MULTIMEDIA UNIVERSITY

FACULTY OF COMPUTING AND INFORMATICS

BACHELOR IN COMPUTER SCIENCE (HONS)

SOCIAL MEDIA COMPUTING - CDS5344

TRIMESTER III, SESSION 2024/2025

US Airline Sentiment Analysis

By:

Ooi Li Yoong (1211306826)

Heng Peng Yong (1211306514)

Acknowledgment

We are immensely thankful to our instructor, Dr. Mohammad Shadab Khan, whose insightful guidance has been pivotal throughout this project's journey.

Table of Content

Chapter 1: Project Overview

1.1 Background

Twitter, as one of the popular social media platforms nowadays, allows its users to post their opinions regarding almost any topic or idea. The large amount of posted data enables researchers to extract and analyze it through various methods, such as sentiment analysis [1]. Sentiment analysis revolves around identifying users' sentiments. Therefore, it is helpful to understand customers' satisfaction levels with services provided by business brands. Consequently, the business side can continuously improve its services [2].

In this study, we are focusing on sentiment analysis for the US airline industry using related tweets. The airline industry has grown at a breakneck pace in the last two decades due to its importance in providing an efficient approach to traveling worldwide. However, besides its primary functionality, customers nowadays emphasize comfortability when rating the flying experience. These expectations can be met by optimizing different services or requirements requested by the customers. Consequently, sentiment analysis is beneficial for understanding such requirements.

Over the years, researchers have attempted different approaches to perform sentiment analysis. Specifically, sentiment analysis can be performed at three levels: document-level, sentence-level, and aspect-level [3].

However, most of the sentiment analysis did not extend to the aspect-level due to its relatively higher difficulty. As a result, their findings only reveal the overall sentiment expressed by the sample and the sentiment of each comment related to the brand, while the specific topics of the brand discussed in the comments are undiscovered. Consequently, business providers find it challenging to improve their services effectively, as precise business strategies cannot be formulated easily.

Moreover, due to the large sample size, manual labeling of sentiment for each sentence is impractical. In conjunction with that, most studies leveraged machine learning algorithms such as SentimentIntensityAnalyzer to automate this process [4]. Yet, the reliability of the automated labeling should be questioned and scrutinized. Most studies overlooked this part and proceeded to train one or more classifiers based

on this labeled data so that it can be used to predict the sentiment of a new sentence in the future. As a result, the classifier may not perform well because the training data itself might be problematic.

1.2 Problem Statement and Project Objective

In this study, we are utilizing a dataset describing the sentiments posted on Twitter towards the US airline industry during February 2015. The dataset has been pre-processed and contains sentence-level sentiments in which the generation process is unknown. Unfortunately, [1] shows that the reliability of these sentence-level sentiments is worrying.

In addition, the given dataset also provides topics mined from negative sentiments. However, this column is incomplete, in which some of the topics are labeled as "Can't Identify". As a result, some complaints might be ignored.

Therefore, this study aims to achieve three objectives, as shown in the following:

1) To create an optimized set of sentence-level sentiments
2) To mine a set of unique topics discussed in the negative comments
3) To identify an optimal classifier for predicting sentiments of unseen tweets related to the airline industry

1.3 Chapter Outline

The rest of this study is organized as the following: Chapter 2 presents a literature review of related studies. Chapter 3 describes the methods used to achieve the objectives. Chapter 4 explains the findings derived. Chapter 5 summarizes this study's achievements and future directions.

Chapter 2: Literature Review

2.1 Introduction

This chapter provides a thorough literature review encompassing methods for sentiment analysis of different levels and supervised learning of sentiment classification. In addition, we will mainly focus on studies leveraging the dataset of Twitter comments for US airline service.

2.2 Sentiment Analysis

2.2.1 Document-level

Document-level sentiment analysis reveals the overall sentiment expressed in one whole document. If there is only one dataset involved, then it implies the overall sentiment that can be extracted from its data. [5] presents the count for each unique sentiment available in the dataset, and the largest group is “Negative”. In other words, the central tendency of sentiments towards the US airline indusrty is negative. Most studies including but not limited to [5, 6] uses mode as the central tendency measure.

2.2.2 Sentence-level

Sentence-level sentiment analysis focuses on determining the sentiment expressed in an individual sentence rather than an entire document [3]. Sentence-level sentiments are available in the US airline dataset, but the generation process is unknown. [1] points out that the reliability of these sentiments should be questioned, and further displays a few examples of falsely labeled sentiments. For instance, one tweet with its context as “flying @virginamerica” is given as negative sentiment. In this case, this sentiment should be neutral instead of negative. Therefore, [1] deploys the TextBlob method to generate a new set of sentiments. Later, they showed a few examples of the fixed sentiment, which seems to be logical from a human perspective. On the other hand, most studies including but not limited to [5, 6] proceed with their sentence-level sentiment analysis with the given sentiments.

2.2.3 Aspect-level

Aspect-level sentiment analysis emphasizes a more granular dimension to identify the specific aspects or topics discussed in each sentence [7]. [2] utilizes the Latent Dirichlet Allocation (LDA) algorithm to identify topics related to negative tweets about the US airline service. As a result, some main topics identified are "scheduling", "crew service", "meal", and others. This can be beneficial for airline companies to understand the specific areas of service to improve.

2.2.4 Classification

Most classifiers only accept numerical input. Therefore, textual data must be transformed into their numerical format. This process is known as vectorization [13]. For this approach, some approaches are Bag of Words (BoW), Term Frequency - Inverse Document Frequency (TF-IDF) and Word2Vec. Table 2.2.4.1 summarizes the approaches used for vectorization.

Reference	BoW	TF-IDF	Word2Vec
[8]	✓	✓	
[10]			✓
[12]		✓	
[11]	✓	✓	
[13]		✓	✓
[15]	✓	✓	

Table 2.2.4.1: Vectorization Approaches

The authors in [11] create a simple BoW model for document classification and text modeling. They also incorporate this model into their approach for later use in their proposed classifier model to analyze tweets. The authors in [15] use feature extraction technique, which is BoW used for text modeling, particularly in machine learning algorithms for tweet sentiment classification. The process involves creating a vocabulary from the tweets, tracking the frequency of each word, and then passing this numerical matrix to the model for training.

