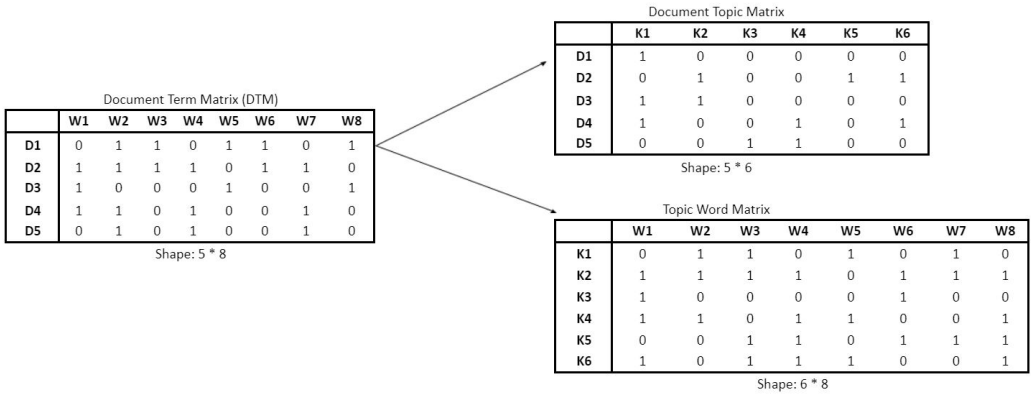

Figure 3.3: Matrices Representation (Seth, 2021)

These matrices already provide topic-word and document topic distributions. However, the fundamental goal of this model is to improve the distribution of this data. To improve these matrices, LDA employs sampling techniques. It goes through each word *'w'* for document *'d'* and tries to replace the present topic assignment with a new one. With a probability P that is the product of two probabilities *p1* and *p2*, a new topic *'k'* is assigned to the word *'w'*. After the data is represented based on the matrices, the words are transformed into vectors. In order to detect word similarities, vectors convert words from a vocabulary to a corresponding vector of real numbers (Navigli and Martelli, 2019). After the creation of the vector, the LDA vector space is represented as shown in the following Figure 3.4.

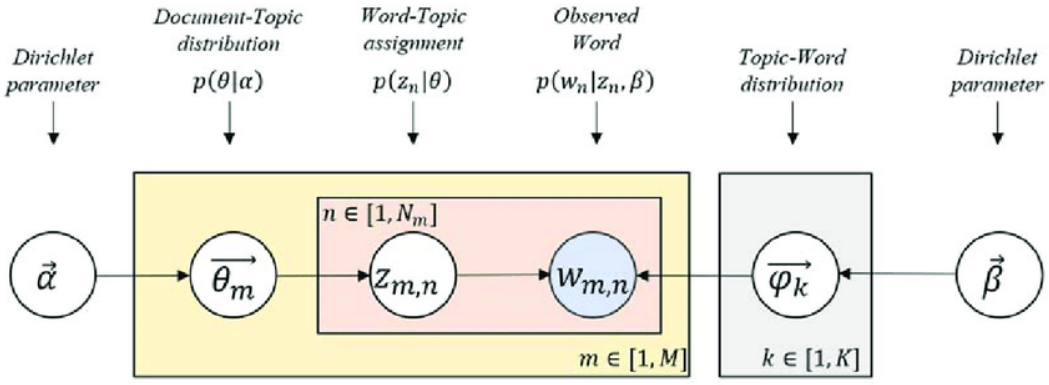

Figure 3.4: Vector Space LDA (Seth, 2021)

The yellow space represents all the documents (M), and the peach colour is the number of words within a document (N). According to the vector space, each term is associated with a hidden topic (Z); this topic is assigned to a number of topic words based on word distribution. The main goal of LDA is to find the optimal representation of the document-topic and topic-word distribution. This distribution can be influenced based on the parameters implemented when training the model.

3.4.1 LDA Parameters

This topic modeling approach can be implemented in various ways, but the model's performance comes down to estimating one or more parameters. This topic modeling algorithm consists of different parameters[5]. The most crucial parameters, in this case, are *Number of topics*: The optimal number of topics extracted from the corpus, *Passes:* number of passes through the corpus during training, *Alpha:* controls the prior distribution over the topic weights across each document, and *Eta* controls prior distribution over the word weights across each topic.

For this project, a grid search was implemented to identify the optimal parameters for the domain and data type given. Grid search is a method of exhaustively searching the hyperparameter space of a given algorithm. Table 3.1 presents the parameter values chosen based on the grid search.

Parameters	**Value**
Number of Topics	10
Passes	10
Alpha	0.31
Eta	0.90

Table 3.1: LDA Parameters

As mentioned, these parameters can help improve the model's performance, but some limitations need to be considered. Therefore, the following sub-section presents the limitations of the model.

3.4.2 LDA Limitations

The first drawback of this generative model is that it fails to cope with large vocabularies. Based on previous research, executors had to limit the vocabulary used to fit and implement a good topic model. This can lead to consequences for the performance of the model. To restrict the vocabulary usually, the most and least frequent words are eliminated; this trimming may remove essential terms from the scope (Dieng et al., 2020). For this project, this does not apply as the size of the data is not restricting. Another significant limitation is that the core premise of LDA is that documents are considered a probabilistic mixture of latent topics, with each topic having a probability distribution over words, and each document is represented using a bag-of-words model (BOW). Based on this approach, topic models are adequate for learning hidden themes but do not account for a document's deeper semantic understanding. The semantic representation of a word can be an essential element in this procedure. For example, for the sentence 'The man became the king of England', the representation of a bag of words will not be able to identify that the word 'man' and 'king' are related. Finally, when the training data sequence is altered, LDA suffers from 'order effects' which means that different topics are generated. This is the case due to the different shuffling order of the training data during the clustering process. Any study with such order effects will have a systematic inaccuracy. This inaccuracy can lead to misleading results, such as erroneous subject descriptions.

[5] `https://radimrehurek.com/gensim/models/ldamodel.html`

3.5 Non-Negative Matrix Factorization (NMF)

NMF is a decompositional, non-probabilistic matrix factorization algorithm within the linear-algebraic group (Egger and Yu, 2022). It was first introduced as positive matrix factorization by Paatero and Tapper (1994) and is used to reduce dimensionality and analyze data (Kuang et al., 2015). This approach uses the factor analysis method to give less coherent terms a lower weighting. NMF has gotten a lot of attention in the last two decades, and it has been effectively used for a wide range of tasks within various fields, including the Natural Language Processing field and text mining.

This algorithm operates by decomposing/factorizing high-dimensional vectors into lower-dimensional representations. In the case of topic modeling, the input is used is high-dimensional vectors such as TF-IDF (Egger and Yu, 2022). This algorithm transforms TF-IDF data by splitting a matrix into two lower ranking matrices (Egger and Yu, 2022) and has the representation of BoW. TF-IDF is a metric for determining the relevance of a word in a group of documents. Figure 3.5 presents the topic modeling procedure using NMF.

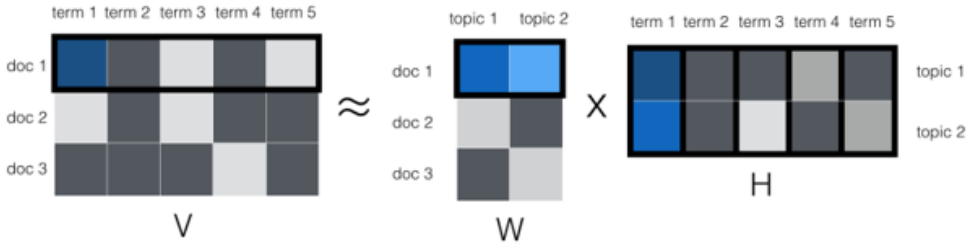

Figure 3.5: NMF Representation (MacMillan and Wilson, 2017)

NMF will generate two matrices from the original matrix (A) (W and H). W stands for the topics it discovered, and H stands for the coefficients (weights) associated with those subjects (Mifrah and Benlahmar, 2020). W and H have a special interpretation: Wij quantifies the relevance of topic j in document i, and Hij quantifies the relevance of term j in topic i (MacMillan and Wilson, 2017). Compared to probabilistic models such as LDA, this model essentially gives the same output by producing two forms of output: a keyword-wise topic representation (W columns) and a topic-wise document representation (W columns) (the columns of H) (Kuang et al., 2015).

3.5.1 NMF Parameters

Like any other model, NMF takes some parameters [6]. In this case, three parameters were set: **Number of topics :** The optimal number of topics extracted from the corpus, **Passes:** number of passes through the corpus during training as well as **Minimum Probability:** If set to True, topics with smaller probabilities are filtered out. For this project, these parameters were set with the following values:

[6] `https://radimrehurek.com/gensim/models/nmf.html`

Parameters	**Value**
Number of Topics	10
Passes	10
Minimum Probability	True

Table 3.2: NMF Parameters

In combination with these parameters, there were some limitations taken into consideration. The following section presents the challenges that occur when implementing this model.

3.5.2 NMF Limitations

The main limitation of NMF is the separability assumption made when generating the topics. The separability assumption is identical to the anchor-word hypothesis, which states that each topic has a unique anchor word that does not exist in other topics. The anchor word assumption may not always be accurate due to words and phrases having numerous meanings. Another limitation is the vocabulary size; based on previous research, it was recorded that NMF works better with small texts, such as tweets and titles. Finally, just like LDA, this algorithm does not consider the semantic relation between words, as it uses a BoW representation.

Due to these limitations, many generative and deep learning models have been generated using these traditional approaches as a base to improve the topic quality. The following section will present a deep learning model developed for topic modeling.

3.6 BERTopic

Devlin et al. (2018) presented Bidirectional Encoder Representations from Transformers (BERT) as a fine-tuning approach in late 2018 . BERT is a pre-training strategy for Natural Language Processing (NLP) that successfully exploits a sentence's deep semantic information (Hosseini and Varzaneh, 2022). A variation of Bidirectional Encoder Representations from Transformers (BERT) has been developed to tackle topic modeling tasks. BERTopic was developed in 2020 by Grootendorst (2020) and is a topic modeling technique that uses transformers and class TF-IDF to produce dense clusters that are easy to understand while maintaining significant words in the topic description. This deep learning approach supports sentence-transformers model for over 50 languages for document embedding extraction (Egger and Yu, 2022). This topic modeling technique follows three steps: document embeddings, document clustering and document TF-IDF .

