



Data Analysis 2 (Task 3)

US Airline Sentiment on Twitter

Name	ID
Joud Ahmad Al-huthaly	444002970
Nehal Hamed Al-zahrani	444001073

DEPARTMENT OF (INFORMATION SCIENCE-DATA SCIENCE)
COLLEGE OF COMPUTER AND INFORMATION SYSTEMS UMM
AL-QURA UNIVERSITY

Table of content

Introduction...(page 3)

Exploratory Data Analysis...(page 4-6)

Data cleaning..(page 7)

Data Preprocessing..(page 8)

Feature Extraction..(page 9-10)

Dividing Data into Training and Test Sets and logistic model
(page 11)

Link to Google collaboration : <https://colab.research.google.com/drive/1sQQz9XpdPrBy95ffp0YQNt1gvdJe1can?usp=sharing>

Link to dataset : <https://www.kaggle.com/datasets/crowdflower/twitter-airline-sentiment>

Introduction

Sentiment analysis refers to the process of analyzing opinions or emotions about a certain topic using data such as text or images. It helps businesses make informed decisions based on public sentiment. For example, if a product receives negative feedback, the company may decide to modify or discontinue it to avoid losses.

In this project, we will analyze tweets about six major U.S. airlines to classify whether the sentiment expressed is positive, negative, or neutral. This analysis helps companies understand and respond effectively to customer feedback.

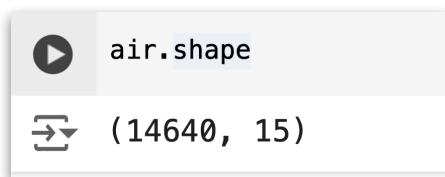
The Dataset

	tweet_id	airline_sentiment	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	airline_sentiment_gold	name	negativereason_g	
0	570306133677760513	neutral	1.0000	NaN	NaN	Virgin America	NaN	cairdin		I
1	570301130888122368	positive	0.3486	NaN	0.0000	Virgin America	NaN	jnardino		I
2	570301083672813571	neutral	0.6837	NaN	NaN	Virgin America	NaN	yvonnalynn		I
3	570301031407624196	negative	1.0000	Bad Flight	0.7033	Virgin America	NaN	jnardino		I
4	570300817074462722	negative	1.0000	Can't Tell	1.0000	Virgin America	NaN	jnardino		I

reason_confidence	airline	airline_sentiment_gold	name	negativereason_gold	retweet_count	text	tweet_coord	tweet_created	tweet_location	user_timezone
NaN	Virgin America	NaN	cairdin	NaN	0	@VirginAmerica What @dhepburn said.	NaN	2015-02-24 11:35:52 -0800	NaN	Eastern Time (US & Canada)
0.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica plus you've added commercials t...	NaN	2015-02-24 11:15:59 -0800	NaN	Pacific Time (US & Canada)
NaN	Virgin America	NaN	yvonnalynn	NaN	0	@VirginAmerica I didn't today... Must mean I n...	NaN	2015-02-24 11:15:48 -0800	Lets Play	Central Time (US & Canada)
0.7033	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica it's really aggressive to blast...	NaN	2015-02-24 11:15:36 -0800	NaN	Pacific Time (US & Canada)
1.0000	Virgin America	NaN	jnardino	NaN	0	@VirginAmerica and it's a really big bad thing...	NaN	2015-02-24 11:14:45 -0800	NaN	Pacific Time (US & Canada)

Exploratory Data Analysis

- **tweet_id**: Unique identifier for each tweet.
- **airline_sentiment**: Sentiment of the tweet (positive, neutral, negative).
- **airline_sentiment_confidence**: Confidence score for the sentiment.
- **negativereason**: Reason for a negative sentiment (only for negative tweets).
- **negativereason_confidence**: Confidence score for the negative reason.
- **airline**: The airline mentioned in the tweet.
- **airline_sentiment_gold**: Gold standard sentiment (if available).
- **name**: Twitter username.
- **negativereason_gold**: Gold standard negative reason (if available).
- **retweet_count**: Number of retweets.
- **text**: The tweet text.
- **tweet_coord**: Coordinates where the tweet was sent from.
- **tweet_created**: Timestamp of when the tweet was created.
- **tweet_location**: User-specified tweet location.
- **user_timezone**: User-specified timezone.

