Abstract:

The airline industry has come a long way thanks to several aviation-related and technological innovation, all while catering to the needs of the modern traveler. As most large businesses that have anything to promote, airline companies market to prospective customers through Twitter, by keeping travelers aware of relevant travel information, the latest deals they may have, or a unique offering that may distinguish them from other airlines. This progress report provides an update on this project, which seeks to use topic modeling techniques such as Latent Dirichlet Allocation (LDA) to discover how airline companies use Twitter to drive consumer and traveler engagement; the resulting analysis will seek to explain the uncovered topics that may reflect airlines’ marketing strategies, ultimately visualizing these results for all airlines on a custom-built dashboard.

Introduction:

Airlines are interested in how they can best support their customer base and prospective travelers by establishing a two-way communication between them on social media. Through these efforts, one scrolling on their Twitter feed can be converted into an engaged traveler and activated member of the airline’s customer base. Incorporating innovative ways to help customers better understand what an airline offers and how it can serve them during their travel experience can help airlines build customer loyalty as well as promote their offerings via online reviews/forums or simple word-of-mouth exchanges between engaged customers and prospective travelers. For example, airlines may do any of the following or more:

* Build a near-real-time customer service pipeline to answer typical travel questions quickly (<https://thriftytraveler.com/guides/airlines/airlines-twitter/>).
* Tailor their social media presence to attract specific customer segments (i.e. younger generations, residents of nearby countries) (<https://www.washingtonpost.com/travel/2022/10/04/ryanair-twitter-strategy-gen-z/> ; <https://www.sciencedirect.com/science/article/pii/S1877042814039366>).
* Discover new opportunities for marketing campaigns to ultimately build brand awareness of the airline in the customer (<https://www.kambr.com/articles/how-airlines-embrace-of-social-media-is-evolving-after>).

Therefore, deriving the (extent to which) certain topics make up airline companies’ tweets can tell us what topics certain airlines focus on more than others; from this, we can determine what is important to prospective travelers on an airline-by-airline basis. For example, one airline might especially promote their exclusive club membership benefits, while another may focus on marketing to customers that are looking for the cheapest airfare.

One way to understand where airline companies stand in their marketing strategies is by aggregating their airline-to-customer communications (which, in this project, are represented as organic tweets and retweets) to perform topic modeling. Topic modeling is a text mining technique used in research to derive hidden topics and meaningful semantic structures from text across a corpus of documents.

This paper discusses related motivations and research regarding marketing strategies in the airline industry, where the research has widely focused on customer-to-airline engagement, while the project discussed in this paper is more aligned with airline-to-customer engagement.

What distinguishes this project is it seeks to assess the *beginning* of the marketing strategy – not customer engagement/experience, as in these examples, but rather *company* engagement. As such, airlines can benefit from visual results this paper’s work derives to better visualize where their social media strategy focus is currently directed, in order for them to, for example:

* Justify creation or adjustment of marketing strategy: Determine where resources supporting airlines’ marketing/social media strategies need to be strengthened or relocated, which can be seen as necessary to do before infusing customer feedback analyses into campaigns, which needs directional and financial backing before approval.
* Perform competitor-topic analyses: See where competitor airline companies lie on spectrums pertaining to similar or new/emerging topics in their tweets.

Related work:

* Background
* Analysis of airline company tweets

A variety of academic research involve analyzing Twitter accounts or tweets related to airline companies. [INSERT AUTHOR] segment customers that follow or interact with an airline’s Twitter account to better understand the airline’s customer base (<https://www.sciencedirect.com/science/article/abs/pii/S0969699718302072>). Such an analysis can be layered in with this project on topic-modeling tweets from airline account, by aggregating customer engagement with specific topics at the segment level. [INSERT AUTHOR] perform sentiment analysis on airline-specific customer feedback from tweets (<https://ieeexplore.ieee.org/document/8377739>). This can help airlines understand travelers’ pain points, which they can potentially address as part of the travel experience or as making a commitment to improving their social media strategy.

