Beyond: Undergraduate Research Journal

Volume 7 Article 3

July 2023

Analysis of US Airline Stocks Performance using Latent Dirichlet Allocation (LDA)

Amina Issoufou Anaroua
Embry Riddle Aeronautical University, anarouaa@my.erau.edu

Recommended Citation

Issoufou Anaroua, Amina (2023) "Analysis of US Airline Stocks Performance using Latent Dirichlet Allocation (LDA)," *Beyond: Undergraduate Research Journal*: Vol. 7 , Article 3.
Available at: https://commons.erau.edu/beyond/vol7/iss1/3

Analysis of US Airline Stocks Performance using Latent Dirichlet Allocation (LDA)

Amina Issoufou Anaroua, Hari Adhikari, Ph.D.

Abstract

Various events, such as changes in the interest rate or the hijacking of a commercial aircraft, can lead to significant shifts in airline stock performance. This study aimed to measure the impact of aviation-related news announcements on the stock performance of US airlines, focusing on different topics. The dataset included aviation news covering airlines, airports, regulations, safety, accidents, manufacturers, MRO, incidents, aviation training, general aviation, and others obtained from Aviation Voice. To uncover patterns that could explain the movements of US airline stocks, a natural language processing technique called Latent Dirichlet Allocation (LDA) was employed. The process involved text mining and topic modeling of the aviation-related data. The LDA model was then utilized to identify and capture specific topics mentioned in the news releases from Aviation Voice. By investigating the links between stock returns and the identified topics, the study revealed significant variations in financial performance across different topics. Notably, topics related to technology, fuel, and training positively impacted the short and long-term moving averages of US airline stocks. On the other hand, topics related to defense and travel costs only influenced the medium-term run. These findings shed light on the factors that influence US airline stock performance and provide valuable insights for investors and industry stakeholders.

Introduction

Online platforms such as blogs, news, and social networks are fed with a plethora of daily information. These platforms quickly spread large amounts of information and are useful in providing evidence or answering questions if they can be systematically analyzed in a less costly and more scientific way. On this ground, this research uses digitized information related to the aviation industry to analyze the stock performance of United States (US) airlines. Determining factors that influence stock returns has always been a critical research subject. Feuerriegel et al. (2016) state that it is appealing to include quantitative and qualitative information to explain stock returns due to financial events. News announcements contain an enormous amount of qualitative details that can be useful in explaining stock price movements. For example, the wording released by a corporate press might indicate the firm's future performance. When judging the fair price of a stock, news plays a crucial role for investors. Feuerriegel and Pröllochs (2021) discuss that knowing the news stories that matter to the audience and are more relevant than others is valuable.

A literature review revealed that prior studies (Humpherys et al., 2011; Huang et al., 2014; Loughran and McDonald, 2016; Chen et al., 2017) have identified hidden signals in news announcements. These studies have also established the significant influence of news announcements on stock performance (Hendershott et al., 2015). To enhance the accuracy of stock price predictions, previous research employed deep learning models like Recurrent Neural Network (RNN) and Facebook Prophet, utilizing substantial time series data. Inspired by these findings, this study investigates the impact of aviation-related news announcements on the performance of US airline stocks. Latent Dirichlet Allocation (LDA) is employed to analyze the magnitude, content, and trends in textual aviation news announcements. Specifically, the study focuses on identifying the topics within the aviation news corpus that affect the stock returns of US airlines. Finally, the influence of these extracted topics on the stock returns of US airlines is examined.

Aviation plays a vital role in promoting tourism, fostering economic growth, and facilitating international trade. It serves as a crucial link that connects businesses, cultures, and people worldwide. In today's globalized world, the convenience, efficiency, and speed of airlines are essential for seamless connectivity. However, airlines stand apart from other businesses in several aspects, and are particularly vulnerable to economic conditions. Firstly, they require substantial capital for aircraft acquisition or leasing, maintaining robust information systems, and managing crews. Secondly, the industry is exposed to various external factors like fuel price fluctuations, market volatility, disease outbreaks, and terrorism. During a recession, passengers tend to opt for cheaper alternatives like cars or trains, resulting in reduced travel and fewer planned business trips, leaving airlines with empty seats and facing liquidity challenges.

Despite deregulation, the airline industry still encounters numerous regulatory and legal constraints. Consequently, the choice to study the airline industry and its correlation with news in influencing stock prices offers a unique contribution to the literature, holding managerial implications.

The paper makes two main contributions. Firstly, it establishes a novel link between news and its impact on airline stock performance. Additionally, natural language processing techniques, such as LDA, are utilized to investigate these relationships. As a result, this research opens a new avenue for exploration in the field. Furthermore, the study's discovery that investors exhibit preferences for specific news themes when investing in airline stocks may offer valuable insights into short-term movements in airline stock prices.

This paper is organized into different sections as follows. Section two reviews previous studies using Latent Dirichlet Allocation to discover topics and how they explain the relationship between news disclosures and stock market return. Section three describes the research methodology to analyze the cumulative abnormal stock returns over three, five, seven, 15, and 21 days corresponding to the topics from aviation news articles. Section four presents the findings from applying the methodology discussed in Section three. Section five describes the importance and contributions of the study. And finally, Section six summarizes the findings and provides suggestions for future research.