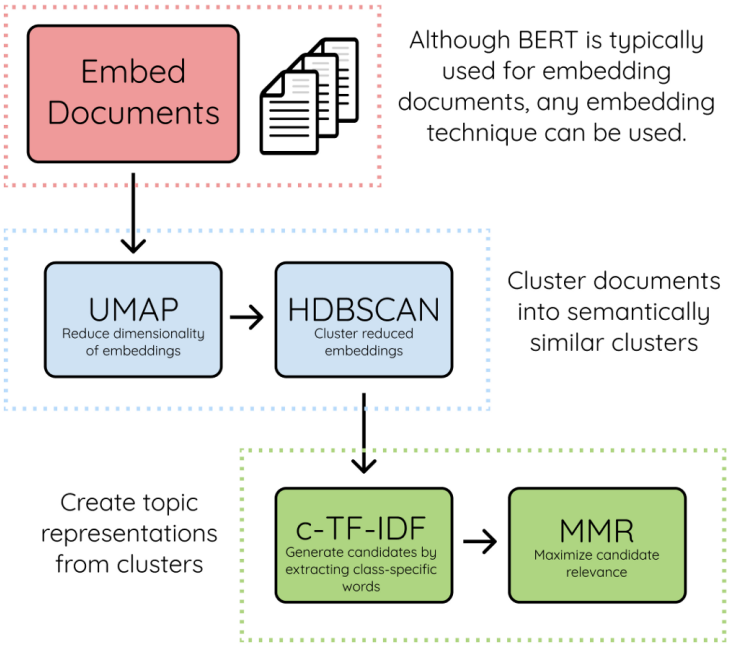

Figure 3.6: BERTopic Representation by Grootendorst (2020)

3.6.1 First Step: Document Embeddings

The first step of the model is responsible for converting the corpus to document and word embeddings. This is a way to map the words into numerical vector spaces while considering the semantic meaning of words. To generate the embeddings, BERT uses Sentence-BERT, also known as SBERT. This framework uses pre-trained language models which allow users to convert text input to dense vector representations. For this project, these embeddings will then be used as an input for clustering to group semantically similar documents (Grootendorst, 2022). Different types of embeddings can be used for this step.

3.6.2 Second Step: Document Clustering

Before clustering the embeddings, a dimensionality reduction is implemented. Reducing the dimensions has its benefits; for example, it takes care of redundant features, takes less computation and training time, and helps visualise data. In this case, the external library UMAP (Non-linear) is used as it can preserve the local and global features in lower dimensions. After the reduction is made with the help of HDBSCAN, the embeddings are clustered. To reduced and represent the noise as outliers the soft-clustering technique, HBSCAN is use to represent clusters (Grootendorst, 2022). This eliminates the assignment of unrelated documents to any cluster and is likely to improve subject representations. The following Figure presents clustering with the use of HDBSCAN.

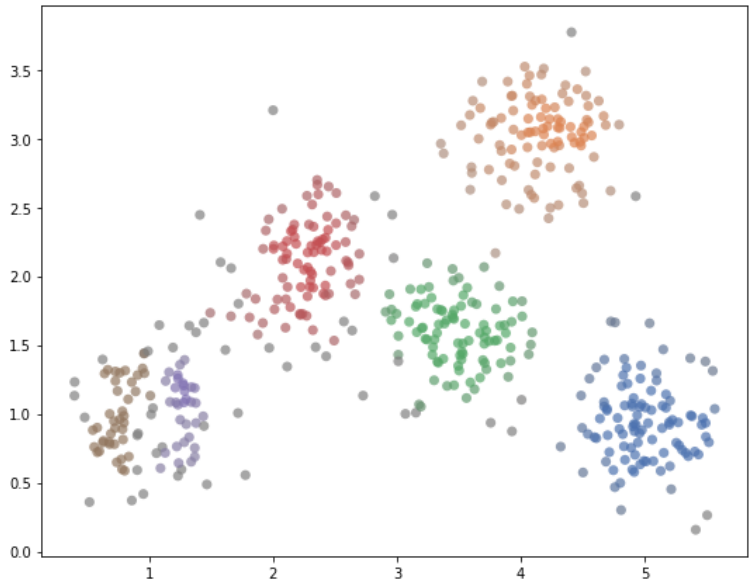

Figure 3.7: Clusters identified with HDBSCAN (Smith, 2021)

3.6.3 Third Step: Document C-TF-IDF

The topic representations are based on the documents in each cluster, with one topic given to each cluster. To discover what distinguishes one topic from another based on its cluster words, a class-based TF-IDF is implemented. The original formula concerns measuring the representation of the importance of a word to a document, while the adaptation concerns the representation of a term's significance to a topic instead (Grootendorst, 2022). The following Figure 3.6.3 presents the adapted formula by Maarten Grootendorst to perform a class-based TF-IDF.

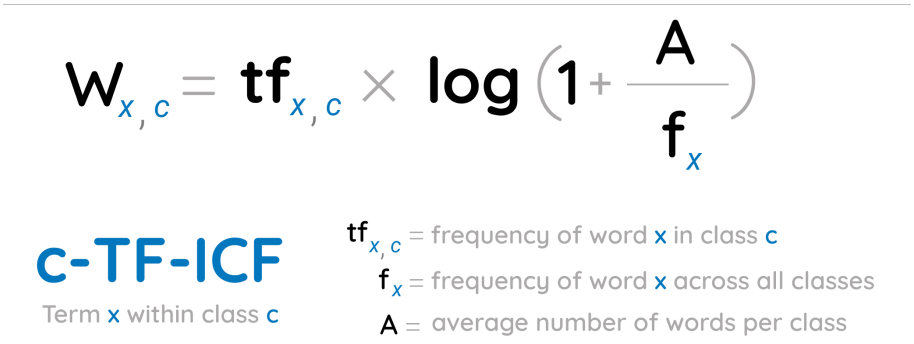

Figure 3.8: c-TF-IDF Formula by Grootendorst (2020)

The purpose of the class-based TF-IDF is to provide the same class vector to all documents inside a single class. The frequency of each word t is extracted for each class i and divided by the total number of words w. This is a form of regularization of frequent words in the class, then the total number of documents m is divided by the total frequency of word t across all classes n. As a result, rather than modeling the value of individual documents, this class-based TF-IDF approach models the significance of

words in clusters. This enables us to create topic-word distributions for each document cluster (Grootendorst, 2022).

3.6.4 BERTopic Parameters

For this deep learning approach, there were also some parameters [7]that needed to be taken into consideration, such as **Language:** The primary language used in the data, **Number of Topics:** The optimal number of topics extracted from the corpus, and **Embedding model:** The embedding model used. In this case, these parameters were set to the following values:

Parameters	**Value**
Number of Topics	10
Language	Multilingual
Embedding model	SentenceTransformers[8]

Table 3.3: BERTopic Parameters

For comparison purposes, the number of topics was set to 10 for all topic models. The following section presents some limitations that need to be considered when implementing the last model.

3.6.5 BERTopic Limitations

Compared with other topic modeling techniques, BERTopic, when it comes to topic representation, does not consider the cluster's centroid. A cluster centroid is 'a vector that contains one number for each variable, where each number is the mean of a variable for the observations in that cluster. The centroid can be thought of as the multi-dimensional average of the cluster' (Zhong, 2005). While BERTopic takes a different approach, it concentrates on the cluster as a whole, attempting to simulate the cluster's topic representation. This provides for a broader range of subject representations while ignoring the concept of centroids. Depending on the data type, ignoring the cluster's centroids can be a disadvantage. Moreover, even though BERTopic's transformer-based language models allow for contextual representation of documents, the topic representation does not directly account for this because it is derived from bags-of-words. The words in a subject representation illustrate the significance of terms in a topic while also implying that those words are likely to be related. As a result, terms in a topic may be identical to one another, making them redundant for the topic's interpretation (Grootendorst, 2022). Finally, an essential disadvantage of BERTopic is the time needed for fine-tuning.

After an in-depth research was conducted on the possible parameters (Tables 3.1, 3.2 and 3.3) and limitations that can occur all models were implemented. The code used for implementing the models can be found here: repository [9]. The following chapter presents the results generated from the three topic modeling approaches and the evaluation. First, an overview of the results is provided, and then a comparative

[7] `https://maartengr.github.io/BERTopic/api/bertopic.html`
[8] `https://www.sbert.net/docs/pretrained_models.html`
[9] `https://github.com/cltl-students/Andronikou_Konstantina_text_mining_thesis`

evaluation will be presented. Finally, an error analysis is discussed to investigate the topic predictions.

Chapter 4

Results

4.1 LDA Results

The output generated from the LDA algorithm contains topics, per-document topic assignments, and topic proportions. The results of LDA analysis on the customer conversations in the airline domain are presented in the following Figure 4.1. As mentioned, the topics are distributed over words; the top ten words with the highest frequency derived from posterior distribution are provided for each topic.

Topic 0	
Keywords	**Frequency**
test	0.049
information	0.031
amsterdam	0.025
document	0.020
netherlands	0.020
embassy	0.019
country	0.017
authority	0.016
covid	0.016
check	0.016

Topic 1	
Keywords	**Frequency**
voucher	0.154
receive	0.073
email	0.068
request	0.033
refund	0.029
address	0.022
check	0.019
book	0.018
delay	0.012
wait	0.011

Topic 2	
Keywords	**Frequency**
voucher	0.065
refund	0.060
open	0.042
respond	0.039
notify	0.037
message	0.035
webcare	0.035
policy	0.027
situation	0.020
business	0.017

Topic 3	
Keywords	**Frequency**
reservation	0.145
message	0.129
request	0.075
social	0.061
advisor	0.060
processing	0.060
facilitate	0.060
book	0.020
check	0.018
confirmation	0.011

Topic 4	
Keywords	**Frequency**
message	0.087
reply	0.053
book	0.042
accept	0.041
virus	0.040
apology	0.038
answer	0.038
work	0.035
depart	0.034
valid	0.032

Topic 5	
Keywords	**Frequency**
website	0.089
receive	0.083
voucher	0.079
refund	0.046
schedule	0.046
update	0.044
answer	0.044
board	0.041
regulation	0.040
health	0.040

Topic 6	
Keywords	**Frequency**
seat	0.044
check	0.040
airport	0.037
available	0.029
luggage	0.028
message	0.017
service	0.015
claim	0.013
website	0.012
earlier	0.011

Topic 7	
Keywords	**Frequency**
change	0.127
date	0.054
ticket	0.044
book	0.029
fare	0.022
confirm	0.016
passenger	0.015
luggage	0.014
amsterdam	0.013
option	0.012

Topic 8	
Keywords	**Frequency**
refund	0.127
receive	0.046
card	0.040
book	0.034
email	0.030
request	0.028
account	0.027
money	0.020
credit	0.019
confirmation	0.018

Topic 9	
Keywords	**Frequency**
service	0.071
answer	0.052
rebook	0.049
reservation	0.047
wait	0.047
check	0.044
colleague	0.043
update	0.042
center	0.040
longer	0.038

Figure 4.1: LDA Results

A label can be generated for a more straightforward observation of the topics rather than presenting them as combination of words. It takes human judgment to assess the coherence of the topics to give a label (Blei, 2012). This project will not take the further step and assign labels as this is not the primary focus of the report. However, in a further stage of this research an error analysis is presented where some potential labels are suggested. Unfortunately, automatic topic labeling is sometimes difficult to generalize due to the topic discovery of the unsupervised learning process (Bastani et al., 2019).