The authors in [13] use the TF-IDF method to assign weight to the word based on its appearances. Authors in [14] use TF-IDF as text feature extraction

method to transform words to vector of 1 row and same number column of the words that need to be transformed.

The author in [13] uses Word2Vec which is a type of word embedding method for text classification tasks. The dataset used was about COVID-19 article in Indonesia [13]. [14] utilized the Word2Vec method which is composed of two pieces of algorithm. The two algorithms were named Continuous BoW (CBoW) and Skip-Gram. The dataset that authors in [10] have used is the Arabic tweets which are related to terrorism activities. They use skip-gram of Word2Vec method which is a neural network implementation to generate their proposed model [10].

After completing the vectorization process, the numerical data now can be inputted into classifiers. Table 2.2.4.2 shows some classification algorithms used by the authors including Decision Tree (DT), Logistic Regression (LR) and Random Forest (RF).

Reference	DT	LR	RF
[9]			✓
[10]			✓
[11]		✓	
[13]		✓	
[16]	✓		✓
[17]	✓	✓	✓
[20]			✓
[21]	✓	✓	✓
[22]	✓		✓
[23]		✓	

Table 2.2.4.2: Classification Algorithm Used by Researchers

In [15], the authors use DT which is a flexible algorithm used to assign labels based on the highest score to classify the tweets. The study [21] uses various machine learning algorithms like DT to classify and compare sentiment data from Twitter posts about Facebook Marketplace, Instagram Shop, and TikTok Shop. The DT algorithm performed best, showing TikTok had the highest positive sentiment (93.07%) compared to Facebook and Instagram [21]. DT algorithm is applied on

COVID-19 tweets to be classified into positive, neutral, and negative sentiments [22].

The authors in [17] use the extension of LR which is SoftMax regression for multiple classes also known as multinomial logistic regression. The LR algorithm was applied to predict whether the sentiment in each tweet social commerce platforms was positive or negative [21]. In [23], the LR model achieved an accuracy of 74% when applied to shorter tweets, but the accuracy decreased with longer tweets, indicating that it is more effective in sentiment classification when dealing with concise text.

The authors in [15] also use RF which is a supervised algorithm to classify the tweets. RF also has been used for classifying high negative and high positive tweets [17]. The researchers in [10] use a dataset of Arabic tweets and generated feature vectors using Word2vec and Word2vec by weighted average methods. The RF algorithm is employed to predict the sentiment of each tweet based on these feature vectors [10].

Chapter 3: Methodology

3.1 Introduction

This chapter explains the methods employed in this study. Figure 3.1.1 shows our workflow. Firstly, we perform an exploratory data analysis (EDA) on our dataset to understand its properties. Then, we carry out necessary data pre-processing to prepare it for the upcoming sentiment analysis of three levels: document-level, sentence-level, and aspect-level. Furthermore, we attempt some classification algorithms to predict the sentiments of unseen tweets related to US airline service. We evaluate the constructed classifiers with suitable metrics and eventually decide the best-performing classifier. It is notable that the random states are always set to zero whenever applicable to ensure that the results are reproducible.

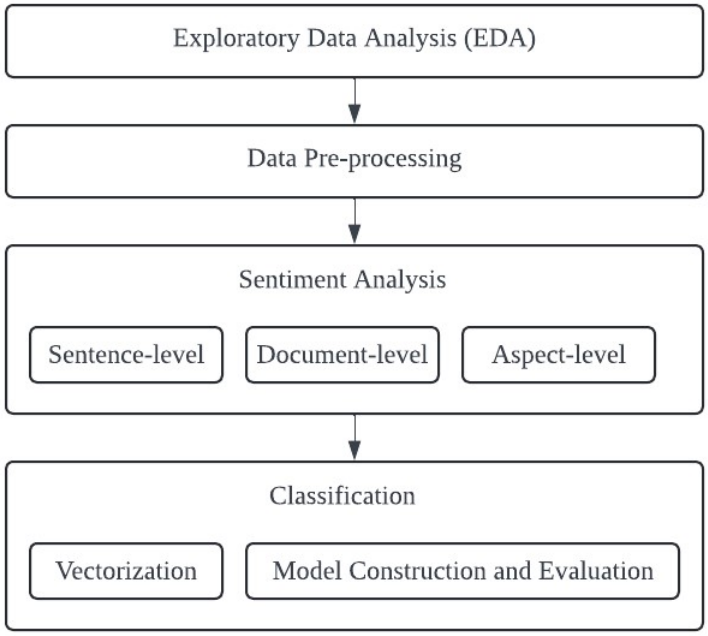

Figure 3.1.1: Workflow

3.2 Exploratory Data Analysis (EDA)

EDA is crucial to understand more about the dataset. At this stage, we check for any missing value and duplicated row. Furthermore, we visualize the most frequently used words using a Word Cloud and the distribution of the given sentence-level sentiments using a bar chart.

3.3 Data Pre-processing

The raw dataset is transformed into a more readily used format through different tasks such as dropping unused columns, removing missing value and duplicated row if any.

In addition, since previous studies show that the given sentence-level sentiments' reliabilities can be improved, we attempt to do so. We transform the tweets into their suitable forms for automated sentiment identification through a series of tasks like removing non-letters, conversion to small letters, removing stopwords, and lemmatzation. Specifically, removing stopwords implies eliminating frequently occuring words that often do not carry significant meaning such as "is", "in", "are", and many more. On the other hand, lemmatization reduces each word to their base form while ensuring the resulting word is valid. For instance, the word "flying" is converted to "fly" after lemmatization.

3.4 Sentence-level Sentiment Analysis

This study employs four algorithms for the identification of sentence-level sentiment of each tweet. The algorithms are SentimentIntensityAnalyzer, BERT, roBERTa, and Multilingual distilBERT. Their performances are evaluated and compared with their respective statistics of sentiment confidence. For each sentiment predicted, a corresponding confidence value will be provided by the algorithm. In general, a higher confidence value implies that the prediction might be more accurate. Furthermore, the confidence values of the given sentiment are available as well, hence involved in the comparison. Eventually, we decide the optimal set of sentiments.