Literature Review

Fang et al. (2018) used the LDA to obtain the latent topic from an extensive collection of research abstracts. They performed a regression analysis on the document-topic to categorize the topics into cold and hot ones. One of their most important findings is that LDA discovered topics that are highly consistent with the ones indexed by human experts. Recent research by Xu et al. (2020) used machine learning techniques to analyze more than one million tweets to understand the psychological reactions and users' discourse related to COVID-19. They identified and categorized topics into themes. The results of that study did not reveal messages on symptoms and treatments as dominant topics. However, the analysis demonstrated that fear related to the unknown nature of COVID-19 is prevalent in all topics. In addition, Glasserman et al. (2020) evaluated methods for selecting topics in news articles to justify stock returns. Their empirical and theoretical results are that supervised Latent Dirichlet Allocation (sLDA) often overfits returns to the detriment of the topic model. The plain LDA models provided better results out of random sample searches. After testing thousands of news articles related to Standard & Poor's (S&P) 500 index firms, they found supervised topics to explain volatility, stock returns, and other types of labels. Their research approach improved the prediction performance of a considerable corpus of news articles related to business.

Feuerriegel et al. (2016) presented the

effects of topics found from ad hoc announcements on stock market returns in the German market. They successfully extracted many topics into different groups. Their paper showed that certain topics do not affect abnormal returns of stocks while others have a significant effect. Even though Feuerriegel and Pröllochs (2021) use the same approach as data mining using regulatory 8-K filings from U.S. companies, their empirical evidence indicates a considerable inconsistency among news stories in terms of their impact and relevance on financial markets. However, they found a statistically meaningful abnormal return for disclosures regarding business strategy, earning results, mergers and acquisitions, credit rating, and the health sector. Furthermore, Mahajan et al. (2008) proposed a stock market analysis system that studied financial news to find and classify the main events that affect the market. The study argued that the system predicts whether the stock market will rise, or fall based on current news articles.

The application of the LDA technique for topic modeling and discovering topics from the document collection aligns with the methodologies discussed in previous articles. Nonetheless, this paper adopts a slightly different approach to elucidate the relationship between news and stock returns. For instance, Fang et al. (2018) utilized regression analysis, which is similar to the approach taken in this study. However, there are distinctions between their work and this paper. While they classified topics as hot or cold, this paper combines regression analysis with LDA to examine how topics influence stock returns over time. Furthermore, this research delves into the magnitude of the impact of topics on airline stock returns, making a valuable contribution to the existing literature.

Research Methodology

The study employed LDA, a technique initially developed by Blei et al. (2003). Topic modeling was utilized to analyze the content of aviation news announcements and identify topics that could explain stock performance. LDA was chosen as it can encompass a wide range of news releases related to aviation topics. The research methodology involved text mining, followed by the application of LDA, and finally conducting multiple regression analyses using cumulative abnormal returns corresponding to publications. Figure 1 below provides a detailed overview of the methodology used.

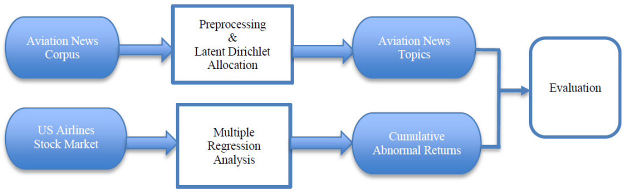

Aviation News Corpus
Preprocessing & Latent Dirichlet Allocation
Aviation News Topics
Evaluation
US Airlines Stock Market
Multiple Regression Analysis
Cumulative Abnormal Returns

Figure 1: Methodology for studying variations of airlines stock returns across topics in aviation news

Mining Process

Text Mining

Text mining is a widely used technique that assesses a sizable collection of documents to extract the information one seeks or learn new information (Linguamatics, n.d.). This technique was used to identify relationships, facts, and assertions within the set of data, which is too large for manual processing. Text mining is a preferred method to efficiently analyze such an extensive data set. The extracted texts from news related to the aviation industry were transformed into a better structure for analysis. Before continuing to the topic modeling, the following preprocessing steps were applied:

- **Data cleaning:** The news articles contain some information and words at the beginning and end of most of the articles. This general information was removed as it is not pertinent to the underlying events. The main body text of the articles was kept as it is the focus. Moreover, new line characters were removed along with distracting single quotes. Then, a regular expression library was used to lowercase the text.

- **Tokenization:** Each sentence was tokenized into a list of words, removing punctuation and unnecessary characters altogether. In the process of tokenization, some characters like punctuation marks are discarded. *Gensim* was used to remove punctuation.

- **Stemming:** Words were reduced to their word stems that affixes to suffixes and prefixes, or to the roots of words known as a lemma. In this way, the total number of unique words in the dictionary was reduced. In the next step, *CountVectorizer* creates the number of columns in the document-word matrix that are denser with lesser columns. The process helps to get well-classified topics at the end. The *Spacy* package was used as it is better than *PorterStemmer* or *Snowball*.

- **Creation of document-word matrix:** The LDA topic model algorithm needs a document word matrix as the principal input. One was created using *CountVectorizer*. It was configured to consider words that have occurred at least ten times (min_df), then remove built-in English stopwords, and convert all words to lowercase. To be qualified as a word, the words must contain numbers and alphabets of at least a length of three.

Topic Modeling

Topic modeling is the most appropriate method for developing a model to accurately predict future stock performance (Tong and Zhang, 2016). The topic modeling method is effective for this study to automatically search, understand, organize, and summarize the extensive electronic archives. In addition, the current study attempted to search for information of specific interest, aviation news announcements. The word "topic" represents implicit themes to be estimated in documents (Tong and Zhang, 2016). Topic modeling is used to find the probability of occurrence of specific topics in a collection of aviation-related news. Then, from those topics, words are extracted. Finally, a document coverage distribution of topics is generated to explore topics' perspectives on the data.