Based on the results generated for all ten clusters, it can be seen that the highest frequency scores were assigned to the following words: 'voucher' with a score of 0.154, 'reservation' with a value of 0.145, 'change' and 'refund' with a frequency of 0.127. It can also be seen that for all topics there is a decrease in terms of the frequency scores. **Topic 0** recorded the highest occurring term to be 'test' with a frequency score of 0.049. This topic had a frequency score range between 0.049 and 0.016 for all ten words. In the case of **Topic 1** the most frequent term is 'voucher' with 0.154, and the values assigned ranged between 0.154 and 0.011. For the next cluster, **Topic 2** the word 'voucher' has the highest scoring term with 0.065. The terms have a rapid decrease in frequency, leading to a range between 0.065 and 0.017. The same case can be reported for **Topic 3**, as the highest frequency score is 0.145 for the term 'reservation' and the frequency value ranges between 0.145 and 0.011. For the following cluster, **Topic 4**, the highest frequency term is 'message' with a score of 0.087. The range of this cluster is between 0.087 and 0.032. **Topic 5** reports the term 'book' as the highest frequency with a score of 0.108 and a range between 0.108 and 0.040. The following cluster, in comparison with the previous topics has a low-frequency score for the first keyword. The highest occurrences were reported for the word 'seat' with a score of 0.044 and the lowest score set to 0.011. For the following topics, **Topic 7** and **Topic 8**, the highest frequency score was 0.127 for the terms 'change' and 'refund'. The former cluster ranges between 0.127 and 0.012 and the latter, between 0.127 and 0.018. Finally, the last topic, **Topic 9** has the highest occurrence score assigned for the term 'service' with a score of 0.071 and the lowest with 0.038.

To evaluate the coherence and semantic similarity of the terms within the topic, a further investigation will be conducted. The sections 4.4 and 4.5 describes this procedure. The following part of the report presents the results generated from the second approach, NMF.

4.2 NMF Results

The outcome generated from the NMF algorithm is also per-document topic assignments and topic proportions. For this approach, the topics are also distributions over words. The top ten words with the highest frequency derived from posterior distribution are provided for all ten clusters. Figure 4.2 presents the results recorded when implementing the second traditional approach.

Topic 0	
Keywords	**Frequency**
message	0.139
reservation	0.138
request	0.075
social	0.061
advisor	0.060
processing	0.060
facilitate	0.060
test	0.010
ticket	0.009
available	0.009

Topic 1	
Keywords	**Frequency**
voucher	0.376
receive	0.039
longer	0.013
offer	0.012
situation	0.012
notify	0.012
open	0.012
policy	0.012
respond	0.012
email	0.010

Topic 2	
Keywords	**Frequency**
check	0.159
airport	0.049
available	0.031
information	0.026
amsterdam	0.025
confirm	0.024
test	0.021
website	0.019
invite	0.012
document	0.013

Topic 3	
Keywords	**Frequency**
departure	0.195
date	0.140
arrival	0.088
propose	0.086
gonalves	0.086
money	0.083
account	0.082
passenger	0.040
ticket	0.032
voucher	0.011

Topic 4	
Keywords	**Frequency**
service	0.053
answer	0.047
webcare	0.042
wait	0.038
colleague	0.037
rebook	0.034
reservation	0.032
update	0.028
center	0.028
traveler	0.028

Topic 5	
Keywords	**Frequency**
book	0.108
email	0.104
receive	0.085
confirmation	0.038
address	0.034
confirm	0.025
request	0.020
payment	0.018
reply	0.017
passenger	0.016

Topic 6	
Keywords	**Frequency**
change	0.219
date	0.045
ticket	0.032
fare	0.029
book	0.027
confirm	0.018
amsterdam	0.014
option	0.013
free	0.012
return	0.011

Topic 7	
Keywords	**Frequency**
website	0.107
receive	0.101
answer	0.055
refund	0.052
update	0.051
health	0.047
declaration	0.046
visit	0.046
schedule	0.045
wait	0.045

Topic 8	
Keywords	**Frequency**
refund	0.277
request	0.046
voucher	0.031
card	0.029
passenger	0.026
money	0.025
account	0.024
credit	0.014
book	0.013
situation	0.012

Topic 9	
Keywords	**Frequency**
luggage	0.110
seat	0.061
hold	0.038
book	0.020
extra	0.020
ticket	0.020
baby	0.020
bring	0.019
passenger	0.018
piece	0.014

Figure 4.2: NMF Results

Across all the topics generated, it can be seen that the following terms scored the highest frequency scores:'voucher' with 0.376, 'refund' with a score of 0.277, 'change' with a value of 0.219 and 'departure' with a frequency of 0.195. **Topic 0** has a range of frequency scores between 0.139 and 0.009, with the highest occurring term 'message'. The next cluster, **Topic 1** contains the word 'voucher' with 0.376 and the distribution of the terms ranges between 0.376 and 0.010. In the case of **Topic 2**, 'check' scored the highest frequency score with 0.159. This topic had a distribution range between 0.159 and 0.013. The following topic, **Topic 3**, is a similar case to the previous cluster as the highest occurrence score was recorded for the term 'departure' with 0.195. The distribution of all ten terms ranges between 0.195 and 0.011. In the case of **Topic 4**, the highest scoring word is 'service' with a score of 0.053, and the cluster ranges between the highest score and 0.028. For the following topic, **Topic 5**, the highest score recorded was 0.108 for the term 'book'. The frequency values for this case were between 0.108 and 0.016. **Topic 6** generated the term 'change' as the highest occurring word with a score of 0.219 and a range reaching 0.011. A similar decrease can also be reported for **Topic 7**, as the highest value was 0.107 for the word 'website', and the lowest distribution was set to 0.045. Finally, for the last two topics, the highest scoring words were 'refund' with a score of 0.277 and 'luggage' with 0.110. For **Topic 8** the distribution range is between 0.277 and 0.012 while for **Topic 9** is between 0.110 and 0.014.

It can be seen that the clusters generated by NMF are similar to the topics reported for the first traditional approach, LDA. The following section presents the results of the final model implemented, the deep learning approach.

4.3 BERTopic Results

As mentioned before, this topic modeling approach generates topic representations through three steps. The following Figure 4.3 presents the result clusters recorded when implementing the deep learning approach. It is worth mentioning that some

words are censored from the results due to the privacy regulations of the company.

Topic 0	
Keywords	**Frequency**
respond	0.189
notify	0.122
open	0.114
message	0.097
earlier	0.095
webcare	0.094
policy	0.093
business	0.089
voucher	0.081
cancel	0.066

Topic 1	
Keywords	**Frequency**
test	0.102
pcr	0.060
greece	0.054
message	0.049
embassy	0.048
information	0.045
reservation	0.045
flight	0.044
spain	0.043
[redacted]	0.042

Topic 2	
Keywords	**Frequency**
luggage	0.100
dog	0.089
kg	0.069
hold	0.067
pet	0.058
cabin	0.048
add	0.043
flight	0.043
seat	0.42
reservation	0.041

Topic 3	
Keywords	**Frequency**
website	0.120
receive	0.103
voucher	0.095
flight	0.093
declaration	0.086
relate	0.085
regulation	0.085
visit	0.084
heath	0.84
board	0.082

Topic 4	
Keywords	**Frequency**
website	0.080
receive	0.079
flight	0.071
voucher	0.069
refund	0.056
relate	0.052
update	0.052
regulation	0.052
schedule	0.051
board	0.051

Topic 5	
Keywords	**Frequency**
updatesnl	0.104
traveller	0.103
center	0.101
keep	0.094
rebook	0.089
colleague	0.089
webcare	0.077
service	0.076
busy	0.072
wait	0.072

Topic 6	
Keywords	**Frequency**
message	0.162
assistant	0.150
digital	0.149
relevant	0.146
virus	0.138
depart	0.135
valid	0.120
accept	0.119
way	0.116
apology	0.115

Topic 7	
Keywords	**Frequency**
voucher	0.103
cancel	0.100
flight	0.097
receive	0.090
website	0.085
relate	0.080
regulation	0.080
schedule	0.077
board	0.076
refund	0.067

Topic 8	
Keywords	**Frequency**
survey	0.488
cooperation	0.445
complete	0.408
service	0.389
brief	0.266
feedback	0.265
experience	0.247
center	0.240
hear	0.389
[redacted]	0.177

Topic 9	
Keywords	**Frequency**
daughter	0.057
flight	0.054
wife	0.051
charge	0.037
mother	0.036
book	0.034
message	0.033
reservation	0.033
[redacted]	0.030
ticket	0.029

Figure 4.3: BERTopic Results

Based on all ten topics that have been generated, it can be seen that the highest frequency words were: 'survey' with a distribution score of 0.488, 'cooperation' with 0.445, 'complete' with a score of 0.408 and 'service' with a score of 0.389. In the case of **Topic 0**, the highest frequency value was reported for the term 'respond' with 0.189, while the distribution ranges between 0.189 and 0,066. For **Topic 1** the frequency scores ranged between 0.102 and 0.042, with the highest distribution occurrence assigned to the term 'test'. **Topic 2** presents the word 'luggage' as the highest frequency within the topic. The distribution of the cluster ranges between 0.100 and 0.041. A similar decrease can be seen in the next cluster, **Topic 3**, as the highest value is 0.120 with a distribution range reaching 0.082. In this case, the highest score is assigned to the term 'website'. **Topic 4** also has the word 'website' as the highest frequency scoring word with 0.080 and the lowest being 0.051. The following topic, **Topic 5**, has a range of 0.104 and 0.072, with the highest scoring term being 'updatesnl'. In the case of **Topic 6**, the term 'voucher' has the highest score with 0.103 and 0.067 as the lowest. **Topic 8** includes the highest range out of all topics starting with a distribution score of 0.488 until 0.177. The highest value was assigned to the term 'survey'. Finally, for the last cluster, **Topic 9** the highest distribution occurrence is set to 0.057 for the term 'daughter'. The scores of this topic range between 0.057 and 0.029.

4.4 Evaluation

To evaluate the topic quality and predictive performance of the results generated from all three models, the following methods were implemented: topic coherence c_v, topic

coherence u-mass and cosine similarity.

Topic coherence is a part of the larger subject of what are good topics, what properties of a document collection make it more suitable for topic modeling, and how can topic modeling's potential be utilized for human benefit (Newman et al., 2010). This evaluation method can be defined as the degree of significance between the words inside a topic in terms of how interpretable it is. The goal of the topic coherence metrics employed in this study is to assess the quality of topics from a human-like standpoint. While, the implementation of an overall evaluation based on cosine similarity is to give an insight on the semantic similarity within the topics. These evaluation methods were chosen due to the frequent use in previous research on topic modeling such as Keane et al. (2015),Grootendorst (2022), Kapadia (2019), etc. For this project, the following evaluation approaches were used:

- **C_v:** this measure 'is based on a sliding window, one-set segmentation of the top words and an indirect confirmation measure that uses normalized pointwise mutual information (NPMI) and the cosine similarity' (Mifrah and Benlahmar, 2020, p.5758)
- **C_umass:** this measure takes into consideration the document co-occurrence counts, one-preceding segmentation, and a logarithmic conditional probability as a confirmation measure (Mifrah and Benlahmar, 2020).
- **Cosine Similarity (CS):** the direction of two vectors is calculated by measuring the cosine of the angle between them. With the use of the inner space the similarity of two vectors is measured (Alodadi and Janeja, 2015). In this case the following reported cosine similarity score is the overall average of for all topics.

