

Exploring Airline Tweets Topics:
A Comprehensive Analysis of Passenger
Experiences and Aviation Discourse on
Twitter

Table of Contents

Certificate of Approval	Error! Bookmark not defined.
Author's Declaration	Error! Bookmark not defined.
Plagiarism Undertaking	Error! Bookmark not defined.
Acknowledgements.....	Error! Bookmark not defined.
Table of Contents	2
List of Figures	4
List of Tables.....	5
Chapter 1: Introduction	6
1.1. Problem Statement	7
1.2. Research Questions.....	8
1.3. Research Objectives	8
1.4. Scope of the Study	8
1.5. Significance of the Study.....	9
1.6. Definition of Terms	10
1.7. Summary	11
Chapter 2: Literature Review	12
2.1 Introduction	12
2.2 Customer Satisfaction	13
2.3 Sentiment Analysis for Customer Feedback	15
2.4 Customer Satisfaction in the Airline Industry	18
2.5 Sentiment Analysis in the Airline Industry	20
2.6 Theoretical Framework	21
2.7 Summary	23
Chapter 3: Methodology	24
3.1 Introduction	24
3.2 Research Design	24
3.3 Population and Sampling	25
3.4 Validity & Reliability	27
3.5 Summary	28
Chapter 4: Data Collection	29
4.1 Introduction	29

4.2	Data Collection	29
4.3	Description of Data	32
4.4	Summary	33
Chapter 5:	Data Analysis	34
5.1	Introduction	34
5.2	Data Preprocessing	35
5.3	Word Recurrence	38
5.4	Text to Vector Representation	39
5.5	Clustering	40
5.6	Summary	45
Chapter 6:	Results & Conclusions	46
6.1	Introduction	46
6.2	Word Frequency.....	47
6.3	K-Means Clusters.....	54
6.4	Conclusion	67
6.5	Summary	68
Chapter 7:	References.....	69

List of Figures

Figure 1: Airline Tweet Count	32
Figure 2: Null and Non-Null Values Comparison	36
Figure 3: Pre-processing Word Frequency.....	39
Figure 4: Euclidian Distance Formula	41
Figure 5: Elbow curve.....	43
Figure 6: Silhouette Score Graph.....	44
Figure 7: Most Common Words	47
Figure 8: Scatterplot of clusters	55
Figure 9: Word Cloud 1	56
Figure 10: Word Cloud 2	57
Figure 11: Word Cloud 3.....	58
Figure 12: Word Cloud 4	59
Figure 13: Word Cloud 5	60
Figure 14: Word Cloud 6	61
Figure 15: Word Cloud 7	62
Figure 16: Word Cloud 8	63
Figure 17: Word Cloud 9	64
Figure 18: Word Cloud 10	65
Figure 19: Word Cloud 10	65
Figure 20: Word Cloud 11.....	66

List of Tables

Table 1: List of Airlines in the Sample	25
Table 2: Airline Tweet Count	32

Chapter 1:

Introduction

One of the services that offer memorable experiences, emotions, and feelings to customers is tourism. It involves traveling for distinct reasons, such as leisure, social, or professional, within one's country or abroad. The tourism sector depends on the happiness and satisfaction of tourists, who spend more than half of their income on services in the modern era. Therefore, customer satisfaction is vital for the success of this sector (Law et al., 2022).

The main mode of transportation for international tourists is airlines, which significantly impacts the tourism industry and the global economy. Airlines enable people to visit various places and cultures, create business opportunities, and generate resources. They also facilitate global mobility by connecting different regions and continents (Vieira et al., 2019).

Customer satisfaction is a key factor for any business, including airlines. To achieve long-term success, airlines need to understand their customers' needs and expectations and meet them effectively. They also need to establish good relationships with their customers, as they are the ones who determine the business's performance and profitability. Customer satisfaction leads to customer retention, which is essential for a competitive market (Denning, 2021).

The airline industry has undergone significant changes in the last twenty years, especially in the way it collects customer feedback. Traditional methods of feedback, such as surveys and questionnaires, are no longer sufficient or efficient. They are simple and slow to process. Instead, social media comments have emerged as a valuable source of feedback, as they reveal the opinions and perspectives of passengers (Yakut et al., 2015). Online content, such as reviews and feedback, influences customers' purchasing decisions, as they can express their preferences and expectations freely in the digital era (Jain et al., 2019).

Business Intelligence (BI) is a field that provides conceptual frameworks for analyzing and evaluating business processes. One of the techniques used in BI is Topic Modeling, which aims to understand the underlying topics in a corpus of text.

The global air transport business is a prominent service sector that involves many passengers and financial resources. To maintain their market position and improve their services, airlines need to adopt modern methods for collecting and analyzing customer experiences and interactions. These methods include various forms of feedback, reviews, posts, and tweets that customers share on different platforms.

The internet has changed the way customers give feedback to businesses. Customers can now access various online platforms easily and share their opinions anytime and anywhere. A single review or complaint from a customer can reach millions of people, giving customers unprecedented power (Jansen et al., 2009). Airlines must manage and meet multiple expectations to provide the best customer experience.

In the past, human emotions were understandable only by humans. However, recent technological advancements have made it possible to analyze text data on the internet based on emotional values. This data includes user posts on social media sites, product reviews, and inquiry histories. This analysis helps airlines ensure customer satisfaction and meet expectations in today's social media age, where customer feedback plays a crucial role in repurchase decisions.

1.1. Problem Statement

The airline industry is a tough and demanding sector that needs large investments and has a small margin of error. Many airlines have gone out of business or have had to change their business model to survive in the global market, where economic borders have almost vanished. To succeed in this sector, airlines must focus on competitiveness.

One of the key factors that influence the competitiveness of airlines is customer experience. It refers to the perception and feeling that passengers have about the airline's service, from booking to landing. Customer experience affects customer satisfaction, loyalty, and word-of-mouth, which are crucial for the airline's reputation and profitability. Despite the difficulties caused by the coronavirus outbreak, airlines continue to invest in enhancing the customer experience. The fierce competition and elevated expectations require a strong commitment to delivering high-quality experiences.

To measure and improve customer experience, airlines must know how passengers feel about their services. One way to do this is to analyze the emotional value of text data from passenger feedback. By analyzing copious amounts of data from passenger feedback, such as tweets and other social media posts, airlines can get valuable insights into customer sentiment. However, analyzing tweets can be hard because of the informal language used, such as slang, abbreviations, emojis, and sarcasm. These factors can make it hard to accurately determine sentiment.

Tweets about airlines are a valuable source of passenger feedback for analyzing sentiment. They offer real-time insights into customer satisfaction, opinions, and expectations. However, it is important to consider that several types of airlines operate under different business models, such as legacy airlines, low-cost carriers, or hybrid. Each business model has distinct characteristics, such as service quality, pricing, network coverage, and target customer segments. Therefore, analyzing tweets specifically related to airlines operating under a certain business model can help managers identify the dimensions of customer satisfaction that are most relevant.

Analyzing sentiment and topics can also provide valuable insights for other aspects of airline operations, such as sales, marketing, and customer support. For example, sales and marketing departments can use sentiment analysis to measure customer reactions to their offers and adjust their strategies accordingly. In customer support, recognizing the emotional state of users can help prioritize and manage critical support requests more effectively and efficiently.

1.2. Research Questions

- What are the main topics discussed in tweets related to airlines?
- Are the tweets helpful in making binary decisions such as yes/no, liking/disliking, or positive/negative?
- Can the tweets depict passengers' emotions towards the airlines?
- How helpful is topic modeling in finding airline brand advocates?

1.3. Research Objectives

- Find out and classify the main themes discussed in tweets related to airlines by performing a thorough analysis of a large dataset.
- Measure the sentiment expressed in airline-related tweets and determine how well they can provide binary decision insights, such as positive/negative sentiment or liking/disliking.
- Evaluate the ability of tweets to show passengers' emotions towards airlines and explore the variety and intensity of emotions expressed.
- Examine the effectiveness of topic modeling in finding airline brand advocates within the pool of tweets and study the characteristics and behaviors of these advocates.

To achieve these research objectives, the study will involve gathering and analyzing many tweets related to airlines, using natural language processing techniques, sentiment analysis, and topic modeling. The findings will contribute to a deeper understanding of the themes discussed, sentiment expressed, passenger emotions, and the identification of brand advocates within the context of airline-related tweets.

1.4. Scope of the Study

The rapid growth and diffusion of added information technologies, especially in e-commerce, has sparked increasing interest in consumer behavior in online purchases. Similarly, today's society demands more convenience and less time when buying products or services online, for which they know that it is the quickest way to search for information and shows the importance of companies having increased skills to maintain an effective online presence.

Since the deregulation of the aviation industry, airlines have found ways of reducing costs, which has resulted in a decline in the quality of services onboard. However, it has also made flying affordable to a large population. As "Data is the new black gold"! Many companies have a huge amount of data available and want to use it more to make decisions.

Introduced in 2006, Twitter is a microblogging platform that allows users to post short messages, known as tweets, which are visible to other users. The company has over 450 million active users, a number which is expected to reach 652 million in the next 5 years. Twitter's tweets are also used as a source of sentiment analysis, by identifying and extracting the emotional tone and attitude of customers toward a service or product. Sentiment analysis can help airlines improve their services, marketing strategies, and customer satisfaction.

The use of Latent Semantic Analysis in existing literature is a drawback as it cannot oversee synonymy and polysemy, which are common in natural language (Yang et al., 2012). Synonymy is the existence of multiple words with the same meaning, while polysemy is the existence of multiple meanings for a single word (Gabrilovich & Markovitch, 2005). This limitation is overcome by using more advanced techniques such as Latent Dirichlet Allocation (LDA) (Song et al., 2011) and Neural Network-based models like BERT, which can better capture the context and meaning of words, allowing them to manage synonymy and polysemy more effectively.

The aim of this study is to identify the topics of discussion in the tweets mentioning some of the airlines on twitter. To the best of the authors' knowledge, no previous research has been done using these techniques in the aviation industry, which focuses only on legacy airlines, and uses BERT modeling.

The results of the study will have practical applications and implications for real-world problems of business analytics as it would deal with the analysis of different text data to discover previously unknown relationships or trends from existing data. Airlines will be able to use the information they need to improve processes and make positive changes. In future, carriers that do not use business analytics risk becoming inefficient and outdated, and falling behind competitors who use business analytics to their advantage.

1.5. Significance of the Study

The rapid growth of unstructured customer feedback data and sophisticated analysis tools has opened new avenues for discovering customer preferences using topic modeling methods. However, there is still room for improvement, especially in the fast-changing field of Artificial Intelligence. The main goal of this research is to fill these gaps by focusing on the automatic detection and classification of human emotions from various sources, such as physiological signals and social media texts.

We narrow our focus to the aviation industry, where the main objectives are profitability and revenue optimization. In this context, the study aims to provide a comprehensive analysis of customer feedback, with a special focus on airlines in the post-pandemic era. The use of state-of-the-art deep learning techniques and sentiment analysis helps to identify customer issues and effectively improve the overall customer experience.

Sentiment analysis, a key part of this research, is a powerful tool that allows airlines to extract useful information from the plethora of positive and negative comments on online platforms. This valuable knowledge enables them to make better decisions, improve their services, and increase customer satisfaction. A crucial role is played by machine learning techniques, such as BERT and DistilBERT, which achieve high accuracy in topic modeling and sentiment classification of tweets related to the aviation domain.

The expected contribution of this research goes beyond topic modeling and sentiment analysis. It also involves language model pretraining and natural language processing, advancing the field of interpreting human language with high speed and reliability. As the research progresses, it seeks to show a way towards processing and analyzing human language, transforming communication with unprecedented efficiency.

The results of this study will have important implications for policymakers and airline managers. By identifying the key topics that emerge from passenger concerns and expectations, these insights will guide them in making informed decisions in the dynamic environment of the airline industry. With the help of topic modeling and sentiment analysis, airlines can strive to improve their services and customer experiences, leading the way in customer-oriented excellence.

The combination of advanced topic modeling and sentiment analysis techniques demonstrates the hidden potential within the substantial amounts of customer feedback data. This research project is an exciting exploration of the possibilities of Artificial Intelligence technology, offering to reveal valuable insights into the changing world of customer behavior and satisfaction. As the aviation industry adopts these innovative methods, it embarks on a thrilling journey of continuous improvement, always aiming to exceed customer expectations and ensure excellence in service delivery.

1.6. Definition of Terms

Airline Industry: The airline industry provides transportation services via flights for both people and cargo. To provide these services, airlines may rent or own aircraft, and occasionally they may combine their services with those of another airline for mutual advantage (Belobaba et al., 2015).

Customer: The person or entity who makes the decision to purchase a good or service (Simons, 2014), on an occasional or regular basis, from a supplier. It is also a person who receives from a company, against payment, commercial supplies, or services.

Satisfaction: A psychological state resulting from an experience of purchase and/or consumption and relative” (Vanhamme, 2016). The relative nature of satisfaction refers to the fact that the judgment of satisfaction is the result of the comparison between the subjective experience lived by the consumer and a standard of comparison.

Customer Service: On a general level, customer service consists of all the human and organizational resources mobilized to meet the information and service needs of customers before, during and after the purchase. In the phases preceding a purchase, customer service can therefore also address prospects (Rust & Zahorik, 1993).

Machine Learning: Machine learning is a subgroup of artificial intelligence. In machine learning, the machine improves its performance by learning independently from data without separate operating instructions. Most applications utilizing artificial intelligence are based on machine learning. In commercial use, with the help of machine learning, it is possible to create complex algorithms and models, which can be used, for example, to support decision-making or to make production processes more efficient.

Topic Modeling: Topic modeling is a latent variable statistical modeling technique that discovers the abstract topics that underlie a corpus of documents by clustering words and

phrases that are semantically related into topics. It is a powerful tool for natural language processing, and it has a wide range of applications, such as document clustering, sentiment analysis, and recommender systems (Abdelrazek et al., 2023).

In the context of social media, topic modeling can be used to identify topic clusters in tweets and posts. This can be helpful for understanding the content of social media conversations, identifying trends, and categorizing users. The amount of text required for topic modeling depends on the size and complexity of the corpus. In general, a small corpus of text may be sufficient to show what insights can be gained by applying generative topic models. However, for more complex corpora, a larger corpus of text may be required.

1.7. Summary

The research aims to identify and classify the main themes discussed in tweets about airlines, measure sentiment, evaluate emotions expressed, and find airline brand advocates. This involves analyzing a large dataset of tweets using natural language processing, sentiment analysis, and topic modeling techniques.

The significance of the study lies in its contribution to customer preference discovery using topic modeling, the potential to improve business analytics through sentiment analysis, and the advancement of language model pretraining and natural language processing. The findings have practical applications for airlines, enabling them to make positive changes and remain competitive in the evolving market.

The tourism sector, a service that offers memorable experiences and emotions to customers, relies heavily on customer satisfaction for its success. Airlines, as the main mode of transportation for international tourists, play a significant role in the tourism industry and the global economy. To thrive in the competitive market, airlines must focus on understanding and meeting customer needs and expectations. Traditional feedback methods have given way to social media comments as a valuable source of customer opinions.

Online platforms have revolutionized customer feedback, enabling real-time insights and empowering customers with unparalleled influence. Analyzing sentiment and topics in tweets related to airlines can provide valuable insights for sales, marketing, and customer support departments. To succeed in this sector, airlines need to prioritize customer experience and understand how passengers feel about their services. However, analyzing tweets can be challenging due to the informal language and multiple business models.

Business Intelligence (BI) techniques, including Topic Modeling, are used to analyze and evaluate business processes, making them essential for airlines to collect and analyze customer feedback effectively.

The study also delves into the definition of terms, including the airline industry, customer satisfaction, machine learning, and topic modeling. Understanding customer service, satisfaction, and the role of machine learning helps airlines better cater to customer needs and expectations.

Chapter 2:

Literature Review

2.1 Introduction

This chapter provides a comprehensive review of the relevant literature on the main topics of this research, namely customer service, sentiment analysis, topic modeling, and the airline industry. The purpose of this review is to establish the theoretical background and the current state of knowledge on these topics, as well as to identify the research gaps and opportunities for further exploration.

The chapter is organized as follows. The first section introduces the concepts of customer service and how they have evolved over time. It examines various definitions and perspectives on customer service from different disciplines and contexts. It also summarizes the significant work done on customer service, especially in relation to customer satisfaction and business profitability. Moreover, it explores the concept of co-creation, which refers to the active involvement of customers in creating value with service providers.

The second section discusses the concept of sentiment analysis, opinion mining and topic modeling. It explains what these terms mean, how they are related, and why they are important for analyzing customer feedback. It also describes the increasing use and relevance of sentiment analysis in various domains and applications, such as marketing, social media, e-commerce, and politics. Furthermore, it presents the latest methods and techniques in sentiment analysis and topic modeling, with a focus on natural language processing and machine learning approaches.

The third section reviews the literature on the airline industry. It highlights how the sector is connected to global market trends and how it has been affected by a range of factors, such as deregulation, competition, innovation, and the COVID-19 pandemic. It also discusses how airlines try to maintain profitability and competitiveness in a dynamic and challenging environment. Additionally, it examines how airlines use customer service and topic modeling to understand and improve customer experiences and satisfaction.

The chapter ends with a brief recapitulation of the major insights and outcomes of the literature review, as well as an articulation of the research questions and objectives that steer this study. The literature review has provided a comprehensive and critical overview of the relevant topics and concepts that form the theoretical foundation and the current state of knowledge for this research. It has also identified the existing gaps and opportunities in the literature that warrant further investigation and exploration. The research questions and objectives have emerged from the literature review and reflect the main purpose and scope of this study.

2.2 Customer Satisfaction

The birth of research on consumer satisfaction dates to the 1970s (Hartley, 1976). Since then, the number of published articles has continued to increase as the attainment of quality in products and services has become a pivotal concern for the people. It is no longer an epiphenomenon but a trend that characterizes research in the field of marketing.

While the concept of customer service and its quality was not precisely defined by the researchers, a review of the existing literature showed that the research approached the concept either from a production-based or customer-oriented viewpoint. The production-based or manufacturing-based approach focuses on the objective or technical quality of standardized product measurement, the customer-oriented approach focuses on customer perceptions or Quality in the eye of the customer. Due to the complex nature of quality, the customer perspective has become the dominant approach in defining service quality in research (Akter et al., 2019).

The concept of customer satisfaction is defined in diverse ways in the marketing and finance sector. Considering the scarcity of reliable literature, the results of the study, (Oliver & Bearden, 1983) stated that satisfaction is a function of the performance of the product or service relative to the customer's expectations, this performance being considered as the degree of presence of the attributes sought, wanted or expected in a product or service by the customer. The authors admit that customer satisfaction is a function of the evaluation of service quality and price.

(Parasuraman et al., 1985) showed that consumers used similar criteria in evaluating the quality of service that fall within ten main categories that are determinants of service quality, namely: reliability, responsiveness, efficiency, competence, Access, Courtesy, Communication, Credibility, Security, Understanding/knowing the customer, Tangibles.

Then the same researchers returned in 1988 (Parasuraman et al., 1988) to develop a multi-element scale to measure the quality of service (SERVQUAL). Tangible: Physical facilities, equipment, and an individual's appearance. Reliability: The ability to perform the promised service reliably and accurately. Responsiveness: Willingness to assist customers and prompt delivery of services. Affirmation: the knowledge and courtesy of employees. Empathy: The care and individual attention that the company provides to customers. This study defined the quality of service as seen by the customer, as it stems from comparing what the customer feels that the service companies should provide (i.e., their expectations) with his perceptions of the company's performance in providing services. The perceived service quality is seen as the degree and direction of discrepancy between customer perceptions and expectations.

As customer satisfaction is attracting increased interest from companies that have understood that the key to success lies in a customer-centered philosophy of action, the interest shown by researchers in the study satisfaction is also significant. (Vanhamme, 2016) defines customer satisfaction as “a psychological state resulting from a purchase and/or consumption experience and a relative”. (Rahmoune et al., 2022) defined customer satisfaction as the customer’s comparison between his expectations about the product and its perceived performance, with the actual product or service.

It is sometimes perceived as an emotion, sometimes as a rational appreciation. Three situations can appear in a customer’s experience; either the performances are below perceived expectations (dissatisfaction), or at their level (satisfaction) or beyond expectations (enthusiasm).

In relational marketing, where providing value to the customer is a particularly important criterion, the aim is to engage the customer in the long term. In a market with long-term mutual relations with the customer, businesses will be successful, and the customer will be satisfied with the goods or services they receive (Hennig-Thurau et al., 2002). The theory of the positive correlation between relational marketing and customer loyalty, customer satisfaction, and customer value was further proven in several other cases, such as (Şendur, 2009), and (Sheth & Parvatiyar, 1995).

Customer satisfaction and brand loyalty can be measured with customer feedback. In the light of this information obtained, companies need to make more customer-oriented decisions on products, services, pricing, and delivery, or they are in a risk of losing customers due to poor customer service experiences (Jamal & Bucklin, 2006).

Businesses strive to earn lifelong consumers by delivering personalized experiences and excellent service across all channels. Customer service activities help to build loyalty by providing reliable and relevant solutions to their problems. Getting feedback from customers can be done in many ways. Businesses can both solve and prevent customer loss by being aware of the negative experiences they face.