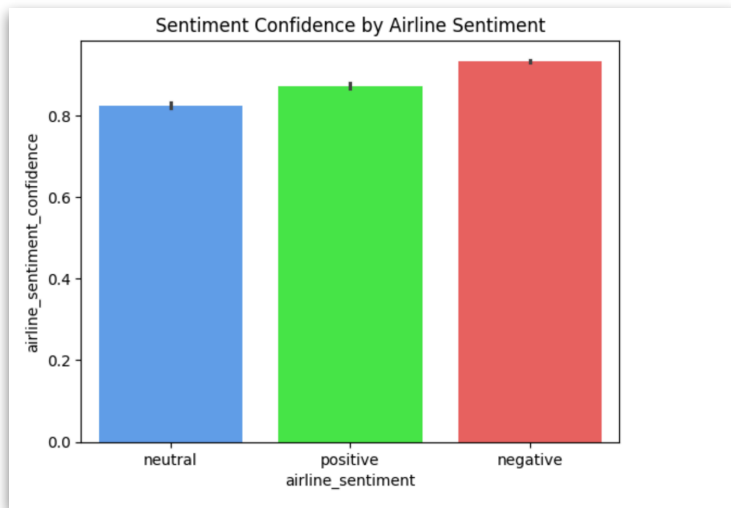


```
air.shape
```

```
(14640, 15)
```

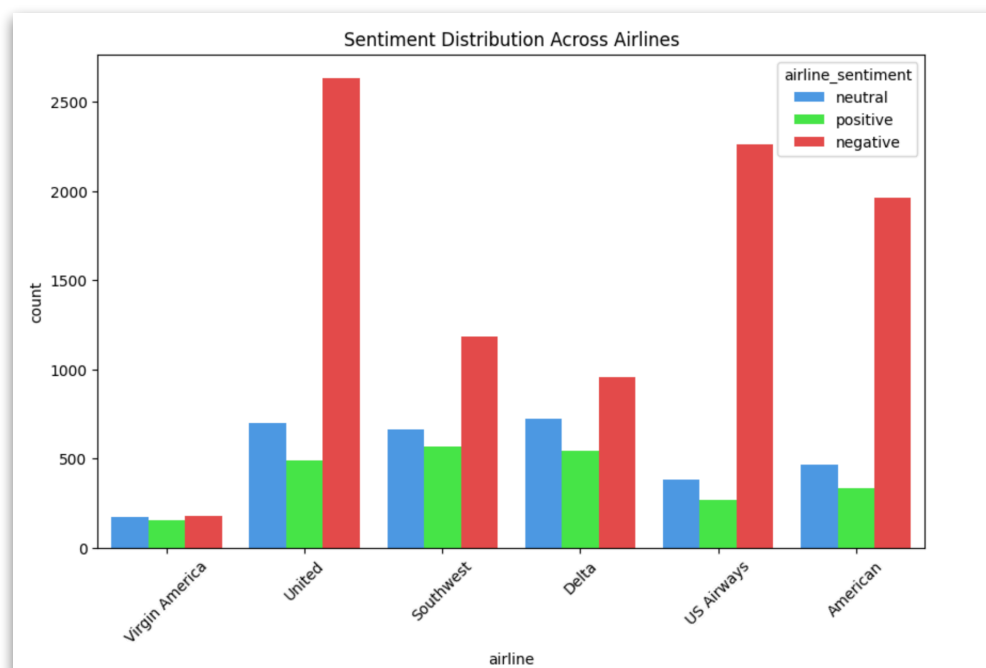
This code displays the number of columns and rows in the dataset