It is worth mentioning that the highest the score of the c_v, the more coherent and understandable a topic can be to a human. While in the case of c_umass, the closer the score is to 0, the better. Finally, the highest the score for cosine similarity the better semantic correlation across the words within a topic. The following Table 4.1 presents the results generated when implementing the chosen evaluation methods.

Topic Models	**C_V**	**C_umass**	**CS**
LDA	0.58	-1.58	0.19
NMF	0.55	-1.74	0.20
BERTopic	0.74	-0.01	0.23

Table 4.1: Evaluation Results

Based on the scores reported, it can be seen that the highest performing topic model is BERTopic, with a c_v coherence score of 0.74, c_umass -0.01. In the case of the traditional models, it can be seen that the scores do not have a substantial difference. LDA scored 0.58 for the first coherence measure and -1.58 for the second one, while NMF scored 0.55 for c_v and -1.74 for c_umass. Even though NMF scored the lowest for both topic coherence evaluation methods in the case of cosine similarity it slightly outperformed LDA. BERTopic scored the highest score out of all models for also this measurement. However, the scores reported for cosine similarity do not have a remarkable difference.

To investigate in greater detail the results provided by the different topic modeling algorithms an error analysis was conducted. This analysis is aiming to give more details on the performance of the models and what needs to be taken into consideration for future studies. The following section presents the error analysis of the results obtained by all three topic modeling algorithms.

4.5 Error Analysis

Based on the evaluation results, we can get an idea of the performance of the models in terms of predictive performance and topic quality. However, further analysis is required in order to understand the scores evaluated. It can be seen that the scores of the topic coherence for the traditional models were not high compared to the deep learning approach. This section aims to investigate the procedure that resulted in the reported scores. A crucial factor that can influence the performance of these topic models is the coherence between the terms within a topic. It might be the case that due to low semantic relation between the words, the overall topic coherence decreases.

Pairwise cosine similarity is implemented to analyze the semantic correlation between the combined words within a topic. Based on previous research such as Keane et al. (2015), Luo et al. (2022) evaluation was conducted based on an overall cosine similarity score. However, this project aims to investigate a level deeper by measuring the cosine similarity based on all possible word pairs within a topic. With this method, terms for each topic are measured in pairs to find the most semantic correlated and less correlated words. In this case, we aim to investigate the topic quality in greater detail by analyzing the word combinations generated from each topic model. This method was inspired by Belford and Greene (2020) and O'callaghan et al. (2015). However, to our knowledge, this additional evaluation step of pairwise cosine similarity on all possible pairs in a topic modeling output has not been implemented before. In addition to pairwise cosine similarity scores, an observation-based human judgment is combined for a better intuitive comprehension of topic quality and predictive performance. The following methods were chosen to analyze the content of each topic individually.

- **Pairwise Cosine Similarity (CS):** the semantic similarity is measured based on all possible pairs within a topic to determine whether the angel of two vectors is pointing in the same direction.
- **Observation-based Human Judgment:** evaluating by looking at the most probable words in the topic. In this case, the top 10 words of each topic are used for this approach.

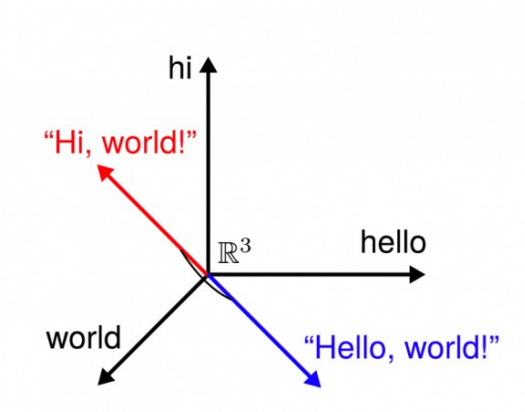

Figure 4.4: Cosine Similarity (Al Ghamdi and Khan, 2022)

For example, in Figure 4.4, it can be seen that 'Hello World' is closer in the vector space to the terms 'Hello' and 'world' compared to 'hi'. In this case in order to further investigate the terms within the topics, an embedding model was created based on the data provided for the project. It was decided to generate an embedding model based on the data for this project. As the data is from a specific domain, it was hypothesised that embeddings trained on general data will not reflect the peculiarities of this domain. Moreover, there was lack of resources in terms of pre-trained embedding models on the airline domain. The data set included conversations such as 'Hello, I would like to book an additional cabin luggage but I am not able to do so on the booking portal.' [1]. As it can be seen the provided example contains specific expressions used in the conversational data within this domain. Therefore, this embedding model was used for the error analysis.

To generate the embeddings, Word2Vec was used. Mikolov et al. (2013) published Word2Vec in 2013, which is a word embedding methodology for encoding words as numerical vectors in a high-dimensional space while maintaining their semantic and grammatical links. This model can be trained over a large number of documents and is used to extract concepts like semantic relatedness, synonym recognition, concept classification, etc. A Word2Vec model discovers meaningful relationships and converts them into vector similarities. The usage of this model can reduce the chances of having word ambiguity and have a better representation within the vector space of the data. This embedding model was only used for the error analysis to measure the cosine similarity of each topic. Finally, for this step a variety of different types of embeddings can be used such as pre-trained ones.

The main aim of these two methods is to analyze the terms within topics in greater detail and find a potential correlation between them. Moreover, this error analysis will try and find possible labels for generalizing the topic based on the coherence of each cluster. This project will not execute the further step of topic classification but will analyze the combination of the words and their mutual point and suggest potential labels. The following subsections present an in-depth analysis of all topics individually for each topic model.

[1] This is a demo example inspired by the original data.

4.5.1 LDA

This section presents the results based on cosine similarity and observation-based human judgment. First, an overview of the scores recorded for the cosine similarity is presented, and then an in-depth analysis based on the observation-based human judgment is described. Figure 4.5.1 presents all topics with their corresponding cosine similarity score based on the embedding representations. For the error analysis, three clusters are described in greater detail, the highest scoring, the lowest scoring and one with an average score.

Topic 0	
Keywords	**Frequency**
test	0.049
information	0.031
amsterdam	0.025
document	0.020
netherlands	0.020
embassy	0.019
country	0.017
authority	0.016
covid	0.016
check	0.016
CS	**0.210**

Topic 1	
Keywords	**Frequency**
voucher	0.154
receive	0.073
email	0.068
request	0.033
refund	0.029
address	0.022
check	0.019
book	0.018
delay	0.012
wait	0.011
CS	**0.208**

Topic 2	
Keywords	**Frequency**
voucher	0.065
refund	0.060
open	0.042
respond	0.039
notify	0.037
message	0.035
webcare	0.035
policy	0.027
situation	0.020
business	0.017
CS	**0.085**

Topic 3	
Keywords	**Frequency**
reservation	0.145
message	0.129
request	0.075
social	0.061
advisor	0.060
processing	0.060
facilitate	0.060
book	0.020
check	0.018
confirmation	0.011
CS	**0.294**

Topic 4	
Keywords	**Frequency**
message	0.087
reply	0.053
book	0.042
accept	0.041
virus	0.040
apology	0.038
answer	0.038
work	0.035
depart	0.034
valid	0.032
CS	**0.084**

Topic 5	
Keywords	**Frequency**
website	0.089
receive	0.083
voucher	0.079
refund	0.046
schedule	0.046
update	0.044
answer	0.044
board	0.041
regulation	0.040
health	0.040
CS	**0.403**

Topic 6	
Keywords	**Frequency**
seat	0.044
check	0.040
airport	0.037
available	0.029
luggage	0.028
message	0.017
service	0.015
claim	0.013
website	0.012
earlier	0.011
CS	**0.084**

Topic 7	
Keywords	**Frequency**
change	0.127
date	0.054
ticket	0.044
book	0.029
fare	0.022
confirm	0.016
passenger	0.015
luggage	0.014
amsterdam	0.013
option	0.012
CS	**0.261**

Topic 8	
Keywords	**Frequency**
refund	0.127
receive	0.046
card	0.040
book	0.034
email	0.030
request	0.028
account	0.027
money	0.020
credit	0.019
confirmation	0.018
CS	**0.239**

Topic 9	
Keywords	**Frequency**
service	0.071
answer	0.052
rebook	0.049
reservation	0.047
wait	0.047
check	0.044
colleague	0.043
update	0.042
center	0.040
longer	0.038
CS	**0.086**

Figure 4.5: LDA Cosine Similarity per topic

Topic 5

This topic scored the highest cosine similarity out of all topics, with a score of 0.403. With a relatively high cosine similarity score, it can be assumed that the terms are semantically related to some degree. Within this topic, words such as 'website', 'voucher', and 'refund' were combined. The following figure presents the highest correlation scores between word pairs.

Pairs	**CS**
voucher-receive	0.740
website-receive	0.709
refund-voucher	0.690
receive-voucher	0.652
website-answer	0.652

Table 4.2: Highest Pairwise Cosine Similarity for **Topic 5**

As it can be seen from Table 4.2, the main terms that have the highest score of semantic correlation are: 'voucher', 'refund', 'receive', and 'website'. It can be seen that specifically in the airline domain, these terms are frequently used together. Even though the definitions of these words do not align, the combination of these terms often occurs in the data of this project, and therefore the vectors are close to each other in the embedded space. Moreover, it can be hypothesised that these terms have a great semantic similarity as they do not carry any ambiguity that can influence the sentence meaning. However, there are some terms within the topic that do not have a clear semantic correlation. Table 4.3 presents the lowest scoring word pairs.

Pairs	**CS**
update-board	0.108
regulation-update	0.171
website-board	0.168
regulation-website	0.208

Table 4.3: Lowest Pairwise Cosine Similarity for **Topic 5**

Even though the term 'website' was frequently seen in the highest scoring pairs, in combination with other terms, resulted to the similarity score decreasing. This is can be due to some words having multiple meanings; for example, the word 'board' can be both a noun and a verb depending on the content. For example, 'the passengers boarded on the plane' or 'Is it allowed to bring a surfboard on the plane?'. It might be the case that the term's ambiguity influences the semantic meaning and the correlation score. These words based on the score might not be the best match, but let's look at observation-based human judgment. The following Figure presents a word cloud with the top 10 words of this cluster.