Given the complexity surrounding the holistic nature of customer experience, measuring the has made measuring customer perceptions of interactive service experiences challenging (Ordenes et al., 2014). Nowadays, many organizations are taking advantage of innovative technologies that allow them to collect feedback from their customers and stakeholders through various channels such as e-mail, phone calls, surveys, reviews, and questionnaires. This method delivers insights from the vast amount of data available on the internet and predicts customer sentiments and preferences for online purchases with a high degree of accuracy. This could be extremely useful for various fields such as economics and marketing, as it can help them tailor their products and services to meet the needs and expectations of their target audience.

In today’s digital world, the power of social media platforms should not be underestimated. Many entrepreneurs are using their brand’s social media accounts as a way of engaging with their existing and potential customers and inviting them to share their opinions and stories about their products or services. As we discussed earlier, positive feedback and testimonials on brand’s social media accounts can have a significant impact on the purchasing decisions experience of prospective customers who are looking for reliable and trustworthy information about the brand. (Lăzăroiu et al., 2020).

2.3 Sentiment Analysis for Customer Feedback

In today's world where digitalization has increased, the importance of customer satisfaction is increasing day by day. Many companies care about customer feedback to avoid losing customers with a negative buying experience. In the meantime, the feedbacks received are evaluated correctly and made meaningful with the help of sentiment analysis (Pang & Lee, 2008). The enormous amount of textual data available every day on the internet has encouraged research in several areas that automatically process and analyze these texts. (Keramatfar & Amirkhani, 2019). Understanding user sentiment towards a brand is an essential component in the development of successful marketing strategies. This is where Sentiment Analysis comes into play, and why it is such a crucial tool for businesses.

By regularly analyzing the moods and opinions of customers, businesses can gain valuable insights into their needs and desires. This information can then be used to adapt the range of goods and services offered to better meet the needs of the target audience. As a result, businesses could continuously restructure their processes to maximize sales and improve customer satisfaction. Sentiment Analysis provides businesses with a powerful tool for understanding their customers and making informed decisions that drive success. (Lau et al., 2018).

Sentiment detection has manifold applications in information systems, such as categorizing reviews, synthesizing feedback, and other real time applications. It is observed that sentiment classifiers are profoundly contingent on domains or topics. From the preceding work, it is apparent that no classification model invariably surpasses the other, and several types of features have disparate distributions. It is also observed that distinctive features and classification algorithms are amalgamated in an efficacious way to surmount their individual shortcomings and benefit from each other's virtues and augment the sentiment classification performance (Vinodhini et al., 2012).

Topic modeling is a statistical technique that can be used to identify the underlying topics in a corpus of text. In recent years, topic modeling has been increasingly used for sentiment analysis, as it can be used to extract aspects (i.e., specific topics) from customer reviews and categorize these aspects into similar classes. This has opened new possibilities for sentiment analysis, by allowing for a more fine-grained analysis of user sentiment (Rana et al., 2016).

(Vinodhini et al., 2012) also found that topic modeling can be used to improve the performance of sentiment classifiers in a variety of domains. However, the effectiveness of topic modeling depends on the quality of the topic model, the type of sentiment classifier used, and the domain of the text. This means that the focus should be shifted towards developing better topic models for sentiment detection.

Research by (Mandal et al., 2021) examines the performance of fifty-six different methodologies for computing textual similarity across Indian court case statements when applied on a dataset of Indian Supreme Court Cases. Among the fifty-six different methods, thirty are adaptations of extant methods and twenty-six are novel methods. The methods scrutinized include models such as BERT and Law2Vec. The results of this study suggested that the more traditional methods for computing textual similarity are still the most effective for the legal domain.

This is likely because these methods can capture the semantic similarity between documents, even when the documents are not explicitly linked. They concluded that the conventional methods (such as the TF-IDF and LDA) that depend on a bag-of-words representation outperform the more sophisticated context-aware methods (such as BERT and Law2Vec) for computing document-level similarity.

Based on approximately three hundred research articles, (Kherwa & Bansal, 2020) paper provides a comprehensive review of topic modeling, a statistical technique for uncovering the latent thematic structure of a corpus of text. The authors discuss the various methods of topic modeling such as Latent Dirichlet Allocation (LDA), Latent Semantic Analysis (LSA), and Non-Negative Matrix Factorization (NMF) and their applications in different fields such as social media analysis, text classification, and information retrieval.

The paper then discusses the different evaluation metrics for topic models, such as perplexity, coherence, and interpretability. Their research reviews several empirical studies that have evaluated the performance of topic models on different corpora. The authors also discuss the challenges associated with topic modeling such as the selection of the number of topics, the quality of the topics generated, and the interpretation of the topics. The results of these studies suggest that topic models can be effective for a variety of tasks, including text classification, clustering, and summarization.

Similar work by (George & Birla, 2019) provides an in-depth survey of different topic modeling methods, focusing on three prominent and influential models: Latent Semantic Analysis (LSA), Probabilistic Latent Semantic Analysis (PLSA), and Latent Dirichlet Allocation (LDA). It also discusses some of the extensions and variations of LDA that have been proposed to address specific challenges or applications. The paper compares the major features, advantages, and limitations of these models, and provides examples of how they can be applied to diverse types of textual data.

The paper by (Egger & Yu, 2022) compared four topic modeling techniques: latent Dirichlet allocation (LDA), non-negative matrix factorization (NMF), Top2Vec, and BERTopic, for the purpose of analyzing Twitter data. The paper evaluated the advantages and disadvantages of each algorithm in terms of their accuracy and interpretability in a social science context. The paper found that BERTopic and NMF were the most suitable algorithms for analyzing Twitter data. BERTopic was able to detect topics with high precision, meaning that it assigned words and documents to the correct topics. NMF was able to detect topics with high clarity, meaning that it produced topics that were easy to understand and label. LDA and Top2Vec were less suitable than BERTopic and NMF, but they still had some potential for analyzing Twitter data.

LSA is a method of Natural Language Processing (NLP) that analyzes the association between documents and terms by generating a set of concepts that relate them. It uses a matrix decomposition technique called singular value decomposition (SVD) to reduce the dimensionality of the document-term matrix and capture the semantic similarity between words and documents. LSA can identify the main topics or themes that are present in a corpus of documents and the degree to which each term belongs to each topic.

LSA is a simple and efficient method that can handle large and sparse matrices, but it also has some drawbacks, such as the lack of probabilistic interpretation, the arbitrariness of choosing the number of topics, and the inability to handle polysemy (words with multiple meanings) and synonymy (words with similar meanings) (Deerwester et al., 1990).

PLSA is a generative model that addresses some of the limitations of LSA by introducing a probabilistic framework for topic modeling. It assumes that each document is a mixture of topics, and each topic is a probability distribution over words. This method can infer the latent topics from the observed words and documents using an expectation-maximization (EM) algorithm, which maximizes the likelihood of the data given the model parameters. It can also disambiguate words with multiple meanings and group words with similar contexts by assigning them to different topics based on their co-occurrence patterns (Hofmann, 2001). The algorithm has several advantages over LSA, such as providing a principled way of choosing the number of topics, allowing for model selection and evaluation using likelihood-based criteria, and being able to oversee polysemy and synonymy.

LDA is a widely used generative probabilistic model that extends PLSA by introducing a Dirichlet prior over the topic distributions. This prior encodes the assumption that documents tend to have a sparse set of topics, and topics tend to have a sparse set of words. It can also model the exchangeability of words and documents, which means that their order does not affect their probability. Each document is represented as a bag of words, where each word is drawn from one of the topics assigned to the document according to a multinomial distribution.

LDA can learn the topics and their word distributions from a corpus of documents using various inference algorithms, such as Gibbs sampling or variational inference and can also estimate the topic proportions for each document using posterior inference. Latent Dirichlet Allocation can flexibly manage several types of textual data, such as news articles, web pages, blogs, social media posts, etc.

LDA has several benefits over PLSA, such as avoiding overfitting by regularizing the model parameters, providing a generative process for new documents, and incorporating prior knowledge or assumptions about the topic distributions (Porteous et al., 2008).

BERT (Bidirectional Encoder Representations from Transformers) is a model that represents an encoder block of a transformer and is pre-trained with a huge corpus of untagged text. The power of BERT is that the same pre-trained model can be used, simply by adding an extra layer of output, in a variety of tasks with very good results for all of them (Devlin et al., 2019).

Released by Google in 2018, BERT is an open-source, network-based technique for natural language processing (Natural Language Processing – NLP). It employs a deep neural network with multiple layers of transformers to encode the semantics and pragmatics of words and sentences from both left and right directions. This allows BERT to capture the semantic and syntactic features of a text, which is essential for sentiment analysis.

BERT can conduct sentiment analysis by utilizing its potent representation learning capability to capture the semantic and syntactic attributes of a text. It can also oversee intricate and lengthy sentences by using its attention mechanism to concentrate on the pertinent parts of the text. This model can categorize the sentiment of a text by passing its encoded representation to a softmax layer that outputs a probability distribution over the potential classes, such as positive, negative, or neutral (*BERT vs Other Deep Learning Models for Sentiment Analysis*, n.d.).

2.4 Customer Satisfaction in the Airline Industry

The tourism industries are dynamic, uncertain, and time-sensitive, requiring real-time decision making and updated market insights to deliver value to customers and gain competitive advantage. This is because these industries are constantly evolving, with new destinations, transportation options, and travel experiences emerging all the time. To stay ahead of the curve, tourism and travel businesses need to be able to make quick decisions based on the latest market data. One way to do this is to use artificial intelligence (AI) to analyze enormous amounts of data in real time.

AI has been used to track trends, identify emerging opportunities, and predict customer behavior. This information can then be used to make informed decisions about pricing, marketing, and product development. (Stylos et al., 2021).

Air transport is a time-sensitive industry that encompasses the transportation of goods and passengers, as well as the management and development of air transport infrastructure (Bratović, 2020). In this industry, airline management and researchers have extensively studied passenger satisfaction and loyalty levels, examining the relationships between several factors. For instance, (Gourdin, 1988) identified three key factors of quality in commercial air travel: ticket pricing, airline safety, and on-time performance. Similarly, (Elliott & Roach, 1993) investigated the factors influencing passenger evaluation of airlines and found that perceptions about service quality attributes are influenced by internal factors within individuals. They identified six key elements for assessing airline service quality: food and beverage quality, timeliness, baggage handling, seat comfort, airline check-in, and in-flight service.

(Ali et al., 2015) focused on the relationship between service quality dimensions and customer satisfaction, specifically evaluating the service quality of Pakistan International Airlines (PIA) among Pakistani and foreign nationals. However, the study's limitation is that the results cannot be generalized to other airlines since it only examined perceptions of PIA's service quality. Nonetheless, their work was expanded upon by (Saleem et al., 2017) who found a correlation between trust, service quality, and airline repurchase decisions among Pakistani frequent flyers, with customer satisfaction mediating this relationship.

(Sandada & Matibiri, 2016) conducted research on the South African airline industry and concluded that loyalty programs, special treatment, and frequent flyer programs play a crucial role in retaining loyal customers. (Zhang et al., 2020) reaffirmed the positive association between superior service quality and repurchase intention, mediated by the passenger's perceived value of the airline brand. (Sezgen et al., 2019) observed that passenger satisfaction differs and is dependent on the class of ticket.

Several models have been developed to understand the factors affecting customer satisfaction in the airline industry, including customer value, pricing, and coordination. (Park et al., 2020) used Herzberg's two-factor theory and analyzed many online reviews to conclude that customer value positively correlates with customer satisfaction. They found that in-flight attributes such as cabin neatness, meals, and entertainment systems are crucial for customer value. They also classified pre-flight factors, such as check-in and boarding, as dissatisfiers.

Since the implementation of the airline deregulation act in 1978, the industry has witnessed increased competition. (Spicer, 2018). notes that the competitive environment benefits customers by providing a wide variety of options to compare and choose from. Modern airlines are increasingly investing in customer-oriented business activities to enhance the customer experience at reasonable prices amidst tariff wars (Bratović, 2020) to the degree of satisfaction that is felt is needed (Budi Setiawan et al., 2021).

Additionally, (Boubker & Naoui, 2022) highlight the significant impact of factors such as airline service quality, tangible aspects of the airline and airport, and crew behaviors on the brand image perceived by passengers.

The airline market has set new standards in the last decade, due to the use of digitization and efficiency in their processes. This digital growth is expected to continue its growth as digital devices integrate into our lives and change purchasing behavior. The access to real-time ticket price comparisons, alongside customer reviews means that competition among carriers will become increasingly fierce (Ashour et al., 2023).

The industry is recognizing investment in people, innovation, technology, and infrastructure to transform travel by planning personalized strategies for marketing. They ensure to consider all the specificities that characterize services in relation to products, which include the seating comfort, schedule flexibility, in-flight amenities, safety, and on-time performances (An & Noh, 2009). Consumers can compare the offers of airline companies and choose the company that offers the highest quality at an acceptable price.

With the spread of Coronavirus, the entire world plunged into a crisis of health, economy, and social wellbeing. The global aviation industry was hit hard by the restriction of airline transportation, which was one of the first bans brought with the declaration of the pandemic. The COVID-19 pandemic is unique in terms of its impact on the aviation industry, even compared to earlier crises such as oil embargos (Doganis, 2013), the 9/11 terrorist attacks (Bouwer et al., 2021), and the 2008 global monetary crisis (Morrell, 2021). The authors note that earlier crises had long-term effects on the aviation industry, such as increased fuel efficiency, changes to safety and security measures, and the introduction of optional fees and upgrades.

The extensive travel restrictions caused by the pandemic created a challenging environment for the airline industry and its partners, further complicating the current financial tightening conditions in the insurance market (Nižetić, 2020). Several research were conducted to measure and understand the effect of the restrictions placed on the movement after the coronavirus.

(Monmousseau et al., 2020)'s research on the passenger perspective of travel restrictions highlights the impact on the relationship between passengers and airlines. They argue that these restrictions have created a sense of uncertainty and anxiety among passengers, which has made it more difficult for airlines to build trust and loyalty.

At present, there are approximately five thousand airlines operating across the globe. The airline industry itself is undergoing a recovery after facing tough economic conditions during the Coronavirus pandemic. Growing at an annual rate of 12.7%, the global market for airlines is projected to reach \$767 Billion by the year 2027. Some of the countries that continue to be among the fastest growing global markets for airlines are China, Japan, Canada, Germany, and the United States. (StrategyR - Global Industry Analysts, 2022).

2.5 Sentiment Analysis in the Airline Industry

Sentiment analysis is a technique that aims to identify and extract the opinions, emotions, and attitudes of customers or users from textual data, such as reviews, comments, tweets, etc. Sentiment analysis can help businesses and organizations to understand their customers' needs, preferences, satisfaction, and loyalty, as well as to monitor their reputation, brand image, and market trends.

The airline industry is one of the domains that can benefit from sentiment analysis and topic modeling (Hasib et al., 2021), as it is a highly competitive and service-oriented sector that relies on customer feedback and word-of-mouth.

Topic modeling is a technique that discovers the latent themes or topics that underlie a corpus of documents. The airline industry has employed topic modeling for various purposes, such as: comprehending customer sentiment, by detecting the topics that are most salient to customers and the sentiment pertaining to these topics; discerning trends, by detecting the topics that are influencing the airline industry; and segmenting customers, by detecting the topics that reflect their interests and preferences. These purposes can help airlines to enhance their service, gain a competitive edge, and provide a customized service (Kwon et al., 2021a).

While statistics and numbers can provide information such as the raw performance of an airline profitability and growth trends, however, they do not explain how passengers react to a certain event. The use of Natural Language Processing techniques brings the prospect of providing the deep knowledge as to why and how passengers perceive and things as (Rustam et al., 2019) mentioned, machine learning techniques perform better for classification than traditional approaches because they can automatically extract relevant features from the data, reducing the need for manual feature engineering.

The research study by (Sezgen et al., 2019) analyzed reviews from Tripadvisor.com. This research had a sample of opinions from more than five thousand passengers of fifty different airlines. Their work revealed a deeper understanding of major factors influencing customers' satisfaction and dissatisfaction with full-service and budget airlines, as well as with economy and premium cabins. They concluded that the determinants of customer satisfaction and dissatisfaction can vary slightly based on the airline's business model and service class.

(Kwon et al., 2021b)'s research used online customer reviews to explore the factors that influence a customer's decision to purchase a ticket from an Asian airline. Their approach identified patterns and trends in the reviews, rather than evaluating a specific hypothesis. Similarly, airport management practices such as opening a new passenger terminal can potentially lead to passenger satisfaction at the terminal, which contribute to a significant positive impact on non-aviation revenues (do Nascimento et al., 2022).

The closest research to our subject was done by (Park et al., 2022) which explored multiple relationships between these factors and highlighted the remarkable predictive capabilities of deep learning models. With impressive accuracy rates of 94% for predicting customer churn risk and 90% for determining customer satisfaction, the findings surpassed previous research and significantly advanced the theoretical knowledge in this domain.

The robustness of the deep learning models demonstrated their potential as powerful tools for businesses to optimize their operations and customer-oriented decision-making processes. These groundbreaking results not only demonstrate the efficacy of deep learning in customer analytics but also encourage further exploration of its applications in various other areas to unlock even greater potentials for understanding complex relationships in large datasets.

2.6 Theoretical Framework

The preprocessing phase of the text data involves a series of essential steps. These steps include the removal of punctuation, numbers, repeated characters, special characters, and the conversion of emojis to words. Moreover, contractions are expanded to facilitate further analysis.

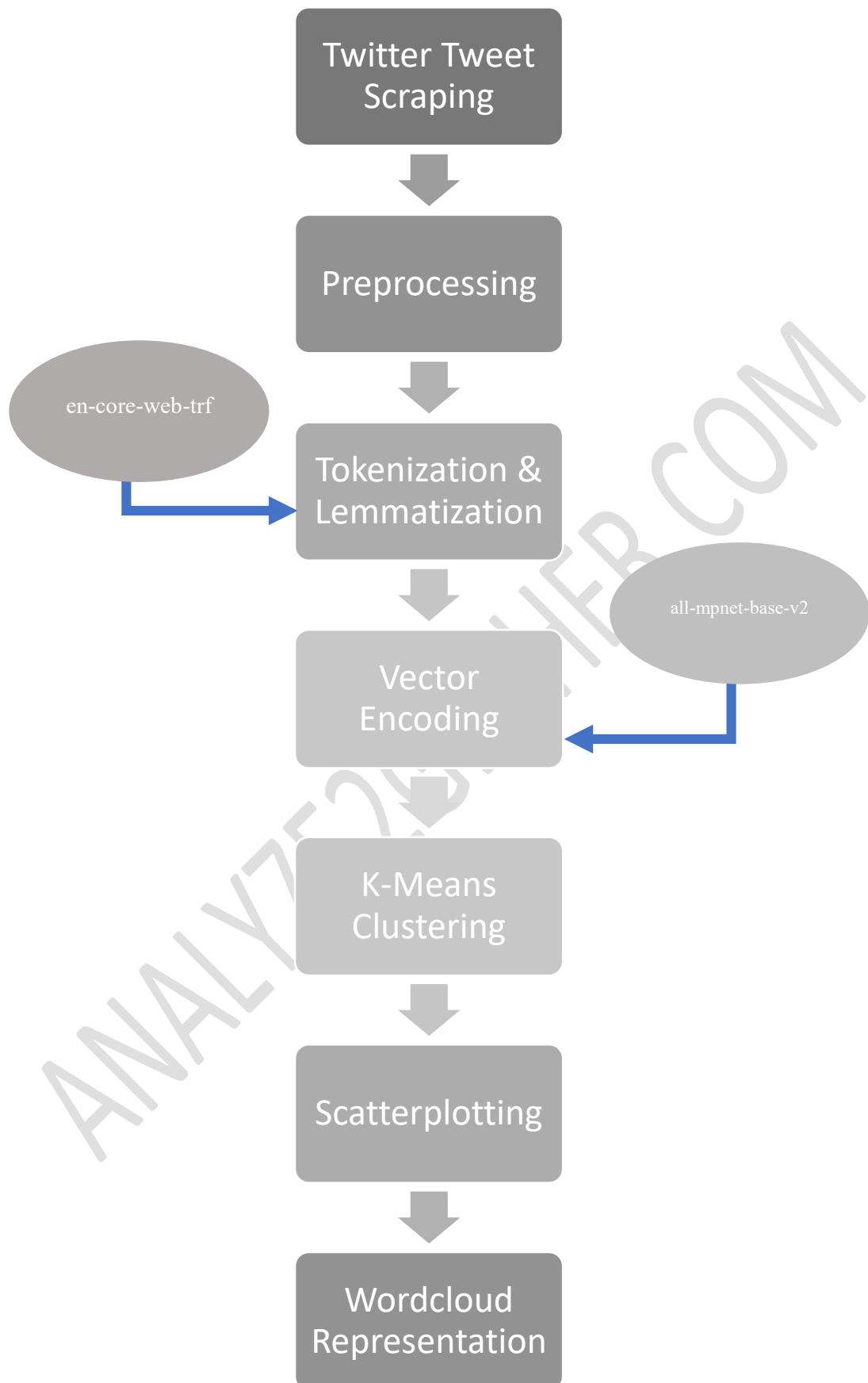
Tokenization and stop word removal are essential tasks undertaken to enhance the quality of the processed text data. The spacy NLP library is utilized to tokenize the cleaned text, and stop words are removed from each tokenized text. The results are stored in a designated column named 'tokens' within the DataFrame. To facilitate the extraction of meaningful features from the tweet data, the code utilizes the Sentence Transformers model to encode the cleaned text into vector representations.