3.5 Document-level Sentiment Analysis

Using the optimal sentiments, the distribution of all sentiments from the dataset is again visualized using a bar chart. From there, the major sentiment can be identified. In addition, a word cloud is created with the pre-processed texts. This word cloud is expected to show more information than the initial word cloud as unnecessary stopwords are removed and all words are reduced to their root forms.

3.6 Aspect-level Sentiment Analysis

For each optimal sentiment, it is given a sentiment score according to its polarity. If it is negative, its sentiment score is -1. The sentiment score is 0 if the tweet expresses neutral sentiment. Else, the sentiment score is 1. Then, the average sentiment score for each airline is computed. Consequently, the sentiment towards each airline can be compared, and airlines that are more problematic can be identified.

Besides that, this study leverages the LDA algorithm to mine topics frequently discussed. Before it, the dataset must be converted into LDA acceptable format. For such a purpose, we attempt the BoW algorithm of varying n-gram like unigram, bigram, and trigram. In addition, the optimal number of topics must be identified. To do so, we run LDA with different experimental settings and observe the respective coherence scores. Number of topics ranging from [3, 14] are tried. Eventually, the optimal model is identified, and its topic mining results are analyzed for extracting insights.

3.7 Classification

A copy of the pre-processed dataset is split into 80% training data and 20% testing data. Then, the complete pre-processed dataset, training and testing dataset are saved respectively.

Classifiers cannot understand textual data. Therefore, vectorization is necessary. We attempted three algorithms: BoW, TF-IDF and Word2Vec. Specifically, for BoW and TF-IDF, a specialized model is first instantiated and fitted into the complete pre-processed dataset for it to understand the data structure. Then, this trained vectorizer is utilized to transform the training and testing data respectively.

The concept behind BoW is to indicate the presence of each word in the vocabulary for every tweet, transforming it into a vector representation. Using 1s and 0s to mark the occurrence of each word [8].

TF-IDF measures the importance of the word then extracts usual information from the text data. This method allocates a word with a special value by following the frequency in which it occurred in a document. Then, the TF-IDF transforms words into

weight values and removes most common words during processing. As a result, after applying TF-IDF to the dataset, a review is converted into a vector of term weights [12].

Word2Vec is a word embedding method that makes words with similar meanings cluster together, and these clusters are arranged in a way that allows certain word relationships to be represented using mathematical vectors [10].

After the vectorization, we attempted three classification algorithms namely DT, LR and RF. Table 3.7.1 shows a summary of different experimental setups attempted.

Model Code	Vectorization	Classification Algorithm
M_1	BoW	DT
M_2	BoW	LR
M_3	BoW	RF
M_4	TF-IDF	DT
M_5	TF-IDF	LR
M_6	TF-IDF	RF
M_7	Word2Vec	DT
M_8	Word2Vec	LR
M_9	Word2Vec	RF

Table 3.7.1: Experimental Setups Attempted

Subsequently, for the evaluation, we use metrics such as accuracy, precision, recall and F_1-score. These metrics also incorporate measures of True Positives (TP), False Positive (FP), False Negative (FN), and True Negatives (TN) to assess the classifier's performance. TP means that when a positive outcome is predicted, and the actual outcome is indeed positive. FP happens when a positive outcome is predicted, but the actual outcome is negative, which is also called Type 1 Error. FN is when a negative outcome is predicted, but the actual outcome is positive, which is known as a Type 2 Error. TN occurs when a negative outcome is predicted, and the actual outcome matches that prediction [19].

Accuracy measures the proportion of correct predictions made by classifier model. However, researchers often avoid using accuracy as a performance metric in

classification tasks involving class imbalance, as it may not effectively highlight the importance of identifying rare cases [18]. Equation 1. shows the formula of accuracy.

1. $Accuracy = \frac{TP+TN}{TP+TN+FP+FN}$

Precision measures the percentage of predicted positive cases that are truly positive, indicating fewer false positives. It can help to control false positives, but it is not sufficient for evaluating models in class imbalance situations, as it is sensitive to the number of false positives [18]. Equation 2. shows the formula of precision.

2. $\Pr e\,cision = \frac{TP}{TP+}$

Recall measures the number of actual positive cases identified by a model. In [18], it has been mentioned that it is not influenced by class imbalance since it focuses only on the positive group. Recall is vital for determining true positive rates and detecting actual positives. Equation 3. shows the formula of recall.

3. $Recall = \frac{TP}{TP+FN}$

F_1-score is an evaluation metric for class imbalance problems, representing the harmonic mean of precision and recall, then combining both measures [18]. Equation 4. shows the formula of F_1-score.

4. $F_1 - score = \frac{TP}{TP+\frac{1}{2}(FP+FN)}$

Finally, we select an optimal classifier that can be used to predict the sentence-level sentiments of new tweets in future.

Chapter 4: Findings

4.1 Introduction

This chapter illustrates findings extracted during the EDA, sentiment analysis and comparison between classifiers.

4.2 Exploratory Data Analysis (EDA)

Upon initial checking, we found that the raw dataset contained fifteen features. Table 4.2.1 describes these features.

Feature	Description
tweet_id	Index key for referencing each unique tweet
airline_sentiment	Describes the sentiment of the tweet. The sentiment can be “positive”, “neutral”, or “negative”
airline_sentiment_confidence	Describes the confidence level of the sentiment
negativereason	Describes one topic discussed from each negative tweet. The topic can be “Customer Service Issue”, “Late Flight”, Can’t Tell”, “Cancelled Flight”, “Lost Luggage”, “Bad Flight”, “Flight Booking Problems”, “Flight Attendance Complaints”, “longlines”, or “Damaged Luggage”
negativereason_confidence	Describes the confidence level of the given topic mined from each negative tweet
airline	Describes the airline company mentioned in the tweet
airline_sentiment_gold	The purpose of this feature is unclear
name	Describes the username
negativereason_gold	The purpose of this feature is unclear
retweet_count	Describes the retweet count for each tweet
text	Describes the tweet
tweet_coord	Describes the coordinate of the user when posting a tweet
tweet_created	Describes the time of a tweet creation
tweet_location	Describes the user-specified location when posting a tweet
user_timezone	Describes the time zone followed by the user

Table 4.2.1: Description of Features in the Raw Dataset

The “tweet_id” was utilized when checking for any duplicated row. As a result, 155 duplicated rows were discovered, and each pair presented the same tweets. Therefore, we considered such duplications to be faulty and removed them. Then, the index key was removed as it was no longer useful.