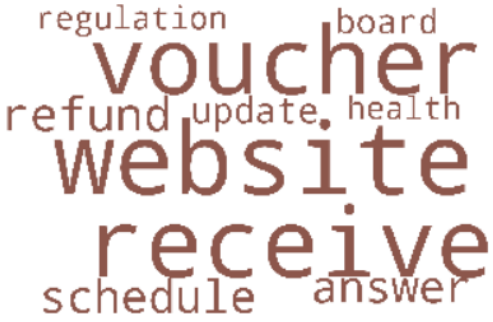

Figure 4.6: Top 10 words for **Topic 5**

Based on the top frequent words of this topic, it can be assumed that the customers were contacting the company to collect information for a variety of matters. For example, how they can find specific information on the website regarding matters such as refund, company's regulations regarding health or boarding, etc. Therefore, this topic based on a human judgment can potentially be labeled with 'FAQ Website'. Overall, even though some of the combined terms do not have a the greatest correlation based on pairwise cosine similarity, a generalised label can be produced with the combination of human judgment.

Topic 0

Based on the results generated for the first topic, it can be seen that the cosine similarity was set to 0.210. Even though this score is not high, there is some semantic correlation between the terms in the topic. For example, this topic combined words such as 'document', 'information, 'covid', and 'test'. Therefore, in terms of semantic similarity, it can be seen that to some degree, there is a correlation between the words, especially for the following terms:

Pairs	**CS**
country-netherlands	0.615
embassy-authority	0.545
document-embassy	0.543
information-document	0.465
test-document	0.458

Table 4.4: Highest Pairwise Cosine Similarity for **Topic 0**

Based on these pairs, the word 'document' is frequently semantically related to a significant amount of the terms included in the cluster. For all of the pairs, the combination and correlation found are logical if we look at the grammatical meaning. For example, the correlation is evident for the first word pair as the Netherlands is a country. In the case of the second pair, the term 'embassy' is defined as 'a group of government officials', usually associated with authority. In the airline domain, many travellers need specific documents from an embassy to be eligible to fly with an airline. Moreover, especially during the pandemic, there was a high demand for additional information before boarding a flight, such as a document certification for COVID-19 testing. However, some combinations of terms have no semantic correlation, and this might be the cause of not a relatively high overall score. The following Table presents some pairs that have scored low cosine similarity.

Pairs	**CS**
country-test	-0.084
covid-check	-0.070
information-amsterdam	-0.021
authority-check	0.012
check-country	0.084

Table 4.5: Lowest Pairwise Cosine similarity for **Topic 0**

In comparison with the previous topic presented, it is evident that there negatively correlated word-pairs. It can be seen that the term 'check' is frequently seen in these pairs. This can be due to the word's ambiguity, as it can be interpreted as both a noun and a verb. For example, in the following sentences, the meaning of the word changes, 'There were issues with the security check at the airport' or 'I paid the ticket with a check'. This word's ambiguity is something that is not considered when measuring cosine similarity. Another interesting observation is that even though the combination of 'covid-check' sounds logical, especially in the past two years regarding the similarity in the vector space, the correlation is negative. This might be the case due to the lemmatization procedure as phrases such as 'checking covid tests' or 'providing covid

documents in check-in' were frequently used. While with the lemma 'check' this correlation is lost. However, the following remarks were reported when observing the topic through human judgment. The following Figure presents a word cloud with the top 10 terms for this topic.

Figure 4.7: Top 10 words for **Topic 0**

Based on the generated word cloud, it can be assumed that the customers were contacting the company to ask questions concerning information about COVID-19-related documents or tests that need to provide when entering a country such as the Netherlands. For this topic, an individual can potentially generate a label for this cluster, such as 'COVID-19 travelling rules Netherlands'.

Topic 4

This topic was one of the clusters that scored the lowest in cosine similarity, with a score of 0.084. Based on only the score, an individual can assume that the combination of these terms from the model cannot be easily generalised. It is expected that the word combination might not be ideal for providing a potential label. This combination contains a diverse word variety, combining terms such as 'virus', 'apology', 'depart', etc. The following Table presents the most semantically correlated terms within the topic.

Pairs	**CS**
accept-virus	0.435
accept-valid	0.357
apology-reply	0.271
message-apology	0.264
answer-reply	0.248

Table 4.6: Highest Pairwise Cosine Similarity for **Topic 4**

Based on the results of these pairs, it can be seen that the highest scoring word pair is 'accept-virus'. Even though this pair scored an average score, there is no clear indication of the semantic correlation. This is also the case for most word pairs, except the last pair, as the terms 'answer' and 'reply' can be considered synonyms. It can be assumed that in this specific case, these words were frequently used in the same conversation within the data and therefore are close to each other in the embedded space. If we combine human perspective to analyse these pairs further, some assumptions can be drawn for the correlation of the words. On various occasions, big companies have

to apologise for inconvenient situations; customers contact them for the word pairs 'message-apology', and 'apology-reply' is logical in this case. Within this topic, there was a significant amount of negative values for some terms.

Pairs	CS
virus-work	-0.085
depart-answer	-0.066
answer-book	-0.053
accept-answer	-0.028
valid-depart	-0.16

Table 4.7: Lowest Pairwise Cosine Similarity for **Topic 4**

In this case, the terms that frequently appeared in the low correlated pairs are: 'answer' and 'depart'. Even though these words are commonly used in the airline domain, the semantic correlation is low. Therefore, let's observe the terms based on an observation-based human judgment, without considering the cosine similarity score. The following Figure presents the word cloud generated for topic 4.

Figure 4.8: Top 10 word for **Topic 4**

In the case of topic 4, the combination generated by the model is not ideal, especially for classification purposes. This cluster contains some outliers that prevent an individual from generalizing the concept of the topic. In this case, terms such as 'virus', 'work' and 'depart' can be considered the outliers as they are not correlated with the rest of the assigned terms.

LDA Conclusion Remarks

Based on the topics assigned and the overall results, it can be concluded that topics with a cosine similarity ranging from 0.403-to 0.208 can be generalized as the terms correlate to some degree. This leads to an overall rate of 60% of the results that can be generalized and be used for the next step of topic classification. In this case, the six topics that can potentially be classified are Topic 0,1,3,5,7,8. On the other hand, the topics with a lower assigned value can not be generalized as there are clusters with a high number of outliers that will make the classification step harder.

4.5.2 NMF

This section presents the error analysis of the results of the second traditional approach, NMF. For the in-depth analysis, not all clusters are described only three topics, the highest scoring, average and low scoring.

Topic 0	
Keywords	**Frequency**
message	0.139
reservation	0.138
request	0.075
social	0.061
advisor	0.060
processing	0.060
facilitate	0.060
test	0.010
ticket	0.009
available	0.009
CS	**0.230**

Topic 1	
Keywords	**Frequency**
voucher	0.376
receive	0.039
longer	0.013
offer	0.012
situation	0.012
notify	0.012
open	0.012
policy	0.012
respond	0.012
email	0.010
CS	**0.101**

Topic 2	
Keywords	**Frequency**
check	0.159
airport	0.049
available	0.031
information	0.026
amsterdam	0.025
confirm	0.024
test	0.021
website	0.019
invite	0.012
document	0.013
CS	**0.109**

Topic 3	
Keywords	**Frequency**
departure	0.195
date	0.140
arrival	0.088
propose	0.086
gonalves	0.086
money	0.083
account	0.082
passenger	0.040
ticket	0.032
voucher	0.011
CS	**0.132**

Topic 4	
Keywords	**Frequency**
service	0.053
answer	0.047
webcare	0.042
wait	0.038
colleague	0.037
rebook	0.034
reservation	0.032
update	0.028
center	0.028
traveler	0.028
CS	**0.078**

Topic 5	
Keywords	**Frequency**
book	0.108
email	0.104
receive	0.085
confirmation	0.038
address	0.034
confirm	0.025
request	0.020
payment	0.018
reply	0.017
passenger	0.016
CS	**0.266**

Topic 6	
Keywords	**Frequency**
change	0.219
date	0.045
ticket	0.032
fare	0.029
book	0.027
confirm	0.018
amsterdam	0.014
option	0.013
free	0.012
return	0.011
CS	**0.294**

Topic 7	
Keywords	**Frequency**
website	0.107
receive	0.101
answer	0.055
refund	0.052
update	0.051
health	0.047
declaration	0.046
visit	0.046
schedule	0.045
wait	0.045
CS	**0.427**

Topic 8	
Keywords	**Frequency**
refund	0.277
request	0.046
voucher	0.031
card	0.029
passenger	0.026
money	0.025
account	0.024
credit	0.014
book	0.013
situation	0.012
CS	**0.184**

Topic 9	
Keywords	**Frequency**
luggage	0.110
seat	0.061
hold	0.038
book	0.020
extra	0.020
ticket	0.020
baby	0.020
bring	0.019
passenger	0.018
piece	0.014
CS	**0.251**

Figure 4.9: NMF Cosine Similarity per topic

Topic 7

It can be seen from the generated results that the highest cosine similarity was reported for topic 7, with a score of 0.427. The following Table presents five-word pairs with the highest semantic correlation based on pair-based cosine similarity.

Pairs	**CS**
website-receive	0.709
receive-refund	0.652
website-answer	0.652
visit-website	0.638
receive-answer	0.630

Table 4.8: Highest Pairwise Cosine Similarity for **Topic 7**

Based on these word pairs, it can be seen that topic 7 contains similar terms to topic 5 from LDA. In this case, the term 'website' and 'receive' frequently appear in the highest correlated pairs. Even though these terms have entirely different definitions, it is the case that these combinations do often occur in this data and domain. This is one of the reasons why these words are highly correlated in the embedded space, as the vector space was created specifically with the data provided for the project. If the human eye observes these terms, there is a logical and clear correlation between them. For example, many travellers contact a company to inform if they received a refund requested or request to receive an answer for a matter. Moreover, it is the case many customers visit the website of the company to find specific information. As LDA clusters, this model also contains some less correlated words. Even though the scores are significantly low, this topic has no negative values in terms of cosine similarity.

Pairs	CS
schedule-wait	0.032
health-wait	0.203
visit-schedule	0.232
wait-declaration	0.247
declaration-refund	0.261

Table 4.9: Lowest Pairwise Cosine Similarity for **Topic 7**

In this case, it can be seen that the term most frequently not correlating with the rest of the words within the cluster is 'wait'. It can be assumed that for this topic, 'health declaration' was a bi-gram, but due to the tokenization step, the words were separated. The individual components of this bi-gram do not have a high semantic relation with the rest of the terms. If we look at this combination of words through human judgment, a potential label can be generated to generalize the topic.

Figure 4.10: Top 10 words for **Topic 7**

For this topic, it can be hypothesized that the customers contacted the company on how to access or request specific information on the website. As mentioned, this topic is similar to topic 0 from the first model; therefore, the same label can be assigned, 'FAQ Website'.

Topic 9

This topic scored an overall cosine similarity of 0.251; however, the combination of terms has a great semantic similarity. For example, the following pairs have the highest correlation within the cluster.

