11-2020

Exploring the impact of COVID-19 on aviation industry: A text mining approach

GOTTIPATI Swapna
Singapore Management University, SWAPNAG@smu.edu.sg

Kyong Jin SHIM
Singapore Management University, kjshim@smu.edu.sg

Weiling Angeline JIANG
Singapore Management University, wljiang.2019@mitb.smu.edu.sg

Sheng Wei Andre Justin LEE
Singapore Management University, andre.lee.2019@mitb.smu.edu.sg

Exploring the Impact of COVID-19 on Aviation Industry: A Text Mining Approach

Swapna Gottipati, Kyong Jin Shim, Angeline Weiling Jiang, Andre Justin Sheng Wei Lee
School of Information Systems
Singapore Management University
swapnag, kjshim, wljiang.2019, andre.lee.2019@smu.edu.sg

***Abstract*— Our study presents a comprehensive analysis of news articles from FlightGlobal website during the first half of 2020. Our analyses reveal useful insights on themes and trends concerning the aviation industry during the COVID-19 period. We applied text mining and NLP techniques to analyse the articles for extracting the aviation themes and article sentiments (positive and negative). Our results show that there is a variation in the sentiment trends for themes aligned with the real-world developments of the pandemic. The article sentiment analysis can offer industry players a quick sense of the nature of developments in the industry. Our article theme analysis adds further value by summarizing the common key topics within the positive and negative corpora, allowing stakeholders in the aviation industry to gain more insights on areas of concerns or aspects that are affected by the pandemic.**

Keywords—analytics, COVID-19, text mining, aviation industry

I. Introduction

The COVID-19 is an infectious disease which spread to 213 countries and territories around the world. As of August 2020, the pandemic resulted in over 24 million infection cases and over 830,000 deaths [1]. This sudden outbreak has brought upon unprecedented challenges to many industries in varying degrees. In particular, the pandemic has resulted in a significant impact on the commercial aviation industry globally due to severe travel restrictions imposed by numerous countries as a measure to combat the virus. It has been reported that as of June 2020, the industry witnessed a drastic 65% drop in the number of scheduled flights compared to the same month in the past year [2]. This has prompted bankruptcies and massive layoffs in many aviation companies - with many stakeholders in the aviation industry seeking financial support from governments.

As the deadly pandemic is still very much on-going and the aviation industry struggling to recover from the setback amidst constant changes and looming uncertainties in the market outlook, it is difficult for stakeholders in the aviation industry to continuously monitor the market changes in their effort to plan for long term measures to tide through the pandemic. As such, the different stakeholders in the aviation industry must understand how the industry outlook has evolved and continues to evolve in order to identify areas of concern or potential growth in the industry.

News articles provide insights on market [3, 4], politics [5] and society matters [6]. They are the source of industry experts to analyse the trends and predicts the changes and opportunities. In our project, we use news articles related to the aviation industry to analyse the impact of Covid-19. Article theme analysis is the task of finding the key aviation themes in the articles. Some examples include, “Flight designs”, “Safety regulations”, “Flight accidents” and so on. Knowing the theme might help us to know the trends in the aviation industry over time but the sentiment of the articles will provide more insights in terms of whether there is negativity or positivity in the articles. For example, when the revenue drops or the flight design fail, the sentiment of the article is negative whereas when the industry is expanding or purchasing new flights, there is a positive sentiment in the articles. The overall sentiment analysis of the articles about all the companies provides us with the general statistics about the industry.

In our project, we propose two tasks to analyse news articles about the aviation industry; article sentiment analysis and article theme analysis. We apply techniques from text mining and NLP areas to analyse content of news articles. The first challenge is that the sentiment of the business articles depends on business terminology. The second challenge is that themes of the articles are embedded in the text.

We apply sentiment analysis models [7] to extract the sentiment of the articles. The findings aid the stakeholders in the industry to evaluate the existing market situation and make long term decisions. Topic modelling techniques [8] are applied to the outputs of sentiment analysis to delve deeper and understand the trending themes among the positive and negative sentiment articles. The findings aid aviation authorities to prioritize support for parts of the aviation sectors depending on their current outlook and look out for any unexpected or newly emerging themes under each sentiment.

II. Methodology

A. Dataset

The dataset used for this project consists of global daily news articles related to aviation and aerospace industries, which are scrapped from FlightGlobal [9]. It is a media platform specialising in aerospace. The scope of the project aims to focus on the COVID-19 period, and hence we collected the articles which are published between 1st January and 30th June 2020. Final dataset consists of 2,389 Global news articles.