Latent Dirichlet Allocation

LDA is a probabilistic model which generates explanation(s) to sets of data as to why there are similarities, if there is any, by uncovering underlying implicit topics from the documents (Tong and Zhang, 2016). With LDA, the researcher assigned a label to a topic found by the computer based on words. LDA topic modeling is employed to analyze the content of aviation news announcements. LDA estimates the number of topics in the entire announcement as well as the proportion of those topics. For topic extraction, LDA considers K pre-specified topics, $\beta_{(1:K)}$, in the collection of documents for research, where a distribution over a fixed vocabulary is defined for every topic, $\beta_{(k)}$ (Feuerriegel et al. ,2016). In addition, a two-stage process is created for each document as shown in Figure 2.

1. For each document, a distribution over topics is selected randomly:
   - $\theta_{(d)}$, topic distribution for document d
   - $\theta_{(d,k)}$, proportion of topic k in document d

2. Next, the process continues for each word in the document:
   - From the distribution, $\theta_{(d)}$, a topic is randomly selected
     - $z_{(d,n)}$, topic assignment for the nth word in document d
     - $z_{(d)}$, topic assignment for all the words in document d
   - From the distribution, a word is randomly selected over the fixed vocabulary corresponding to the topic assignment, $z_{(d,n)}$
     - $w_{(d,n)}$, nth word in document d
     - $w_{(d)}$, words observed in document d

Based on this notation, the allocation generative process for LDA can be represented as Equation 1 below, from Feuerriegel et al. (2016).

$$P(\beta_{1:K}, \theta_{1:D}, Z_{1:D}, W_{1:D}) = \prod_{i=1}^{K} P(\beta_i) \prod_{d=1}^{D} P(\theta_d) \left(\prod_{n=1}^{N} P(Z_{d,n} | \theta_d) P(W_{d,n} | \beta_{1:K}, Z_{d,n}) \right) \quad (1)$$

From the above equation, the probability of observing a word, $w_{(d,\ n)}$, depends on the topic assignment, $z_{(d,\ n)}$, and all the topics, $\beta_{(1:\ K)}$.

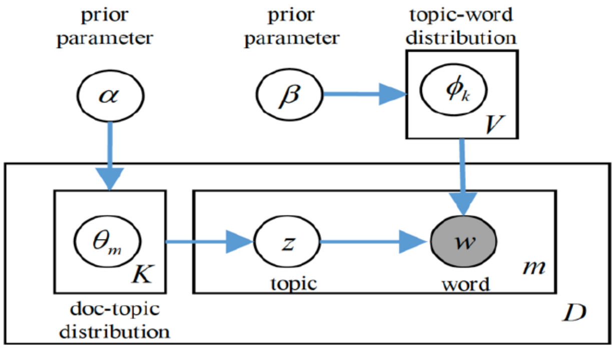


Figure 2: LDA graph model from Jelodar et al. (2019)

As shown in Figure 2, LDA views each document, d_m, containing N_m words in a news corpus, D, as a mixture of K different topics that are represented by a K-dimensional "document-topic" distribution, θ_m. Each topic, k, is considered as a mixture of V words that are represented by a V-dimensional "topic-word" distribution, ϕ_k. The generative process of LDA is as follows:

1. For each document, d_m, LDA draws a "document-topic" distribution, θ_m, over all K topics, where α describes the prior observation for the "document-topic" count

2. For each topic k, LDA draws a "topic-word" distribution, ϕ_k, over all V words, where β describes the prior observation for the "topic-word" count.

3. For each word. w_i in d_m, for i between 1 and N_m, LDA samples a topic, k, and a word, V.

The goal of using LDA is to find the hidden structure topics by naming topic distribution over each document, and naming topic assignment and each word over each document.

After using the document probability distribution over each topic, the topics primarily discussed were found and the number of times each topic is involved in the document.

Multiple Regression Analysis

A multiple regression model was used to analyze the relationship between the single dependent variable (cumulative abnormal returns) and several independent variables (topics from news articles). This model attempts to discover the significance of the relationship between the stock price and the associated type of news. Cumulative abnormal returns over certain periods were computed to investigate stock price reactions to news disclosures. The significance of events and news associated with the aviation industry (e.g., regulations, accidents, incidents, etc.) to airlines' stock prices are analyzed using this method.

The process is as follow:

- **Normal returns:** The expected return in the absence of a news announcement based on the market model (Konchitchki and O'Leary, 2011). The market model is defined as shown in Equation 2 from Feuerriegel et al. (2016):
  - $R(\tau) = \alpha + \beta R_m(\tau) + \varepsilon(2)$
  - Where α, firm's recent performance track record
  - β, sensitivity to general market movements
  - R_m, market return using S&P 500 Index
  - R, airlines stock returns using Dow Jones U.S. Airlines Index, DJUSAR
  - ε, zero mean disturbance term
  - τ, time

- **Abnormal returns:** The difference between the actual and expected return on a given day, shown as Equation 3:
  - $AR(\tau) = R(\tau)- E[R(\tau)](3)$
  - Where $R(\tau)$, actual return, measured by change of price of a security
  - $AR(\tau)$, abnormal return at time τ
  - $E\ [R\ (\tau)]$, expected return using the market model

- **Cumulative Abnormal Returns (CAR):** The sum of the abnormal returns over three, five, seven, and 15 days, as shown in Equation 4.