Pairs	CS
hold-piece	0.636
bring-piece	0.561
piece-luggage	0.519
hold-extra	0.492
baby-piece	0.478

Table 4.10: Highest Pairwise Cosine Similarity for **Topic 9**

Based on Table 4.10, it can be seen that the most correlated pair is 'hold-piece'. In the airline domain, the most frequently discussed matter is in terms of baggage arrangements. Many of the terms can be interpreted in multiple ways. For example,

'hold' can occur as a verb and a noun. In terms of the words 'hold' in the airline domain, it is usually used to refer to 'hold-luggage'. This term concerns baggage that will be stored in the aeroplane's cargo hold. Therefore, the score for the combination of 'hold-piece' and 'hold-extra' is high due to the high correlation of the words in the domain. This is also the case for the term 'piece'; in this domain, the word is used as a noun, for example, 'The company pieced multiple destinations together' or 'The passenger had one piece of luggage which can also be seen from the high cosine similarity between 'piece-luggage'.In the case of the final word-pair, it can be assumed that the correlation is high as passengers need to report when travelling with a baby in a similar way to reporting for baggage. However, some terms do not have a significant correlation with each other.

Pairs	**CS**
book-bring	-0.230
bring-ticket	-0.088
book-baby	-0.082
luggage-passenger	-0.024
book-piece	-0.022

Table 4.11: Lowest Pairwise Cosine Similarity for **Topic 9**

It can be seen that the term that frequently appears in the lowest correlation pairs is 'book'. As also mentioned, this can be due to the word's ambiguity, as it can occur as a noun and a verb depending on the content. Moreover, it is interesting that the terms 'luggage-passenger' have a low correlation score in this domain. From a human perspective, this pair is logical as all passengers are carrying some form of luggage when travelling. The following word cloud presents the top 10 words that contribute to generating a potential label.

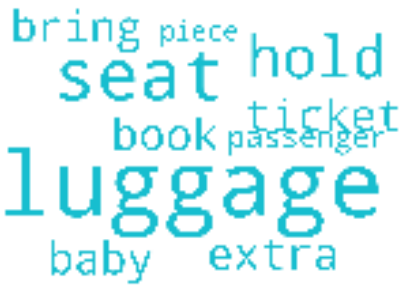

Figure 4.11: Top 10 words for **Topic 9**

Based on an observation-human judgment, it can be seen that the customers, in this case, were contacting the company to receive information about extra flight details such as the arrangement of a piece of luggage or a seat. Therefore, a potential label, such as 'Fare options' can be assigned for this topic.

Topic 4

The last topic presented for the NMF topic model is topic 4; this cluster reported the lowest cosine similarity between the terms. The overall score was 0.078. The following Table presents five-word pairs that correlate to some degree.

Pairs	CS
answer-update	0.437
rebook-traveler	0.426
answer-wait	0.366
traveler-colleague	0.340
webcare-update	0.300

Table 4.12: Highest Pairwise cosine Similarity for **Topic 4**

In this case, it can be seen that the highest correlated word pairs are mainly about customer service. The term 'webcare' is defined as the online interactions between companies and customers about questions, complaints, and experiences concerning the organization's products or services. All word pairs have a visible correlation except the fourth one. For this pair, 'traveller-colleague' seems to have a significant correlation but is not visible. It can be hypothesized that these terms correlate specifically to this domain. It is the case that many people travel for business and are accommodated by colleagues. In the case of the word 'answer,' many customers are requesting an answer from the company for an update, or the customers have to wait for an answer. Even though this topic can hint at the conversations executed, some terms do not fit the overall topic.

Pairs	CS
colleague-center	-0.296
webcare-reservation	-0.140
traveler-center	-0.127
traveler-service	-0.119
service-webcare	-0.088

Table 4.13: Lowest Pairwise Cosine Similarity for **Topic 4**

It can be seen from Table 4.13 that there is a variety of different terms that have a negative semantic similarity. Even though these pairs might not have a connection semantically based on human judgment, there is a connection. For example, the word pair 'traveller-service' is logical as many travellers contact a company's service for potential concerns or questions. This is also the case for 'service-webcare' as both are concerned about contacting the company for help. For the following instances, it can also be the case that the words 'service' and 'centre' were once a bi-gram. In the bi-gram form, 'service center' can be considered a synonym for 'webcare', but due to the tokenization procedure, the terms are separated and do not carry the same meaning. The following word cloud presents the top 10 most frequent words within the topic, which can help produce a potential label.

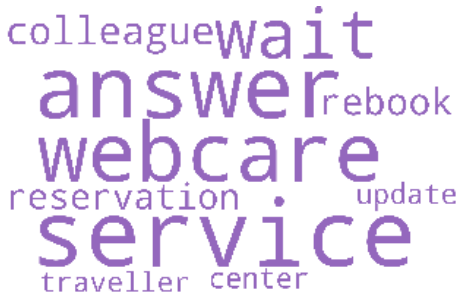

Figure 4.12: Top 10 words for **Topic 4**

Even though the cosine similarity scores for this topic were not significantly high, a potential label can be generated based on observation-based human judgment. Due to the combination of words such as 'webcare', 'service', 'traveler', it can be assumed that this topic is related to the customer center, or service option that the company has. Therefore a potential label that can describe topic 4 is 'Customer Service'.

NMF Conclusion Remarks

Based on the topic generated, in this case, the cosine similarity is not the main factor in determining if a topic can be generalized. Evaluating in terms of cosine similarity can give some insight into how semantically correlated the terms are within the topic. However, in combination with observation-based human judgment, the evaluation procedure can be more accurate. In this case, it can be seen that 60% of the cluster can be generalized despite the similarity score. Seven topics can be easy to classify, topics 4,5,6,7,8,9. The rest of the topics have a high number of outliers which makes the assignment of a potential label harder.

4.5.3 BERTopic

The following section presents the error analysis of the deep learning approach's last topic model. The error analysis will be executed in terms of cosine similarity and observation-based human judgment. The following Table presents the cosine similarity scores for each topic individually. Some of the words within the topic are covered due to privacy reasons concerning the company providing the data. As mentioned before, this topic model uses embeddings which consider the semantic relation of the terms before combining terms and generating a topic. Therefore, the words within the topics are expected to correlate better than the traditional approaches. The following Figure gives an overview of the cosine similarity for each topic.

Topic 0	
Keywords	**Frequency**
respond	0.189
notify	0.122
open	0.114
message	0.097
earlier	0.095
webcare	0.094
policy	0.093
business	0.089
voucher	0.081
cancel	0.066
CS	**0.084**

Topic 1	
Keywords	**Frequency**
test	0.102
pcr	0.060
greece	0.054
message	0.049
embassy	0.048
information	0.045
reservation	0.045
flight	0.044
spain	0.043
[redacted]	0.042
CS	**0.185**

Topic 2	
Keywords	**Frequency**
luggage	0.100
dog	0.089
kg	0.069
hold	0.067
pet	0.058
cabin	0.048
add	0.043
flight	0.043
seat	0.42
reservation	0.041
CS	**0.321**

Topic 3	
Keywords	**Frequency**
website	0.120
receive	0.103
voucher	0.095
flight	0.093
declaration	0.086
relate	0.085
regulation	0.085
visit	0.084
heath	0.84
board	0.082
CS	**0.388**

Topic 4	
Keywords	**Frequency**
website	0.080
receive	0.079
flight	0.071
voucher	0.069
refund	0.056
relate	0.052
update	0.052
regulation	0.052
schedule	0.051
board	0.051
CS	**0.397**

Topic 5	
Keywords	**Frequency**
updatesnl	0.104
traveller	0.103
center	0.101
keep	0.094
rebook	0.089
colleague	0.089
webcare	0.077
service	0.076
busy	0.072
wait	0.072
CS	**0.089**

Topic 6	
Keywords	**Frequency**
message	0.162
assistant	0.150
digital	0.149
relevant	0.146
virus	0.138
depart	0.135
valid	0.120
accept	0.119
way	0.116
apology	0.115
CS	**0.121**

Topic 7	
Keywords	**Frequency**
voucher	0.103
cancel	0.100
flight	0.097
receive	0.090
website	0.085
relate	0.080
regulation	0.080
schedule	0.077
board	0.076
refund	0.067
CS	**0.415**

Topic 8	
Keywords	**Frequency**
survey	0.488
cooperation	0.445
complete	0.408
service	0.389
brief	0.266
feedback	0.265
experience	0.247
center	0.240
hear	0.389
[redacted]	0.177
CS	**0.206**

Topic 9	
Keywords	**Frequency**
daughter	0.057
flight	0.054
wife	0.051
charge	0.037
mother	0.036
book	0.034
message	0.033
reservation	0.033
[redacted]	0.030
ticket	0.029
CS	**0.156**

Figure 4.13: BERTopic Cosine Similarity per topic

Topic 7

This topic has scored the highest cosine similarity across all ten topics generated by the model. Therefore, the score was set to 0.415, with the following terms having the highest semantic correlation.

Pairs	**CS**
voucher-receive	0.740
refund-voucher	0.709
website-receive	0.690
receive-refund	0.652
cancel-flight	0.574

Table 4.14: Highest Pairwise Cosine Similarity for **Topic 7**

Based on the Table, it can be seen that the highest scoring word pairs mainly concern matters such as ‘voucher’, ‘refund’, etc. The most correlated verb, in this case, is ’receive’ and cancel ’as it is frequently used in the airline domain. It is the case that many customers contact airline companies concerning a flight cancellation or for any potential outstanding matters such as receiving a refund or a voucher. Moreover, these

terms are frequently used in customer conversations as the first four word pairs have also been reported as results from the two traditional approaches. However, like any other model, some terms do not significantly correlate with the overall combination of words.

Pairs	CS
website-board	0.168
website-regulation	0.208
relate-board	0.245
refund-board	0.248
regulation-flight	0.299

Table 4.15: Lowest Pairwise Cosine Similarity for **Topic 7**

It can be seen that the primary term that does correlate with a great number of words is 'board'. As mentioned, this term is ambiguous, which can lead to controversial meaning; it can be both a noun and a verb. Even though these pairs have a low semantic correlation based on cosine similarity, observing them through human judgment might find a potential correlation. For example, 'website-regulation' can be a logical combination as many customers search for the company's regulation online. Another logical correlation is 'regulation-flight' as it can be the case that many passengers search for what regulations they need to follow when boarding a flight. Based on the following cloud, this topic can be generalised with the following label 'FAQ website'.

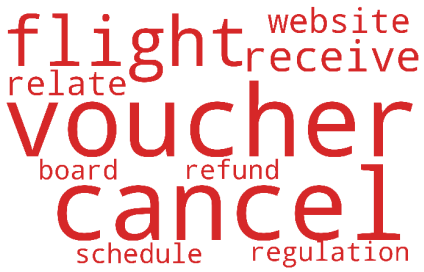

Figure 4.14: Top 10 words for **Topic 7**

Topic 2

This topic had an average cosine similarity score with an overall semantic correlation score of 0.321. The following Table presents five-word pairs that have the highest semantic correlation.