K-means clustering, a widely used unsupervised learning algorithm is applied to group the tweet data into distinct clusters. A silhouette score is utilized to determine the optimal number of clusters by calculating the within-cluster sum of squares (inertia) for various cluster numbers.

The final K-means clustering is executed using the KMeans class from the scikit-learn library. Each tweet sample is assigned to a specific cluster, and the cluster labels are recorded in the 'clusters' column of the DataFrame.

To enhance the visual representation of the clustering results, the chapter employs Uniform Manifold Approximation and Projection (UMAP) for dimensionality reduction of the vector representations. UMAP embeddings are calculated to visualize the clusters in a 2D scatter plot. Each data point represents a tweet sample, and the color of the points indicates their respective assigned clusters.

To gain further insights into the main topics of each cluster, the code generates word clouds. By grouping text samples based on their assigned clusters and employing the Word Cloud library, the chapter presents the most frequent words in each cluster. These word clouds offer an overview of the primary themes or topics within each cluster, aiding in the interpretation and analysis of the clustering results.



2.7 Summary

The literature review chapter delves into the concept of customer satisfaction, tracing its evolution over the years. It underscores the significance of customer service and satisfaction as fundamental aspects of successful businesses and explores how companies are embracing novel techniques to capture and enhance the customer experience. Building upon this foundation, the focus narrows down to the realm of airline customer perception and satisfaction.

Within the context of airline customer satisfaction, the literature review highlights key determinants that contribute to customer satisfaction, including aspects such as seating, crew behavior, and on-time performance. These factors have been extensively studied and recognized as influential components in shaping the satisfaction levels of airline passengers.

Furthermore, the literature review ventures into the domain of airline sentiment analysis and topic modeling techniques, examining the outcomes and insights garnered from these approaches. By exploring these literatures, a comprehensive understanding of how sentiment analysis and topic modeling can inform the assessment and understanding of customer satisfaction within the airline industry is developed.

Combining the insights derived from these diverse sources of literature, a theoretical framework is formulated. This framework aims to elucidate how airlines can effectively align their service quality with the expectations of customers. By synthesizing the literature on customer satisfaction, airline customer perception, sentiment analysis, and topic modeling, the theoretical framework provides a structured approach for airlines to optimize their services and meet the evolving demands and preferences of customers.

Chapter 3:

Methodology

3.1 Introduction

The information age is already here. The rate of data production today is higher than ever, and it is constantly increasing. Most of the time, people who deal with data daily work mostly with unstructured text data. Some of this data includes elements like images, videos, sounds, etc. Some of the sources of this data are websites, daily blogs, news sites and more. Analyzing all this information faster is necessary and often crucial.

This chapter outlines the key aspects of the tweets' topic modeling study, which includes various elements. The research design, population and sampling procedures, validity and reliability considerations, research approaches, and clustering and dimensionality reduction techniques are some of these elements. They ensure the robustness and accuracy of the study's findings by providing the overall framework, selecting the data sources, ensuring the credibility and consistency, guiding the data analysis, and uncovering patterns and reducing the complexity of the dataset.

3.2 Research Design

Nature

This study employs a topic modeling approach, which falls under the category of quantitative methodology. Topic modeling is a computational technique that has the objective of discovering hidden thematic structures that exist within an ample collection of textual data. This technique involves analyzing the frequency and co-occurrence patterns of words and documents, as well as the relationships among them, to identify latent topics or clusters that represent the main themes or concepts in the data. To perform this technique, the data is converted into numerical representations, such as vector embeddings, which are mathematical expressions that capture the semantic and contextual information of words and documents. These numerical representations enable quantitative analysis and computational processing, such as clustering and dimension reduction, which are used to identify and visualize the topics.

Type

This study adopts a case study approach, which involves an in-depth examination of a specific phenomenon or case, often within its real-life context. This research approach is suitable for investigating and analyzing specific instances or cases within the airline industry, such as passenger satisfaction. The aim is to explore and analyze a particular situation or event to gain a comprehensive understanding of the subject matter.

Case studies allow researchers to gather rich and detailed data, often using multiple sources of evidence, such as interviews, observations, and documentation. By conducting detailed examinations of real-life cases, researchers can delve into the complexities and nuances of passenger satisfaction, identifying key determinants and exploring the relationships between several factors. This can lead to valuable insights and in-depth knowledge about the subject of interest.

3.3 Population and Sampling

The population of this study consists of all airline tweets that were posted between 22nd January 2022 and 31st March 2023. The sample of airlines was selected based on their Skytrax ratings, with seven airlines being selected from the top 10 and 3 airlines being selected from the top twenty. This sample is representative of the population of airline tweets, as it includes a mix of both high-rated and lower-rated airlines. The use of a sample allows for the study to be conducted more efficiently, while still providing reliable results. The list of the airlines is as follows:

Airline	Top	Handle	Followers	Account creation
Qatar Airways	10	@qatarairways	41.3M	March 10, 2009
Emirates	10	@emirates	33.7M	March 12, 2009
Singapore Airlines	10	@singaporeair	22.6M	July 26, 2009
Cathay Pacific Airlines	10	@cathaypacific	17.3M	February 22, 2009
All Nippon Airways	10	@flyANA_official	14.7M	May 11, 2009
Etihad	10	@etihad	10.2M	March 11, 2009
Lufthansa	10	@lufthansa	10.1M	February 25, 2009
Japan Airlines	20	@JAL_Official_jp	13.5M	February 10, 2009
Garuda Indonesia	20	@IndonesiaGaruda	10.9M	March 9, 2009
Swiss Air	20	@FlySWISS	9.6M	March 11, 2009

Table 1: List of Airlines in the Sample

The data, which was collected on 2023-07-06 08:55:35 Pakistan Standard Time shows that the airline with the most active followers is Qatar Airways, with 41.3 million followers. Emirates follows closely behind with 33.7 million followers. Singapore Airlines is in third place with 22.6 million followers, followed by Cathay Pacific Airlines with 17.3 million followers. All Nippon Airways, Japan Airlines, Garuda Indonesia, Etihad, Lufthansa, and Swiss Air round out the list, with follower counts ranging from 14.7 million to 9.6 million.

The account creation dates for the airlines' Twitter accounts also vary. Qatar Airways was the first airline to create a Twitter account, on March 10, 2009. Emirates followed suit on March 12, 2009, and Singapore Airlines created its account on July 26, 2009. The remaining airlines created their accounts in February or March of 2009. These are the insights into the popularity of these airlines on Twitter. Qatar Airways is the most popular airline on Twitter, followed by Emirates and Singapore Airlines.

The data presented in this table can be used to inform marketing and social media strategies for airlines. Airlines with many active followers on Twitter can use their platform to reach a large audience and promote their products and services. Airlines with newer Twitter handles can use the data to see how other airlines have grown their followings and develop strategies to achieve comparable results.

Qatar Airways is one of the most powerful airlines in the world, with a fleet of latest-generation aircraft. Its hub at Hamad International Airport in Doha is one of the world's best hubs. As the airline proudly celebrated its 25th anniversary in 2022, Qatar Airways has been named Airline of the Year by international airline rating organization Skytrax for an unprecedented seventh time.

Emirates is a major global airline of the United Arab Emirates. The company is based in Dubai. The airline is part of the Arab Air Carriers Organization and belongs to the conglomerate The Emirates Group, which is owned by the Emirate of Dubai. The company started its operations in 1985 with two chartered planes. Since then until today, the management has not hesitated to strive for continuous development, and today it owns more than 265 aircraft flying to and from approximately 155 destinations within eighty countries. The story of the launch of Emirates Airlines and its continuous successes for 30 years has been linked to the story of Dubai's renaissance and the vision of its leaders.

Singapore Airlines is widely regarded as an example of excellence in the airline sector. Based in the tiny Singapore island, the airline considered to be one of the best and largest in the world in recent years. The company was voted 2nd Company of the Year at the Skytrax World Airline Awards behind Asiana in 2010. According to Skytrax again, Singapore Airlines is second in the world for the "Best Business Class Award" after Qatar Airways, second in the world for the Award for world's Best First Class behind Etihad again for the year 2010.

Cathay Pacific was founded in Hong Kong in 1946 by Australian Sydney H. de Katznov and American Roy C. Today, Cathay Pacific is one of the leading airlines in the world. It remains intricately linked to the Hong Kong location, which flies to 110 destinations in thirty-five countries. There are also destinations on codeshare flights with other One-world airlines alliances.

All Nippon Airways was established in 1952 and has become the largest airline in Japan and the best 05 airline for the year according to the "SKYTRAX" website, with its headquarters and operations center at Narita Airport in Tokyo. It operates eighty-two international routes, and the company also owns more than two hundred aircraft. Its services are distinguished by diversity and competition with the preservation of cultural heritage For Japan. The company organizes trips between various cities in North America, Asia, and China.

Etihad Airways was founded by Royal Decree (Amiri) of the United Arab Emirates in July 2003. It began operations in November 2003 and has grown to become one of the fastest growing airlines in the world. It is based in Abu Dhabi, from where it offers non-stop flights to around 40 destinations in the Middle East, Asia, Oceania, Europe, Africa, North America, and South America.

Lufthansa is one of the oldest major airlines. It was founded on January 6, 1926, in Berlin - for comparison, the implementation of Air France only dates from 1933 - following the merger of two German companies: Deutsche Aero Lloyd and Junkers Luftverker.

The airline is a founding member of the Star Alliance, the largest airline alliance in the world, the company was founded in 1926, and is the fifth largest airline in the world in terms of passenger volume; It flies more than 70 million passengers annually from the company's hubs at two main airports: Frankfurt International Airport and Munich Airport. Lufthansa flights fly to 209 destinations in eighty-one countries and serve with its partners about 410 destinations around the world. The company has the third largest fleet of passenger flights around the world in cooperation with its subdivisions located all over the world.

Japan Airlines is a member of the JAL Group, which is engaged in domestic and international air transportation business (passenger and cargo), with seven consolidated subsidiaries and one affiliated company. The airline has over 224 aircraft in its fleet which fly to 199 domestic and international routes.

Garuda Indonesia is Indonesia's flag carrier airline, serving over ninety destinations worldwide. It has a fleet of 202 aircraft and is known for its "Indonesian Hospitality" service concept. Garuda Indonesia has won numerous awards, including "The World's Best Economy Class" from TripAdvisor and "Five Star Airline" from Skytrax.

Swiss International Air Lines (SWISS) is the flag carrier airline of Switzerland, connecting the country to Europe and the world with direct flights from Zurich and Geneva. The premium airline operates one of the most modern fleets in Europe and offers a comprehensive range of airport-to-airport services for high-value, time-sensitive, and care-intensive freight. The airline is committed to the highest standards of product and service quality, as well as to environmental sustainability. The airline has set ambitious CO₂ reduction targets, aiming to reduce its net emissions by 50% by 2030 and achieve carbon neutrality by 2050. One of SWISS's key focus areas is the promotion of sustainable fuels.

3.4 Validity & Reliability

In any model development, it is important to assess reliability and validity. We enact several strategies and considerations to ascertain these for the topic modeling approach delineated in the methodology. These measures endeavor to augment the precision, coherence, and credibility of the derived topics from airline tweets. The following are the salient aspects to sustaining validity and reliability:

Validity

Since the data is being scraped from the web, and in the cases where the amount of data grows in real-time, there is less time to review and verify the information. This can lead to inaccuracies in the data, such as incorrect ratings or irrelevant information. These inaccuracies can affect the accuracy of the model, especially if they make up a huge portion of the data. We meticulously select a representative dataset to encompass diverse airlines, regions, and time periods. We administer preprocessing techniques, such as noise removal, lowercasing, stemming, and stop-word removal, to curtail irrelevant information and augment the quality of the topics.

We observed that a substantial number of tweets within our dataset emanated from the airlines themselves, which posed a potential source of bias. To ensure unbiased data analysis, we decided to exclude these tweets from our dataset, as they could have skewed the analytical outcomes by introducing a positive bias. Airlines tend to post positive content about their services, which could have compromised the objectivity and integrity of our data analysis process. We ameliorated the impartiality and veracity of our data analysis process by skipping the tweets from the airline handles.

Since the research is based on the reviews and tweets on airline passengers This results in a coherent model that can work well or be mediated on several areas. To include other business models could on the other hand, lead to a varied result. The amount of data is at some point also important for the validity of the result.

Reliability

Reliability is essential for establishing trust in the topic modeling results. By applying K-Means clustering on the vectors, which is a common clustering algorithm that groups data points based on their similarity, we aimed at detecting distinct clusters or topics within the data. The reliability is preserved by setting the random state parameter to ensure consistent results, which enables reproducibility in the clustering process.

Similarly, the use of UMAP (Uniform Manifold Approximation and Projection), a dimension reduction technique, we preserved the global structure and relationships of the data points in a lower-dimensional space. It eases the visualization of the clusters and topics in a two-dimensional space. Reliability is maintained by specifying the parameters (e.g., `n_neighbors`, `n_components`, `metric`) to ensure meaningful representations.

An interactive scatter plot is created using Plotly, a powerful visualization library. The plot shows the clusters/topics in distinct colors, enabling a visual understanding of the data distribution. The `hover_name` parameter is set to display the corresponding tweet text and appropriate labeling of the clusters is ensured to maintain the reliability.

3.5 Summary

This chapter outlines the methodology for topic modeling of airline tweets, addressing the challenges posed by the abundant unstructured text data in the information age. Key aspects covered include research design, population and sampling, validity and reliability, research approaches, and clustering and dimensionality reduction techniques.

The research design employs quantitative methodology, utilizing topic modeling as a computational technique to uncover hidden themes within the tweet data. By analyzing word and document frequency patterns, topic modeling identifies latent topics or clusters. A case study approach is adopted to examine specific airline-related phenomena, drawing on data from Twitter, interviews, and observations for comprehensive analysis.

The population comprises all airline tweets posted within specific periods, while a sample of ten airlines, selected based on Skytrax ratings, represents this population, ensuring efficiency and reliability. To maintain validity, preprocessing techniques are employed to reduce noise and bias, with tweets from airline handles excluded to ensure impartiality. Additionally, reliability is upheld by using consistent parameters for K-Means clustering and UMAP dimensionality reduction techniques.

Chapter 4:

Data Collection

4.1 Introduction

This chapter delineates the comprehensive data collection process. It commences with the rationale of utilizing Twitter for the sentiment analysis, highlighting the advantages and challenges of this platform. This is succeeded by the explication of preference in terms of programming language and the Integrated Development Environment that is being employed for this thesis, elucidating the features and benefits of these tools. Subsequently, the thesis expounds the process of tweet extraction from Twitter, describing the steps and methods involved in this task. Once the data is extracted, it must undergo an extensive pre-processing phase where the noise is eradicated from the data, to ensure that the subsequent analysis and modeling procedures proceed smoothly.

4.2 Data Collection

Twitter keeps real-time data through which companies can find out users' thoughts on any topic. They can apply filters such as user, location, time, keyword, hashtag and customize the data they require. Twitter is one of those platforms where firms can understand their customers' needs and honest opinions. Organizations can instantly monitor what users think about their product or and those of their competitors. Pulling data from Twitter and bringing together industry's influencers and tweets helps companies in understanding and analyzing the most shared content. It gives the companies more awareness of tweet activity and reaches their customers with confidently by knowing their interests and practices. After having the tweet data scraped, companies perform sentiment analysis to classify them as positive, neutral, or negative.

Programming Language

Tweet scraping is the process of extracting data from Twitter, such as tweets, user profiles, and topics. It can be done for various purposes, such as sentiment analysis, market research, or trend detection. There are separate ways to scrape Twitter data using Python. The use of Python is justified due to several reasons.

Firstly, Python has an ease of use as it saves time by not requiring extra steps such as semicolons at each line break or braces at each start of condition. The notion of indentation also makes it possible to have a clean and readable code.

Secondly, Python offers a rapidly growing ecosystem, especially in terms of machine learning and natural language processing (NLP). It has a vast collection of libraries like NumPy, Matplotlib, Pandas etc. Something that helps to provide the user with several methods and services for varied reasons. In addition to that, it will be especially useful in the processing of data extracted from the site. Another quality of Python is that it has predefined types of variables, which do not need to be defined every time.

Thirdly, the syntax of Python is easily understandable as it is remarkably like the English language. The notion of indentation also facilitates the task since it makes it possible to separate the blocks of the code. Web scrapping aims to save time. In Python, small programs perform huge tasks.

Moreover, in case of an error, there is a large community of helpers available at the disposal in various platforms such as GitHub, Stack Overflow and Kaggle. These websites are excellent platforms for machine learning and data analysis projects, and their communities provide guidance on best practices.

Scraping Tool

Twitter data is a rich source of information that can be used for various data modeling applications, such as natural language processing, machine learning, or social network analysis. However, accessing and collecting Twitter data is not a trivial task, as Twitter imposes certain limitations and challenges on its data availability and accessibility. Therefore, it was important to understand the features and trade-offs of different tools for scraping Twitter data. To collect the data, we had to explore and evaluate different Python packages for scraping Twitter data.

One way is to use the Twitter API, which provides access to various endpoints and functionalities, but has some limitations on the number of requests and the time range of tweets. Another way is to use other libraries, such as Tweepy, Snscape, or Twint, which do not rely on the API and can bypass some of its restrictions. However, they may not offer the same features or stability as the API.

Twitter had developed an Application Programming Interface (API) for. This tool allows the collection of data from the social network, and through this API, one can access information such as username, real name, location, posted text, publication address, number of retweets, likes, replies, followers, account creation date, tweet creation date, verification status, and total number of tweets.

The Twitter API is divided into different versions and products, each with its own specifications and use cases. For example, the standard v1.1 API provides basic access to recent tweets and users; the premium v1.1 API provides enhanced access to historical tweets and users; the enterprise v1.1 API provides full access to historical tweets and users; and the new v2 API provides more detailed and flexible access to tweets and users2. The Twitter API requires users to apply for a developer account and obtain an API key and secret, which are used for authentication and authorization.

The Twitter API also imposes certain rate limits and quotas on the number of requests that can be made per time window or per day. For example, the standard v1.1 API has a rate limit of 15 requests per 15 minutes for most endpoints; the premium v1.1 API has a rate limit of 30 requests per minute for most endpoints; the enterprise v1.1 API has a rate limit of 50 requests per second for most endpoints; and the new v2 API has a rate limit of 300 requests per 15 minutes for most endpoints.

Moreover, the Twitter API has some limitations on the amount and type of data that can be accessed or retrieved. For example, the standard v1.1 API only returns tweets from the last seven days; the premium v1.1 API only returns tweets from the last 30 days; the enterprise v1.1 API only returns tweets from March 2006 onwards; and the new v2 API only returns tweets from July 2018 onwards.

Furthermore, the Twitter API does not provide access to some data fields or features that may be of interest to researchers or analysts, such as deleted tweets, protected tweets, direct messages older than 30 days, or tweet engagement metrics.

Considering the limitations, our choice for twitter data scraping fell on Snscape, which is another unofficial Python scraper for Twitter. Snscape was created in 2019 by Just Another Archivist as an open-source project to scrape data from various social networking sites, such as Twitter, Facebook, Instagram, Reddit, Telegram, YouTube, and more. Snscape supports some of the features and endpoints of the Twitter API, such as reading tweets, searching for tweets and users, and accessing historical data.

Snscape also provides some additional functionalities that are not available in the native Twitter API, Tweepy, or Twint, such as scraping without an API key or authentication, scraping without any rate limit or quota, scraping any tweet from any time period, scraping tweet replies and threads, scraping tweet quotes and mentions, scraping tweet URLs and links, scraping tweet hashtags and keywords, scraping tweet languages and locations, scraping tweet sources and devices, scraping tweet coordinates and places, and scraping tweet retweeters and likers.

Scraping Procedure

Two essential Python libraries, Snscape and Pandas, are utilized by the implemented code, which provides powerful tools for data manipulation and analysis. The code plays a crucial role in scraping tweets specifically related to selected airlines, thereby gathering a comprehensive dataset for sentiment analysis and topic modeling. Dynamic queries are created through a systematic approach to collect tweets that mention the chosen airlines within a predefined time. A structured data frame is then used to organize and store these tweets, offering a detailed view of various aspects such as the date of posting, user information, tweet content, geographical location (if available), and the airline mentioned. With its multiple columns, this structured data frame serves as a valuable resource for conducting in-depth analysis and extracting meaningful insights regarding customer perspectives, sentiments, and experiences within the airline industry. The acquired dataset becomes readily accessible for subsequent topic modeling and sentiment analysis by saving the data frame as a CSV file, facilitating a deeper understanding of customer satisfaction, and contributing to the identification of key themes and trends.