- Number of columns: 15
- Number of rows: 14640

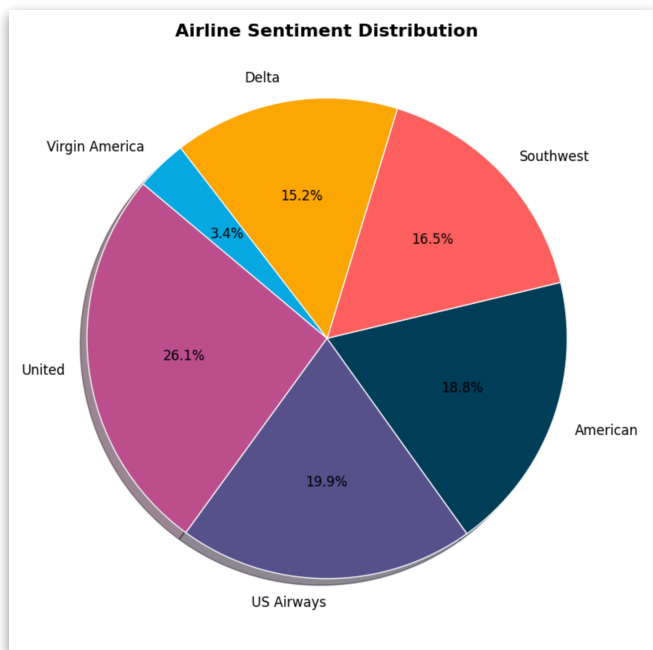


The confidence in neutral sentiment is slightly below 0.8, while positive sentiment confidence is around 0.85. The highest confidence is in negative sentiment, slightly exceeding 0.85.

We conclude that users express negative sentiments with higher confidence compared to positive or neutral sentiments, which may suggest that complaints or negative feedback are more clear and assertive when expressed.



The chart shows that negative sentiment is most prevalent across most airlines, especially United and US Airways, which may indicate challenges in customer experience or a low level of satisfaction. On the other hand, we note that Southwest and Delta have a relative balance between negative and positive emotions, reflecting an acceptable level of performance. Virgin America has the lowest number of ratings overall, with a low percentage of negative ratings, which may indicate higher satisfaction or a smaller customer base compared to other companies.



The pie chart reflects the relative distribution of ratings across airlines, with United leading at 26.1%, indicating that it receives the most customer attention, whether positive or negative. This is followed by US Airways at 19.9% and American at 18.8%, showing that these companies also experience a significant amount of interaction.

On the other hand, Southwest and Delta, which account for 16.5% and 15.2% respectively, reflect a moderate level of customer engagement. Companies like Virgin America hold only 3.4% of the ratings, which may indicate a smaller scale of operations or a more stable experience that does not prompt customers to interact as much.

Data cleaning

```
tweet_id          0
airline_sentiment 0
airline_sentiment_confidence 0
negativereason    5462
negativereason_confidence 4118
airline           0
airline_sentiment_gold 14600
name              0
negativereason_gold 14608
retweet_count     0
text              0
tweet_coord       13621
tweet_created     0
tweet_location    4733
user_timezone     4820
dtype: int64
```

We check for missing values in the DataFrame named **air**, then remove columns that contain a significant amount of missing values, specifically **airline_sentiment_gold**, **negativereason_gold**, and **tweet_coord**.

Next, we replace NaN values in the **negativereason** column with **"unknown"** and confirm that there are no missing values in that column.

Following that, we replace **NaN** values in the **tweet_location** column with **"unknown"** and verify that there are no missing values remaining.

Afterward, we calculate the mean of the values in the **negativereason_confidence** column and replace the missing **NaN** values in this column with the calculated mean confidence **mean_confidence**. This helps maintain the data and improve the accuracy of the analysis by filling in the missing values in a way that relies on the existing information.

Finally, we check again for any remaining missing values after the replacements

t	airline_sentiment_confidence	negativereason	negativereason_confidence	airline	name	retweet_count	text	tweet_created	tweet_location	user_timezone
al	1.0000	unknown	0.638298	Virgin America	cairdin	0	@VirginAmerica What @dhepburn said.	2015-02-24 11:35:52 -0800	unknown	Eastern Time (US & Canada)
e	0.3486	unknown	0.000000	Virgin America	jnardino	0	@VirginAmerica plus you've added commercials t...	2015-02-24 11:15:59 -0800	unknown	Pacific Time (US & Canada)
al	0.6837	unknown	0.638298	Virgin America	yvonnalynn	0	@VirginAmerica I didn't today... Must mean I n...	2015-02-24 11:15:48 -0800	Lets Play	Central Time (US & Canada)
e	1.0000	Bad Flight	0.703300	Virgin America	jnardino	0	@VirginAmerica it's really aggressive to blast...	2015-02-24 11:15:36 -0800	unknown	Pacific Time (US & Canada)
e	1.0000	Can't Tell	1.000000	Virgin America	jnardino	0	@VirginAmerica and it's a really big bad thing...	2015-02-24 11:14:45 -0800	unknown	Pacific Time (US & Canada)