Pairs	CS
dog-pet	0.848
cabin-dog	0.778
pet-cabin	0.739
luggage-kg	0.658
hold-kg	0.634

Table 4.16: Highest Pairwise Cosine Similarity for **Topic 2**

Based on the Table, it can be seen that the terms 'dog-pet' is strongly related to the highest cosine similarity that has been reported so far for all three models. The semantic correlation between these terms is visible as the word 'dog' is a hyponym of the 'pet' hypernym group. For the last two-word pairs, it can be seen that the terms 'luggage', 'kg' and 'hold' are highly related as many passengers do contact the airline for matters concerning hold luggage or the regulations concerning the appropriate kilos for baggage. In this case, it can be assumed that the passengers were contacting the company for matters concerning baggage and pet regulations. In this cluster, some terms do have a great semantic correlation with the rest of the combined words. The following Table presents the lowest cosine similarity scoring terms.

Pairs	**CS**
cabin-flight	-0.033
pet-flight	-0.030
hold-flight	-0.019
add-flight	0.658
seat-flight	0.096

Table 4.17: Lowest Pairwise Cosine Similarity for **Topic 2**

This topic contains a significant amount of terms that have a negative cosine similarity score. These pairs might not have a high correlation, but the terms are correlated in terms of human judgment, especially in the airline domain. It is the case that many travellers contact an airline company to arrange different details for their flight. For example, reserving a seat for the flight or adding a piece of hold luggage or arranging to bring their pet on the flight. For this topic, the following word cloud is observed to generate a potential generalisation label.

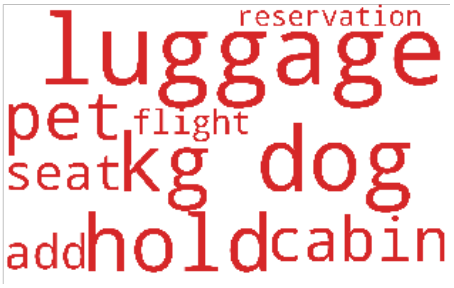

Figure 4.15: Top 10 words for **Topic 2**

Based on the combination of the words, a potential label can be generated; in this case, it can be 'Fare arrangements' or 'Business policy'. This label can be used to generalize the combination of terms from the model.

Topic 0

The last topic presented for the third approach is the cluster that scored the lowest cosine similarity out of all ten topics. The measured similarity for topic 0 was set to 0.084. The following five word-pair combinations scored the highest pair-wise cosine similarity.

Pairs	CS
voucher-cancel	0.522
business-policy	0.471
respond-open	0.426
business-open	0.389
notify-earlier	0.358

Table 4.18: Highest Pairwise Cosine Similarity for **Topic 0**

Based on Table 4.19, it can be seen that the terms that have the highest correlation in this combination of words are 'voucher' and 'cancel'.It can be the case that these terms were used frequently in the same conversation and therefore are correlated and close to the embedded space of the data. For example, in the airline domain, a significant concern is the cancellation of flights, and the reimbursement of the passengers is some way, such as a voucher. Moreover, it can be seen that terms such as 'business', 'policy' and 'open' are correlated. Based on a linguistic perspective, these words' semantic meaning and definition do not correlate clearly. However, from the standpoint of human judgment, a connection can be found especially for the domain investigated. For example, it can be assumed that these terms are based on frequent questions such as 'What is your business policy for pets?' or 'What are the opening hours of your business?'. This can also be the case for 'respond-open'; it can be assumed that the company's agent is informing about the specific hours they respond to customer questions. Finally, the combination of the last pair is not visible, but the correlation score might be due to the embeddings chosen for this analysis. The following Table presents five-word pairs that scored a negative cosine similarity.

Pairs	CS
voucher-earlier	-0.269
cancel-earlier	-0.161
webcare-policy	-0.122
webcare-earlier	-0.082
notify-message	-0.049

Table 4.19: Lowest Pairwise Cosine Similarity for **Topic 0**

Based on the lowest scoring pairs, it can be that the term 'earlier' does not have a great semantic correlation with a significant amount of terms within the topic. Even though the score of these terms is low based on observation-based human judgment, some correlation was found. In the case of 'webcare-policy', it can be assumed that the customer is contacting the webcare-customer service to obtain information about the company's policy. This can also be the case for 'notify-message' as it can be the case that the agent is contacting customers through a message or email to notify them of matters such as the cancellation of a flight. The following word cloud was used to generalize the topic and potentially provide a label.

Figure 4.16: Top 10 words for **Topic 0**

Based on the top 10 words, a label can be generated to some degree. It can be assumed that a great amount of the words refer to customer service and webcare contact. However, the generalisation process is not accessible due to outliers such as 'earlier' and 'notify'.

BERTopic Conclusion Remarks

Based on the topic generated from the deep learning approach, it can be seen that a great number of topics are coherent and can be potentially generalized with a label. In this case, topics with a cosine similarity lower than 0.185 are difficult to classify due to outliers. For this model, 70% of the topics can be generalized, topics 1,2,3,4,7,8 and 9. It might be assumed that high cosine similarity scores were generated due to the embedding usage during the training of the model. The highest scoring word-pair was reported for this case out of all three topic modeling approaches.

The following chapter presents the discussion section, suggestions for future research, and limitations that need to be considered. Finally, the conclusion section will briefly describe what has been reported in greater detail in the previous chapters.

Chapter 5

Conclusion & Discussion

5.1 Discussion

With the problems of extracting relevant information from unstructured texts in mind, this study aimed to compare the predictive performance and topic quality of three topic modeling methods. Two traditional approaches, LDA and NMF and a deep learning one, BERTopic. Considering the results and the error analysis, some observations were made. If we briefly compare the models, it can be seen that based on coherence, the topics generated by the LDA model are slightly better than the ones generated by NMF. However, the deep learning approach scored the highest for both types of topic coherence. In the case of cosine similarity, NMF slightly scored higher than LDA, with BERTopic reporting the best score.

The evaluation scores based on topic coherence and cosine similarity can provide insight into the degree of significance between the words and how interpretable they are. However, these methods measure something objective. Even if the score is high, it does not mean that the topic quality is accurate, and therefore an observation-based human judgment was combined. Based on human judgment, when comparing BERTopic to the other approaches, it can be seen that the traditional methods are unable to detect embedded meanings within a corpus. This lack of semantic meaning resulted in the evaluation scores of both traditional models being lower than the deep learning approach. It is worth mentioning that 50% of topics generated from LDA and NMF were similar in terms of word combinations. For example, it can be seen from the Figures 4.1 and 4.2 that LDA topics 2,3,5,8,9 are similar to the topics 1,0,7,8,4 from NMF, respectively. Even though, these topic models use different approaches to extract the topics, the overall results are similar. While in the case of BERTopic only 10% of similarity was found with the other models. **Topic 3** from the deep learning approach contains a similar combination of terms to **Topic 5** from LDA and **Topic 7** from NMF. With the deep learning approach, many new terms have been generated. For example, the terms 'board', 'dog', 'mother', etc., are words that were not present in the results of the traditional models. Finally, the highest correlated pairs in terms of semantic similarity were found in BERTopic. **Topic 2**, in Section 4.5.3, reported two-word pairs with the highest pairwise cosine similarity across all topics from the three topic modeling approaches. This can be due to the use of an embedding model during the model's training process.

This report made it clear from the start that the data used for this project originated from a specific domain. However, the topics' content is colloquial and does not necessarily have an airline-specific jargon. Furthermore, readers do not need prior knowledge as the topics generated contain universal terms. Finally, many aspects influenced the results reported. The following section presents the limitations and recommendations for future research.

5.1.1 Limitations and Future Research

There were several limitations that influenced the study's results. The main limitation of this report was the time frame given to execute the project. This study was conducted in three months, which was not enough to for a detailed analysis and for exploring more approaches. This limitation prevented an elaborative experimental phase as a great amount of the time was used to pre-process the data. Another limitation was in terms of resources; during the starting stage of this report, it was decided to implement a K-means clustering in combination with an embedding model, but due to the limited resources, that was not possible. Due to the data size, this topic model took a tremendous amount of training time. In the working environment used for this project, a Central Processing Unit (CPU) was responsible for training the models; however, based on the data size, this option can be time-consuming. Moreover, other resources were not allowed to be used due to the privacy rules concerning the data. For example, a Graphics processing unit (GPU) through a different environment such as Google colab could be a solution, but that was not possible due to privacy reasons. Finally, a challenging aspect of this project was the unsupervised nature of the task. Due to the lack of gold labels, the evaluation and error analysis of the project was difficult to approach. As the intrinsic evaluation methods could give an insight in terms of the coherence of the topics, they do not necessarily report the accuracy of the topic quality. Therefore observation-based human judgment was combined, but this raises more elements that need to be considered when reading the error analysis. The human judgment evaluation is based on the author's perspective and intuition, and therefore biases might arise. As it is a subjective interpretation of the topics.

With this limitation in mind, there are some suggestions for future research. One of the pre-processing steps implemented for this project was a language detection library in order to deal with code-switching. Therefore it would be interesting for future research to re-conduct this project with code-switching to see the combinations the models generate between the different languages. Moreover, the output of the models for this project is token level, and it might be interesting to implement this process in terms of sentence or conversation level. With this approach, the token's context will be visible for further analysis. Moreover, it can be the case that some conversations contain more than one topic. Therefore, a multi-topic approach will be interesting to investigate the possibility of more than one topic. An interesting aspect of the model approaches is implementing the traditional model combined with an embedding model. With this implementation, we can investigate and analyse if the performance is similar to a deep learning approach such as BERTopic. For the error analysis of this report, an embedding model based on the data was used. It will be interesting to implement other embedding models that are trained on general data or on a larger amount of data. To see if this scores would better represent the topics. Finally, as this project used specific data from the airline domain, it will be interesting to use these models on

different domain data to see if the performance will remain the same.

5.1.2 Conclusion

This research focused on the automatic retrieval of topics using topic modeling techniques from customer conversations in the airline domain. The main goal behind the in-depth insight was to retrieve the hidden topics and evaluate them in terms of predictive performance and topic quality. Furthermore, this automatic topic retrieval can provide the company with information about the most frequent matters that the customers talk about. The primary approach consisted of comparing three topic models, two traditional and one deep learning. For this project, the models chosen were: LDA, NMF and BERTopic. Henceforth, after implementing and comparing the performance of the models, the following answers to the proposed research questions were reported.

How does the performance of different topic modeling techniques differ in terms of predictive performance and topic quality in customer conversations in the airline domain?

It can be seen that based on all the evaluation methods implemented that the model with the best performance when it comes to predictive performance and topic quality is BERTopic. This deep learning approach scored the highest topic coherence and cosine similarity scores compared to the other topic models. Moreover, as mentioned in the error analysis, 70% of the topics generated from this approach can be generalised and used for the next step, topic classification. While in the case of the traditional methods, the coherence scores were not significantly high. Moreover, based on the pair-wise cosine similarity, it can be seen that the combinations used for the topics were sometimes abstract. This leads to a lower generalisation rate for these models, with 60% of the topics being suitable for classification. Finally, it is worth mentioning that it is surprising that the overall semantic similarity of NMF is slightly higher than LDA, as both of the models are a bag-of-words approach.

