Group 8: Sentiment Analysis for Top Airlines in the USA

Overview

In this project, we focus on leveraging Natural Language Processing (NLP) techniques to analyze sentiments expressed in a Twitter dataset, specifically within the airline industry domain. The primary objectives include sentiment analysis, the construction of a precise tweet classification model, and the development of a chatbot capable of responding to customer feedback and directing queries to the appropriate resolution teams. By accomplishing these goals, we will create solutions that enable airline companies to extract valuable insights from social media data which will play a big role in enhancing their customer service, improving their service offering and empowering these organizations to make data-driven decisions.

Problem statement

The advent of social media has generated an abundance of data, presenting both opportunities and challenges for organizations. This vast pool of data offers unparalleled insights into customer perceptions, preferences, and feedback. However, many organizations are yet to develop frameworks and strategies to effectively analyze and interpret such data. Insights from this data holds the potential to benefit various domains, including business operations, marketing strategies, public opinion analysis, and more.

Our stakeholders (top American airline companies) have requested us to analyze social media raw data and showcase the customer sentiment as either positive, neutral or negative while identifying the top drivers for these sentiments.

Our dataset is sourced from Twitter, capturing a wide array of tweets, and our primary focus is analyzing and visualizing drivers for key & top public & customer sentiment. We aim to address critical questions and challenges faced by airlines, such as understanding passenger sentiments from unstructured data and predicting engagement metrics. By doing so, we strive to provide airlines with the tools and knowledge needed to enhance customer experiences, optimize operations, and make data-driven decisions in an ever-evolving and competitive industry.

Business Understanding

Our dataset includes data from X for the top 5 airline companies in the USA and these are our primary stakeholders for the project. These airlines include; United Airlines, US Airways, American Airlines, Delta Airlines and Southwest Airlines.

The airline industry is a vital sector that provides air transport services for passengers and cargo. It plays a crucial role in connecting people, businesses, and regions across the world. The industry comprises a diverse range of airlines, including full-service carriers, low-cost carriers (LCCs), regional airlines, and charter airlines. Major players are often categorized into international, national, and regional carriers, each serving specific markets.

Global airline industry trends:
The trends below were useful in developing our understanding and hence optimal analysis of our dataset.

1. *Digital transformation and technology adoption*: While late adopters, airlines are increasingly leveraging technology to enhance operational efficiency, customer experiences, and overall service quality. This includes implementing mobile apps, self-service kiosks, AI-powered chatbots, and data analytics for personalized marketing and improved decision-making.
2. *Sustainable and environmentally friendly practices:* Environmental sustainability has become a significant focus within the industry. Airlines are investing in more fuel-efficient aircraft, exploring biofuels, and implementing eco-friendly practices to reduce carbon emissions and mitigate their environmental impact.
3. *Demand for personalized travel experiences:* Travelers now seek personalized experiences, leading to a shift in airline strategies. Airlines are customizing services, offering ancillary products, and tailoring loyalty programs to meet individual preferences and needs.
4. *Partnerships and alliances*: Collaborations, partnerships, and alliances among airlines have become prevalent. These agreements help airlines expand their networks, improve cost-efficiency, and offer travelers more seamless travel options.
5. *Health and safety measures post-pandemic*: The COVID-19 pandemic has significantly impacted the industry. Airlines are implementing stringent

health and safety protocols to regain traveler confidence. Measures include enhanced cleaning procedures, health screenings, and contactless processes.

Airline industry profit margins:

- The net profit margin for airlines is highly influenced by market conditions, fuel prices, operational efficiency, competition, and economic trends.
- On average, net profit margins for airlines typically range from 2% to 5%. However, it's important to note that individual airline net profit margins may fluctuate, and some airlines may experience periods of losses due to various factors affecting the industry.
- Airlines with a positive brand tend to have lower customer churn and in return higher pricing and higher revenues.

Objectives

Main Objective

Perform analysis on raw tweets & extract public sentiment & as well as build a chatbot solution

Specific Objectives

1. To build a model that classifies raw tweets into the 3 sentiment classes for future use
2. To visualize the top drivers for each sentiment to help management target service delivery improvement
3. To create & deploy a chatbot to monitor customer feedback on X that provides real time responses

Research Questions

The project will be answer the below research questions:

1. What are the major airlines represented in tweets per the dataset?
2. What are the predominant sentiments expressed regarding major U.S. airlines?
3. How does the sentiment compare between the various airlines in our dataset?
4. What are the most common reasons for negative sentiments?