Overall, these analyses can promote ideas to help airlines improve their business and marketing strategies by tailoring to their travelers’ needs and attracting more customer segments in the process.

* Usage of LDA for topic modeling on tweets

LDA (<https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf>) is a topic modeling method that supports analyzing a large set of documents. It takes as input a document corpus, from which a document-term matrix representation is derived for further computation. LDA then produces two different matrices: a document-topic matrix that describes each document’s composition with respect to the derived topics, and a topic-word matrix that describes the likelihoods of each word being associated with each topic. In light of our context, the resulting document-topic matrix from the LDA process can help to understand to what extent a particular tweet/set of tweets is/are associated with each topic, and the resulting topic-term matrix can help to dig deeper into what terms and to what extent terms make up a topic.

Topic modeling has been used in a wide variety of tweet analysis research as a way to uncover hidden meanings beneath tweets, which are the documents in this context (https://www.cs.toronto.edu/~jstolee/projects/topic.pdf). In particular, [INSERT AUTHOR] have explored topic modeling in relation to airline online reviews, lending more insight into customer-to-airline engagement and satisfaction (<https://www.mdpi.com/2078-2489/12/2/78>).

* [Also look at JUST united’s topics? And include that in the visual dashboard maybe]
* Create embeddings from airline tweets
* Use embeddings to create topics via LDA topic modeling
* For each airline, average(?) the tweet embeddings to produce a score for each topic that the LDA discovered
* Now, each airline has a set of features, where each feature comes from a topic’s vectorized embedding
* [COMPARE each airline on the topics]
  + Visualization:
    - airline by topic heatmap (<https://pub.towardsai.net/tweet-topic-modeling-part-4-visualizing-topic-modeling-results-with-plotly-66d5dbaaf7fb>)
      * Toggles to select up to X airlines
      * Need to assign human-interpretable names to topics
    - Bubble chart showing top words per topic1
  + Other visualization ideas:
    - could be a set of number lines, or could be a 2D chart where you can pick two topics to visualize at a time and each airline’s scores on those topics
    - Each airline-topic can be represented by a bubble
      * Bubble color:
        + Average engagement [could be adding up likes, retweets, etc.] with such tweets [see below]
        + average sentiment of tweets whose highest topic score is for that topic (should there be a threshold for how high the topic score needs to be to be classified in that topic, or any topic, in the first place?)
      * Bubble size: number of tweets that fall into this airline-topic combo
    - By able to toggle for airlines that have at least 50, 100, 200, 500 tweets
* Data analysis to include in the paper:
  + <https://www.machinelearningplus.com/nlp/topic-modeling-visualization-how-to-present-results-lda-models/#4.-Build-the-Bigram,-Trigram-Models-and-Lemmatize>
    - Distribution of document (tweet) word counts
    - Distribution of document (tweet) word counts BY TOPIC
  + <file:///Users/raiha/Downloads/information-12-00078-v2.pdf>
    - Page 5: Improved diagram of system design
    - Page 7: Frequency analysis of uni/bi/tri-grams found across entire set of airline tweets
* These features can be used to predict a financial target value about the airline company
* This can tell us whether or not tweeting about a particular topic may make more sense for the overall marketing strategy of the airline company, therefore contributing to the company’s financial success, although there are many other factors contributing to a company’s overall performance.

Current progress, including data, methods, and system:

All code development for this project so far has been done in Python and utilizing the PyCharm and Jupyter Lab IDEs; all code so far can be found on GitHub: (<https://github.com/rkhan15/airline-topic-modeling>).