  - $\mathbf{CAR_j = \sum_{i=0}^{j}(AR_{i,j})(4)}$
  - Where j, simple moving average over three, five, seven, 15, and 21 days
  - i, sum of the daily abnormal returns

- **Multiple Regression:** Used to find the relationship between the cumulative abnormal returns and the different topics across the news articles over time, shown as Equation 5.
  - $\mathbf{Log(CAR_j) = \alpha + \sum_{i=0}^{j}(\beta_{i,j}Topic_{i,j})+u_{i,j}(5)}$
  - Where β, α, unknown parameters
  - u, error terms
  - i, from one to ten for the ten news topics
  - j, for three, five, seven, 15, and 21-days rolling average of abnormal returns
  - Topic, news topics

Evaluation

This section presents the research findings regarding the analysis of how major aviation events affected the performance of US airline stocks. The previously introduced approach was used to conduct topic extraction from the data. It was hypothesized that some aviation news would affect the performance of US airline stocks. Topics were identified that triggered significant stock price changes. In the next section, the aviation news corpus used in the analysis is introduced. Furthermore, the extracted topics from Latent Dirichlet Allocation are presented and investigated for their impact on US airlines' market prices.

Corpus

A corpus of aviation news was created as follows: First, aviation news was downloaded from Aviation Voice. Then, the news from the beginning of January 2016 to the end of December 2019 was considered. A total of 1716 news articles was obtained from scraping the website. This time frame was chosen to avoid accounting only for news affected by a single market event. This period provided the possibility of exploring significant events in aviation. Many filtering steps were applied to the news corpus. All articles with fewer than 100 words were removed from the data. Only news articles that were in English were kept. Some unnecessary characters were removed from the texts. A total of 1549 articles were kept at the end. Understanding the data and being on the right track was vital to see if more preprocessing was needed before training the model.

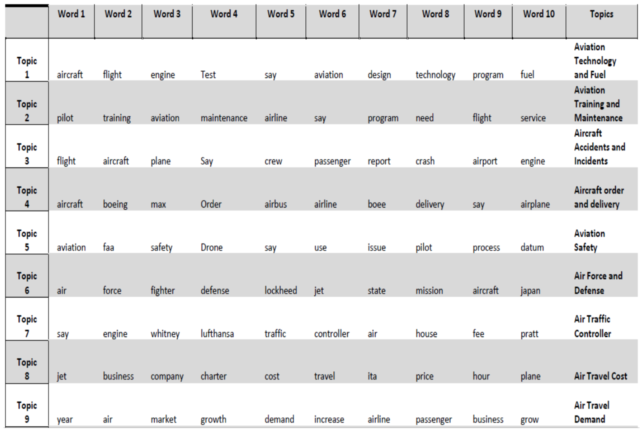
For the financial data, it was collected from Investing.com. Dow Jones US Airlines Index (DJUSAR) and S&P 500 were retrieved as the benchmark corresponding to the aviation news dates from 2016 to 2019 to be accurate, coherent, and consistent in the analysis. DJUSAR was chosen to account for all the airlines included in the index, and the research concerned mainly the airlines contained in it. The data was restricted up to 2019, to exclude the COVID-19 pandemic that affected the entire market. Daily stock returns were collected for all trading days in the sample.

Topic Extraction from Aviation News

The title of an aviation news article did not include a topic label or code clarifying the theme of its content. For this purpose, it was part of the analysis to find the corresponding topics using LDA on the created document word matrix from the aviation news corpus. All data processing and modeling were performed with the python programming language in Jupyter notebook.

The most crucial tuning parameter for LDA models was n_components, representing the number of topics. To determine the optimal number of topics, several methods existed, such as empirical likelihood (Li and McCallum, 2006), marginal likelihood (Newton and Raftery, 1994, Griffiths and Steyvers, 2004, and Wallach, 2006), perplexity (Blei et al., 2003), and hierarchical Dirichlet processes (Teh et al., 2006), among others. In this study, a grid search was employed to discover the optimal number of topics for the model. Subsequently, ten topics were identified, and the top ten words per topic were extracted.

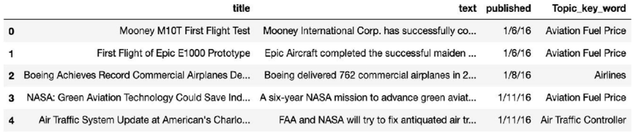
Before validating the topics obtained from topic modeling, the topics had to be labeled. There were certain automatic labeling methods (Mei et al., 2007). However, these methods were not convenient for this study, where the labeling needed domain knowledge, being aviation knowledge. To ensure quality labeling, most topic model researchers (Chang et al., 2009) inferred the topic and labeled manually. A manual labeling procedure for each topic was performed, and each was assigned a meaningful name based on the terms and the content of each document. For example, topic one was named Aviation Training and Maintenance because its top ten stemmed terms were "pilot," "training," "aviation," "maintenance," "airline," "say," "program," "need," "flight," and "service". It was reasonable to choose the topic with the highest probability as selected topics had a probability of close to 100% for most news. The topics were inferred according to their keywords and put into the data frame. Table 1 below shows the top ten words with assigned topics.