What additional pre-processing steps can be combined with traditional ones to improve the performance of topic modeling in the airline domain?

Three main pre-processing steps had to be implemented: language detection, removing highly frequent domain-specific words, removing emojis, and removing automatically generated responses. As the customer and agent conversations were executed on WhatsApp and can be considered an informal environment, code-switching was a phenomenon that needed to be taken into consideration. Moreover, the airline company that the data originated from is a Dutch company; therefore, it is expected that the Dutch language will appear in the data. In this case, a language detection library was used only to retrieve the English conversations. After further investigation of the data, it was seen that some of the most frequent terms used did not contribute to any potential topics. Therefore a frequent check was implemented to find these terms used in the conversations and exclude them. Words such as 'thanks' and 'welcome' were frequently used. Additionally, like any company that offers customer support, this company automatically generated responses to inform customers about matters such as business hours or asking for feedback. These responses did not influence the potential topics, and therefore it was decided to remove them. Finally, as the conversational data was

executed on an online platform, the use of emojis was frequent. As these pictograms can cause 'noise' in the data they were excluded. These additional pre-processing steps reduced the noise in the data and provided an output suitable for the topic modeling task.

Will existent evaluation methods reflect the topic quality in user-generated content in this domain?

The evaluation methods chosen for this project were mainly focused on the topic coherence and semantic similarity of the word combination within a topic. To some extent, these methods can reflect the topic quality. However, evaluation of an unsupervised environment is challenging, and many issues can occur. For example, the intrinsic evaluation methods can reflect the coherence but not necessarily the accuracy of the topic quality. Therefore, an additional technique of evaluation was implemented for this project for an in depth understanding of the output. Pairwise cosine similarity was able to analyze the terms of each topic in pairs and find the most semantic correlated and less correlated word combinations. Moreover, an observation-based human judgment was combined for a better intuitive comprehension of the topic quality and predictive performance. However, the human judgment component can result to a biased interpretation based on the subjective interpretation. Therefore, as mentioned in Section 2.3, the evaluation process of topic models is still a subject under development.

Appendix A

Evaluation Overview

Table 5 Types of comparisons for model evaluation

Types of comparisons	Sources
Human ratings/scores	E.g., Tirullinai and Tellis (2014, pp. 470);
Comparing results of an automated process to human ratings/scores and evaluations	
External reports and categories	E.g., Tirullinai and Tellis (2014, pp. 470);
Comparing the clustering output of topic models to external reports (e.g., consumer reports) or already present categories	
Traditional clustering techniques	E.g., Trusov et al. (2016, p. 417);
Comparing the output of topic models to traditional customer segmentation and clustering techniques	
Specific metrics, associated with a field	E.g., Weng et al. (2010, pp. 267);
Comparing a topic model to specific algorithms, associated with a research field (like page rank, in degree, etc.)	
Topic models[a]	E.g., Christidis and Mentzas (2013, p. 4377); Hruschka (2014, p. 270); Jacobs et al. (2016, pp. 397); Tirullinai and Tellis (2014, p. 471); Trusov et al. (2016, p. 417); Wang et al. (2012, pp. 127);
Comparing a topic model to other topic models	
Comparing (mathematical, componential, parameterwize (like the number of topics)) variations of the same topic model	

[a]Comparison on the same data, on in sample & predictive (hold-out data) and on different datasets

Figure A.1: Evaluation Methods on customer/marketing domain (Reisenbichler and Reutterer, 2019)

Bibliography

A. Abu-Jbara and D. Radev. Umichigan: A conditional random field model for resolving the scope of negation. In ** SEM 2012: The First Joint Conference on Lexical and Computational Semantics–Volume 1: Proceedings of the main conference and the shared task, and Volume 2: Proceedings of the Sixth International Workshop on Semantic Evaluation (SemEval 2012)*, pages 328–334, 2012.

N. M. Al Ghamdi and M. B. Khan. Assessment of performance of machine learning based similarities calculated for different english translations of holy quran. *International Journal of Computer Science & Network Security*, 22(4):111–118, 2022.

M. Alodadi and V. P. Janeja. Similarity in patient support forums using tf-idf and cosine similarity metrics. In *2015 International Conference on Healthcare Informatics*, pages 521–522. IEEE, 2015.

K. Bastani, H. Namavari, and J. Shaffer. Latent dirichlet allocation (lda) for topic modeling of the cfpb consumer complaints. *Expert Systems with Applications*, 127: 256–271, 2019.

M. Belford and D. Greene. Ensemble topic modeling using weighted term co-associations. *Expert Systems with Applications*, 161:113709, 2020.

D. M. Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

D. M. Blei, A. Y. Ng, and M. I. Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

J. Chang, S. Gerrish, C. Wang, J. Boyd-Graber, and D. Blei. Reading tea leaves: How humans interpret topic models. *Advances in neural information processing systems*, 22, 2009.

Churchill and Singh. The evaluation of topic modeling. *ACM Computing Surveys (CSUR)*, 2021.

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.

A. B. Dieng, F. J. Ruiz, and D. M. Blei. Topic modeling in embedding spaces. *Transactions of the Association for Computational Linguistics*, 8:439–453, 2020.

A. Dorantes, G. Sierra, T. Y. D. Pérez, G. Bel-Enguix, and M. J. Rosales. Sociolinguistic corpus of whatsapp chats in spanish among college students. In *Proceedings of the Sixth International Workshop on Natural Language Processing for Social Media*, pages 1–6, 2018.

M. Dutta. Topic modelling with lda -a hands-on introduction. 2021.

R. Egger and J. Yu. A topic modeling comparison between lda, nmf, top2vec, and bertopic to demystify twitter posts. *Frontiers in Sociology*, 7, 2022.

T. Ganegedara. Intuitive guide to latent dirichlet allocation. *Towards Data Science (blog). Medium. March*, 27, 2019.

C. Goyal. Part 15: Step by step guide to master nlp – topic modelling using nmf. 2021.

M. Grootendorst. Bertopic. 2020. URL `https://maartengr.github.io/BERTopic/`.

M. Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. 2022.

T. Hofmann. Probabilistic latent semantic indexing. In *Proceedings of the 22nd annual international ACM SIGIR conference on Research and development in information retrieval*, pages 50–57, 1999.

S. Hosseini and Z. A. Varzaneh. Deep text clustering using stacked autoencoder. *Multimedia Tools and Applications*, 81(8):10861–10881, 2022.

P. Huilgol. Quick introduction to bag-of-words (bow) and tf-idf for creating features from text. *Analytics Vidyha*, page 2020, 2020.

H. Jelodar, Y. Wang, C. Yuan, X. Feng, X. Jiang, Y. Li, and L. Zhao. Latent dirichlet allocation (lda) and topic modeling: models, applications, a survey. *Multimedia Tools and Applications*, 78(11):15169–15211, 2019.

S. Kapadia. Evaluate topic models: latent dirichlet allocation (lda). *Towards Data Science*, 2019.

N. Keane, C. Yee, and L. Zhou. Using topic modeling and similarity thresholds to detect events. In *Proceedings of the The 3rd Workshop on EVENTS: Definition, Detection, Coreference, and Representation*, pages 34–42, 2015.

N. Korfiatis, P. Stamolampros, P. Kourouthanassis, and V. Sagiadinos. Measuring service quality from unstructured data: A topic modeling application on airline passengers' online reviews. *Expert Systems with Applications*, 116:472–486, 2019.

D. Kuang, J. Choo, and H. Park. Nonnegative matrix factorization for interactive topic modeling and document clustering. In *Partitional Clustering Algorithms*, pages 215–243. Springer, 2015.

H.-J. Kwon, H.-J. Ban, J.-K. Jun, and H.-S. Kim. Topic modeling and sentiment analysis of online review for airlines. *Information*, 12(2):78, 2021.

B. Y. Liau and P. P. Tan. Gaining customer knowledge in low cost airlines through text mining. *Industrial management & data systems*, 2014.

Y. Luo, T. Wan, and Z. Qin. Topic modeling of political dynamics with shifted cosine similarity. In *International Symposium on Integrated Uncertainty in Knowledge Modelling and Decision Making*, pages 267–278. Springer, 2022.

K. MacMillan and J. D. Wilson. Topic supervised non-negative matrix factorization. *arXiv preprint arXiv:1706.05084*, 2017.

J. B. MacQueen. Some methods for classification and analysis of multivariate observations. 1967.

S. Mifrah and E. Benlahmar. Topic modeling coherence: A comparative study between lda and nmf models using covid'19 corpus. *International Journal of Advanced Trends in Computer Science and Engineering*, pages 5756–5761, 2020.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. Distributed representations of words and phrases and their compositionality. *Advances in neural information processing systems*, 26, 2013.

R. Navigli and F. Martelli. An overview of word and sense similarity. *Natural Language Engineering*, 25(6):693–714, 2019.

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*, pages 100–108, 2010.

D. O'callaghan, D. Greene, J. Carthy, and P. Cunningham. An analysis of the coherence of descriptors in topic modeling. *Expert Systems with Applications*, 42(13):5645–5657, 2015.

P. Paatero and U. Tapper. Positive matrix factorization: A non-negative factor model with optimal utilization of error estimates of data values. *Environmetrics*, 5(2): 111–126, 1994.

F. Pascual. Topic modeling: An introduction. volume 5, pages 1–4, 2019.

F. Rahutomo, T. Kitasuka, and M. Aritsugi. Semantic cosine similarity. In *The 7th International Student Conference on Advanced Science and Technology ICAST*, volume 4, page 1, 2012.

M. Reisenbichler and T. Reutterer. Topic modeling in marketing: recent advances and research opportunities. *Journal of Business Economics*, 89(3):327–356, 2019.

M. E. Roberts, B. M. Stewart, and E. M. Airoldi. A model of text for experimentation in the social sciences. *Journal of the American Statistical Association*, 111(515): 988–1003, 2016.

G. Salton and M. J. McGill. Introduction to modern information retrieval. 1986.

N. Seth. Part 2: Topic modeling and latent dirichlet allocation (lda) using gensim and sklearn. 2021.

Sinivasan. Topic modeling in python : Using latent dirichlet allocation (lda). In *2020 International Conference on Electrical Engineering and Computer Science (ICECOS)*, pages 386–390. IEEE, 2020.

N. Smith. Hdbscan clustering with neo4j. *Towards Data Science*, 31, 2021.

S. Srinivas and S. Ramachandiran. Discovering airline-specific business intelligence from online passenger reviews: an unsupervised text analytics approach. *arXiv preprint arXiv:2012.08000*, 2020.

I. Vayansky and S. A. Kumar. A review of topic modeling methods. *Information Systems*, 94:101582, 2020.

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*, pages 1105–1112, 2009.

S. Zhong. Efficient online spherical k-means clustering. In *Proceedings. 2005 IEEE International Joint Conference on Neural Networks, 2005.*, volume 5, pages 3180–3185. IEEE, 2005.