B. Analytics Tasks

To analyse the news articles and discover insights on the impact of COVID-19 on the aviation industry, two main text analytics tasks are performed; article theme analysis and article sentiment analysis. Article theme analysis task aims to identify topics or themes in articles. These are not similar to the keywords for an article displayed for reader navigation but the key aviation topic or theme that the article belongs to. In our project, we also aim to study the changes in theme distributions over time from the outputs of the task [10]. We apply topic modelling [11] on the articles extract themes and analyse the outputs qualitatively and quantitatively.

Complementing topic modelling is sentiment analysis that has been widely used for instance to understand the public's sentiments on political issues [12]. In our preliminary analysis on a sample of articles, we observed articles dataset contains negative, positive and neutral sentiments. Neutral sentiment articles usually share the facts without subjectivity such as; financial numbers, business plans, fleet management, operations etc. In our main analysis, we filter the articles with a neutral sentiment. Combined with themes, the sentiment of the articles enables us to study the subjectivity across topics over six months and infer the impact of COVID-19 on the aviation industry.

In our preliminary analysis, we observed that themes for the positive sentiment articles and negative sentiment articles may vary. There are some common themes as well as sentiment specific themes. For example, an aviation accident is a negative theme whereas the flight design with focusses on the technology is a positive theme. Therefore, we first separated the articles by the sentiment polarity and then extracted the themes from the separated corpus.

III. SOLUTION DESIGN

The overview of the solution approach to analyse the news articles is depicted in Fig. 1 and the details are in provided in sub-sections. We first pre-process the news articles content and apply topic modelling and sentiment modelling on the processed data to identify the underlying themes together with the sentiment across the six months.

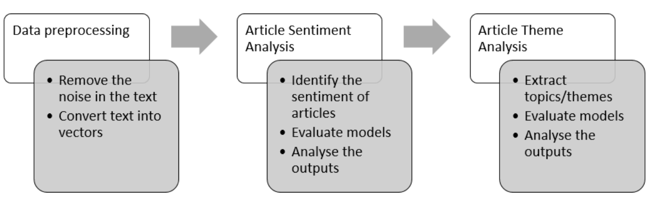


Fig. 1. Solution Overview

A. *Data Pre-Processing*

Fig. 2 summarizes the data preprocessing steps [13]. First, the noise in the articles such as HTML tags is cleaned. The text is tokenized and converted to lower case. At this point, the pre-processed text is used for article sentiment analysis as removal of stop words (e.g. not in "not happy") could affect the sentiment analysis. For article theme analysis, we proceed to remove stop words that are listed in NLTK package [14] and user-defined from the manual analysis.

Through an iterative process of manually identifying the top words using word frequency distribution and the topic-word distribution output from the initial rounds of topic modelling, we identify words that are unimportant and add them to the stopword list used for topic models. This is followed by lemmatization to generate high-quality topics from LDA models. Lastly, the text is converted into vectors which will be the inputs for the next steps in solution design.

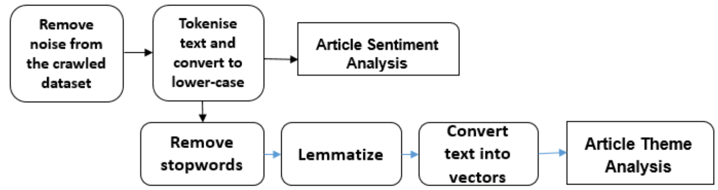


Fig. 2. Data Preprocessing Steps

B. *Article Sentiment Analysis*

Sentiment analysis is a process where information is analyzed using natural language processing (NLP) and the goal is to discover the polarity of a document; negative, positive, or neutral sentiment [15]. Using sentiment analysis, we try to understand the overall sentiment of the article but labelling the article as positive, negative or neutral. The impact of the COVID-19 can be detected from the sentiment of the article. We also can analyse how the sentiment changed for a given aviation theme for each month. Fig. 3 provides an overview of the article sentiment analysis steps.

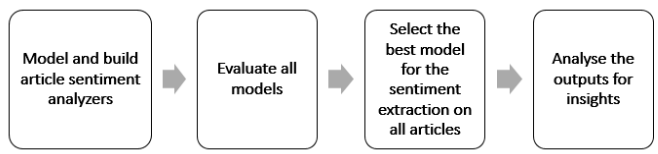


Fig. 3. Sentiment Analysis Steps

Recall that we interested in classifying global news articles in the aviation and aerospace industries into 3 broad sentiments, namely (i) positive, (ii) negative, and (iii) neutral. As pre-labels are not available, the unsupervised and lexicon-based approach will be required to classify the articles into their respective polarities. We consider four sentiment analysis techniques, namely (i) Lexicon [12], (ii) SentiWordNet [16], (iii) TextBlob [17] and (iv) VADER [18] for this task.