Tooling, languages, packages:

* Tooling: PyCharm, Jupyter Notebook, Plotly/Dash for visualization
* Languages: Python
* Packages: gensim, nltk (for lemmatizing)

Data stuff:

Data downloading and acquisition:

Organic tweets and retweets were collected from 85 Twitter accounts associated with airline companies headquartered worldwide using Twitter API v2 (<https://developer.twitter.com/en/docs/twitter-api>). For a given Twitter handle corresponding to an airline (i.e. @united, the Twitter handle for United Airlines), the following steps were taken to download its tweets:

1. Get the user ID of the airline company’s Twitter account (unique ID associated with one Twitter account) using the user lookup endpoint (<https://developer.twitter.com/en/docs/twitter-api/users/lookup/api-reference/get-users-by>) (i.e. 260907612, the user ID for Twitter handle @united)
2. Get up to the 800 most recent tweets (based on API request caps) from the airline company’s Twitter account, where the request response includes tweets include organic tweets and retweets and exclude replies (<https://developer.twitter.com/en/docs/twitter-api/tweets/timelines/api-reference/get-users-id-tweets>). [Footnote: Replies were excluded from requests, because some airline Twitter accounts perform customer service actions via tweet replies to customer Twitter accounts.]

In total, 67,540 tweets were downloaded on December 4, 2022; non-English tweets were filtered to leave 57,404 in total for analysis. The following shows the distribution of number of tweets (in English language only) per airline account: [INSERT HISTOGRAM]. 74% of airline Twitter accounts returned at least 700 tweets. Generally, the Twitter V2 API allows for downloading up to the 800 most recent tweets (and retweets, which are exclusive of the 800 number) when requesting for tweets without replies, so accounts for which less than exactly 800 tweets were downloaded are accounts that do not have more than 800 tweets.

Chart, histogram

Description automatically generated

Data pre-processing:

All tweets were converted into a tabular dataset, consisting of the text and the metadata associated with each tweet. Examples of metadata include date and time created, number of likes/retweets/replies, etc (<https://developer.twitter.com/en/docs/twitter-api/tweets/timelines/api-reference/get-users-id-tweets>). Preprocessing of the tweets was done to clean the raw text by removing text related to links, retweets (i.e. tweet text beginning with “RT”), mentions, hashtags, punctuation, and numbers. Following this, tokenization was done to split each tweet into individual terms. These terms are then lemmatized to then create a set of unigrams, bigrams, and trigrams that will make up all of the terms for which the LDA topic model calculates document-term and term-topic probabilities.

Before building the language model, it is useful to understand the distribution of tweet length; this translates into observing the distribution of tweet tokens:

Chart, histogram

Description automatically generated

The number of tweets per token has a slight positive skew; based on the interquartile range of the distribution, most documents (tweets) being fed into the LDA model will have anywhere from 6 to 16 tokens.

Methods:

Topic modeling using Latent Dirichlet Allocation (LDA):

Before creating the LDA model, a dictionary that maps each term in the corpus to an id and a corpus that describes the term-tweet frequency are generated. In order to determine the best number of topics to ask the model for, a grid-search style check will be conducted to calculate the resulting perplexity and coherence measures from each model. Perplexity serves as an intrinsic evaluation metric to easily measure the quality of a language model. It is applied on a holdout set of the document dataset and is a function of the probability that a language model assigns to the test corpus (https://arxiv.org/pdf/1601.00248.pdf). Topic coherence, or simply coherence, measures how similar words with high probabilities of belonging to a topic are to one another, serving as a measure to define a language model’s semantic interpretability (<http://svn.aksw.org/papers/2015/WSDM_Topic_Evaluation/public.pdf>).

[after and outside the methods section] Planned experiments:

After assessing numbers of topics according to the abovementioned metrics, topics that emerge from the LDA model can be observed in the following ways:

* Examine the key terms in each topic to see how each term contributes to the model given the weights assigned to them.
* Assign each topic with a human-interpretable label.
* Assign each tweet to its dominant topic and create average topic scores for each airline. This will be visualized in the final interactive dashboard within a heatmap that shows airline-topic scores, where the user will be able to select up to a certain number of airlines to visualize their topics at one time.
* Visualize the top terms per topic across all airlines, by plotting each topic-term’s importance to the topic against the count of tweets in the topic that contain the term, using a bubble chart. A bubble’s color will represent an aggregated measure of user engagement with the tweet (i.e. number of likes, retweets, etc.) and a bubble’s size will represent the term’s frequency out of the top terms being visualized for the topic.