Data Preprocessing

Natural Language Toolkit (nltk): The **nltk** library is used for natural language processing tasks, particularly for handling stopwords. Stopwords are common words in a language, such as "and," "the," and "is," that are often filtered out during text processing. In our code, we download the stopwords list and create a set of English stopwords.

We preprocess the text in the **text** column of the DataFrame **air** by downloading stopwords and defining common English words to filter out. We normalize the text through a function that performs the following steps:

We preprocess the text in the **text** column of the DataFrame **air** by downloading stopwords and defining common English words to filter out. We normalize the text through a function that performs the following steps:

- **Convert to Lowercase:** All text is converted to lowercase to ensure uniformity.
- **Remove Special Characters:** A regular expression is used to remove any character that is not a lowercase letter or space.
- **Remove Stopwords:** Common words (stopwords) are filtered out from the text.
- **Remove Excess Whitespace:** Any extra spaces between words are trimmed.

After applying this normalization, we create a new column named **normalized_tweet**, which stores the processed text. We then verify the results by displaying the original and cleaned text side by side, preparing the data for more effective analysis.

Feature Extraction using TF-IDF (Term Frequency-Inverse Document Frequency)

TF-IDF (Term Frequency-Inverse Document Frequency) is a statistical measure used in text processing and information retrieval to evaluate the importance of a word in a document relative to a collection of documents.

- **Term Frequency (TF)** measures how often a word appears in a specific document, reflecting its significance within that document.
- **Inverse Document Frequency (IDF)** assesses how common or rare a word is across all documents.

The TfidfVectorizer is initialized with specific parameters:

- Maximum of 5000 words.
- Requires words to appear in at least 3 documents.
- Excludes words present in over 90% of documents.
- Removes common English stopwords.

The process involves:

1. Converting tweets into a TF-IDF matrix.
2. Transforming the matrix into a DataFrame.
3. Calculating the average TF-IDF score for each word.
4. Sorting words based on average TF-IDF values.

TF-IDF Matrix:

	aa	aadvantage	aarp	aas	abc	ability	able	aboard	abq	absolute	\
0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
1	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
3	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
4	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
...	
14635	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
14636	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
14637	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
14638	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
14639	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	
	...	yr	yrs	yuma	yup	yvr	yyz	zero	zkatcher	zone	zurich
0	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
1	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
2	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
4	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
...
14635	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14636	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14637	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14638	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0
14639	...	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0

[14640 rows x 3980 columns]
Top words by mean TF-IDF:

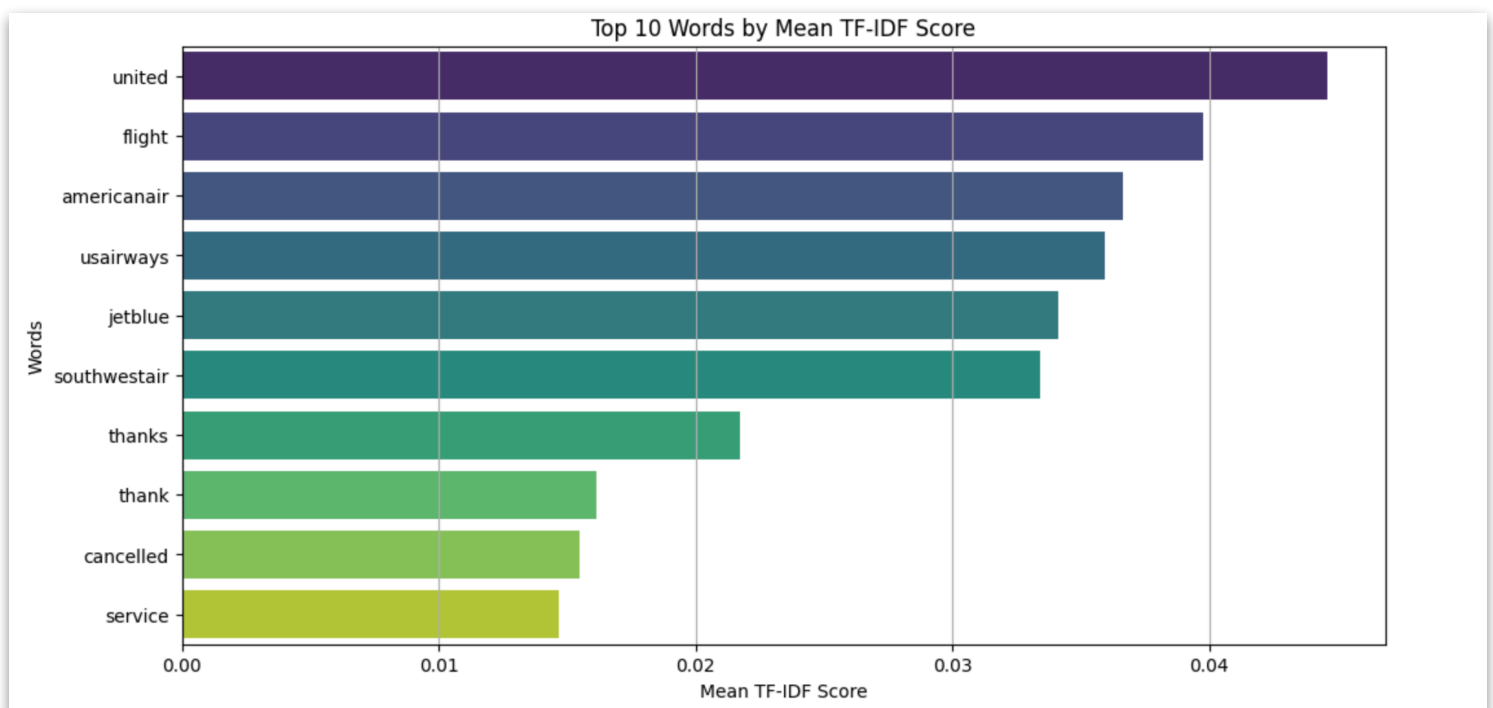
	word	mean_tfidf
3700	united	0.044610
1357	flight	0.039775
121	americanair	0.036647
3738	usairways	0.035932
1904	jetblue	0.034140
3288	southwestair	0.033394
3515	thanks	0.021738
3512	thank	0.016147
491	cancelled	0.015461
3149	service	0.014663
1645	help	0.014249
3555	time	0.012221
860	customer	0.012204
1768	im	0.012103
1715	hours	0.011741
1672	hold	0.011644
1371	flights	0.010866
3778	virginamerica	0.010590
2620	plane	0.010378
2326	need	0.009813

Results include:

- Displaying the top 20 words weighted by their average TF-IDF values.
- Creating a bar plot showcasing the top 10 words with the highest TF-IDF scores.

Notable high-weighted words, such as "united," "flight," and "americanair," These results indicate that users interact significantly with these words, highlighting their importance in the context of tweets. Therefore, these words can be considered important indicators for understanding conversations and trends within tweets.

- Here's a bar plot using Matplotlib and Seaborn to visualize the **top 10 words with the highest mean TF-IDF scores**, showing the scores on the x-axis and the words on the y-axis.



Dividing Data into Training and Test Sets and logistic model

Here, we built a model for text analysis using logistic regression to classify the sentiments of tweets related to airlines. The confusion matrix reflects the model's performance by showing the number of correct and incorrect predictions.

The results indicate that:

- The model successfully classified **1726** samples as positive sentiments, with some errors (**120 and 24**).
- For negative sentiments, the model correctly classified **284** samples, but there were some errors (**295 and 35**).
- For neutral sentiments, the model achieved **256** correct classifications with some errors (**127 and 61**).

Accuracy Rate: The model's accuracy rate was **77.39%**, indicating good performance but also suggesting the possibility of improving the model for better results.

This analysis reflects the model's ability to classify the sentiments of tweets, opening up opportunities to enhance the model's performance through various methods, such as improving data processing, using additional features, or trying other models.