	Word 1	Word 2	Word 3	Word 4	Word 5	Word 6	Word 7	Word 8	Word 9	Word 10	Topics
Topic 1	aircraft	flight	engine	Test	say	aviation	design	technology	program	fuel	Aviation Technology and Fuel
Topic 2	pilot	training	aviation	maintenance	airline	say	program	need	flight	service	Aviation Training and Maintenance
Topic 3	flight	aircraft	plane	Say	crew	passenger	report	crash	airport	engine	Aircraft Accidents and Incidents
Topic 4	aircraft	boeing	max	Order	airbus	airline	boee	delivery	say	airplane	Aircraft order and delivery
Topic 5	aviation	faa	safety	Drone	say	use	issue	pilot	process	datum	Aviation Safety
Topic 6	air	force	fighter	defense	lockheed	jet	state	mission	aircraft	japan	Air Force and Defense
Topic 7	say	engine	whitney	lufthansa	traffic	controller	air	house	fee	pratt	Air Traffic Controller
Topic 8	jet	business	company	charter	cost	travel	ita	price	hour	plane	Air Travel Cost
Topic 9	year	air	market	growth	demand	increase	airline	passenger	business	grow	Air Travel Demand

Table 1: Inferred dominant topics by keywords and highest probability

The topic LDA was built and took the text through the same routine of transformations done for the words before predicting the document topic. Therefore, each document within the data frame was assigned the dominant topic, a topic with the highest probability. Table 2 below shows the predicted topics for each article of the aviation news within the original dataset. In the section of Stock Market Response across topics, each news document was connected to the corresponding DJUSAR abnormal return.



	title	text	published	Topic_key_word
0	Mooney M10T First Flight Test	Mooney International Corp. has successfully co...	1/6/16	Aviation Fuel Price
1	First Flight of Epic E1000 Prototype	Epic Aircraft completed the successful maiden ...	1/6/16	Aviation Fuel Price
2	Boeing Achieves Record Commercial Airplanes De...	Boeing delivered 762 commercial airplanes in 2...	1/8/16	Airlines
3	NASA: Green Aviation Technology Could Save Ind...	A six-year NASA mission to advance green aviat...	1/11/16	Aviation Fuel Price
4	Air Traffic System Update at American's Charlo...	FAA and NASA will try to fix antiquated air tr...	1/11/16	Air Traffic Controller

Table 2: Prediction result of the original dataset

A popular visualization package was used to help interactively understand the relationships between the topics and interpret individual topics. The topic visualization helped to select each topic to view its topmost frequent terms using different values of the λ parameter. Figure 3 below shows topic visualization with parameter λ=1 and topic two selected. The Intertopic Distance Plot was also explored to help learn about how topics related to each other and the possibility of higher-level structure between groups of topics. The importance of each topic over the entire news corpus was represented by the area of the circle. The distance between the center of circles revealed the similarity among topics. For each topic, the histogram on the right side listed the top 30 most important terms with their estimated level of frequency for the selected topic two.

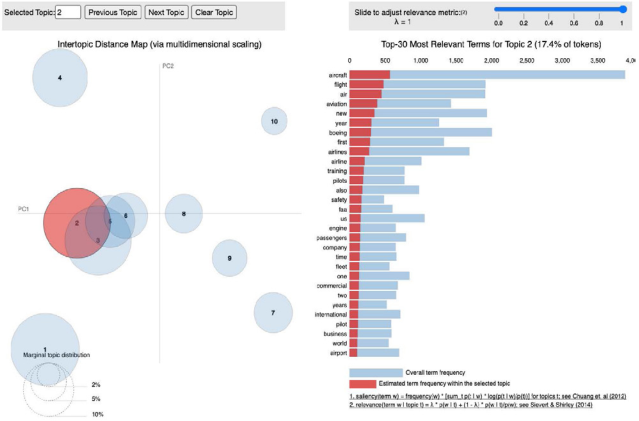
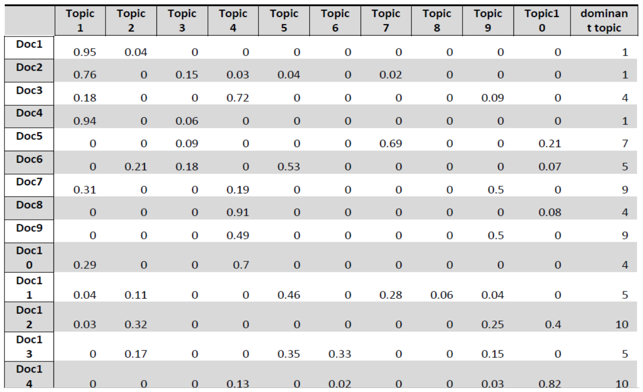

Selected Topic: 2
Previous Topic
Next Topic
Clear Topic
Slide to adjust relevance metric:(2)
λ = 1
0.0
0.2
0.4
0.6
0.8
1
Intertopic Distance Map (via multidimensional scaling)
PC2
PC1
4
10
2
5
6
3
8
9
7
1
Marginal topic distribution
2%
5%
10%
Top-30 Most Relevant Terms for Topic 2 (17.4% of tokens)
0
500
1,000
1,500
2,000
2,500
3,000
3,500
4,0
aircraft
flight
air
aviation
new
year
boeing
first
airlines
airline
training
pilots
also
safety
faa
us
engine
passengers
company
time
fleet
one
commercial
two
years
international
pilot
business
world
airport
Overall term frequency
Estimated term frequency within the selected topic
1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)
2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