To determine the most suitable article sentiment analyser, we choose the dataset that is a subset of all articles. We obtain the news articles tagged with "Coronavirus" (comprised of 708 articles i.e. ~30% of the full dataset). We believe that by analyzing aviation news related to "Coronavirus", the human evaluation will most likely be able to identify any misclassification more accurately as intuitively, the airline industry was badly affected by travel restrictions etc.

Lexicon: A prior study [13] compiled a list of words for sentiment detection. It consists of around 6,800 positive and negative sentiment words in the English language. We first extract the positive and negative words in each article. Each positive word received a score of 1 while each negative word received a score of -1, and words not found in the lexicon did not receive any score. Eventually, an article's score is the sum of the scores of all its words. The article is judged to have Positive sentiment if the overall score is greater than 0, Negative sentiment if the overall score is less than 0, and Neutral sentiment if the overall score is equal to 0.

SentiWordNet: This is a lexical resource that contains a list of English terms which have been assigned a score of positivity, negativity and objectivity (or neutrality). Each synset of WordNet gets a score between 0 and 1 for each of the three categories mentioned above, and the total score of the three categories will sum up to 1. For our context, if the average article score is ≥ 0.01, the article is judged to have Positive sentiment. If the article score is ≤ -0.01, the article is judged to have Negative sentiment. Otherwise, it is judged to have Neutral sentiment.

TextBlob: This sentiment module contains two implementations, namely (i) PatternAnalyzer based on the pattern library – default implementation and (ii) NaiveBayesAnalyzer – an NLTK classifier trained on movie reviews. Using the default implementation, TextBlob returns 2 values, 'Polarity' and "Subjectivity". The polarity score, a value between -1 to 1, is assigned to the text based on the most commonly occurring positive and negative adjectives. While, the subjectivity score will be assigned to the text with a value between 1 (opinion, emotion or judgement) to 0 (factual information). For our study, we consider the polarity score to identify the overall sentiment of the article.

VADER: This is a rule-based and lexicon sentiment analysis tool which was built to detect sentiments expressed in microblog-like contexts. In addition, VADER is sensitive to both the polarity (positive, negative, neutral) and the intensity of the sentiments expressed in social media contexts such as Twitter, Facebook, Instagram etc. A prior study [18] indicated that VADER also works well on texts from other domains. Unlike typical bag-of-words model, VADER also implements the grammatical and syntactical rules and hence is sensitive to word-order between terms in the sentence-level text. Using the lexicon, each word will be tagged to a valence score, and thereafter summed and normalized to obtain a compound score. For our study, we consider the compound score to identify the overall sentiment of the article.

To evaluate the classifiers, we use a specific test data set of 708 news articles. We use both the visual and human analysis to choose the best model.

C. *Article Theme Analysis*

Topic modelling, also known as topic analysis, is the discovery of topics within the documents where each topic is a set of words frequently co-occurring together. For our project, we are interested to cluster the articles within the positive and negative corpora classified using sentiment scores from sentiment analysis to discover the broad topics of the articles in each corpus and understand the aspects of the aviation industry which were performing better or suffering over the pandemic period.

As each article may contain multiple topics, the unsupervised Latent Dirichlet Allocation (LDA) method [8] is used for extracting themes in the entre corpus first. We trained three topic modelling techniques, namely (i) Gensim [19], (ii) LDA Mallet [20] and (iii) scikit-learn (SKLearn) [21] to determine the most optimal model for our dataset. The overall approach is summarized in Fig. 4.

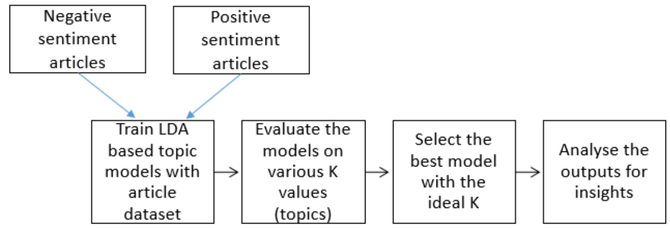


Fig. 4. Theme/Topic Analysis Steps

We compare the performance using the coherence and perplexity scores for varying number of themes (K). The ideal number of topics in the positive and negative corpora is determined separately using both quantitative (e.g. coherence score) and qualitative assessment (e.g. visual assessments). After the best model is selected, the themes are generated by adjusting the number of topics K for both sets of articles. We choose a dominant topic for each article in our analysis.