Apart from that, some features were less crucial for our further analysis thus removed, including “airline_sentiment_gold”, “name”, “negative_reason_gold”, “tweet_coord”, “tweet_created”, “tweet_location”, and “user_timezone”.
Subsequently, the remaining features except “negativereason” and “negativereason_confidence” contained no missing value. For these two specific features, their missing values were found to be irrelevant because those rows did not represent negative sentiments. In other words, all remaining features were completed.
Besides that, some features were renamed into more meaningful representations. For instance, “text” was renamed to “tweet”.

Furthermore, a word cloud was generated to visualize the frequently occurring words in the unprocessed tweets. From this word cloud as shown in Figure 4.2.1, we observed that the words with higher occurrence were airline brands such as “American Air”, “SouthWest Air”, and others. Other interesting words included but not limited to “bag”, “time”, and cancelled”. These words suggested some topics mentioned in the tweets might be “luggage handling”, “scheduling arrangement”, etc. However, this study observed that a significant number of words that were less meaningful appeared as well, including but not limited to “will” and “really”. These words were considered to be stopwords and could be removed in further analysis.

Figure 4.2.1: Raw Word Cloud

Finally, a bar chart was used to visualize the distribution of the given sentiments. As shown in Figure 4.2.2, most of the tweets were expressing negativity after an untold analysis. This delivered a message such that the overall document-level sentiment is negative. In other words, most customers were dissatisfied with the US airline industry.

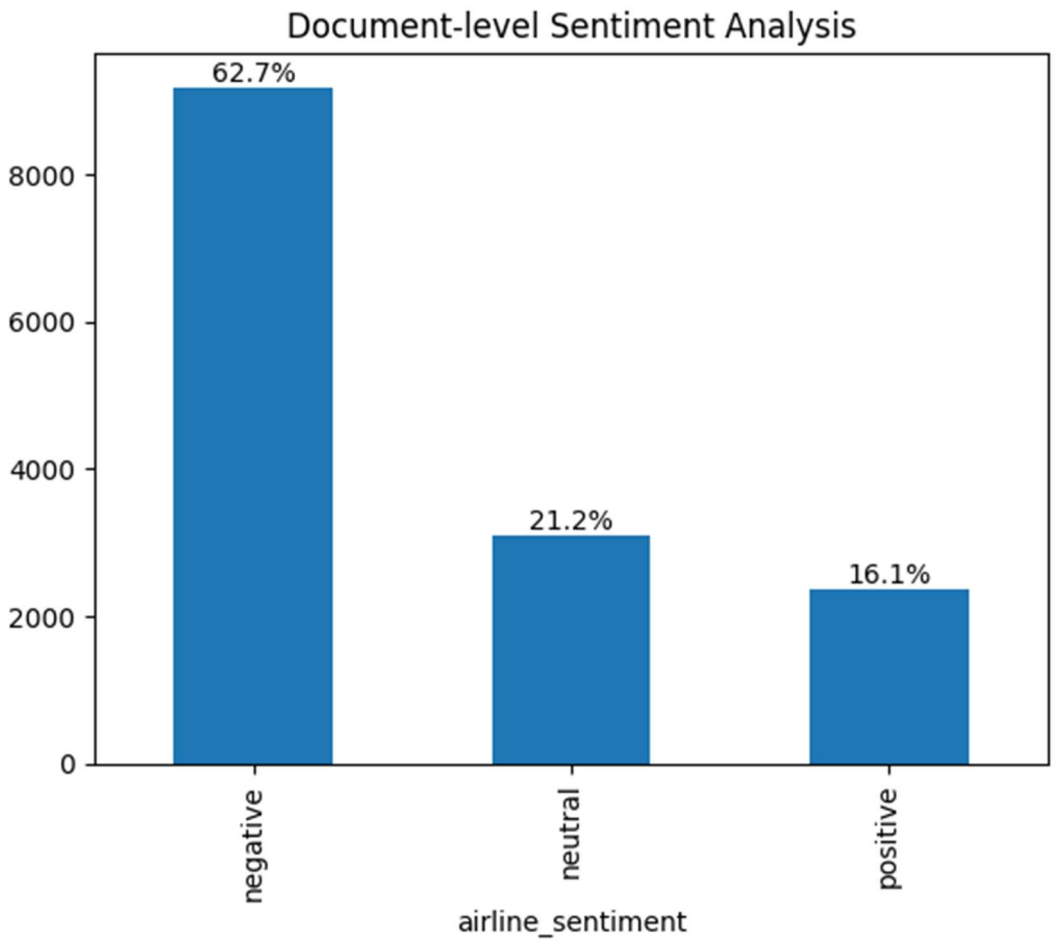

Figure 4.2.2: Distribution of Given Sentiments

However, since previous study pointed out that the given sentiments lacked reliability, this part of findings would be considered only to a certain extent. Examples of faulty given sentiments are provided in Table 4.2.2.

Index	Tweet	Given Sentiment
82	@VirginAmerica you're the best!! Whenever I (begrudgingly) use any other airline I'm delayed and Late Flight :(	negative
86	@VirginAmerica Can't bring up my reservation online using Flight Booking Problems code	neutral
1235	@United Thank you for your response!	neutral
6017	@SouthwestAir Hi Guys good morning how are you doing	positive

Table 4.2.2: Examples of Faulty Given Sentiments

4.3 Sentence-level Sentiment Analysis

Initially, this study attempted four algorithms for the identification of the sentence-level sentiments. They are SentimentIntensityAnalyzer, BERT, roBERTa, and distilBERT. Their confidence scores are compared with the given set. However, this study later discovered that none of the four methods outperformed the given set. Among them, BERT's performance came close to the given set and roBERTa's performance was acceptable.