4.3 Description of Data

Using the step above, 62478 tweets were scraped. However, after removing the airline tweets, the number significantly dropped to 59,514 tweets. The breakdown of the tweets in the descending order was as follows:

Rank	Airline	Tweet Count
1	Emirates	20385
2	Qatar Airways	9852
3	Etihad Airways	7788
4	Singapore Airlines	7306
5	Lufthansa	6689
6	Cathay Pacific	2686
7	All Nippon Airways	1802
8	Swiss International Air Lines	1744
9	Japan Airlines	823
10	Garuda Indonesia	439

Table 2: Airline Tweet Count

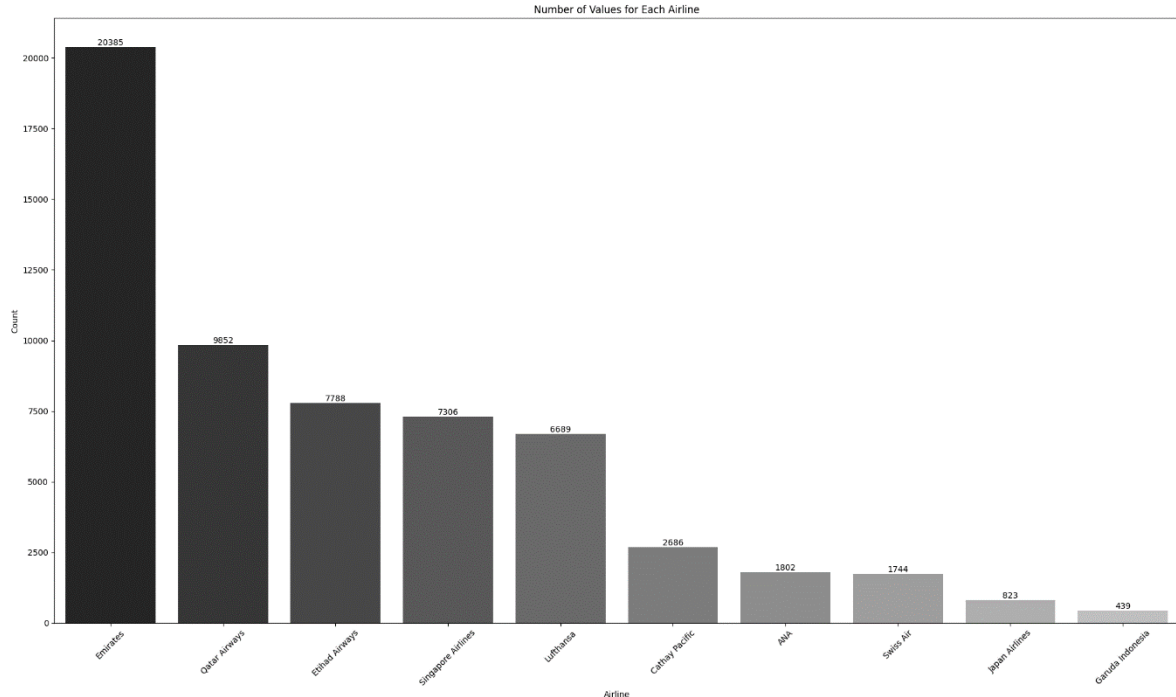


Figure 1: Airline Tweet Count

The table above and the visualization graph furnished below exhibit the rankings and corresponding tweet counts for a selection of airlines. The data accentuates the popularity and engagement levels of these airlines based on the number of tweets they elicited.

Emirates clinched the top position with an immense tweet count of 20,385, signifying a remarkable amount of online activity and public interaction pertaining to the airline. Qatar Airways trailed in the second position with 9,852 tweets, manifesting an elevated level of online presence and engagement as well.

Etihad Airways occupied the third position with 7,788 tweets, evincing a considerable amount of social media activity. Singapore Airlines followed closely with 7,306 tweets, indicating a notable level of online deliberations and interactions regarding their services.

Lufthansa secured the fifth position with 6,689 tweets, implying a substantial presence and engagement on social media platforms. Cathay Pacific ranked sixth with 2,686 tweets, manifesting a lower but still significant level of online activity and discussions.

All Nippon Airways clinched the seventh position with 1,802 tweets, indicating a moderate level of engagement and online presence. Swiss International Air Lines trailed in the eighth position with 1,744 tweets, manifesting a similar level of online activity.

Japan Airlines occupied the ninth position with 823 tweets, indicating a comparatively lower level of engagement and online discussions. Lastly, Garuda Indonesia ranked tenth with 439 tweets, demonstrating a lower online presence and engagement compared to the other airlines in the list.

These statistics furnish valuable insights into the popularity and public perception of these airlines based on the number of tweets they elicited. However, it is imperative to note that tweet count alone does not furnish a comprehensive evaluation of an airline's overall performance, as factors such as sentiment, context, and the nature of the tweets themselves are not considered in this analysis.

4.4 Summary

We focus on the crucial importance of social media in this chapter, especially Twitter, which stands out among various social media platforms. Twitter allows users to voice and share their views, which makes it essential to use tweets for analyzing different topics of discussion. We also explain how tweets are extracted, highlighting the key role of the Twitter API and relevant libraries that enable this extraction process, and allow researchers to customize their tweet extraction queries as per their needs. Finally, we present a collection of statistics related to the collected data, which provide useful information and empirical evidence for the further analyses in this research area.

Chapter 5:

Data Analysis

5.1 Introduction

The main objective of this section is to employ and discuss various techniques to preprocess the text data, which is a rich source of information, but also poses several challenges for extracting useful insights from it. Tweets is a form of Text data, which are usually derived from various sources, such as social media posts, customer reviews, or scientific articles, often lacks a clear and consistent structure, with opinions and sentiments being scattered and hindered using emojis, punctuation marks, and spelling errors. Consequently, cleaning the text data before performing any analysis on it becomes a formidable and essential task that requires careful attention and consideration.

Text cleaning assumes a paramount role within the data pre-processing pipeline, as it involves applying various techniques to clean, organize, and standardize the text data to enable its effective utilization with data science and machine learning methods. These methods are optimized for working with numerical and vectorized data, which means that the text data must be converted into a suitable encoded format that can be understood by machines.

The quality of the text data profoundly impacts the outcomes and results of text mining and sentiment analysis tasks, as it affects the accuracy and reliability of the analysis. Additionally, these techniques aim to streamline and improve the classification process by increasing its efficiency and accuracy, as well as reducing the computational complexity and cost. Subsequent sections will delve into a detailed explanation of these techniques and their applications in different scenarios and domains.

5.2 Data Preprocessing

Libraries

To prepare our data for analysis, we first had to perform some data cleaning steps. The first step in our data cleaning pipeline was to install and import several important libraries into the integrated development environment (IDE) that we were using.

Spacy: The spaCy library is a high-performance NLP library that provides pre-trained models for English and other languages. The “en-core-web-trf” model from spaCy is designed for English text analysis, and the toolkit has extensive capabilities for splitting words, labeling parts of speech, finding syntactic relations, identifying entities, and more.

Hugging Face Transformers: The Transformers toolkit is for natural language processing (NLP) operations and has a variety of ready-made models, such as “all-mpnet-v2”. It has features for adjusting, predicting, and creating text in different languages.

UMAP: A potent technique for shrinking the dimensions of data is the UMAP (Uniform Manifold Approximation and Projection) toolkit in Python. It can display data with many dimensions in fewer dimensions while keeping the original patterns and connections. UMAP has a fast and adaptable execution, enabling users to use it for various kinds of data and examine intricate datasets in a lower-dimensional space.

Pandas: The pandas library is a widely used and powerful data manipulation and analysis tool for the Python programming language. It offers a comprehensive set of data structures and functions that enable users to efficiently work with structured data, such as data that is organized in tables or time series. With the help of pandas, users can perform various tasks on data, such as loading, cleaning, transforming, analyzing, and visualizing data, making it an essential tool in data science and analysis workflows.

Fuzzywuzzy: Fuzzywuzzy is a library tool that facilitates fuzzy string matching and similarity analysis. It implements a range of string-matching functions, including the Levenshtein distance algorithm. The Levenshtein distance algorithm enables Fuzzywuzzy to quantify the similarity between two strings based on the number of insertions, deletions, and substitutions required to transform one string into another, providing a numerical measure of string similarity.

Exploratory Data Analysis

The data preprocessing phase consisted of several steps, and the one that followed the previous step was Data Loading and exploratory analysis. To preserve the quality of the data and ensure its integrity, the DataFrame, which was referred to as df, underwent a thorough and comprehensive examination and cleansing process. This process entailed identifying the presence of missing values and duplicate rows in the data and addressing them accordingly.

One of the functions that was employed to detect missing values in the data was the `isnull()` function. This function created a boolean mask that indicated which values in each column of the DataFrame were null or missing. Another function that was utilized in conjunction with the `isnull()` function was the `any()` function. This function helped to select only those columns that contained at least one missing value in them. The columns that were selected by this function were stored in a variable that was named `null_cols`.

For further analysis of the missing values in the data and to present more information about them, a new DataFrame was created and named `null_df`. This DataFrame had two columns that computed and displayed different statistics about the missing values in each selected column. The first column was named Null Count and it showed the count of missing values for each column that was selected by the `any()` function. The second column was named Null Percent and it showed the percentage of missing values for each column that was selected by the `any()` function. This percentage was calculated by dividing the count of missing values by the total number of rows in the respective column.

To see which columns had missing values and how many, the `null_df` DataFrame was sorted by the Null Count column in descending order. This was done with the `sort_values()` function. The result was a new DataFrame called `null_df_sorted`. It showed the columns with missing values, the number of missing values in each column, and the percentage of missing values out of the total values in each column. For example, the Location column had 13484 missing values, which was about 22.7 % of all the values in that column.

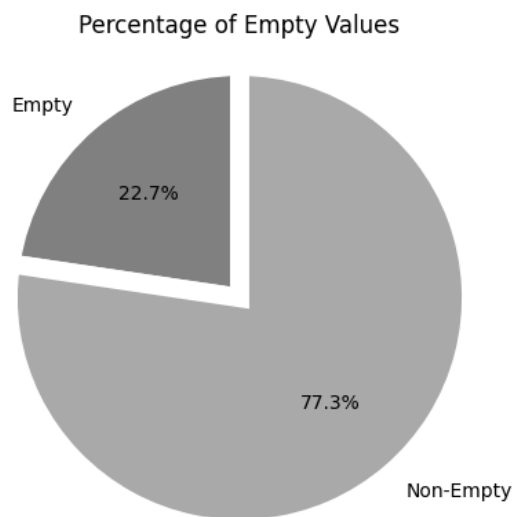


Figure 2: Null and Non-Null Values Comparison

Selecting the Columns

The Location column in the dataset was initially included in the data analysis. However, it was subsequently discovered that several of the values in this column did not correspond to real countries or cities. Instead, these values were made up, such as "150 Miles" and "35000 Feet above sea level." The inclusion of these made-up locations in the Location column would have introduced significant noise into the data analysis and would have skewed the results. As a result, it was decided to exclude the Location column from the data analysis to ensure that the results of the analysis were not adversely affected by the presence of noise.

Text Standardization

Text standardization plays a pivotal role in data preprocessing by ensuring consistent and uniform formats. This involves removing irrelevant characters such as punctuation marks, special symbols, and alphanumeric combinations. Lowercasing all text is another important standardization method to address case sensitivity issues during subsequent analysis.

In the preprocessing pipeline for tweets, the first step involves converting the text to lowercase, enhancing consistency and mitigating case-based duplication of topics. The transformation of uppercase letters to lowercase has demonstrated improved results over the baseline (HaCohen-Kerner et al., 2020). Orthographic correction and lowercase conversion were found to be optimal for this dataset.

The second step of the pipeline focuses on removing URLs from the tweet text, as they are extraneous to sentiment analysis. Employing the 'apply()' function with a lambda function and regular expressions, URLs are replaced with an empty string to minimize noise and irrelevant data.

Subsequently, HTML tags are eliminated from the tweet text using the 'apply()' function, a lambda function, and the BeautifulSoup library. By parsing the tweet text with BeautifulSoup and the "lxml" parser, the plain text without any HTML tags is extracted.

These preprocessing steps are crucial for tweet topic modeling as they improve machine understandability. Lowercasing text ensures words with different casing are treated identically, avoiding topic duplication. Removing URLs, HTML tags, and converting chat words to standard English equivalents enhance sentiment analysis and topic modeling accuracy by focusing on relevant content.

Additional preprocessing steps involve removing punctuation, numbers, extra spaces, and repetitions of punctuation. Punctuation marks are excluded as they provide limited information for sentiment analysis and topic modeling. Numerals are irrelevant and do not contribute to sentiment or topic identification. Extra spaces within tweets are eliminated to ensure text consistency for proper interpretation by natural language processing algorithms. Repetitions of punctuation marks are removed to standardize text and avoid biased interpretations.

Furthermore, emojis embedded in tweets are converted into words using the emoji library for standardization and compatibility with NLP techniques. Special characters, often invisible anomalies in tweets, are eliminated to prevent computational complications and errors. Expanding contractions is necessary to understand text data and assign accurate sentiment during analysis. Contractions can also increase the size of the document-term matrix, which may hinder training NLP models.

To preserve emoticons, a regular expression is devised to remove alphanumeric and extraneous punctuation characters while retaining sequences that signify emoticons. Non-English tweets are excluded from the dataset to maintain data integrity.

Tokenization

Since machines cannot understand human spoken language, programmers must break the texts down into the smallest possible fragments, just like words, in a process called Tokenization. It is a crucial technique used in natural language processing (NLP) which involves breaking text into smaller pieces called tokens (A. Mullen et al., 2018). Since tokens are the basic building blocks of natural language and can represent words, punctuation marks, numbers, or symbols, tokenization becomes a fundamental process before topic modeling. This step allows machines to process it using NLP techniques, such as part-of-speech tagging, stemming, lemmatization, or sentiment analysis as well

as larger projects that involve transforming or generating text, such as machine translation or text summarization.

Stop Words Removal

Stop words are the most frequently used words in a language. They are typically function words, such as articles, prepositions, pronouns, and conjunctions. These words do not add much meaning to the text. For example, the word "the" can be used as an article, a demonstrative pronoun, or a conjunction. This ambiguity can make it difficult for natural language processing algorithms to understand the meaning of the text.

We implemented stop word removal within the lambda function. It used a conditional statement to filter out tokens that are stop words. The conditional statement checks each token generated by 'nlp(x)' with the expression 'if not token.is_stop.' Only tokens that do not satisfy this expression are included in the final list. The code assigns the final list of non-stop word tokens to the 'tokens' column in the DataFrame 'df'..

Lemmatization

When dealing with text processing, for example, when performing a simple search, we are more interested in the root form of the word than the suffix. There are two methods for accessing the root word: stemming and lemmatization. The purpose of both operations is the same: to strip each word of its suffixes and turn them into a common base or root.

Lemmatization considers the morphological analysis of words to extract the correct lemma, which requires having dictionaries for each language for which such analysis is to be provided (Gualberto et al., 2020). The process of lemmatizing the tweets was performed using WordNetLemmatizer from the NLTK library. We first created an instance of the WordNetLemmatizer using `lemmatizer = WordNetLemmatizer()`. WordNet is a lexical database that provides information about word meanings, including their lemmas.

A dictionary called `wordnet_map` was defined to map the Parts of Speech (POS) tags from the NLTK POS tagger to the WordNet POS tags. This mapping is necessary to ensure that the lemmatizer recognized the correct POS of each word. After the words were reduced to their base forms, lemmatization on the tweets corpus was performed.

With all these steps, the dataset becomes more consistent and dependable, as it only keeps tweets written that have gone through the pre-processing mechanisms. This strict structure validation process strengthens the quality of later text analysis procedures, as it removes the noise and confusion that could come from untreated tweets.

5.3 Word Recurrence

Word frequency identification is one of the initial steps which helps to identify the most important words in a piece of text and provides insights into the most frequently occurring words in the dataset. The graph below shows the initial results of the frequency. The graph below shows the initial results of the frequency. As visible from the graph, a lot of these high frequency words were the handles of the airlines and hence, they could not contribute to our sentiment analysis. Hence, there was a need to remove them from our list. For this reason, in the next step, we filtered out values based on their similarity to values in the 'Airline' column.

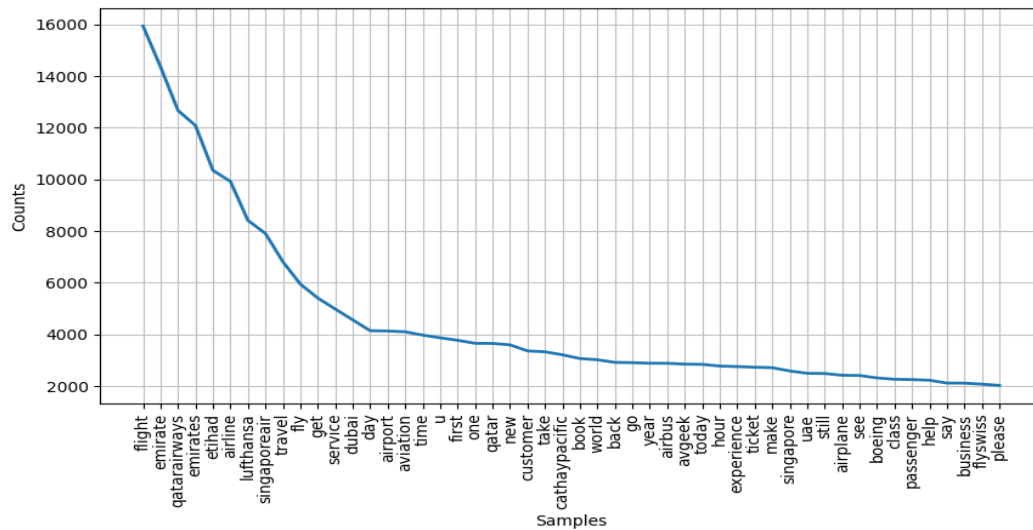


Figure 3: Pre-processing Word Frequency

The Levenshtein library was used to measure the similarity between strings. The first step was to create a list of airline names. Then, a similarity threshold was set, that determines the minimum similarity score required for a word in `word_freq_df` to be considered like any value in the `airline_list`. The similarity score was calculated using the partial ratio method from the `fuzz` library, which measures the similarity between two strings based on the number of matching characters.

The threshold value could be modified but it was set to 80 for optimum results. In the third step, we filtered the DataFrame `word_freq_df` using a lambda function applied to the 'Word' column. The lambda function iterated through each word in the 'Word' column and checks if any value in the `airline_list` had a partial ratio similarity score greater than or equal to the similarity threshold. If a match was found, the lambda function returned True, otherwise it returned False. After the filtering out these results, the data was ready for the next step of analysis.

5.4 Text to Vector Representation

Word embeddings were essential for natural language processing (NLP) tasks, where the aim was to enable machines to comprehend and process human language. Word embeddings were numerical representations of words or texts and their meaning and relationships. Textual data had to be converted into numerical form to utilize machine learning or deep learning models effectively. Models were not able to interpret text data as humans did. Therefore, vectorization techniques or word embeddings were required to transform text into numbers.

Vectorization, or word embedding, was the process of converting text data into numerical vectors. Each word or text was assigned a numerical representation that reflected its meaning and usage. These representations were derived from sophisticated algorithms that analyzed substantial amounts of text data and learned the information and relationships between words based on their co-occurrence patterns within a given corpus. Word embeddings captured the similarities and differences between words based on their context, which assisted models in making more meaningful inferences based on language.

These numerical representations, or word embeddings, served as features for various machine learning models. They provided a compact and meaningful representation of the text data, which enabled the models to understand and reason with language more effectively. They facilitated the conversion of text data into numerical form, which was essential for machine learning and deep learning models.

Powerful NLP models that performed tasks such as sentiment analysis, machine translation, question-answering, and more were built with word embeddings. These tasks involved interpreting and processing human language effectively, which word embeddings helped with by capturing the meaning and relationships between words or texts. Word embeddings were essential for natural language processing (NLP) tasks, where the aim was to enable machines to comprehend and process human language.

We then initialized a SentenceTransformer model called 'all-mpnet-base-v2'. Among the prevalent models in natural language processing (NLP), the all-mpnet-base-v2, all-MiniLM-L6-v2, and all-distilroberta-v1 models demonstrate remarkable results. However, the all-mpnet-base-v2 model surpasses the others for several reasons. First, it employs MPNet, a large-scale masked language model that exhibits superior performance in contextual and semantic comprehension of text. Second, the all-mpnet-base-v2 model leverages its size and capacity, enabling it to tackle intricate NLP tasks efficiently. Third, the all-mpnet-base-v2 model has undergone extensive fine-tuning and achieves high scores across various benchmarks, making it a dependable and versatile choice for NLP applications. Its robustness, language understanding abilities, and validated performance render all-mpnet-base-v2 a compelling option.

The goal of the contrastive learning approach was to train the model to identify which sentence from a pair belonged to it, among a set of randomly chosen alternative sentences. This process helps the model learn meaningful representations by contrasting sentences within the dataset. We applied the encoding process to the 'tokens' column to form a new column called 'text_vector_' that stored the encoded text. This is followed by the encoding process, which took a list of text inputs and returned their corresponding vector representations.

5.5 Clustering

Clustering is a type of unsupervised learning method, which is different from supervised learning method. Supervised learning method aims to learn a correlation relationship between a set of features X of an observation and a target value Y to be predicted.

For example, Unsupervised learning can be used to find airline market segments based on their flight behavior, preferences, and characteristics. Unsupervised learning does not need labeled data, but only the features of the observations. One way of finding patterns in the data is by grouping together airlines that have similar characteristics. This is called clustering, which is the process of dividing the data into clusters or groups based on some measure of similarity or distance. Clustering can help discover hidden structures and relationships in the data that are not obvious or known beforehand.

K-Means Clustering

Among the various unsupervised machine learning algorithms, the k-means algorithm is a popular one. This algorithm can discover distinct and non-overlapping clusters of data points, such that each point belongs to one group. The algorithm uses the minimum squared distance technique to assign each point to the closest cluster or subgroup, based on the distance from the cluster centroid (Zubair et al., 2022).