IV. EXPERIMENTS & EVALUATIONS

A. *Article Sentiment Analysis*

1) *Evaluations*

The summary of the polarity scores on the specific dataset of random 708 articles is shown in Fig. 5. Performance of Lexicon and VADER are similar in terms of the proportion of articles labelled as Negative sentiment. However, compared to Lexicon, VADER had only classified one article of having neutral sentiment and 10% more articles with Positive sentiment than Lexicon (Fig. 5).

Given that VADER is built specifically on social media such as Twitter which has a maximum of 140 characters, the technique may not perform as well as expected on news articles which has more than 140 characters. As a result, Lexicon is determined to be the most suitable technique for our context.

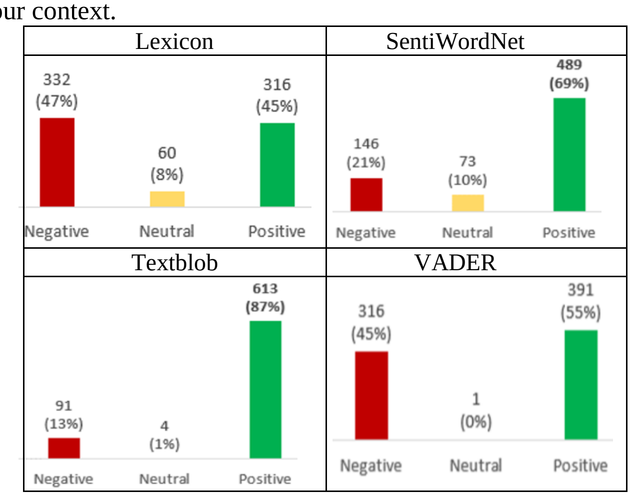
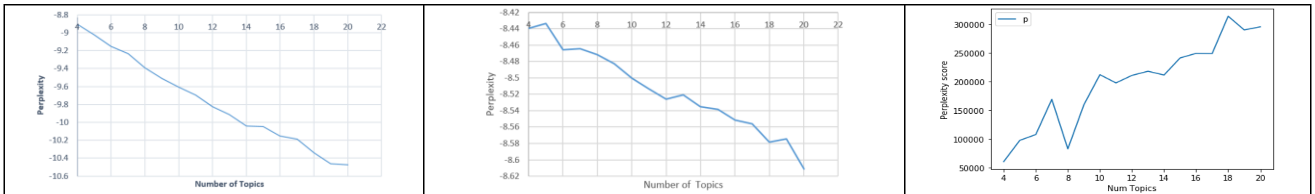


Fig. 5. Evaluations of sentiment analysis on articles with "Coronavirus" tag



(a) LDA Gensim (b) LDA Mallet (c) LDA SKLearn

Fig. 6. Perplexity Score & Number of Topics

2) Misclassification Analysis

Fig. 7 shows examples of misclassification of article sentiment by the two algorithms; SentiWordNet and TextBlob. For the first article, the adjective 'financial' has a first synset of 'fiscal', for which it has a 0.25 Positive score. However, in the context of the article, the word was just part of the term 'financial restructuring', and thus should not have a polarity score at all.

For the second article, the adverb 'well' has a 0.375 Positive score. This word was used in the sentence 'The carrier has also reduced frequencies to eight other Chinese cities, including Guangzhou, Kunming, Nanjing, Tianjin as well as Macau.' 'As well as' is a conjunction that should not be assigned any polarity score. The aggregation of multiple incorrectly labelled words thus leads to the article being wrongly classified as a Positive article.

Article Title	Article Overview	Polarity Scoring
Norwegian cancels 97 Boeing aircraft orders [22]	The article is about how Norwegian, a low-cost carrier from Norway, has cancelled orders for 97 Boeing aircraft, some of which included the now-grounded 737 Max. The article also detailed how the airline had suffered a long hiatus of flights following the coronavirus pandemic.	The article got a net Positive sentiment score of 0.053 (SentiWordNet) and 0.024 (TextBlob).
Singapore Airlines and SilkAir reduce capacity to China over coronavirus outbreak [23]	The article is about how Singapore Airlines and SilkAir suspended flights to multiple destinations in China as the coronavirus outbreak caused "weak demand and operational constraints". The article also highlighted examples of other airlines which had also cut flights to China.	The article got a net Positive sentiment score of 0.048 (SentiWordNet) and 0.045 (TextBlob).