~~These endpoints only allow 800 tweets~~

* ~~Data acquisition and downloading:~~

1. ~~Use the Twitter V2 API to download tweets (excluding replies, so including organic user tweets and retweets) in reverse chronological order from 85 Twitter accounts belonging to airline companies worldwide. (Maybe show breakdown of airlines according to the continent they belong to, or domestic airlines vs international airlines)~~
   1. ~~Documentation:~~ [~~https://developer.twitter.com/en/docs/twitter-api/tweets/timelines/api-reference/get-users-id-tweets~~](https://developer.twitter.com/en/docs/twitter-api/tweets/timelines/api-reference/get-users-id-tweets)
   2. ~~Starter code:~~ [~~https://github.com/twitterdev/Twitter-API-v2-sample-code/blob/main/User-Tweet-Timeline/user\_tweets.py~~](https://github.com/twitterdev/Twitter-API-v2-sample-code/blob/main/User-Tweet-Timeline/user_tweets.py)
2. ~~Retrieved all or the most recent 3200 tweets from each airline’s Twitter user timeline~~
3. ~~In total: 67,532 tweets (before culling down to English tweets), sourced on December 4, 2022.~~ 
   1. ~~English tweets: 59,848~~
4. ~~How many airlines had exactly 800 tweets come in? How many had less, and what was the distribution of number of tweets for those airlines?~~

Method stuff:

* Methods

Evaluation

* Coherence score: <https://www.cs.toronto.edu/~jstolee/projects/topic.pdf>

System:

* Updated system diagram:
  + Timeline scraper (output -> set of jsons with tweets and tweet info for each airline’s most recent 800 self-made tweets or retweets)
    - Retweets can matter too because they reflect something that they would like a customer to see from following the airline
    - This is different from replies, which the customer does not usually see unless they are involved in the tweet conversation or if they are mentioned in the tweets
  + Tweet information parser (output -> tabular dataset)

~~References:~~

* ~~Something about airline related topic modeling (~~[~~https://doi.org/10.3390/info12020078~~](https://doi.org/10.3390/info12020078)~~)~~
* ~~Sentiment analysis on airline tweets (~~[~~https://ieeexplore.ieee.org/document/8377739~~](https://ieeexplore.ieee.org/document/8377739)~~)~~
* ~~Something about topic modeling on Twitter (~~[~~https://www.cs.toronto.edu/~jstolee/projects/topic.pdf~~](https://www.cs.toronto.edu/~jstolee/projects/topic.pdf)~~)~~
* ~~Something about LDA (~~[~~https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf~~](https://www.jmlr.org/papers/volume3/blei03a/blei03a.pdf)~~)~~
* ~~Tailoring tweets or Twitter personality to what may interest potential customer base (~~[~~https://www.washingtonpost.com/travel/2022/10/04/ryanair-twitter-strategy-gen-z/~~](https://www.washingtonpost.com/travel/2022/10/04/ryanair-twitter-strategy-gen-z/)~~)~~
* ~~Marketing campaigns for building brand awareness in customers (Delta BLM example;~~ [~~https://www.kambr.com/articles/how-airlines-embrace-of-social-media-is-evolving-after~~](https://www.kambr.com/articles/how-airlines-embrace-of-social-media-is-evolving-after)~~)~~
* ~~Analysis of customers that follow/interact with an airline’s Twitter account: (~~[~~https://www.sciencedirect.com/science/article/abs/pii/S0969699718302072~~](https://www.sciencedirect.com/science/article/abs/pii/S0969699718302072)~~)~~