To be able to group a dataset into K distinct clusters, the K-Means algorithm needs a way to compare the degree of similarity between different observations. Thus, two data that resemble each other will have a reduced dissimilarity distance, while two different objects will have a greater separation distance. This method enables the partitioning of the data set into distinct and non-overlapping clusters, based on the similarity of the observations. Therefore, data points that are alike will be assigned to the same cluster.

Additionally, each observation can only belong to one cluster at a time (exclusivity of membership). It is not possible for an observation to be part of two different clusters. Mathematical and statistical literature is full of definitions of distance, the best known for cases of clustering is the Euclidean distance. Consider a matrix X with n quantitative variables. In vector space E^n . The Euclidean distance d between two observations x_1 and x_2 is calculated as follows:

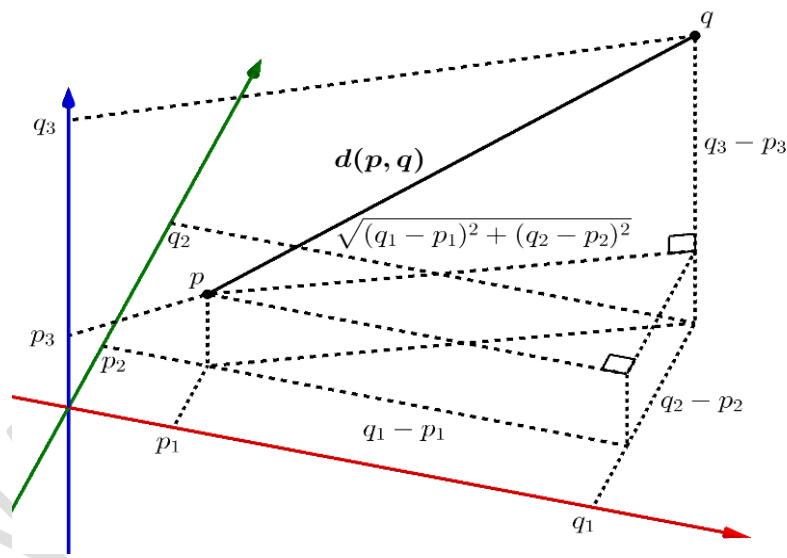


Figure 4: Euclidean Distance Formula

Mini Batch K-Means Clustering

K-means algorithm is one of the fastest clustering algorithms. But when the data gets bigger, it takes more time and space to run K-means because it needs to keep all the data in memory. Some ways have been suggested to make K-means faster and use less space. One of them is the Mini Batch K-means method.

The Mini Batch K-means method is a separate way of doing K-means in machine learning. It solves the space problem by using small, random, fixed-size groups of data. In each step, a random group from the data is picked and used to change the clusters. This goes on until the clusters stop changing.

Each group changes the clusters by mixing the values of the prototypes and the data, using a learning rate that gets smaller as the steps go on (Alonso, 2013). The learning rate is smaller when more data points are in a cluster during the process. The clusters are done when they are the same in two steps in a row. Tests show that the Mini Batch K-means method makes K-means much faster, but the clusters are not as good.

The method works well, but not much research has been done on how the data affects the clusters. For example, how many clusters or how big the data is. More research is needed to see how these things affect the clusters.

In the Mini Batch K-means method, only some data and k centroids need to be kept in memory at each step, not all the data. The distance between the group and the k centroids is found at each step to change the clusters and finish them.

Elbow Method

The Elbow Method is a technique that aims to find the optimal number of clusters without prior knowledge of the true answer. It involves running K-Means for various values of K and evaluating the results. The underlying principle is that as K increases, the differences between clusters tend to decrease, while the differences within clusters tend to increase. Thus, a balance is sought where clusters are as homogeneous as possible within themselves while being as heterogeneous as possible from each other. The Elbow Method refers to identifying the point in the plot that resembles an elbow, indicating the optimal number of clusters.

Since K-Means measures the distance of each observation to the centroid of its assigned cluster, the ideal scenario is to minimize this distance. Mathematically, the goal is to find a value of K that minimizes the intra-cluster sum of squares, also known as within-cluster sum-of-squares (WCSS), where the optimal value is zero.

One of the main difficulties in cluster analysis is to determine the optimal number of clusters, K , that best represents the data. This task becomes even more challenging when the data is large and complex, and when there is no prior knowledge or assumptions about the data structure or distribution. Different clustering algorithms may produce different results for the same dataset, depending on the choice of K and other parameters. Therefore, it is important to find a value of K that reveals meaningful patterns and insights in the data, rather than arbitrary or random groupings. However, there is no universal or objective criterion to determine the correct number of clusters, and the decision often depends on the specific context and goals of the analysis.

Selecting a high value of K can lead to overly detailed partitioning of the data, potentially obscuring meaningful patterns. On the other hand, choosing a low value of K may result in coarse clusters that fail to reveal fine-grained patterns. The Elbow Method helps in finding the best number of clusters by examining the distances between data points and their cluster centers for different values of K . The optimal number of clusters is determined by identifying the point in the curve where there is a significant bend or elbow.

In some cases, the curve may exhibit a subsequent decrease beyond the elbow point, indicating that further increasing the number of clusters leads to a deterioration in clustering quality, resulting in overfitting or fragmented data. When faced with such situations, determining a clear elbow point becomes challenging. The ambiguous nature of the curve may reflect the complexity or overlapping nature of the underlying data structure, making it difficult to identify a distinct number of clusters.

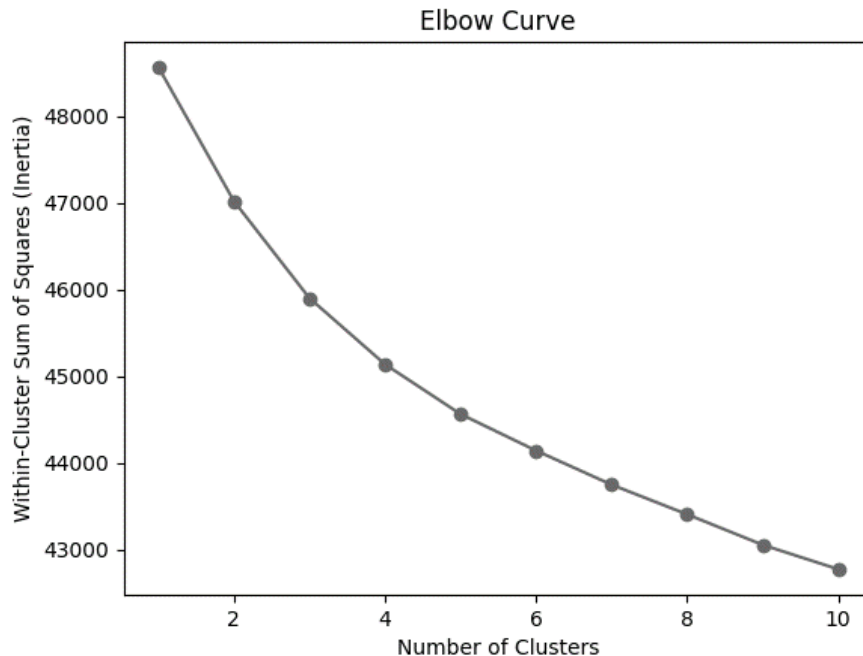


Figure 5: Elbow curve

The plotted graph illustrates the behavior of the curve, revealing a gradual flattening after reaching points 3 and 4. Prior to these inflection points, the curve exhibited a significantly steeper slope, indicating a more pronounced change in the underlying data. However, beyond point four, the gradient becomes less steep.

The choice between 3 or 4 clusters should be made by carefully examining the data and aligning it with the research goals. The aim is to identify the number of clusters that effectively capture the patterns and structure within the data.

Given the absence of a distinct elbow point in the curve, alternative methods are employed to determine the optimal number of clusters. In this regard, the Silhouette curve method is utilized as an effective approach. This technique assists in evaluating the quality and appropriateness of different clustering solutions, considering the intrinsic characteristics.

Silhouette Method

To assess the separation distance between clusters, the Silhouette method provides a robust approach. It uses silhouette plots that visually show how close each point is to neighboring clusters, which helps to determine the optimal number of clusters. The closeness of data points to their respective clusters is quantified by the Silhouette coefficient, which ranges from -1 to 1. Higher coefficients mean stronger similarity within clusters and distinctiveness from other clusters.

This method makes the decision-making process clearer and more meaningful by offering a criterion for selecting the number of clusters. However, the computational costs of calculating the silhouette coefficient for each case should be considered and an appropriate metric should be chosen based on the specific needs of the application.

The silhouette coefficient for a sample is calculated by finding the intra-cluster mean distance “a” and the closest mean cluster distance “b” for each sample. Then, $(b - a) / \max(a, b)$ is computed. A metric for assessing the quality of clustering is provided by the average silhouette coefficient across all samples. The Silhouette method evaluates how well objects are allocated within groups, reflecting the homogeneity of each group. The index values range from -1 to 1, with higher indices meaning greater similarity within the group and dissimilarity from other groups.

In this case, both the silhouette curve scores and the elbow curve graph were considered to determine the optimal number of clusters. The highest silhouette score, indicating well-defined and distinct clusters, was shown by point four in the silhouette curve. Point four was also close to the bend in the elbow curve graph, indicating a significant change in the data structure. These factors were carefully evaluated and four clusters were selected, which balanced within-cluster similarity and inter-cluster dissimilarity, allowing for meaningful pattern recognition and data segmentation.

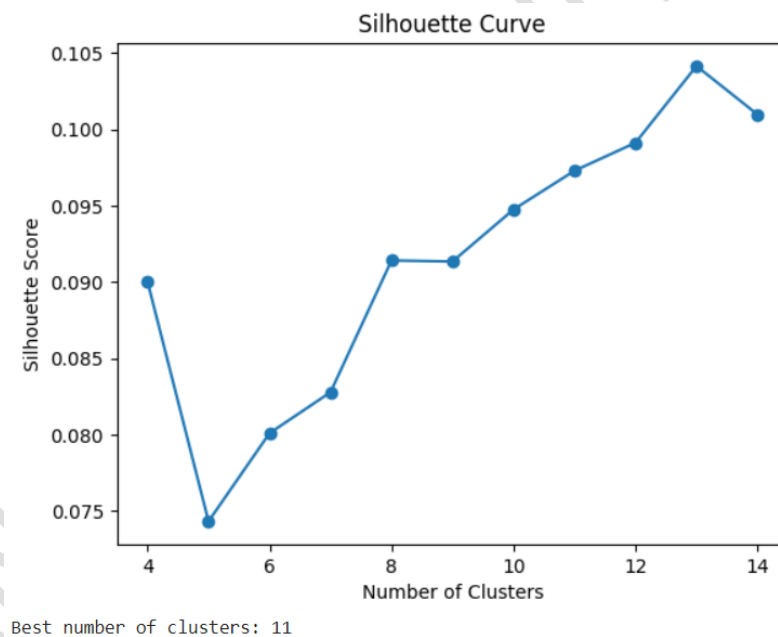


Figure 6: Silhouette Score Graph

5.6 Summary

The application of clustering techniques for topic modeling in tweets is the focus of this thesis chapter. It examines how K-Means and Mini Batch K-Means algorithms can be used in the clustering process. The chapter aims to identify the most effective approach for grouping tweets into coherent topics, which can reveal the underlying themes and discussions on social media platforms. Moreover, the chapter determines the optimal number of clusters, denoted as K , by using a combination of the elbow and silhouette methods.

Clustering Techniques: K-Means and Mini Batch K-Means: The chapter starts by presenting the K-Means algorithm as a common clustering technique. The algorithm partitions data points into K clusters based on their similarity. It can group tweets that share common themes effectively, which is useful for tweet topic modeling. The algorithm updates the centroids and assigns data points to clusters iteratively until convergence. It is a computationally efficient approach that can work on large datasets.

The chapter also investigates the use of Mini Batch K-Means in tweet clustering. This technique is a variant of the K-Means algorithm that uses random subsets, or mini-batches, of the data to update the centroids. This technique reduces the computational burden by processing subsets of the data at each iteration, which is beneficial when dealing with large-scale datasets. Mini Batch K-Means converges faster and can manage streaming data scenarios.

Determining the Optimal Number of Clusters, K : A key aspect of the clustering process is finding the optimal number of clusters, denoted as K . The chapter uses a combination of the elbow and silhouette methods to find the best value for K . The elbow method runs the K-Means algorithm with different values of K and computes the resulting sum of squared distances between data points and their cluster centroids. The point where the decrease in the sum of squared distances becomes negligible, forming an “elbow” shape in the graph, is the optimal K . However, the elbow method may not provide an obvious choice for K .

The chapter also uses the silhouette method, which computes the average silhouette coefficient across all data points. The silhouette coefficient measures the cohesion within clusters and the separation between clusters, with values ranging from -1 to 1. A higher silhouette coefficient indicates that the data points are well-clustered and distinct, suggesting a better choice of K . By combining the elbow and silhouette methods, the chapter concludes that $K=11$ is the optimal number of clusters for the tweet topic modeling task.

Chapter 6:

Results & Conclusions

6.1 Introduction

Using topic modeling techniques, this chapter presents the outcomes of the data analysis conducted on tweets about airlines. This section builds upon the data preprocessing and clustering algorithms discussed in the previous chapter and explores the findings in-depth. The aim of this chapter is to provide a comprehensive understanding of the discussions surrounding airlines on social media platforms, shed light on customer sentiment and emotions expressed in the tweets, and offer insights into the prominent topics and clusters.

The results of the topic modeling analysis are presented at the beginning of the chapter. The identified topics, their distribution, and the corresponding keywords that represent the essence of each topic are highlighted. The topics extracted from the tweets provide a comprehensive overview of the major themes discussed by users, ranging from aspects such as flight experiences, customer service, and pricing to specific features, amenities, and destinations. This analysis serves as a foundation for understanding the underlying trends and concerns within the dataset.

The chapter then delves into the topic modeling of the tweets following the discussion of topics. This study aims to gauge the overall topics of discussion regarding airlines by analyzing the content expressed in the tweets, ranging from positive to negative. The sentiment analysis provides insights into the customers' emotional responses, enabling airlines to understand the factors that influence customer satisfaction and dissatisfaction. It also helps to find ways to improve and make strategies that focus on customers.

Furthermore, the chapter presents the results of the clustering analysis, which groups the tweets into distinct clusters based on their similarities using clustering algorithms like K-Means and Mini Batch K-Means. The cluster analysis provides a more detailed view of the discussions, enabling the identification of specific user segments and their preferences. This information empowers airlines to tailor their marketing strategies and communication efforts to different customer segments, resulting in improved customer engagement and satisfaction.

6.2 Word Frequency

The word frequency results have been updated based on the latest data to provide a detailed understanding of the most frequently cited words and their respective frequencies in the tweets about the top ten airlines. These results enable us to gain insights into the prevalent topics and sentiments conveyed within the tweet dataset. In this section, we will delve into some noteworthy observations that can be derived from the provided word frequency list:

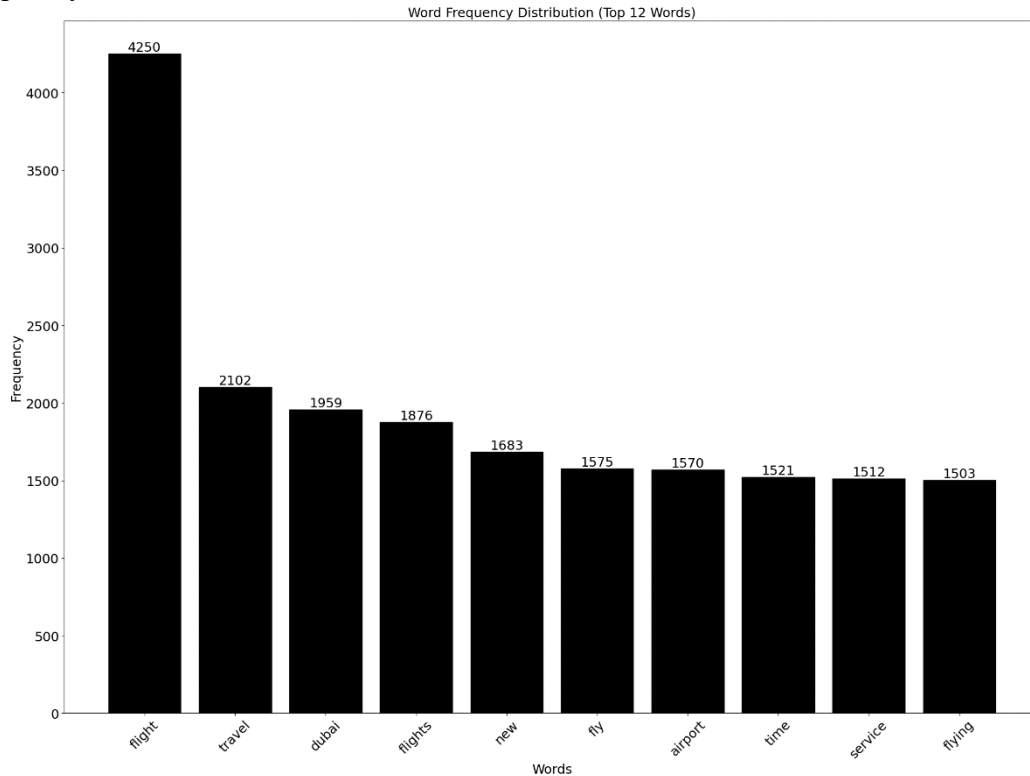


Figure 7: Most Common Words

Flight, Flights & Fly: The analyzed tweets reveal that the word "flight" appears most frequently, with a frequency of 4,250. Additionally, the term "flights" has been used 1,876 times. These numbers, totaling 7,701, indicate a significant focus on flight experiences.

Passengers who flew on these airlines wrote both positive and negative reviews about their trips. For example, passengers tweeted about their flights that got delayed or cancelled, or about a flight about which they were excited. Airlines also used these terms alongside holiday operators and news agencies to market their products and services in the past. For example, there were tweets about airlines starting 'flights' to new destinations, and the launching of special promotions for attracting more travelers. This contributed to the popularity and relevance of these terms, which increased their likelihood of being used by passengers when they tweeted about their own flights in the past.

Several factors could explain why passengers tweeted about their flights in the past. One factor was that passengers shared their experiences with other travelers or with the public. Flying could be a stressful experience, so passengers often tweeted about their flights to

convey their feelings or opinions. They might have also tweeted about their experiences to inform other travelers about their own flights.

Another factor was that passengers sought help from the airlines when they encountered a problem with their flight. If a passenger could not get help from the airline through other channels, such as phone or email, they might have tweeted about it in the hope that the airline would notice their tweet and resolve the issue. A third factor was that passengers connected with other travelers who had similar experiences. Twitter was an effective platform for passengers to communicate with other travelers who offered their suggestions, recommendations, or stories about flying.

Travel: Another frequently appearing term in the dataset is "travel," which occurs 2,102 times. This high frequency indicates Twitter users' inherent interest in discussing various aspects of travel experiences.

Airlines play a pivotal role as a mode of transportation, allowing people to travel by air. When individuals tweet about airlines, they often express their opinions, share experiences, or express expectations related to air travel. A general lookup into these tweets showed that the context ranged from booking flights to writing reviews about their experiences. Twitter serves as a significant platform for discussing airlines, with users frequently expressing their curiosity about new airlines, comparing fares, and seeking insights from others' travel experiences. It has become a subject of widespread interest and curiosity among many Twitter users.

Airlines utilize Twitter as a marketing tool to share company news, announce new destinations, and offer discounts. This strategic use of the platform contributes to increasing brand awareness and attracting potential customers. Twitter, as a social media platform, provides a suitable space for people to communicate with others who share their travel interests. It has become a hub where users freely share their travel stories, seek advice, and discover new destinations.

Dubai: One of the terms that appears frequently in the dataset is "travel," which has a high frequency of 1,959 occurrences. This high frequency reflects the inherent interest of Twitter users in discussing various aspects of travel experiences. Airlines are an essential mode of transportation that enables people to travel by air. When individuals tweet about airlines, they often express their opinions, share experiences, or express expectations related to air travel.

A general lookup into these tweets revealed that the context varied from booking flights to writing reviews about their experiences. Twitter serves as a significant platform for discussing airlines, with users frequently expressing their curiosity about new airlines, comparing fares, and seeking insights from others' travel experiences. It has become a topic of widespread interest and curiosity among many Twitter users.

Airlines also use Twitter as a marketing tool to share company news, announce new destinations, and offer discounts. This strategic use of the platform helps to increase brand awareness and attract potential customers. Twitter, as a social media platform, provides a suitable space for people to communicate with others who share their travel interests. It has become a hub where users freely share their travel stories, seek advice, and discover new destinations. Additionally, Twitter allows users to interact with airlines directly,

providing feedback, asking questions, or requesting assistance. This enhances customer service and loyalty, as well as creating a sense of community among travelers.

New: The word “new” appeared 3,266 times in the analysis of tweets, indicating a considerable focus on various aspects of novelty within the aviation industry. Several factors account for this phenomenon, such as the introduction of new routes, aircraft, technologies onboard, meals, and other aviation-related news. It is noteworthy that websites and blogs rather than individual users generated a sizable proportion of tweets about “new,” suggesting a prominent role of these platforms in disseminating information and news updates within the aviation community.

A prominent factor contributing to the frequency of “new” in tweets is the introduction of new routes. Airlines constantly seek to expand their networks, access new destinations, and improve connectivity. The announcement of new routes sparks considerable interest and excitement among travelers and aviation enthusiasts. The occurrence of “new” in the tweets implies that users are actively interacting with this information, sharing their expectations, opinions, or requesting more details about the newly introduced routes.