With respect to the first objective to identify an optimal set of sentence-level sentiments that outperforms the given set, we have to find alternatives. Therefore, this study leveraged the ensemble framework along with suitable weightages. The weightages were decided as follows: {Given: 0.5, BERT: 0.3, roBERTa: 0.2}. Table 4.3.1 summarizes the performances of all methods.

Method	Statistics of Sentiment Confidence Score	
	Mean	Standard Deviation
Given	0.90	0.16
SentimentIntensityAnalyzer	0.48	0.29
BERT	0.87	0.14
RoBERTa	0.78	0.14
DistilBERT	0.58	0.15
Ensemble	**0.96**	**0.09**

Table 4.3.1: Performances of Different Sentence-level Sentiment Analysis Method

The ensemble method achieved the best performance. Upon checking, 1317 sentiments had been changed. Table 4.3.2 shows a few examples of the changed sentiments. It is notable that the text had been pre-processed to a more compact format. The examples showed that the sentiments had been corrected to a more logical answer. The high mean confidence score and low standard deviation of confidence implied that the model was more certain about its classification.

Tweet	Given Sentiment	Ensemble Sentiment
virginamerica didnt today must mean need take another trip	neutral	negative
virginamerica yes nearly every time fly vx ear worm wont go away	positive	negative
Americanair cant ahold advantage reservation need ticket reservation cancelled flight soon help	neutral	negative

Table 4.3.2: Comparison between the Given and Ensemble Sentiments

4.4 Document-level Sentiment Analysis

Figure 4.4.1 presents the word cloud for the pre-processed tweets. It displays the most frequently occuring words in the whole document. Like the raw word cloud, we

observed the most frequently occuring words were airline brands such as “americanair”, “united”, and others. However, this word cloud delivered more insights regarding possible topics from the tweets. Besides “luggage handling” and “scheduling arrangement”, other topics might include “seat comfortability”, “ticket booking system”, and many more.

Figure 4.4.1: Pre-processed Word Cloud

Figure 4.4.2 showcases the distribution of ensemble sentiments. Compared to the distribution of given sentiments, the ensemble sentiments contained relatively less negative category. Such a decrease was almost 5% and they were mostly converted into the neutral category. The proportion of positive sentiments remained almost unchanged. Overall, the mode is still negative sentiment. In other words, most customers were dissatisfied with the US airline service at that time.

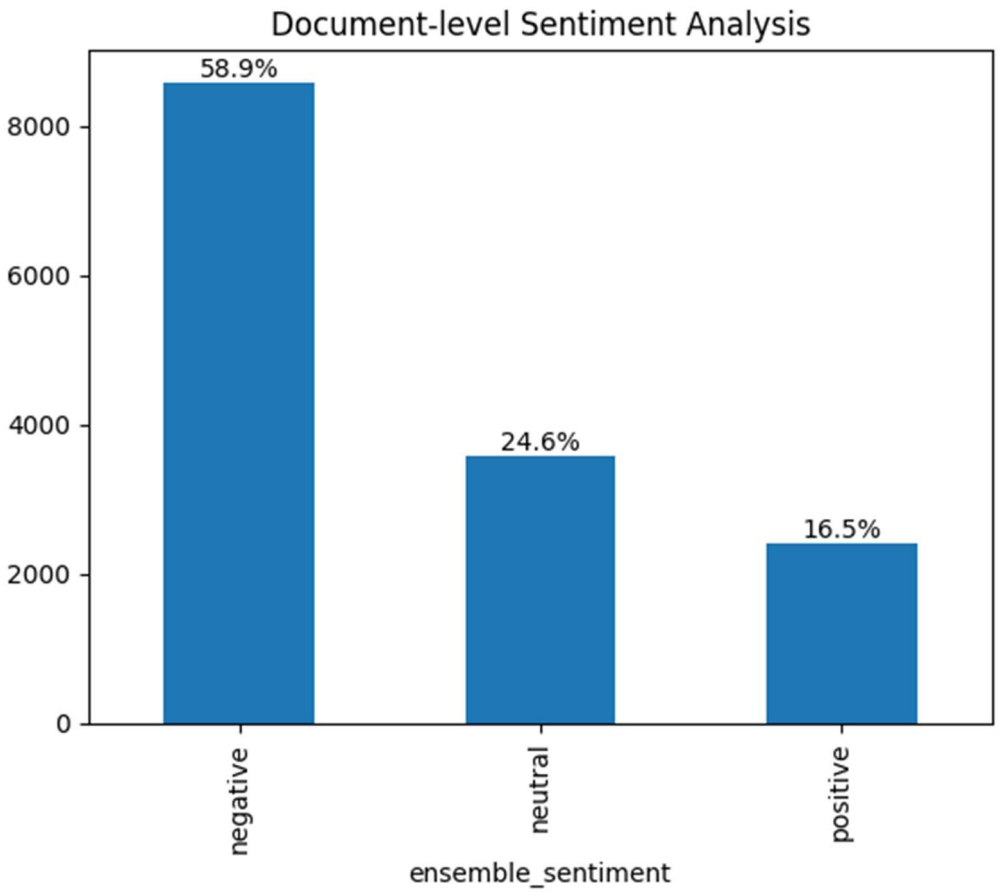

Figure 4.4.2: Distribution of Ensemble Sentiments

4.5 Aspect-level Sentiment Analysis

Considering airline brands as one of the possible topics, we attempted to identify satisfaction level associated with each brand. For each optimal sentiment, it is given a sentiment score according to its polarity. If it is negative, its sentiment score is -1. On the other hand, the sentiment score is 0 if the tweet expresses neutral sentiment. Else, the sentiment score is 1 for the positive tweet. Then, the average sentiment score for each airline is computed. The results were shown in Figure 4.5.1. From that, we understood that US Airways received the most complaints while Virgin America received the least. However, it was notable that all airlines achieved negative average sentiment scores. This was coherent with the overall document-level sentiment being negative. Most customers were hoping for the airline services to be improved.