Figure 3: Data visualization for topic modeling

Stock Market Response Across Topics

The reaction of DJUSAR's cumulative abnormal returns to the disclosed news across the extracted topics is now analyzed. CAR over three, five, seven, and 15 days was analyzed to investigate the magnitude of the relationship. The classification of a news article belonging to a particular topic followed a logical approach in the topic modeling. The dominant topic was assigned by looking at the topic with the highest contribution to the news article. Each document or news article had a probability distribution per topic. Table 3 below shows the topic probability distribution per document. For example, topic one has 95% contribution to document one while other topics had nearly 0% probability contribution, so the model assigned topic one as a dominant topic to document one. The model found the probability of each topic belonging to a document and assigned those topics with their probability to all the document from the news corpus.



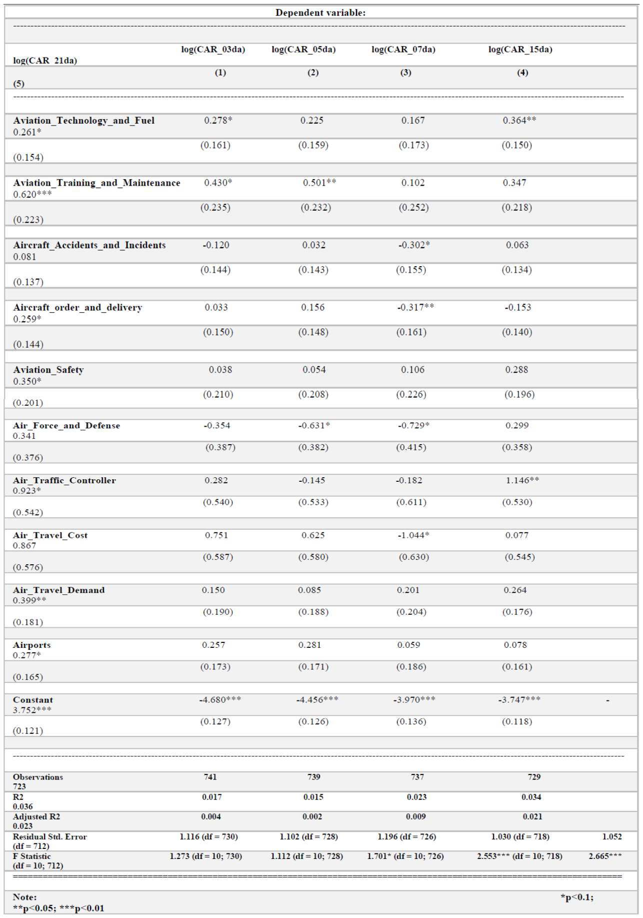
	Topic 1	Topic 2	Topic 3	Topic 4	Topic 5	Topic 6	Topic 7	Topic 8	Topic 9	Topic10	dominant topic
Doc1	0.95	0.04	0	0	0	0	0	0	0	0	1
Doc2	0.76	0	0.15	0.03	0.04	0	0.02	0	0	0	1
Doc3	0.18	0	0	0.72	0	0	0	0	0.09	0	4
Doc4	0.94	0	0.06	0	0	0	0	0	0	0	1
Doc5	0	0	0.09	0	0	0	0.69	0	0	0.21	7
Doc6	0	0.21	0.18	0	0.53	0	0	0	0	0.07	5
Doc7	0.31	0	0	0.19	0	0	0	0	0.5	0	9
Doc8	0	0	0	0.91	0	0	0	0	0	0.08	4
Doc9	0	0	0	0.49	0	0	0	0	0.5	0	9
Doc10	0.29	0	0	0.7	0	0	0	0	0	0	4
Doc11	0.04	0.11	0	0	0.46	0	0.28	0.06	0.04	0	5
Doc12	0.03	0.32	0	0	0	0	0	0	0.25	0.4	10
Doc13	0	0.17	0	0	0.35	0.33	0	0	0.15	0	5
Doc14	0	0	0	0.13	0	0.02	0	0	0.03	0.82	10

Table 3: Topics Probability Distribution Matrix

As shown in Table 3, the Excel file was downloaded from the topic modeling with all the documents and their topic's probability distribution. The dominant topic variables were transformed as indicator variables to represent each topic. R was used to calculate the cumulative abnormal returns over three, five, seven, and 15 days. To run the multiple regression models, the topic indicators were used as the explanatory variables and the CAR as the dependent variable to explore how CAR can be explained by dominant topics selected.

As shown in the results from Table 4, a logarithmic transformation was used for the dependent variable in the models as it is one of the most used transformations in statistical analysis and it also makes the data more symmetric. For robustness, similar results were obtained by taking the logarithmic transform of dependent and independent variables as shown in Table A1. Five models were estimated using CAR over each of three, five, seven, and 21 days to determine the impact and the magnitude of topic indicators.

As shown in Table 4, Aviation Technology and Fuel (Topic$_1$) and Aviation Training and Maintenance (Topic$_2$) were statistically significant at a 10% significance level to explain the CAR over day three. For example, the Aviation Technology and Fuel topic had a coefficient of 0.278. For every one percent increase in the topic in the press, CAR increases by about 32%. The Aviation Technology and Fuel topic was positively significant in explaining CAR over 15 and 21 days.