The incorporation of innovative technologies onboard flights is also a crucial factor influencing conversations about “new.” Airlines continuously adopt innovative technologies to improve passenger comfort, upgrade in-flight entertainment systems, and optimize operational processes. The frequent occurrence of “new” in tweets suggests that users are actively engaged in conversations about the latest technological advancements and their implications for the aviation experience.

Besides routes, aircraft, and technologies, the word “new” also covers discussions related to new meals and other services offered onboard flights. Airlines frequently refresh their menus, collaborate with celebrity chefs, or offer unique dining experiences to differentiate their offerings and cater to diverse passenger preferences. The prevalence of “new” in tweets underscores the attention given to such culinary innovations, as users share their impressions, recommendations, or express excitement about trying the latest meals.

Additionally, it is important to note that many tweets mentioned the capital of India, New Delhi, which is home to one of the busiest airports in the world. The city's high population and status as a major travel hub contribute to its frequent mentions in discussions about aviation. New Delhi serves as a crucial connection point for domestic and international flights, attracting a significant volume of air travelers. Consequently, the city becomes a focal point of conversations regarding flight experiences, services, and airline operations. The bustling nature of New Delhi and its airport underscores its relevance within the aviation discourse on social media platforms like Twitter.

The finding that websites and blogs produced a considerable proportion of tweets about “new” implies the vital role of these platforms as sources of aviation-related news and updates. Websites and blogs often serve as authoritative and reliable sources of information, enabling users to stay informed about the latest developments in the aviation industry. Their extensive coverage and in-depth analysis contribute to the prevalence of “new” in the tweets, as users rely on these sources for comprehensive and up-to-date information.

Aviation: Twitter users showed a keen interest in the aviation industry, as evidenced by the term “aviation” occurring 4,049 times in the tweets analyzed. This term was

frequently used in relation to diverse topics concerning aviation, such as airlines, aircraft, industry events, and news. This indicated that users were well-informed and engaged in conversations about the aviation sector and its developments.

As discussed earlier, air travel is an essential part of the wider civil aviation industry. Hence, it is not surprising that many tweets focused on aviation-related topics. The term “aviation” appearing in many tweets further supported this observation. Users discussed several aspects of aviation, including airlines, aircraft, industry news, and events.. Users also talked about industry news and developments. They shared and commented on breaking news, updates on new aircraft models, technological innovations, regulations, and industry trends. This showed users’ interest and awareness of the latest happenings and developments in the aviation sector.

Moreover, the use of hashtags might have influenced the prominence of the term “aviation” in the tweets that were analyzed. Hashtags are often used on social media platforms like Twitter to categorize and increase the exposure of specific topics. By using hashtags related to aviation, users could make sure that their tweets are noticed by a larger audience interested in the industry. This clever use of hashtags might have boosted the visibility and frequency of the term “aviation” in the analyzed dataset.

Service: With a frequency of 4,040, the word “service” signifies the significance of discussions related to service experiences and customer service within the context of travel or aviation. The tweets may reflect opinions and experiences regarding the quality of service provided by airlines or other service-related aspects. This may indicate the importance and influence of service quality on customer satisfaction and loyalty.

The results indicate how customer service is especially important in the airline industry and how it affects how happy and loyal passengers are. Travelers like to share their stories, thoughts, and feedback on social media platforms, like Twitter, and show how good or bad the service they get from airlines is. The word ‘service’ is used a lot by travelers to talk about different things related to customer service, like how airline staff treat them, how fast and well they answer their questions and problems, what services they get on the plane, and how happy they are with the service they get.

Travelers use the word ‘service’ to say good and terrible things that happen to them during their trips. For example, they may say thank you for great service from airline staff, like nice and helpful cabin crew members, quick and good baggage handling, or clear and honest communication when flights have problems. But travelers may also complain about bad service, like customer support that does not listen, long waits to fix problems, or poor services on the plane.

The frequent use of ‘service’ shows that passengers value service quality from airlines. Customer service is a key part of their travel experience, and their tweets reflect their expectations, opinions, and emotions about the service they receive from airlines. The high occurrence of ‘service’ also implies that airlines should avoid poor service to their passengers. The tweets indicate that excellent customer service throughout the travel journey is crucial. Passengers use ‘service’ in their tweets to show that they appreciate airlines that communicate well, and provide an enjoyable experience.

Airlines should know that how passengers feel about service can change how happy, loyal, and likely to choose them again they are. By looking at the tweets that talk about

customer service experiences, airlines can learn what they need to do better and how to make their service quality better. This may mean spending more money on staff training, making communication better, solving passenger problems faster, and giving good service all the time during the travel process.

Airport: The term ‘airport’ was another important topic with a frequency of 3,266 times in the dataset that was analyzed. This frequency indicated that users participated in discussions and conversations about various aspects of airport experiences.

One sub-topic that emerged in these tweets was about their overall experiences regarding airport services. Users shared their satisfaction or dissatisfaction with services offered at airports, such as the availability of amenities, cleanliness, staff behavior, or the overall quality of customer service. They also discussed their experience of passing through airport security and any mishandling involved in the security procedures.

Some users discussed the crowd and congestion experienced at airports. They shared their feelings about the degree of crowding, long queues, or delays in different areas of the airport, such as check-in counters, security checkpoints, or boarding gates. This included discussions about actions taken by airports to manage crowds, improve flow, and enhance passenger experience.

There were also tweets about airport closures due to weather conditions or other factors occurred. Users shared information, updates, and their experiences related to airport closures, flight cancellations, or disruptions caused by severe weather events. These conversations on Twitter provided real-time insights into the impact of airport closures and disruptions on travelers and their opinions about the handling of such situations.

Time: The frequency analysis of tweets revealed a strong focus on time themes, especially flight schedules, delays, and time management during travel. The word “time” occurred 3,069 times, indicating Twitter users’ attention and sensitivity of their travel experiences, especially the effects of delays and layover durations on their plans.

Flight delays are common in the airline industry and can disrupt travel plans, causing dissatisfaction and inconvenience for passengers. Twitter serves as an outlet for individuals to express their discontent, seek information, vent their frustrations, or share their experiences when flights are delayed. Hence, the frequent use of “time” with flight delays implies that passengers are actively expressing their concerns.

Layover durations, which refer to the time between connecting flights, are another principal factor of travel that influences travel experience. Longer layovers can be seen as an opportunity to explore the connecting airport or nearby attractions, while shorter layovers can induce stress due to the limited time to navigate through security, reach the gate, and ensure a smooth transition.

The frequent use of “time” in relation to travel discussions might indicate passengers sharing their experiences, seeking advice, or expressing their satisfaction or dissatisfaction with the layover durations.

Furthermore, Twitter provides a convenient and accessible way to communicate and get real-time updates about flight schedules and delays. Passengers can interact with airlines, airports, and fellow travelers, exchanging information, seeking assistance, or voicing their

concerns. The prevalence of “time” in these discussions may highlight the importance of timely updates and effective communication from airlines and airports to enable better time management and minimize the negative impact of delays.

Airbus: Finally, the term “airbus” has a frequency of 2,946, suggesting a notable focus on discussions related to Airbus, including aircraft manufactured by Airbus, airline fleets that include Airbus planes. To elaborate, several factors account for the frequent occurrence of the term “Airbus” in the top ten most discussed tweets. Airbus, as a leading aircraft manufacturer, offers a wide range of aircraft models that cater to different market segments and operational requirements. Their aircraft, renowned for their efficiency, comfort, and advanced features, are highly demanded by airlines around the world. Airlines like Qatar Airways, Etihad Airways, and Lufthansa use Airbus aircraft extensively, signifying their confidence in the reliability and performance of these aircraft.

The choice of Airbus aircraft for shorter and busier routes such as by airlines like Qatar Airways, Etihad Airways, and Lufthansa can also contribute to the frequent mentions in tweets. Airbus aircraft, known for their versatility and reliability, are well-adapted for such operations, which require operational efficiency and quick turnaround times. These qualities assist in making them a preferred choice for airlines operating on these routes.

Furthermore, the prominence of Airbus in the most discussed tweets also reflects the fascination and interest surrounding aviation-related topics. Aircraft enthusiasts often engage in conversations about different aircraft manufacturers, comparing performance, features, and overall flying experience. Airbus, being one of the major players in the industry, naturally becomes a topic of interest and discussion.

Flying: In the analysis of tweets related to airlines, the term 'flight' gained significant prominence, ranking as the 10th most frequently occurring and widely discussed term. Throughout the dataset, it appeared a total of 2946 times, indicating the substantial engagement of users in expressing their flight-related experiences, opinions, and expectations via various social media platforms.

The term ‘flight’ appears frequently in the analyzed tweets, indicating that the users have a strong interest and engagement in various aspects of their flight experiences. They use social media platforms to share their opinions, stories, and expectations regarding their flights, as well as to seek information and advice from other users. The term ‘flight’ covers a wide range of subjects and discussions that offer valuable insights into the diverse and multifaceted nature of passenger experiences in the domain of air travel. These discussions include topics such as connecting flight times, which reflect the convenience or challenges associated with making smooth transitions between flights; positive and negative flight experiences, which capture the satisfaction or dissatisfaction of the passengers with the service quality, amenities, and occurrences during their flights; and inquiries about flights to new destinations, which demonstrate the curiosity and exploration of the passengers who want to travel to different places.

Social media platforms function as a valuable medium for passengers to articulate their viewpoints, divulge personal stories, and solicit information pertaining to these specific dimensions of their flight experiences. For instance, users may convey their gratification or discontent with connecting flight times, accentuating the ease or challenges encountered during transitions. They may also narrate noteworthy or problematic flight

experiences, recollecting instances of exceptional service or voicing concerns regarding delays, cancellations, or other issues faced during their journeys.

ANALYZE2GETHER.COM

6.3 K-Means Clusters

Topic modeling using K-mean produced clusters that reflect various themes or topics in the tweets data. Each cluster contains documents or data points that have similar features, such as keywords, context, or semantic meaning.

The first step in the data analysis process was to load the dataset using the Pandas library, which is a popular tool for data manipulation. Next was the cleaning process, which involved removing matching strings with airline names using the Levenshtein similarity ratio, which is a measure of how similar two strings are based on the number of edits required to transform one string into another.

The data cleaning process also included removing URLs, HTML tags, punctuation, numbers, repeated characters, and converting emojis to words. These elements were removed because they did not contribute to the semantic meaning of the text and could interfere with the analysis. Special characters were removed, and contractions were expanded using the contractions library, which is a Python module that provides a list of common contractions and their expanded forms. Rows with 'User' values matching airline names were dropped to eliminate self-referencing, as these tweets did not reflect the opinions or experiences of passengers.

The next step was to convert the text to lowercase and perform tokenization and stop word removal using the spaCy library, which has a collection of pre-trained statistical models and word vectors, which are essential for tasks such as tagging, parsing, and entity recognition. These models allowed us to easily and accurately analyze and extract information from text data. Its advanced convolutional neural network models accurate label, parse, and name entity recognition.

To enhance the accuracy of lemmatization, part-of-speech (POS) tags were mapped, and a lemmatization function was implemented to transform the tokens. Using the SentenceTransformer library, which assists in generating sentence embeddings using transformer models, sentence encodings were generated for the cleaned text. These encodings represented the semantic representations of the text, capturing its meaning and context in a numerical form.

To determine the optimal number of clusters for topic modeling, the within-cluster sum of squares (inertia) was calculated for different cluster numbers using K-means clustering. Inertia is a measure of how cohesive or compact each cluster is, based on the distance between each data point and its cluster centroid. The elbow curve method was applied to identify the "elbow" point, which signifies the optimal number of clusters. The elbow point is where there is a significant decrease in inertia, indicating that adding more clusters does not improve much the cluster quality. The curve was plotted, showing the inertia values for different cluster numbers.

Based on the elbow curve analysis, the optimum number of clusters was determined. The elbow point represented the most significant reduction in inertia, indicating the best balance between cluster compactness and separation. The optimum number of clusters was found to be eleven.

Scatter plot

The scatter plot is a type of graphical representation that uses a collection of points placed on a plane using Cartesian coordinates to display the values of two variables. By assigning a variable to each axis, it can be possible to determine whether there is a relationship or correlation between the two variables, or how they vary with respect to each other.

The two dimensions that are shown on the scatter plot were produced by applying dimensionality reduction using the UMAP algorithm. UMAP, which stands for Uniform Manifold Approximation and Projection, is a technique that can be used to transform high-dimensional data into lower-dimensional data while preserving its structure and patterns.

A total of eleven clusters were generated, as determined optimal number of clusters by the silhouette curve. The scatter plot, generated using the Plotly Express library, allowed for an interactive exploration of the clustered text data. By hovering over data points, the corresponding cleaned text content was displayed, for clearer understanding of the textual information represented in each cluster. The cleaned text content was obtained by applying various preprocessing techniques, such as tokenization, lemmatization, stop word removal, and punctuation removal, to reduce noise and enhance the quality of the text data for clustering and visualization purposes.

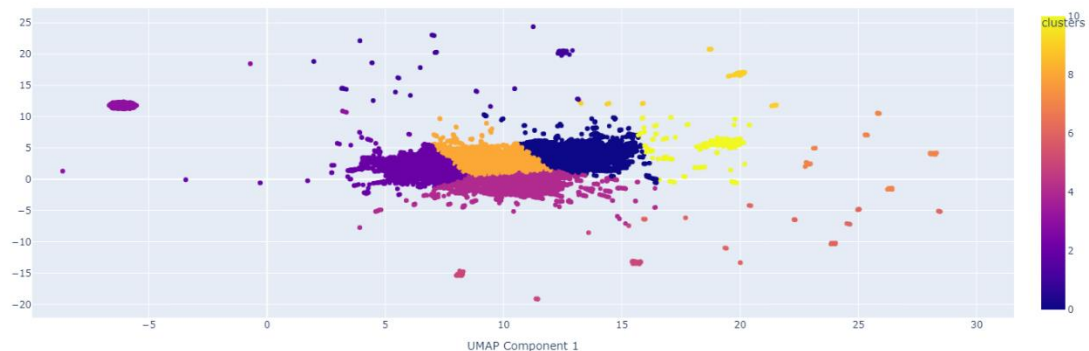


Figure 8: Scatterplot of clusters

Word Clouds

First Cluster

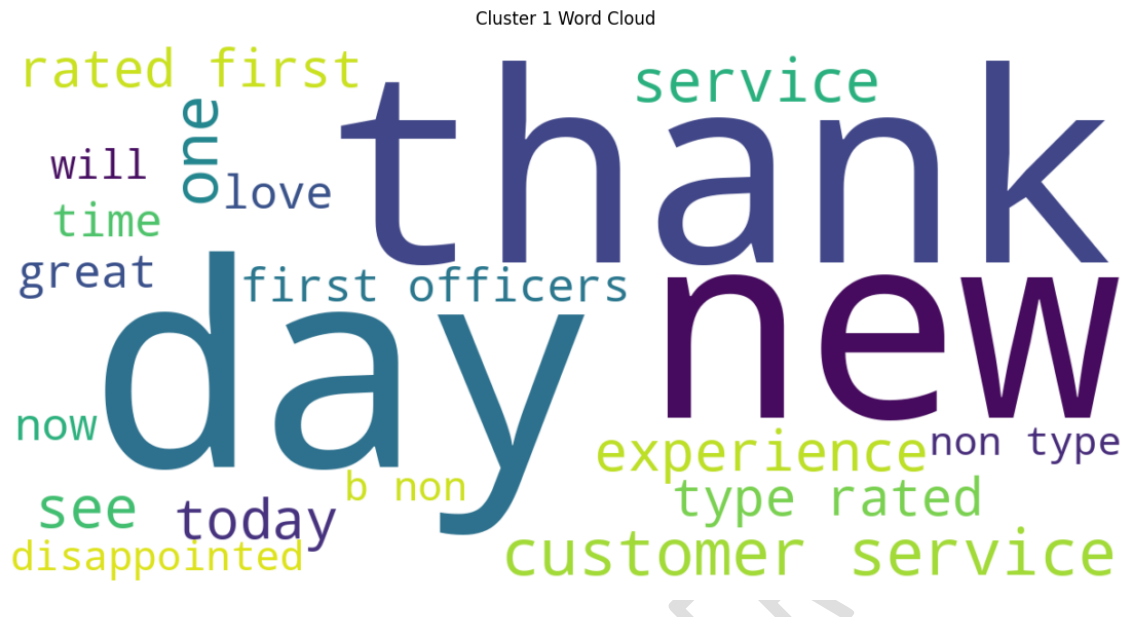


Figure 9: Word Cloud 1

The first cluster, represented by dark blue on the scatter plot, covers a wide area, indicating its diverse range of topics. The word cloud below displays the most frequent words in tweets belonging to this cluster, revealing the themes discussed within it.

Discussions related to aviation professionals' careers and job prospects are evident in the cluster, highlighting the dynamic employment landscape in the thriving aviation industry. Terms like "non-type-rated," and "first officers" refer to diverse types of pilots with different qualifications and experience levels.

The words "service" and "customer service" highlight the significance of the service quality in the airlines. Passengers tend to share their experiences, either positive or negative, concerning the level of assistance, responsiveness, and professionalism exhibited by the airline staff during the travel journey. These discussions can be critical in shaping the airline's reputation and customer loyalty. For instance, passengers appreciated timely and courteous responses from the airline staff, and they have expressed frustration and disappointment over delayed flights or poor communication.

The occurrence of words like "thank you" and "Love" suggests that some tweets express appreciation and positive sentiments towards the airline, for exceptional service or memorable experiences. These tweets may indicate that passengers have received some form of value or benefit from the airline that exceeded their expectations or made them feel special. For example, passengers may thank the airline for providing them with an upgrade, complimentary service, or a friendly gesture.

Conversely, the presence of "disappointed" signifies that other tweets may entail negative feedback or dissatisfaction regarding certain aspects of the airline's service delivery. These tweets may indicate that passengers have encountered some problems or issues during their flight or interactions with the airline that negatively affected their experience or perception.

The term “experience” emerges as a crucial topic in these tweets, signifying the passengers’ overall encounter during their flight and interactions with the airline. Positive experiences encourage brand advocacy, word-of-mouth, and repeat purchases among customers. Negative experiences can prompt customers to share their grievances and potentially influence the decisions of other potential customers. For example, passengers may share their experiences about how they felt during their flight, whether they enjoyed it or not, whether they faced any problems or issues, whether they received any surprises or delights, and whether they would recommend or choose the airline again.

Second Cluster:

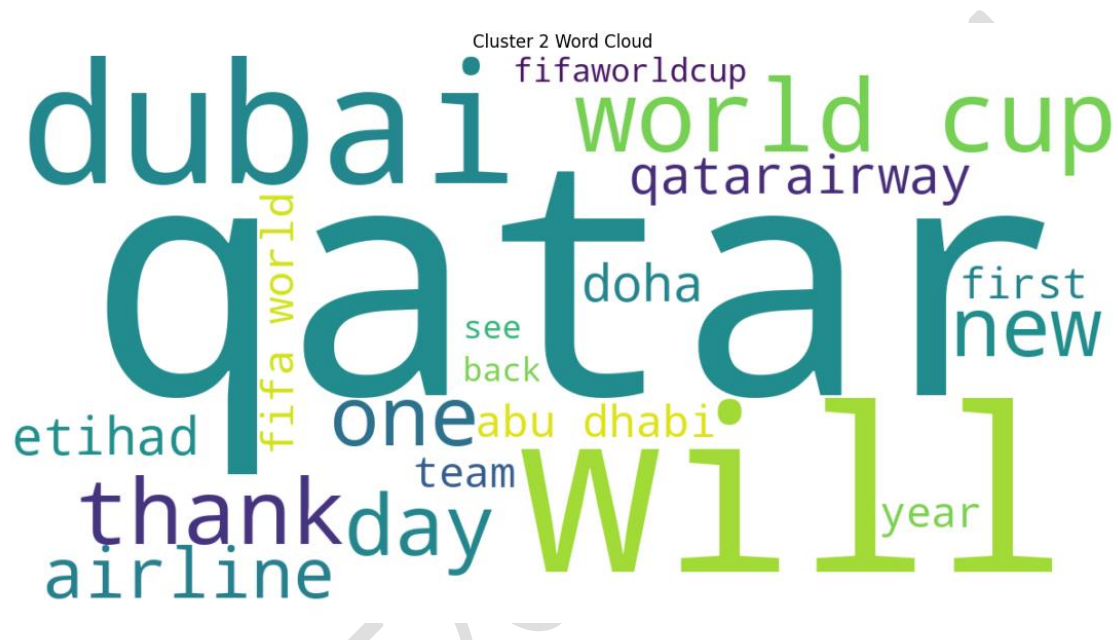


Figure 10: Word Cloud 2

The word cloud for the second cluster provides significant insights into the discussions and sentiments surrounding airlines in the context of the FIFA World Cup, with a specific focus on airlines based in Qatar, such as Qatar Airways. The cluster prominently features large words like "Dubai," "fifaworldcup," "world cup," "Qatar airway," and "qatar," suggesting that these topics are central to the tweets related to airlines.

The term "Dubai" refers to the airlines and travel to Dubai, which is a popular tourist and transit destination known for its iconic skyline. This implies that Dubai-based airlines such as Flydubai and Emirates are a major player in the low-cost and legacy airlines.

The mention of "fifaworldcup" and "world cup" suggests that the tweets which revolved around airlines' involvement in facilitating travel for major sporting events like the FIFA World Cup 2022 in Qatar. These events drew a considerable number of travelers, and airlines played a crucial role in transporting fans from across the World to Qatar.