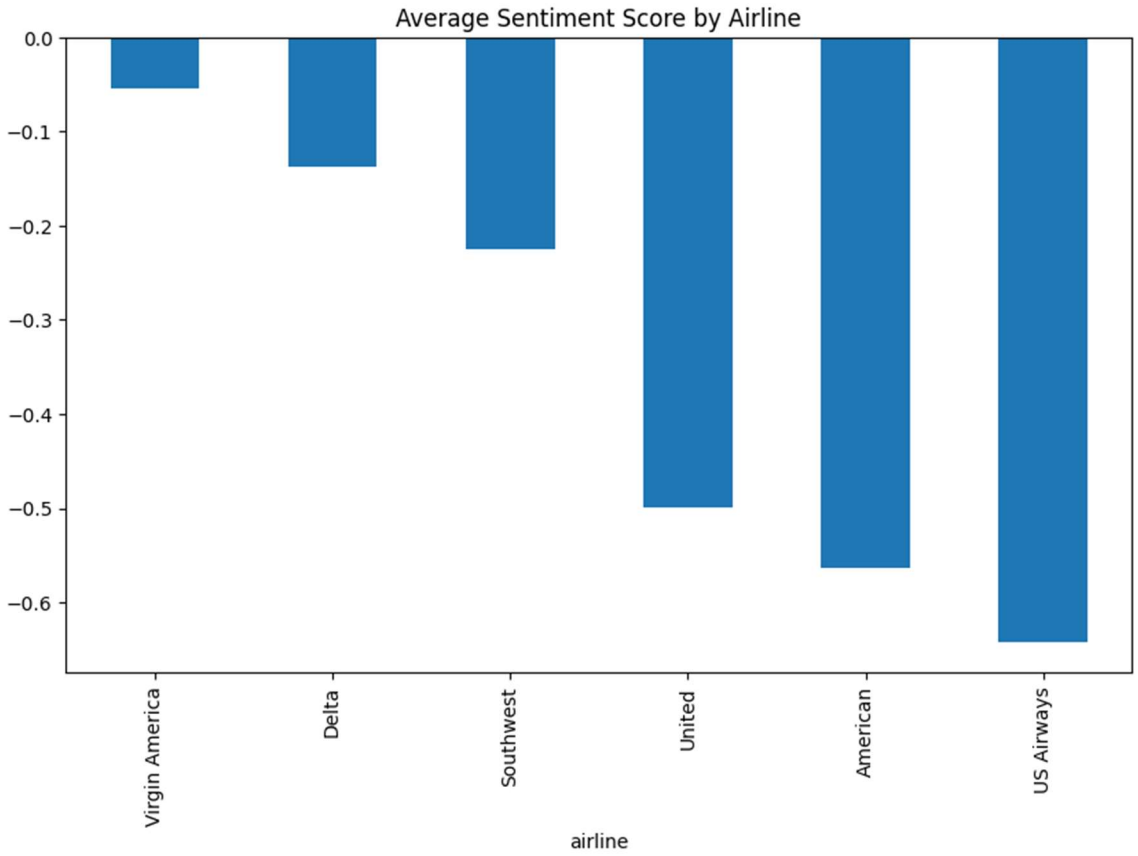

Figure 4.5.1: Average Sentiment Score by Airline Brands

For mining topics from negative tweets, we attempted the LDA algorithm. To identify the optimal number of topics for LDA, we tried values ranging in [3, 14] and checked their coherence score. Figure 4.5.2 visualizes the coherence score of each LDA model with its unique experimental settings. Among them, this study concluded that it was optimal to use a combination of trigram BoW and the LDA algorithm with 11 topics. The increase in coherence score from LDA models utilizing a trigram BoW technique in prior and 11 topics to 13 topics was minimal. Therefore, the slight increase was no longer worth the additional computational resources required. Since the tweets were complex, a trigram BoW was necessary to optimally cover the contexts.

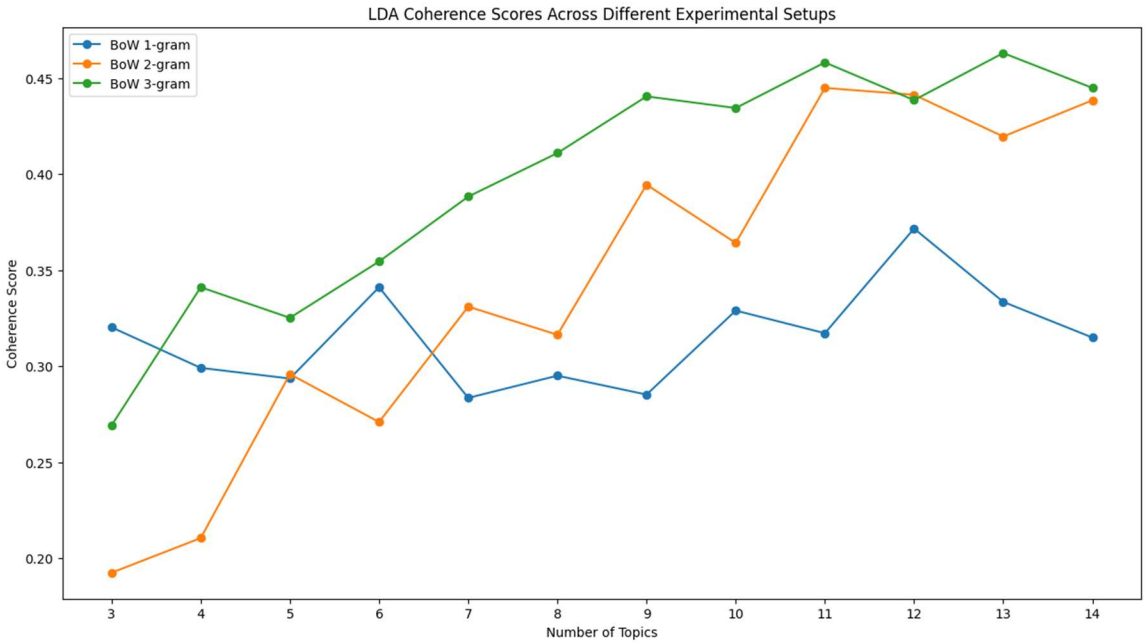

Figure 4.5.2: LDA Coherence Scores Across Different Experimental Settings

To identify topics from the analysis results of LDA, we extract five of the most frequently occurring words from each topic. We then named the topic according to the keywords extracted. Therefore, each topic can be viewed as a cluster and the keywords were cluster members. Ideally, keywords within the same cluster should show higher similarity than those from the other clusters.