	Dependent variable:				
	log(CAR_03da)	log(CAR_05da)	log(CAR_07da)	log(CAR_15da)	log(CAR_21da)
	(1)	(2)	(3)	(4)	(5)
Aviation_Technology_and_Fuel	0.278*	0.225	0.167	0.364**	0.261*
	(0.161)	(0.159)	(0.173)	(0.150)	(0.154)
Aviation_Training_and_Maintenance	0.430*	0.501**	0.102	0.347	0.620***
	(0.235)	(0.232)	(0.252)	(0.218)	(0.223)
Aircraft_Accidents_and_Incidents	-0.120	0.032	-0.302*	0.063	0.081
	(0.144)	(0.143)	(0.155)	(0.134)	(0.137)
Aircraft_order_and_delivery	0.033	0.156	-0.317**	-0.153	0.259*
	(0.150)	(0.148)	(0.161)	(0.140)	(0.144)
Aviation_Safety	0.038	0.054	0.106	0.288	0.350*
	(0.210)	(0.208)	(0.226)	(0.196)	(0.201)
Air_Force_and_Defense	-0.354	-0.631*	-0.729*	0.299	0.341
	(0.387)	(0.382)	(0.415)	(0.358)	(0.376)
Air_Traffic_Controller	0.282	-0.145	-0.182	1.146**	0.923*
	(0.540)	(0.533)	(0.611)	(0.530)	(0.542)
Air_Travel_Cost	0.751	0.625	-1.044*	0.077	0.867
	(0.587)	(0.580)	(0.630)	(0.545)	(0.576)
Air_Travel_Demand	0.150	0.085	0.201	0.264	0.399**
	(0.190)	(0.188)	(0.204)	(0.176)	(0.181)
Airports	0.257	0.281	0.059	0.078	0.277*
	(0.173)	(0.171)	(0.186)	(0.161)	(0.165)
Constant	-4.680***	-4.456***	-3.970***	-3.747***	-3.752***
	(0.127)	(0.126)	(0.136)	(0.118)	(0.121)
Observations	**741**	**739**	**737**	**729**	**723**
R2	**0.017**	**0.015**	**0.023**	**0.034**	**0.036**
Adjusted R2	**0.004**	**0.002**	**0.009**	**0.021**	**0.023**
Residual Std. Error	**1.116 (df = 730)**	**1.102 (df = 728)**	**1.196 (df = 726)**	**1.030 (df = 718)**	**1.052 (df = 712)**
F Statistic	**1.273 (df = 10; 730)**	**1.112 (df = 10; 728)**	**1.701* (df = 10; 726)**	**2.553*** (df = 10; 718)**	**2.665*** (df = 10; 712)**
Note:					***p<0.1; **p<0.05; ***p<0.01**

Table 4: Results of multiple regression models

The results of estimated coefficients of the five different models were summarized in Table 5 for straightforward interpretation.

	CAR03	CAR05	CAR07	CAR15	CAR21
Aviation Technology and Fuel	+			+	+
Aviation Training and Maintenance	+	+			+
Aircraft Accidents and Incidents			-		
Aircraft order and delivery			-		+
Aviation Safety					+
Air Force and Defense		-	-		
Air Traffic Controller				+	+
Air Travel Cost			-		
Air Travel Demand					+
Airports					+

Table 5: Summary of multiple regression models

Table 5 shows that all ten topics were significant determinants of CAR over different moving average periods. Three and five days are considered as short-term moving averages, seven days as medium term, and 15, and 21 days as long-term moving averages. Overall, a positive link between the topics related to technology and fuel, as well as training and maintenance was noted. In the medium term, safety related topics, topics related to defense and costs were negatively associated with cumulative abnormal returns.

Topic$_1$ is related to technology, fuel, aircraft, and engine, while **Topic$_2$** is related to pilot, training, and maintenance. Both topics were positively correlated to the stock returns in the short and long run. The results indicated that a one-unit increase related to these topics in press led to an increase by a certain percentage in cumulative stock returns. It was noticeable that in the long run, these estimated coefficients were slightly higher, showing that a change in those values led to a relatively higher stock return.

Topic$_4$ (delivery, aircraft, Boeing, max), **Topic$_5$** (Safety, FAA), **Topic$_7$** (air, traffic, controller), **Topic$_9$** (market, growth, passenger, demand), and **Topic$_{10}$** (flight, airport, service, route) had prolonged effects on stock returns as these variables were significantly positive only in the long-term moving average estimation.

Topic$_3$ (passenger, aircraft, crew, airport, crash), **Topic$_6$** (defense, fighter, mission, force), and **Topic$_8$** (business, company, travel, cost, price) had negative effects on the stock return in the medium term of the moving averages. For example, Topic$_8$ had a coefficient of -1.044. For every one-unit increase in Topic$_8$ in the press, CAR decreased by about 65%.

Conclusion

Developing a reliable model to predict stock performance is a challenge to many researchers. Investigating how stock returns are influenced or related to events released in the news is another challenge as text data was difficult to quantify or analyze. A technique to extract topics from new announcements in a text format and explore the relationship between aviation-related news announcements and the stock prices of US airlines was employed. The research provides a good understanding of how airline stocks respond to particular news in aviation over a certain period. Therefore, this analysis offers practical insights on investment in airline stocks by understanding how varying critical aviation news influences the financial performance of the airline market. In addition, this research provides empirical evidence on the impacts of news on stock returns, which leads to a new avenue for applying the LDA approach in research using text data.