The words "qatarairway" and "qatar" are central to the cluster, signifying a focus on Qatar's flagship airline, Qatar Airways. It is probable that the tweets discuss the airline's services, routes, customer experiences, or even exclusive offers related to the FIFA World Cup. Qatar Airways, being a prominent airline, would attract considerable attention during such a global event.

In addition to Qatar Airways, "Etihad" and "Abu Dhabi" appear as small words in the cluster. These could indicate mentions of Etihad Airways, another major airline based in Abu Dhabi, and its role in enabling fans to use the airline to attend the FIFA World Cup 2022 and for their regular travels.

The presence of words like "thankday," "airline," and "team" suggests that some tweets may express gratitude or discuss various airlines' teamwork and contributions during the World Cup. These topics could encompass passenger experiences, team sponsorships, or airlines' efforts to support the event.

Third Cluster



Figure 11: Word Cloud 3

The scatterplot's third group has short words and casual phrases that people use a lot on social media, like "heyyyy," "hey," and "guys." These words are called greetings or salutations, and they show that the users are friendly and relaxed with each other. They do not really mean anything specific, but they help start or keep the conversation going in a casual way.

The group also has "man" and "jokers," which are more casual words that people use to talk to each other, sometimes to make jokes or have fun.

The group has some personal words too, like "bestie," "wants," and "oh." These might be about a good friend ("bestie") or something they want or like ("wants" and "oh"). There are also some other words that do not seem to fit together, like "bighead," "clear," "hopper," "okay," and "muller." These might be parts of sentences or random words that people use in everyday talk without saying much.

We must remember that word embeddings or vector representations are good at finding meanings, but they might not get the minute details of casual language or social talk. So, they might put these casual phrases together without knowing what they really mean or how they relate to each other. These words do not tell us much by themselves, so they end up in a separate group that is informal and vague.

This group's words show that the embedding process can group together short, casual, and not important expressions in one group. This makes it hard to understand short and context-dependent texts. We need more analysis and context to get useful information from these groups in tweets about airlines or anything else.

Fourth Cluster



Figure 12: Word Cloud 4

The fourth cluster of the scatterplot contains the term “nan,” which is an acronym for “Not a Number,” a term commonly used in data analysis and text preprocessing to denote missing or undefined values that may occur during various operations on text data. For example, when transforming sentences into numerical vectors or cleaning text data by removing unwanted characters or words, some parts of the text may become invalid or undefined, resulting in “nan” values.

It is worth mentioning here that text preprocessing is a crucial step in data analysis, as it prepares the raw text data for further processing and analysis by applying various techniques to modify the text data according to the specific needs of the analysis. Some of these techniques include removing special characters, punctuation marks, and stop words that do not contribute to the meaning of the text, converting all the text to lowercase to avoid case sensitivity issues, and tokenization, which splits the text into smaller units called tokens. These techniques aim to simplify and standardize the text data and make it more suitable for analysis.

However, these techniques may also introduce some problems or challenges in the text data, such as missing or undefined values. During the text preprocessing or transformation steps, some words or segments of the text may become unrecognizable or lose their original meaning due to several reasons, such as errors in data collection or extraction, incomplete or corrupted data, or inappropriate application of preprocessing techniques. These words or segments are then replaced by “nan” values to indicate that they have no meaningful content or are unavailable.

In the context of Word Cloud generation, which is a graphical representation of the frequency of words in a text corpus, the presence of “nan” in the cluster suggests that there were some cases where the text preprocessing or transformation steps resulted in missing values, leading to the absence of meaningful words in those cases. And hence, when the vectorization took place, it placed them in a cluster which only contains empty values.

Fifth Cluster



Figure 13: Word Cloud 5

The Word Cloud that corresponds to the fifth cluster consists of various terms with varied sizes, which provides valuable insights into the topics and themes that are prevalent and important in the context of air travel, as expressed by users on Twitter.

The largest words in the Word Cloud are “flight,” “airline,” “will,” “flying,” “fly,” and “day.” These words indicate a prominent and dominant focus on air travel, especially flights and airlines, as the core components of the airline experience. Users often tweeted about their experiences and plans related to flying, expressing their anticipation for upcoming flights, or sharing their past travel stories. These words were also part of the tweets where users intended to inquire about flight schedules, routes, prices, availability.

Moving to the medium-sized words, we find terms such as “thank,” “now,” “service.” These words shed light on additional aspects of the airline experience that users also pay attention to and comment on. Expressions of gratitude, such as “thank,” suggest that users often show appreciation for services provided by airlines.

The terms Etihad and Airbus also appear in the Word Cloud. These terms may indicate users’ experiences on Etihad’s airbus fleet, which consists of various models of Airbus A320, Airbus A330, and the now retired Airbus A340 aircraft. Additionally, the words Dubai and Singapore may relate to popular travel destinations or airports in these cities, which may affect users’ preferences or expectations of flights or airlines.

The small-sized words, such as “now,” “travel,” “airport,” “hour,” and “time,” relate to the broader travel experience beyond flights for the passengers. For instance, “travel” may indicate the reason for flying, like leisure or business. “Airport” may involve the facilities, services, or problems at airports. “Hour” and “time” may concern flight durations, layovers, schedules, or punctuality, which may influence users’ comfort or satisfaction. These little keywords can be correlated with the airlines and airports mentioned beside them to have an idea of the subtopics associated with them.

Sixth Cluster:

Figure 14: Word Cloud 6

This cluster primarily centers around a variety of greetings expressed by users in the context of airline-related tweets, as evident from the Word Cloud displaying prominent greeting words and their variations. The prevalence of the largest words, "hello hello," signifies that social media users frequently employ multiple greetings to convey enthusiasm or seek attention, resulting in the repetitive appearance of these terms in the text corpus.

Within this cluster, the smaller words encompass diverse greeting expressions, including "pretty," "hey," "Marhaba," "elonmusk," "hi," and "hello." Each of these words represents distinct types of greetings employed by users in their tweets. "Pretty" may signify expressions of admiration or emphasis, such as "pretty good" or "pretty impressive," adding nuance to greetings. "Hey" and "hi" serve as common informal salutations used to initiate or respond to conversations, fostering a friendly tone in interactions. The inclusion of "Marhaba," an Arabic greeting used to welcome or acknowledge others, highlights the linguistic diversity of users or the destinations they discuss in their travel experiences.

Interestingly, the appearance of "elonmusk" in this cluster is a misinterpretation by the model, where "elon" is misconstrued as "hello" due to phonetic similarities. This may have arisen due to occasional challenges in natural language processing tasks, resulting in misidentifications in the dataset. Moreover, "elonmusk" might also imply discussions related to the renowned entrepreneur and his ventures in the space industry, demonstrating the diverse themes captured by the clustering process.

This cluster captures essential linguistic features that reflect the greeting-related topics and themes in the text corpus. Nonetheless, the presence of occasional misinterpretations serves as a reminder of the complexities involved in natural language processing tasks and highlights the need for ongoing research and refinement in this domain. Understanding the distinct greetings employed by users and the occasional misinterpretations enriches the analysis of airline-related tweets, providing valuable insights into user behavior and communication patterns on social media platforms.

Seventh Cluster

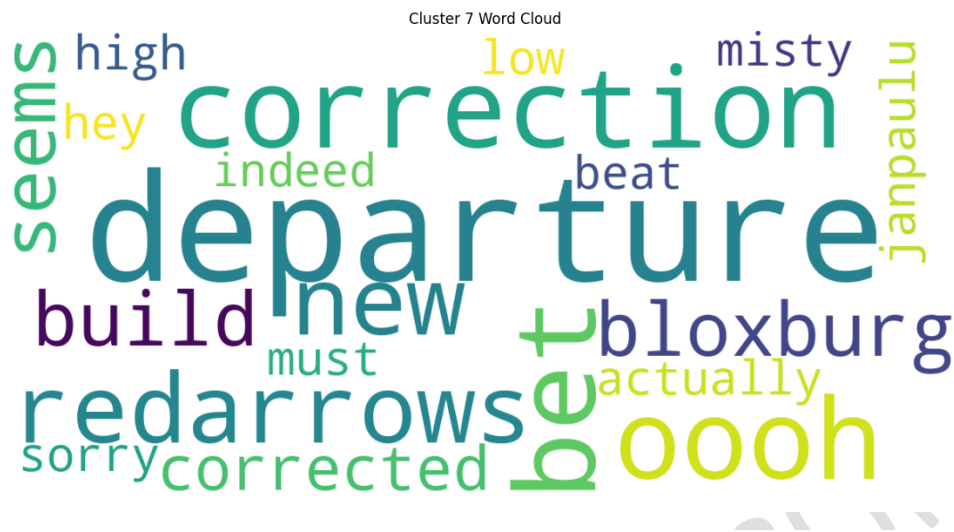


Figure 15: Word Cloud 7

The seventh cluster in the Word Cloud presents a diverse array of terms with varying sizes and orientations. The small words, including "high," "low," "misty," "must," and "actually," appear to be indicative of visibility conditions that affect the airline's decisions on landing, diversions or even to commence a flight. The variations in visibility can have a significant impact on flight operations and safety, necessitating accurate and timely updates to ensure smooth travel experiences for passengers.

The larger words, "departure" and "correction," strongly suggest that this cluster involves information about departure times, with a particular emphasis on updates and corrections related to these timings. Airlines frequently make adjustments to flight schedules due to numerous factors, and this cluster might reflect announcements of such changes, particularly concerning visibility-related issues.

The medium-sized words offer intriguing insights into several topics. "Sorry" suggests the acknowledgment of potential inconveniences or delays faced by passengers, while "corrected" reinforces the notion of updates or improvements to previously announced information.

The vertically oriented words "janpaulu," "bet," "seems," and "beat" offer a distinct pattern that requires further contextual analysis. While the meaning of these words is less evident from the immediate context, they may pertain to specific events or discussions beyond the scope of the topics above.

Moreover, the appearance of "redarrows" in the cluster hints at the presence of the renowned Red Arrows aerobatics team in Qatar before the World Cup. This noteworthy aerial display by the team might have garnered attention and discussion on social media platforms, contributing to its inclusion in this cluster.

Eighth Cluster



Figure 16: Word Cloud 8

The eighth cluster in the Word Cloud highlights the presence of a diverse range of terms and expressions that shed light on the interactions and sentiments within the context of airline-related discussions on social media. The prominence of the word "dear" suggests that users often use this term as a form of informal salutation when addressing airlines or conveying personal emotions related to their travel experiences. This term may encompass a range of sentiments such as gratitude, loyalty, or affection.

The medium-sized words, "ok" and "okay," indicate common responses or acknowledgments frequently employed by users during their interactions with airlines or fellow passengers. These simple expressions typically signify acceptance or agreement with the information presented, such as flight updates, service feedback, or travel tips. The context and tone of the messages can determine whether these expressions reflect satisfaction or dissatisfaction with the airline experience.

Among the smaller words, "diary" suggests that some users utilize social media as a platform to share their travel experiences, documenting their journeys and interactions with airlines. It may serve as a form of personal journaling or storytelling, where users narrate their travel adventures and challenges to their followers or friends.

The term "officer" may be related to interactions with customer service representatives or references to airline personnel responsible for addressing customer inquiries. Users may use this term to recognize the efforts of airline staff who assist them with various issues, such as bookings, check-ins, baggage, or boarding.

Furthermore, the presence of terms like "dearbeautifuls" and "beloved" indicates the use of affectionate language, demonstrating users' admiration or positive sentiments towards airlines or their travel experiences. These terms may reflect a strong attachment or fondness for airlines or flights, or a sense of enjoyment and delight.

Ninth Cluster

Ninth Cluster

Cluster 9 Word Cloud

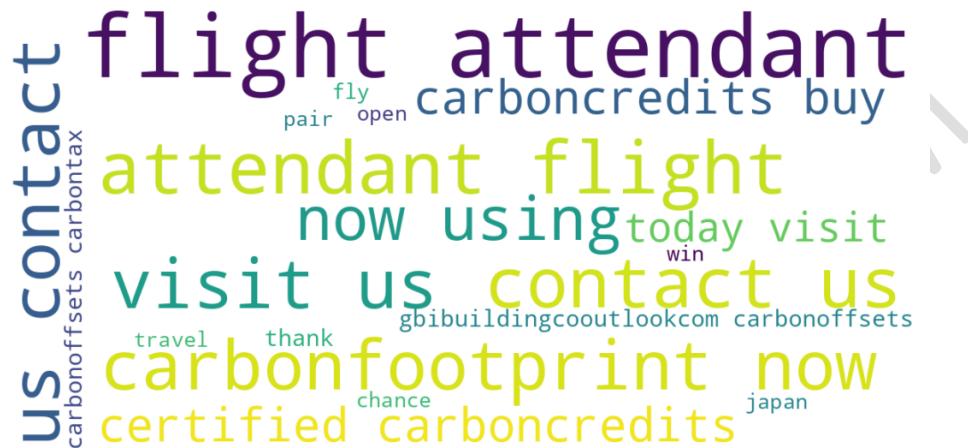


Figure 17: Word Cloud 9

The words that are oriented vertically in the text present phrases such as “contact us,” “carbon offsets,” and “carbon tax.” These words suggest that the text has a focus on topics that are related to environmental awareness and responsibility in the context of airline operations. The phrase “contact us” might imply that the airline is open to communication and interaction with its customers or stakeholders for any inquiries or feedback. Meanwhile, “carbon offsets” and “carbon tax” indicate that the text involves discussions related to environmental sustainability initiatives and policies in the aviation industry. These themes revolve around how airlines are trying to reduce their carbon emissions and implement practices that are more eco-friendly and socially responsible.

The bigger words in this word cloud such as “flight attendant” and “visit us” indicate that the text is relevant to the role and function of the onboard cabin crew and the invitations or offers to visit airline facilities or destinations. The word “now using” suggests that the text is about making announcements or advertising promotions of new services or technologies that the airline has adopted or introduced to enhance the passenger experience and satisfaction.

In the medium-sized words, “carbon footprint” and “certified carbon tax” reinforce the environmentally focused themes prevalent in this cluster. These terms underscore discussions on carbon emissions and carbon tax certifications, highlighting airlines’ commitment to sustainable practices.

The smaller words introduce additional topics that require further exploration for a comprehensive understanding. “Travel” indicates general discussions on the act of traveling or may refer to travel experiences shared by social media users. “Thank” suggests expressions of gratitude or appreciation in the context of interactions with

airlines. “Chance” and “win” might relate to promotional offers or contests run by airlines to engage with their customers.

The ninth cluster of tweets reflects the airline industry's evolving discourse on environmental sustainability, passenger experiences, and promotional activities. The prominence of terms related to carbon offsets, carbon tax, and carbon footprint signifies a growing awareness of environmental responsibility in the aviation sector. Meanwhile, discussions on flight attendants, visiting airline facilities, and new services reveal airlines' efforts to create positive customer experiences and foster passenger engagement. The varied themes captured in this cluster underscore the complexity and dynamic nature of airline-related discussions on social media platforms. Further exploration and analysis of the content associated with these terms would yield valuable insights into the priorities and concerns of both airlines and their customers, contributing to the broader understanding of the aviation industry's endeavors and customer engagement strategies.

Tenth Cluster

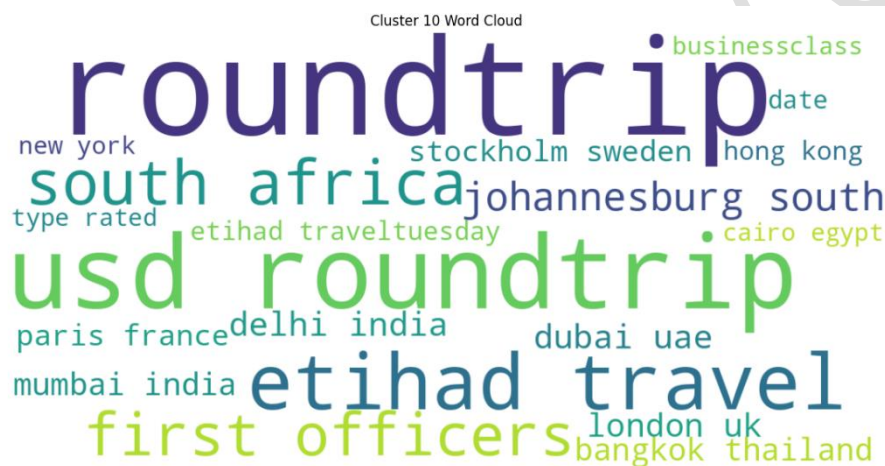


Figure 18: Word Cloud 10

A collection of terms that point to an assortment of destinations frequently served by airlines, contributing significantly to their revenue streams, is displayed in the Word Cloud. These words reflect the global reach and business impact of the aviation industry, as they focus on various aspects related to travel experiences and popular destinations.

The prevalence of round-trip travel arrangements, a common choice for passengers seeking to visit a destination and return to their starting point, is indicated by the term “roundtrip.” “Date” could refer to discussions about travel dates, itinerary planning, or booking processes.

A specific day of the week associated with travel promotions or exceptional deals may be suggested by “TravelTuesday,” potentially indicating airlines’ marketing strategies to attract customers on this day. In addition to the travelers who posted about Tuesday flights using the term as a hashtag.

The significance of the United States Dollar as a standard currency used for pricing flights and travel services, emphasizing the international nature of these routes and transactions, is highlighted by the mention of “USD.”

The geographical diversity of the routes commonly served by airlines is illustrated by the inclusion of specific destinations such as “Stockholm Sweden,” “Hong Kong,” “New York,” “South Africa Johannesburg,” “Cairo Egypt,” “Paris France,” “Delhi,” “Dubai UAE,” “Mumbai India,” and “Bangkok Thailand.” These well-known international destinations attract substantial passenger traffic and contribute to airlines’ revenue generation.

Eleventh Cluster

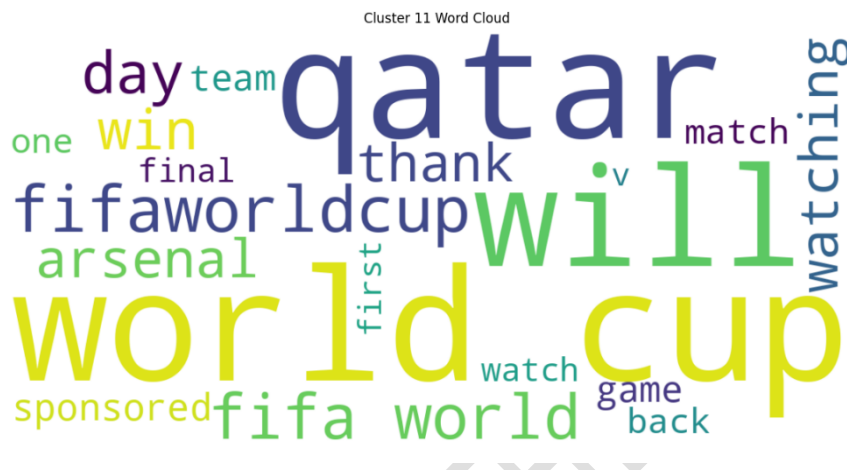


Figure 20: Word Cloud 11

Significant terms related to Qatar and its association with the World Cup are revealed by the final cluster displayed in the Word Cloud. A strong emphasis on Qatar’s role as the host country for the prestigious FIFA World Cup event is indicated by the largest words, “qatar,” “world cup,” and “will.” The World Cup is one of the most significant sporting events globally, and Qatar’s hosting of the tournament has attracted immense attention and interest from sports enthusiasts worldwide.

The focus on the FIFA World Cup is reinforced by the medium-sized words, “fifaworldcup” and “fifa world,” underscoring discussions or content related to the tournament. These terms may be used in hashtags, mentions, or references to the official FIFA World Cup. Further insights into the cluster’s themes are offered by the smaller words, “sponsored,” “arsenal,” “final,” “win,” and “team.” The term “sponsored” suggests the involvement of sponsorships, which are integral to major sporting events like the World Cup. Brands and companies often sponsor teams, events, or specific aspects of the tournament, gaining visibility and promoting their products or services globally.

The inclusion of “arsenal” may indicate discussions related to the renowned English football club, which may have participated in the tournament or had players representing their national teams. This highlights the diverse range of topics related to football and its broader impact on the World Cup.

Discussions or excitement surrounding the tournament’s final match, the pursuit of victory, and the performance and strategies of participating national teams are signified by the words “final,” “win,” and “team.” These terms may encompass analysis, predictions, or emotional reactions to match outcomes and team performances during the tournament.

The vertically oriented term “watching” implies that users may be actively engaged in watching the FIFA World Cup matches, discussing real-time events, or sharing their experiences and emotions during the tournament.

6.4 Conclusion

In addition to emotions, topic modeling is instrumental in identifying airline brand advocates among the tweets. By finding the topics that advocates use to praise and endorse an airline, topic modeling can help understand the characteristics and behaviors of these advocates and their impact on airline reputation and marketing.

The tweets prove to be valuable in making binary decisions, offering insights into positive and negative sentiments towards airlines, and providing indications of liking or disliking certain aspects of the travel experience. Binary decisions are decisions that have two outcomes or alternatives. For example, a binary decision could be whether to fly with a certain airline or not, or whether to recommend an airline to others or not. Tweets can help users make such decisions by providing them with relevant and timely information and opinions from other users.