However, our clustering solution was complicated, where some of the keywords from different clusters essentially contributed to a common topic. Therefore, some clusters were merged, and a suitable topic name was manually given based on the combination of keywords.

In addition, some clusters were ignored due to failure of identifying a suitable topic. For instance, one of the clusters contained the keywords as {“united”, “usairways”, “tommorow”, “jetblue”, “dfw”}. None of the keywords provided a clue for an appropriate naming of this specific topic.

Table 4.5.1 shows the manually labelled topics and their respective most occurring words.

Negative Topic	Top Five Words by Occurrence
Cancelled Flight	• {"**cancelled_flightled"**, "united", "southwestair", "w", "**americanair_cancelled**"}
Delayed Flight	• {"americanair", "jetblue", "u", "**delay**", "**hold_hour"**} • {"aa", "**weather**", "usairways", "**delay**", "**money**"}
Luggage Handling	• {"usairways", "please_help", "americanair_need", "**bag**", "delayed"}
Customer Service Line	• {"americanair", "usairways_americanair", "usairways", "southwestair", "**phone**"} • {"americanair", "usairways", "**hold**", "southwestair", "**line**"} • {"americanair", "**call**", "usairways", "answer", "going"}

Table 4.5.1: Negative Topics and their Top Five Words by Occurrence

For Southwest airline company, it appeared that most of the complaints were related to the "cancelled flight" and "customer service line" topics. More effort should be made to improve the flight scheduling management and customer service line management. For instance, online customer services should be offered in various ways such as phone calls, online chatbots, and others to improve efficiency.

For United airline company, it seemed that the negative sentiments came from the "cancelled flight" issue. The company should focus on providing appropriate money compensation or ways to allocate alternative flight rapidly. This might include business incorporation with other airline companies.

For American airline company, the complaints were regarding the "cancelled flight", "delayed flight", "luggage handling", and "customer service line" aspects. Specifically, customers might be most probably unsatisfied with the compensation provided due to the delayed flights during bad weather. Often, airline companies would refuse to provide refunds by claiming unqualified eligibility. American airline might have to revise their refunding policy related to delayed flights during bad weather.

For US Airway company, the negative tweets discussed topics such as “delayed flight”, “luggage handling”, and “customer service line”. Luggage handling issues might occur in various forms such as lost luggage, damaged luggage, long waiting time for fetching luggage, and others. Here, we suggest that US Airway should provide compensation for these undesired cases.

Compared to the topics provided in the raw dataset, we observed that similar topics included “late flight”, “cancelled flight”, “lost luggage”, and “damaged luggage”. There were more given topics like “customer service”, “bad flight”, and others. However, these topics were ambiguous. The aspects of service to improve were unclear. For instance, “bad flight” solely informed that the overall flight experience could be improved, but the specific parts of customers’ requirements were untold. Unlike such undesried scenarios, our insights provided a clear description of the services to improve, and suggestions were provided as well.

With respect to the second objective to mine a set of unique topics discussed in the negative comments, we fulfilled it by specifying four aspects: “Cancelled Flight”, “Delayed Flight”, “Luggage Handling”, and “Customer Service Line”.

4.6 Classification

The testing performances of models constructed are summarized as in Table 4.6.1.

Model	Accuracy	Precision	Recall	F_1-score
M_1	0.71	0.67	0.66	0.66
M_2	**0.81**	0.79	**0.77**	**0.78**
M_3	0.78	0.76	0.72	0.73
M_4	0.69	0.64	0.63	0.64
M_5	0.80	**0.80**	0.72	0.75
M_6	0.77	0.77	0.68	0.71
M_7	0.50	0.42	0.42	0.42
M_8	0.63	0.60	0.43	0.42
M_9	0.64	0.58	0.47	0.48

Table 4.6.1 Performances of Classifiers

From Table 4.6.1, we observed that experimental settings involving the use of LR algorithms such as M_2, M_5, and M_8 achieved high metric scores compared to those using other classification algorithms. Specifically, M_2 scored the highest accuracy, recall, and F_1-score. Meanwhile, M_5 scored the highest precision. The precision of M_2 came in the second place, with the difference as minimal as 0.01 when comparing with M_5's precision.

In this case, we claimed that LR was highly suitable for our pre-processed dataset. Since our dataset is small (only 14584 rows), LR was effective because it required fewer parameters to estimate than tree-based classifiers.

However, we observed that DT might be less suitable for our case study. Experimental settings leveraging DT like M_1, M_4, and M_7 achieved relatively lower metric scores than those scenarios involving the other two classification algorithms. This was most probably due to DT not having enough data to make accurate splits, leading to poor generalization and performance.

On the other hand, when comparing setups involving LR, we discovered that BoW shone out, while Word2Vec was the worst. BoW was suitable due to itself being a vectorizer with simple structure, thus can be less prone to overfitting when handling a

small dataset. Meanwhile, Word2Vec might not be necessary in this case because in smaller dataset, some of the words may only appear a few times providing too little information for the model to generate accurate embeddings. This can lead to embeddings that do not capture the true meaning of the word, adding noise to the overall word space.

Overall, with respect to the third project objective, we concluded that the optimal solution for predicting sentiments of unseen tweets was a specialized LR model augmented with vectorization via BoW.