It is evident that the content of news announcements conveys information that are reflected in the stock market prices. However, the news that transmits important messages affects stock prices differently over time. Thus, this paper aimed to contribute to the literature by investigating how news topics have varying impacts on stock performance. In particular, the impact of topics found in aviation news on US airline stocks were analyzed. To perform the analysis, the optimal number of topics based on the aviation news corpus were identified. Then, a topic was determined and assigned to each aviation news release. An LDA approach was used to find the optimal number of topics to be ten from the Aviation Voice corpus. After analyzing the most common words related to the topics, a topic title was inferred and assigned. To estimate the US Airline market effect, the cumulative abnormal returns of DJUSAR corresponding to each news disclosure were calculated. The results indicate that the impact of the topics varies over time. Certain topics have no impact on US Airline stocks in the short and long term while other topics have significant effect. On the other hand, some topics only matter in the medium term for a short period of time.

The findings offer compelling evidence of a news story's impact on the airline's stock performance. These results hold practical and managerial value, as they demonstrate how a recently developed text mining methodology can be utilized to extract crucial information for comprehending an industry's financial performance. Moreover, this study paves the way for future research in two key areas. Firstly, the dataset used for this research comprises news articles collected from Aviation Voice. Expanding the scope to include electronic articles from other aviation sources over a more extended period may yield more comprehensive results. Consequently, additional research could be conducted to validate the robustness of the approach. Secondly, while inferring topics, the LDA model disregards the position of individual words. Thus, exploring alternative methods for topic extraction and comparing them with LDA should be a subject of investigation.

References

[1] Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *The Journal of Machine Learning Research*, 3, 993-1022.

[2] Chang J., Boyd-Graber JL., Gerrish S., Wang C., & Blei DM (2009). Reading tea leaves: How humans interpret topic models. Bengio Y, Schuurmans D, Lafferty JD, Williams CKI, Culotta A, eds. *Adv. Neural Inform. Processing Systems* (Curran Associates, New York), 288-296

[3] Fang, D., Yang, H., Gao, B., & Li, X. (2018). Discovering research topics from library electronic references using latent dirichlet allocation. *Library Hi Tech*, 36(3), 400-410.

[4] Feuerriegel, S., & Pröllochs, N. (2021). Investor reaction to financial disclosures across topics: An application of latent dirichlet allocation. *Decision Sciences*, 52(3), 608-628.

[5] Feuerriegel, S., Ratku, A., & Neumann, D. (2016). Analysis of how underlying topics in financial news affect stock prices using latent dirichlet allocation. *In 2016 49th Hawaii International Conference on System Sciences (HICSS)*, 1072-1081.

[6] Glasserman, P., Krstovski, K., Laliberte, P., & Mamaysky, H. (2020). Choosing news topics to explain stock market returns.

[7] Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences - PNAS, 101*(Suppl 1), 5228-5235.

[8] Hendershott, T., Livdan, D., & Schürhoff, N. (2015). Are institutions informed about news? *Journal of Financial Economics*, 117(2), 249-287.

[9] Jelodar, H., Wang, Y., Yuan, C., Xia, F., Jiang, X., Li, Y., & Zhao, L. (2019). Latent dirichlet allocation (LDA) and topic modeling: Models, applications, a survey. *Multimedia Tools and Applications*, 78(11), 15169-15211.

[10] Konchitchki, Y. & O'Leary, D., E. (2011). Event Study Methodologies in Information Systems Research. *International Journal of Accounting Information Systems.* 12(2), 99–115.

[11] Li, W. & McCallum, A. (2006), "Pachinko allocation: DAG-structured mixture models of topic correlations", *Proceedings of the 23rd International Conference on Machine Learning in Pittsburgh*, PA, ACM, New York, NY, pp. 577-584.

[12] Linguamatics. (n.d.). *What is text mining?* Retrieved February 26, 2021, from https://www.linguamatics.com/what-text-mining-text-analytics-and-natural-language-processing

[13] Mahajan, A., Dey, L., & Haque, S. (2008). Mining financial news for major events and their impacts on the market. Paper presented at the, 1 423-426.

[14] Mei QZ., Shen XH., & Zhai CX. (2007). Automatic labeling of multinomial topic models. Berkhin P, Caruana R, Wu X, eds. *The 13th ACM SIGKDD Internat. Conf. Knowledge Discovery and Data Mining* (ACM, New York), 490-499

[15] Newton, M.A. & Raftery, A.E. (1994). Approximate Bayesian-inference with the weighted likelihood bootstrap. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, Vol. 56 No. 1, pp. 3-48.

[16] Teh, Y.W., Jordan, M.I., Beal, M.J. & Blei, D.M. (2006). Hierarchical Dirichlet Processes. *Journal of the American Statistical Association*, Vol. 101 No. 476, pp. 1566-1581

[17] Tong, Z., & Zhang, H. (2016). A text mining research based on LDA topic modelling. *In International Conference on Computer Science, Engineering and Information Technology*, 201-210.

[18] Wallach, H.M. (2006). Topic modeling: beyond bag-of-words. *Proceedings of the 23rd International Conference on Machine Learning in Pittsburgh*, PA, ACM, New York, NY, pp. 977-984

[19] Xue, J., Chen, J., Chen, C., Zheng, C., Li, S., & Zhu, T. (2020). Public discourse and sentiment during the COVID 19 pandemic: Using latent dirichlet allocation for topic modeling on twitter. *PloS One*, 15(9), e0239441-e0239441.