Passengers' emotions towards airlines can indeed be depicted through tweets. Twitter serves as a platform for travelers to share their feelings, excitement, frustration, and appreciation regarding their flight experiences. Emotions are the specific and intense reactions that users have to certain events or situations. Emotions can be positive or negative, and they can vary in intensity and duration. Tweets can capture the emotions of users towards different airlines and their services, which are influenced by their expectations and perceptions. These emotions are reflected in discussions about customer service, flight delays, flight schedules, and airport experiences.

The main topics discussed in tweets related to airlines encompass various aspects such as flight experiences, travel, airlines' marketing strategies, destinations, and technological advancements onboard flights. These topics capture the diverse and dynamic nature of the airline industry and the customer needs and wants. Users frequently express their opinions, share experiences, seek advice, and explore new travel opportunities through these tweets. These tweets provide rich and authentic information about the users' perspectives, preferences, and feedback on different airlines and their services.

Moreover, by analyzing the patterns and frequencies of specific topics, advocates who consistently promote and positively discuss airlines can be discovered. Topic modeling can help find the most relevant and frequent topics that users talk about when they mention airlines. Among these topics, some may indicate that certain users are advocates for specific airlines. Advocates are users who actively and voluntarily support and recommend an airline to others. Advocates have an important level of trust and loyalty towards an airline, and they can influence other users' opinions and decisions.

Through this research, we aim to gain a comprehensive understanding of the topics discussed in airline-related tweets, passengers' emotions towards airlines, and the identification of airline brand advocates. By linking these aspects together, our study contributes valuable insights into the online discourse surrounding the airline industry and the sentiments of passengers, thereby aiding airlines in enhancing their services and marketing strategies based on customer preferences and feedback.

6.5 Summary

This chapter analyzes the most familiar words in tweets about the top ten airlines. It shows that users are interested in flight experiences, travel aspects, new destinations, and promotions. Airlines are also discussed as marketing tools and sources of news and trends. Users share positive, negative service experiences, and seek help and advice from others. They are sensitive to flight schedules, delays, and airport services. They also talk about aircraft, especially Airbus, and their usage by major airlines.

The chapter uses K-Means clustering with eleven clusters to reveal distinct topics in the dataset. These topics include aviation professionals' careers and qualifications, the FIFA World Cup and airlines' involvement, casual greetings and expressions, data cleaning challenges, air travel discussions and popular destinations, user interactions and sentiments, departure times and visibility conditions, environmental awareness and airline sustainability initiatives, time management during travel, service experiences and feedback, and service quality.

The chapter concludes that users have diverse interests in airline-related topics on social media. K-Means clustering provides a comprehensive view of these topics and their themes, offering valuable insights for airlines and researchers.

Chapter 7: References

- A. Mullen, L., Benoit, K., Keyes, O., Selivanov, D., & Arnold, J. (2018). Fast, Consistent Tokenization of Natural Language Text. *Journal of Open-Source Software*, 3(23), 655. <https://doi.org/10.21105/JOSS.00655>
- Abdelrazek, A., Eid, Y., Gawish, E., Medhat, W., & Hassan, A. (2023). Topic modeling algorithms and applications: A survey. In *Information Systems* (Vol. 112). <https://doi.org/10.1016/j.is.2022.102131>
- Akter, S., Wamba, S. F., & D'Ambra, J. (2019). Enabling a transformative service system by modeling quality dynamics. *International Journal of Production Economics*, 207, 210–226. <https://doi.org/10.1016/J.IJPE.2016.08.025>
- Ali, F., Dey, B. L., & Filieri, R. (2015). An assessment of service quality and resulting customer satisfaction in Pakistan international airlines: Findings from foreigners and overseas Pakistani customers. *International Journal of Quality and Reliability Management*, 32(5), 486–502. <https://doi.org/10.1108/IJQRM-07-2013-0110/FULL/XML>
- Alonso, J. B. (2013). *K-means vs Mini Batch K-means: a comparison*. <https://upcommons.upc.edu/handle/2117/23414>
- An M., & Noh, Y. (2009). Airline customer satisfaction and loyalty: Impact of in-flight service quality. *Service Business*, 3(3). <https://doi.org/10.1007/s11628-009-0068-4>
- Ashour, L. M., Magatef, S. G., & Alzaghal, Q. K. (2023). The Impact of E-Ticketing on Behavioral Intention to Use and the Moderating Role of Information and Communication Technology Tools. *Lecture Notes in Networks and Systems*, 485, 483–507. https://doi.org/10.1007/978-3-031-08093-7_32/COVER
- Belobaba, P., Odoni, A., & Barnhart, C. (2015). *The global airline industry*. <https://books.google.com/books?hl=en&lr=&id=HwBhBgAAQBAJ&oi=fnd&pg=PA99&dq=airline+industry&ots=c2iJqtLog6&sig=VaafZRTpyCYEkzUUgk1AlKeuXZU>

- BERT vs Other Deep Learning Models for Sentiment Analysis*. (n.d.). Retrieved July 6, 2023, from <https://www.linkedin.com/advice/0/how-do-you-compare-contrast-bert-other-deep-learning>
- Boubker, O., & Naoui, K. (2022). Factors affecting airline brand love, passengers' loyalty, and positive word-of-mouth. A case study of Royal Air Maroc. *Case Studies on Transport Policy*, 10(2), 1388–1400. <https://doi.org/10.1016/J.CSTP.2022.05.006>
- Bouwer, J., Saxon, S., & Wittkamp, N. (2021). *Back to the Future?: Airline Sector Poised for Change Post-COVID-19*. https://blog.sodipress.com/wp-content/uploads/2021/08/back-to-the-future-airline-sector-poised-for-change-post-covid-19_vf.pdf
- Bratović, P. (2020). *Marketing u prometu na primjeru kompanije Emirates airline*.
- Budi Setiawan, E., Valdhavessa, D., Bambang, H., Marina, S., Desa Fitri Ratu Bilqis, L., Agusinta, L., Arif Hernawan, M., & Sidjabat, S. (2021). How To Build Customer Loyalty: Through Customer Experience, Perceived Price, And Customer Satisfaction. *Turkish Journal of Computer and Mathematics Education (TURCOMAT)*, 12(4), 1546–1554. <https://doi.org/10.17762/TURCOMAT.V12I4.1410>
- Deerwester, S., Dumais, S., Landauer, T., Furnas, G., & Harshman, R. (1990). Indexing by Latent Semantic Analysis. *J. Am. Soc. Inf. Sci.* [https://doi.org/10.1002/\(SICI\)1097-4571\(199009\)41:6](https://doi.org/10.1002/(SICI)1097-4571(199009)41:6)
- Denning, S. (2021, June 22). *Why Peter Drucker Got Customer Value Right*. <https://www.forbes.com/sites/stevedenning/2021/06/22/why-peter-drucker-got-customer-value-right/?sh=11736ddd4402>
- Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. *NAACL HLT 2019 - 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies - Proceedings of the Conference*, 1.
- do Nascimento, M. V., Borille, G. M. R., Bandeira, M. C. G. S. P., & da Silva, L. M. (2022). An evaluation of non-aviation revenues based on passenger satisfaction and quantity of commercial establishments at terminals: A case study of a Brazilian airport. *Case Studies on Transport Policy*, 10(2), 1457–1461. <https://doi.org/10.1016/J.CSTP.2022.05.011>
- Doganis, R. (2013). *Flying Off Course*. <https://doi.org/10.4324/9780203976197>
- Egger, R., & Yu, J. (2022). A Topic Modeling Comparison Between LDA, NMF, Top2Vec, and BERTopic to Demystify Twitter Posts. *Frontiers in Sociology*, 7. <https://doi.org/10.3389/fsoc.2022.886498>
- Elliott, K. M., & Roach, D. W. (1993). Service Quality in the Airline Industry: Are Carriers Getting an Unbiased Evaluation from Consumers? *Journal of Professional Services Marketing*, 9(2), 71–82. https://doi.org/10.1300/J090V09N02_06

- Gabrilovich, E., & Markovitch, S. (2005). Feature generation for text categorization using world knowledge. *IJCAI International Joint Conference on Artificial Intelligence*.
- George, L. E., & Birla, L. (2019). A Study of Topic Modeling Methods. *Proceedings of the 2nd International Conference on Intelligent Computing and Control Systems, ICI CCS 2018*, 109–113. <https://doi.org/10.1109/ICCONS.2018.8663152>
- Gourdin, K. N. (1988). BRINGING QUALITY BACK TO COMMERCIAL AIR TRAVEL. *Transportation Journal*, 27(3).
- Gualberto, E. S., De Sousa, R. T., De Vieira, T. P. B., Da Costa, J. P. C. L., & Duque, C. G. (2020). From Feature Engineering and Topics Models to Enhanced Prediction Rates in Phishing Detection. *IEEE Access*, 8, 76368–76385. <https://doi.org/10.1109/ACCESS.2020.2989126>
- HaCohen-Kerner, Y., Miller, D., & Yigal, Y. (2020). The influence of preprocessing on text classification using a bag-of-words representation. *PLoS ONE*, 15(5). <https://doi.org/10.1371/JOURNAL.PONE.0232525>
- Hartley, R. F. (1976). Use of customer analysis for better market penetration. *Industrial Marketing Management*, 5(1), 57–62. [https://doi.org/10.1016/0019-8501\(76\)90010-9](https://doi.org/10.1016/0019-8501(76)90010-9)
- Hasib, K. M., Habib, M. A., Towhid, N. A., & Showrov, M. I. H. (2021). A Novel Deep Learning based Sentiment Analysis of Twitter Data for US Airline Service. *2021 International Conference on Information and Communication Technology for Sustainable Development, ICICT4SD 2021 - Proceedings*, 450–455. <https://doi.org/10.1109/ICICT4SD50815.2021.9396879>
- Hennig-Thurau, T., Gwinner, K. P., & Gremler, D. D. (2002). Understanding Relationship Marketing Outcomes: An Integration of Relational Benefits and Relationship Quality. *Journal of Service Research*, 4(3). <https://doi.org/10.1177/1094670502004003006>
- Hofmann, T. (2001). Unsupervised learning by probabilistic Latent Semantic Analysis. *Machine Learning*, 42(1–2), 177–196. <https://doi.org/10.1023/A:1007617005950/METRICS>
- Jain, P. K., Pamula, R., Ansari, S., Sharma, D., & Maddala, L. (2019). Airline recommendation prediction using customer generated feedback data. *2019 4th International Conference on Information Systems and Computer Networks, ISCON 2019*. <https://doi.org/10.1109/ISCON47742.2019.9036251>
- Jamal, Z., & Bucklin, R. E. (2006). Improving the diagnosis and prediction of customer churn: A heterogeneous hazard modeling approach. *Journal of Interactive Marketing*, 20(3–4). <https://doi.org/10.1002/dir.20064>
- Jansen, B. J., Zhang, M., Sobel, K., & Chowdury, A. (2009). Twitter power: Tweets as electronic word of mouth. *Journal of the American Society for Information Science and Technology*, 60(11). <https://doi.org/10.1002/asi.21149>

- Keramatfar, A., & Amirkhani, H. (2019). Bibliometrics of sentiment analysis literature. *Journal of Information Science*, 45(1). <https://doi.org/10.1177/0165551518761013>
- Kherwa, P., & Bansal, P. (2020). Topic Modeling: A Comprehensive Review. *EAI Endorsed Transactions on Scalable Information Systems*, 7(24), 1–16. <https://doi.org/10.4108/eai.13-7-2018.159623>
- Kwon, H. J., Ban, H. J., Jun, J. K., & Kim, H. S. (2021a). Topic modeling and sentiment analysis of online review for airlines. *Information (Switzerland)*, 12(2), 1–14. <https://doi.org/10.3390/INFO12020078>
- Kwon, H. J., Ban, H. J., Jun, J. K., & Kim, H. S. (2021b). Topic modeling and sentiment analysis of online review for airlines. *Information (Switzerland)*, 12(2), 1–14. <https://doi.org/10.3390/info12020078>
- Lau, R. Y. K., Zhang, W., & Xu, W. (2018). Parallel Aspect-Oriented Sentiment Analysis for Sales Forecasting with Big Data. *Production and Operations Management*, 27(10). <https://doi.org/10.1111/poms.12737>
- Law, C. C. H., Zhang, Y., & Gow, J. (2022). Airline service quality, customer satisfaction, and repurchase intention: Laotian air passengers' perspective. *Case Studies on Transport Policy*, 10(2). <https://doi.org/10.1016/j.cstp.2022.02.002>
- Lăzăroiu, G., Neguriță, O., Grecu, I., Grecu, G., & Mitran, P. C. (2020). Consumers' Decision-Making Process on Social Commerce Platforms: Online Trust, Perceived Risk, and Purchase Intentions. In *Frontiers in Psychology* (Vol. 11). <https://doi.org/10.3389/fpsyg.2020.00890>
- Mandal, A., Ghosh, K., Ghosh, S., & Mandal, S. (2021). Unsupervised approaches for measuring textual similarity between legal court case reports. *Artificial Intelligence and Law*, 29(3), 417–451. <https://doi.org/10.1007/s10506-020-09280-2>
- Monmousseau, P., Marzuoli, A., Feron, E., & Delahaye, D. (2020). Impact of Covid-19 on passengers and airlines from passenger measurements: Managing customer satisfaction while putting the US Air Transportation System to sleep. *Transportation Research Interdisciplinary Perspectives*, 7, 100179. <https://doi.org/10.1016/J.TRIP.2020.100179>
- Morrell, P. S. (2021). Airline finance. *Airline Finance*, 1–327. <https://doi.org/10.4324/9781003038191/AIRLINE-FINANCE-PETER-MORRELL>
- Nižetić, S. (2020). Impact of coronavirus (COVID-19) pandemic on air transport mobility, energy, and environment: A case study. *International Journal of Energy Research*, 44(13), 10953–10961. <https://doi.org/10.1002/ER.5706>
- Oliver, R. L., & Bearden, W. O. (1983). The Role of Involvement in Satisfaction Processes. *ACR North American Advances*, NA-10. <https://www.acrwebsite.org/volumes/6098/volumes/v10/NA-10/full>
- Ordenes, F. V., Theodoulidis, B., Burton, J., Gruber, T., & Zaki, M. (2014). Analyzing Customer Experience Feedback Using Text Mining: A Linguistics-Based Approach. *Journal of Service Research*, 17(3). <https://doi.org/10.1177/1094670514524625>

- Pang, B., & Lee, L. (2008). Opinion mining and sentiment analysis. *Foundations and Trends in Information Retrieval*, 2(2), 1–135.
- Parasuraman, A., Zeithaml, V. A., & Berry, L. L. (1985). A Conceptual Model of Service Quality and Its Implications for Future Research. *Journal of Marketing*, 49(4), 41. <https://doi.org/10.2307/1251430>
- Parasuraman, A., Zeithaml, V., & Berry, L. (1988). SERVQUAL: A multiple-item scale for measuring consumer perceptions of service quality. *Journal of Retailing*.
- Park, S. H., Kim, M. Y., Kim, Y. J., & Park, Y. H. (2022). A Deep Learning Approach to Analyze Airline Customer Propensities: The Case of South Korea. *Applied Sciences* 2022, Vol. 12, Page 1916, 12(4), 1916. <https://doi.org/10.3390/APP12041916>
- Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., & Welling, M. (2008). *Fast Collapsed Gibbs Sampling For Latent Dirichlet Allocation*.
- Rahmoune, M., Salem, A. Ben, & Zainine, M. A. (2022). Managing telecommunication customer satisfaction versus perception of service quality. *International Journal of Entrepreneurship and Small Business*, 47(2/3), 239. <https://doi.org/10.1504/IJESB.2022.126714>
- Rana, T. A., Cheah, Y.-N., & Letchmunan, S. (2016). *Topic Modeling in Sentiment Analysis: A Systematic Review*. 10(1), 76–93. <https://doi.org/10.5614/itbj.ict.res.appl.2016.10.1.6>
- Rust, R. T., & Zahorik, A. J. (1993). Customer satisfaction, customer retention, and market share. *Journal of Retailing*, 69(2), 193–215. [https://doi.org/10.1016/0022-4359\(93\)90003-2](https://doi.org/10.1016/0022-4359(93)90003-2)
- Rustam, F., Ashraf, I., Mehmood, A., Ullah, S., & Choi, G. S. (2019). Tweets Classification on the Base of Sentiments for US Airline Companies. *Entropy* 2019, Vol. 21, Page 1078, 21(11), 1078. <https://doi.org/10.3390/E21111078>
- Saleem, M. A., Zahra, S., & Yaseen, A. (2017). Impact of service quality and trust on repurchase intentions – the case of Pakistan airline industry. *Asia Pacific Journal of Marketing and Logistics*, 29(5), 1136–1159. <https://doi.org/10.1108/APJML-10-2016-0192/FULL/XML>
- Sandada, M., & Matibiri, B. (2016). An Investigation into the Impact of Service Quality, Frequent Flier Programs and Safety Perception on Satisfaction and Customer Loyalty in the Airline Industry in Southern Africa. *South East European Journal of Economics and Business*, 11(1). <https://doi.org/10.1515/jeb-2016-0006>
- Şendur, F. (2009). *İlişkisel pazarlama çerçevesinde müşteri değeri yaratmanın önemi: Bankacılık sektöründe bir araştırma*. <http://dspace.balikesir.edu.tr/xmlui/handle/20.500.12462/2013>
- Sezgen, E., Mason, K. J., & Mayer, R. (2019). Voice of airline passenger: A text mining approach to understand customer satisfaction. *Journal of Air Transport Management*, 77, 65–74. <https://doi.org/10.1016/j.jairtraman.2019.04.001>

- Sheth, J. N., & Parvatiyar, A. (1995). Relationship marketing in consumer markets: Antecedents and consequences. *Journal of the Academy of Marketing Science: Official Publication of the Academy of Marketing Science*, 23(4). <https://doi.org/10.1177/009207039502300405>
- Simons, R. (2014). Choosing the Right Customer. *Harvard Business Review*. <https://hbr.org/2014/03/choosing-the-right-customer>
- Song, W., Yang, J., Li, C., & Park, S. (2011). Intelligent information retrieval system using automatic thesaurus construction. *International Journal of General Systems*, 40(4). <https://doi.org/10.1080/03081079.2010.530026>
- Spicer, R. (2018, July 12). *Driven by Competition, The Airline Industry is Taking Off | Aviation Pros*. <https://www.aviationpros.com/airlines/article/12419680/driven-by-competition-the-airline-industry-is-taking-off>
- StrategyR - Global Industry Analysts, Inc. (2022). *Airlines - Global Market Trajectory & Analytics*. https://www.researchandmarkets.com/reports/338343/airlines_global_market_trajectory_and_analytics?utm_source=CI&utm_medium=PressRelease&utm_code=4shc44&utm_campaign=1719227+-+Global+Airlines+Market+Report+2022%3a+Market+to+Reach+%24744+Billion+by+2026+-+Emerging+New+Technologies+for+Urban+Mobility+to+Step-Up+the+Excitement+in+the+Market%2c+Post+Pandemic&utm_exec=chdo54prd
- Stylos, N., Zwiagelaar, J., & Buhalis, D. (2021). Big data empowered agility for dynamic, volatile, and time-sensitive service industries: the case of tourism sector. *International Journal of Contemporary Hospitality Management*, 33(3), 1015–1036. <https://doi.org/10.1108/IJCHM-07-2020-0644/FULL/XML>
- Vanhamme, J. (2016). La satisfaction des consommateurs spécifique à une transaction: définition, antécédents, mesures et modes. *Http://Dx.Doi.Org/10.1177/076737010201700204*, 17(2), 55–85. <https://doi.org/10.1177/076737010201700204>
- Vieira, J., Câmara, G., Silva, F., & Santos, C. (2019). Airline choice and tourism growth in the Azores. *Journal of Air Transport Management*, 77, 1–6. <https://doi.org/10.1016/J.JAIRTRAMAN.2019.02.009>
- Vinodhini, G., Journal, R. C.-I., & 2012, undefined. (2012). Sentiment analysis and opinion mining: a survey. *Researchgate.Net*, 2(6), 2277. https://www.researchgate.net/profile/Vinodhini-G-2/publication/265163299_Sentiment_Analysis_and_Opinion_Mining_A_Survey/links/54018f330cf2bba34c1af133/Sentiment-Analysis-and-Opinion-Mining-A-Survey.pdf
- Yakut, I., Turkoglu, T., & Yakut, F. (2015). Understanding Customers' Evaluations Through Mining Airline Reviews. *International Journal of Data Mining & Knowledge Management Process*, 5(6), 01–11. <https://doi.org/10.5121/ijdkp.2015.5601>

- Yang, Z., Lai, Y. X., Duan, L. J., & Li, Y. J. (2012). Short text sentiment classification based on context reconstruction. *Zidonghua Xuebao/Acta Automatica Sinica*, 38(1). <https://doi.org/10.3724/SP.J.1004.2012.00055>
- Zhang, J., Chen, C., Liu, P., He, C., & Leung, C. W. K. (2020). Target-guided structured attention network for target-dependent sentiment analysis. *Transactions of the Association for Computational Linguistics*, 8. https://doi.org/10.1162/tac1_a_00308
- Zubair, M., Iqbal, M. A., Shil, A., Chowdhury, M. J. M., Moni, M. A., & Sarker, I. H. (2022). An Improved K-means Clustering Algorithm Towards an Efficient Data-Driven Modeling. *Annals of Data Science*, 1–20. <https://doi.org/10.1007/S40745-022-00428-2/FIGURES/9>