5. What are the most common reasons/terms used in positive sentiments?

Data Understanding

Data Source

Our dataset was publicly sourced from crowdflower website and is made up of Twitter users' tweets and retweets. The dataset has 14,640 rows and 20 columns. This Twitter data was collected from February 2015 and contributors were engaged in classifying tweets into categories of positive, negative, and neutral sentiments. Additionally, contributors were tasked with categorizing the reasons behind negative sentiments, such as "late flight" or "rude service."

This dataset serves as the foundation for our analysis, enabling us to gain insights into passenger & general public sentiments, engagement patterns, and other trends within the US airline industry.

Link to the Data Source

Data Description

The dataset has 14,640 rows and 20 columns. Below are the columns and their descriptions:

- _unit_id: A unique identifier for each data unit.
- _golden: A boolean value indicating whether the entry is a golden unit in the dataset.
- _unit_state: The state of the unit (e.g., golden).
- _trusted_judgments: The number of trusted judgments for the entry.
- _last_judgment_at: Timestamp of the last judgment for the entry.
- airline_sentiment: The target variable, which represents the sentiment of the airline tweet (positive, negative, or neutral).
- airline_sentiment:confidence: The confidence level associated with the airline sentiment.
- negativereason: The reason for negative sentiment in the tweet.
- negativereason:confidence: The confidence level associated with the negative sentiment reason.
- airline: The airline associated with the tweet.
- airline_sentiment_gold: Additional information about airline sentiment (gold standard).
- name: The name of the user who posted the tweet.

- negativereason_gold: Additional information about the negative sentiment reason (gold standard).
- retweet_count: The number of retweets for the tweet.
- text: The text content of the tweet.
- tweet_coord: Coordinates of the tweet (if available).
- tweet_created: Timestamp of when the tweet was created.
- tweet_id: The unique identifier of the tweet.
- tweet_location: The location associated with the tweet (if provided).
- user_timezone: The timezone of the user who posted the tweet.

The airline_sentiment column is the target variable, which represents the sentiment of the airline tweet that we may want to predict or analyze.

Data Preparation

Importing libraries

Key libraries were imported to support the analysis of NLP using pandas. The included modules for visualization and various modeling libraries.

Loading the data

The data set was then loaded into the jupyter notebook. The data frame was then displayed to show the data per column, the datatypes per column and total entries (hence nulls) per column.
Using pandas data cleaning methods we viewed and analyzed the contents of various columns to understand the contents and usefulness for our analysis. We also established the value counts for each of the airlines. We inspected the tweet column & anticipated the type of data cleaning & text preprocessing to be carried out as below.

Cleaning data

We checked for duplicates & confirmed that there were none. We handled the nulls earlier identified through data imputation for columns with timestamps and used placeholders such as 'Not specified' or 'Unknown' for other missing values.

Data preprocessing for Natural Language processing(NLP)

Text preprocessing serves as the initial step in our NLP project. We undertook the following steps to prepare the data & convert the raw tweets into tokens which

represent clean, structured data that was used for exploratory data analysis (EDA) and model building.

1. Eliminating punctuation marks such as periods, commas, exclamation points, parentheses, asterisks, percentage signs, and at symbols. '!"#$%&'()*+,-./:;?@[]^_`{|}~'
2. Omitting URLs from the text.
3. Removing common stop words (e.g., "the," "and," "is") that do not carry significant meaning.
4. Converting all text to lowercase for consistency.
5. Tokenization, breaking the text into individual words or tokens.
6. Lemmatization, a more sophisticated technique that reduces words to their base or dictionary form, considering context and meaning.

Exploratory Data Analysis (EDA)

We used EDA to visualize the preprocessed dataset to gain insights and understand the dataset characteristics. EDA enabled us to uncover patterns, relationships, and potential issues in your data before diving into more advanced analyses.

Our EDA visualizations were directly linked to answering the project research questions discussed above. I.e. For every research question, we created an appropriate visualization. We mostly utilized barplots and word clouds. We also utilized bigrams to visualize the top words associated with positive sentiment.