Chapter 5: Conclusion

This study set forth with three objectives: (1) To create an optimized set of sentence-level sentiments, (2) to mine a set of unique topics discussed in the negative comments, and (3) to identify an optimal classifier for predicting sentiments of unseen tweets related to the airline industry. Our findings showed that a specialized weighted ensemble framework involving {0.5*Given, 0.3*BERT, 0.2*roBERTa} outperformed the given set by achieving a mean confidence score of 0.96. The importance of ensemble framework was shown. Subsequently, a set of negative topics were identified, namely the "Cancelled Flight", "Delayed Flight", "Luggage Handling", and "Customer Service Line". Airline companies should improve these areas of services for higher satisfaction levels from the customers. Finally, the findings revealed that a specialized LR model augmented with vectorization via BoW achieved an accuracy score of 0.81, which was optimal for classifying the sentence-level sentiments of unseen tweets. Nevertheless, future work should focus on hyperparameter tuning to enhance performance if possible. Domain experts could be included to aid in identifying negative topics for higher accuracy.

Reference

[1] Aljedaani, W., Rustam, F., Mkaouer, M. W., Ghallab, A., Rupapara, V., Washington, P. B., ... & Ashraf, I. (2022). Sentiment analysis on Twitter data integrating TextBlob and deep learning models: The case of US airline industry. *Knowledge-Based Systems*, *255*, 109780.

[2] Farzadnia, S., & Vanani, I. R. (2022). Identification of opinion trends using sentiment analysis of airlines passengers' reviews. *Journal of Air Transport Management*, *103*, 102232.

[3] Kolkur, S., Dantal, G., & Mahe, R. (2015). Study of different levels for sentiment analysis. *International Journal of Current Engineering and Technology*, *5*(2), 768-770.

[4] Elbagir, S., & Yang, J. (2020). Sentiment analysis on Twitter with Python's natural language toolkit and VADER sentiment analyzer. In *IAENG Transactions on Engineering Sciences: Special Issue for the International Association of Engineers Conferences 2019* (pp. 63-80).

[5] Rane, A., & Kumar, A. (2018, July). Sentiment classification system of twitter data for US airline service analysis. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)* (Vol. 1, pp. 769-773). IEEE.

[6] Hasib, K. M., Habib, M. A., Towhid, N. A., & Showrov, M. I. H. (2021, February). A novel deep learning based sentiment analysis of twitter data for us airline service. In *2021 International Conference on Information and Communication Technology for Sustainable Development (ICICT4SD)* (pp. 450-455). IEEE.

[7] Schouten, K., & Frasincar, F. (2015). Survey on aspect-level sentiment analysis. *IEEE transactions on knowledge and data engineering*, *28*(3), 813-830.

[8] Tusar, M. T. H. K., & Islam, M. T. (2021, September). A comparative study of sentiment analysis using NLP and different machine learning techniques on US airline Twitter data. In *2021 International Conference on Electronics, Communications and Information Technology (ICECIT)* (pp. 1-4). IEEE.

[9] Bouazizi, M., & Ohtsuki, T. (2019). Multi-class sentiment analysis on twitter: Classification performance and challenges. *Big Data Mining and Analytics*, *2*(3), 181-194.

[10] Djaballah, K. A., Boukhalfa, K., & Boussaid, O. (2019, October). Sentiment analysis of Twitter messages using word2vec by weighted average. In *2019 Sixth International Conference on Social Networks Analysis, Management and Security (SNAMS)* (pp. 223-228). IEEE.

[11] Hasan, M. R., Maliha, M., & Arifuzzaman, M. (2019, July). Sentiment analysis with NLP on Twitter data. In *2019 international conference on computer, communication, chemical, materials and electronic engineering (IC4ME2)* (pp. 1-4). IEEE.

[12] Mishra, R. K., & Urolagin, S. (2019, December). A Sentiment analysis-based hotel recommendation using TF-IDF Approach. In *2019 international conference on computational intelligence and knowledge economy (ICCIKE)* (pp. 811-815). IEEE.

[13] Nugroho, N. A., & Setiawan, E. B. (2021). Implementation Word2Vec for Feature Expansion in Twitter Sentiment Analysis. *Jurnal RESTI (Rekayasa Sistem dan Teknologi Informasi)*, *5*(5), 837-842.

[14] Bilgin, M., & Köktaş, H. (2019). Sentiment analysis with term weighting and word vectors. *Update*, *27*, 08.

[15] Neogi, A. S., Garg, K. A., Mishra, R. K., & Dwivedi, Y. K. (2021). Sentiment analysis and classification of Indian farmers' protest using twitter data. *International Journal of Information Management Data Insights*, *1*(2), 100019.

[16] El Rahman, S. A., AlOtaibi, F. A., & AlShehri, W. A. (2019, April). Sentiment analysis of twitter data. In *2019 international conference on computer and information sciences (ICCIS)* (pp. 1-4). IEEE.

[17] Saad, S. E., & Yang, J. (2019). Twitter sentiment analysis based on ordinal regression. *IEEE Access*, *7*, 163677-163685.

[18] Johnson, J. M., & Khoshgoftaar, T. M. (2019). Survey on deep learning with class imbalance. *Journal of big data*, *6*(1), 1-54.

[19] Vujović, Ž. (2021). Classification model evaluation metrics. *International Journal of Advanced Computer Science and Applications*, *12*(6), 599-606.

[20] Karthika, P., Murugeswari, R., & Manoranjithem, R. (2019, April). Sentiment analysis of social media network using random forest algorithm. In *2019 IEEE international conference on intelligent techniques in control, optimization and signal processing (INCOS)* (pp. 1-5). IEEE.

[21] Virgananda, M. A., Budi, I., & Ryan, R. S. (2023). Purchase Intention and Sentiment Analysis on Twitter Related to Social Commerce. *International Journal of Advanced Computer Science and Applications*, *14*(7).

[22] Rustam, F., Khalid, M., Aslam, W., Rupapara, V., Mehmood, A., & Choi, G. S. (2021). A performance comparison of supervised machine learning models for Covid-19 tweets sentiment analysis. *Plos one*, *16*(2), e0245909.

[23] Samuel, J., Ali, G. M. N., Rahman, M. M., Esawi, E., & Samuel, Y. (2020). Covid-19 public sentiment insights and machine learning for tweets classification. *Information*, *11*(6), 314.