The EDA enabled us to generate insights on distribution of each sentiment type across airlines. It further enabled us to identify the top reasons/drivers for each type of sentiment. In so doing we understood the potential areas of improvement and potential conclusions and recommendations to our stakeholders.

Modeling

In model building, our target was to build a model that achieves an accuracy of 80-90% indicating a high degree of correctness in the future prediction or classification of raw tweets into the three sentiment categories (positive, neutral & negative). Our other goal was to create a model that achieves a good balance between precision & recall at the same time maximizing the F1 score.

We chose & used these models SVM, Logistic Regression, Random Forest & Multinomial NB.

Two steps we conducted pre-modeling:

1. Extracted the X & y (target) variables
2. Encoded the target labels
3. Performed Train_Test_split

In the modeling step, we utilized functions in defining the steps, methods, hyperparameters, model tuning approaches & model evaluation metrics to be used in identifying the best performing models per our modeling goals.

In summary the function steps included:

1. Creating a function
    a. Creating a pipeline with parameters Count vectorization, SMOTE resampling & the classifier)
    b. Fitting the pipeline on the training data
    c. Evaluating the model on test data and displaying the results per model
    d. Calculating the confusion matrix and classification report & displaying the results per model
2. Evaluating the model outcomes

We first fitted the baseline models and later added hyperparameters & GridSearch CV to the function above.The analysis revealed that negative sentiments are most prevalent among passengers, accounting for approximately 62.69% of the sentiments in the dataset. This highlights the importance of addressing customer concerns and improving overall satisfaction.

Our tuned models produced better results than baseline models. Our final model was the tuned logistic regerssion model that resulted in accuracy of 80.53%.

Conclusions

- Sentiment Distribution - The analysis revealed that negative sentiments are most prevalent among passengers, accounting for approximately 62.89% of the sentiments in the dataset. This highlights importance of addressing customer concerns and improving overall satistaction

- Airlines' Sentiment Distribution - Different airlines exhibit varying sentiment distributions , with some struggling with predominantly negative sentiments , while others maintain a more balanced distribution of negative, positive and neutral sentiments. Understanding thse variations is crucial for each airline's strategy
- Common Negative Reason - "Customer survive issues" is the most frequently cited reason for negative sentiments across the airlines, followed by "Late Flight" and Cancelled Flight." Addressing these common issues can significantly improve passenger satisfaction.
- Model Performance - The logistic regression model and SVM models exhibit strong performancein classifying sentiment, making them suitable choices for implementing a sentiment analysis system for customer interactions

Recommendations

- Improve Customer service - Given that "Customer service issue" is a predominant reason for negative sentimentys across airlines, it's essential for airlines to invest in enhancing their customer service. This includes better training for staff faster response times and improved communicationwith passengers.
- Address Flight Punctuality - late flights are a major concern for passengers and airlines should work on minimizing delays and providing accurate information to travelers. Implementing efficient flight scheduling and contingency plans can help mitigate this issue.
- Invest in Luggage Handling - Passengers' sentiments are negatively affectedby issus related to lost and damaged luggage. Airlines should focus

on improving baggage handling processes to reduce such incidents and ensure a smoother travel experience.

- Enhance Online Booking System - Flight boking problems are a common complaint among passengers. Airlines should update and streamline their online booking systems to make it easier and more user friendly
- Monitor Social Media - Airlines should actively monitor social mediapaltforms for customer feedback and respond promptly to address concerns or complaints. This can help improve customer satisfaction and brand reputation.

Future Improvement Ideas

- Development of chatbot to react to sentiments which is more quicker and faster hence saving on cost and time

Follow-up Questions

- Did we have the right data? Yes. But was Huge
- Do we need other data to answer our question? No
- Did we have the right question? Yes. The question chosen was correct due to the objectives that we wanted to accomplish in the project.

Deployment

Pickling the Model

We save our model through pickling.

User-Interface Design

The model will be integrated with a user-friendly web interface, designed to simplify the user experience.

Model and UI Integration

The selected model will be integrated with the user interface, enabling users to give their sentiment and receive customised feedback. The model will accept user input data and analyse to predict a tweet sentiment. Additionally, the system will include a chatbot, designed to provide interactive and engaging user support to return a customised message.

Deployment Options

Our deployment plan involved streamlit to ensure that users can easily access the system. By implementing these steps, users will receive tailored reponses to their sentiments, ultimately improving customer feedback.