Fig. 7. Examples of incorrectly labelled news articles by SentiWordNet and TextBlob algorithms

Given the visual and human evaluations results, Lexicon model performs better than other algorithms and we choose this for the best model to implement the remaining solution design.

B. Article Theme Analysis

1) Evaluations

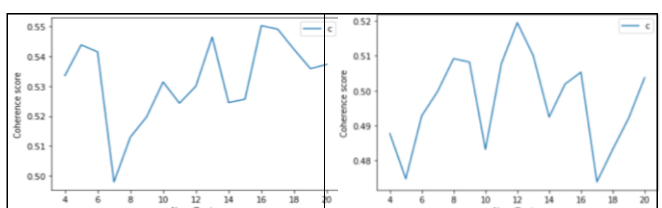
To determine the best performing implementation for our FlightGlobal corpus, we first use a perplexity score – a common supported intrinsic evaluation metric across all three implementations. Perplexity is a measure of the log-likelihood on the test data, where the lower the perplexity, the better the model. To compare the perplexity scores, the evaluation was performed on a test data set containing 108 FlightGlobal articles from 1 to 8 July 2020.

Fig. 6 depicts the comparison of perplexity scores for the number of topics, K ranging from 4 to 20. From Fig. 6, we observe that LDA Mallet gives lower perplexity compared to LDA Gensim. To compare LDA Mallet and SKLearn, we use visuals. Our results show that LDA Mallet has the least overlaps (K=7), whereas SKLearn(K=4) has a major overlap between two of its clusters. Therefore, based on both perplexity scores and visualization analysis, we choose the LDA Mallet for the sentiment-based theme extraction.

2) Best K for Positive and Negative Corpus

Recall that after the article sentiment analysis step, articles assigned with negative sentiment are allocated to the negative sentiment corpus whereas articles assigned with a positive sentiment score are added to the positive sentiment corpus. The most optimal topic modelling implementation (LDA Mallet) is then applied to both corpora separately, using the number of topics evaluations to choose the best number of topics for each corpus. This is necessary as we do not have prior knowledge of the optimal number of topics in each of the new article corpora.

We use coherence scores on each corpus to choose the best number of topics. Coherence scores indicate how relevant the words are in each topic generated by the model. The higher the score, the better the model. Fig. 8 shows the results of coherence scores both corpora between 4 to 20 topics. The optimal number of topics, K, is chosen based on 2 criteria: 1) lowest number of topics with the highest coherence score before the curve flattens out and 2) each topic cluster can be generalized well from a business perspective.



(a) Negative corpus (articles with the negative sentiment) (b) Positive corpus (articles with positive sentiment)

Fig. 8. Coherence scores for negative and positive corpus (LDA Mallet)

We also manually assessed the keywords in the topic clusters by identifying "general areas of interest" applicable to the aviation industry. From our manual analysis, we observe that for negative corpus best K=13 and for positive corpus best K=12. From the evaluations, we choose the best models and implement the end to end solution on the news articles.

V. RESULTS ANALYSIS & FINDINGS

In this section, we perform results analysis of article sentiments, article themes and combined analysis.

A. Article Sentiment Results Analysis

Fig. 9 show that there are significantly more news articles labelled as positive during the first half of 2020, i.e. COVID-19, despite subdued travel demand and tepid outlook for new aircraft orders.

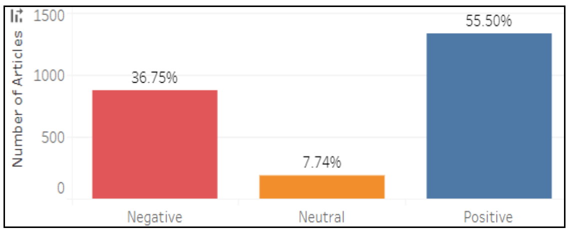


Fig. 9. Sentiment distribution for the FlightGlobal corpus

We further study how the proportion of sentiment changes over time to understand how the impact of the COVID-19 pandemic played out over the months in relation to actual developments in the real world on the aviation domain. Fig. 10 depicts the article sentiment changes over six months.

From Fig. 10, we observe that the proportion of negative sentiments in January and February are very close ~36% and ~35%. During these two months, the brunt of COVID-19 cases was still confined within China, and while countries around the world were slowly experiencing the emergence of local clusters, the aviation industry still did not have a clear picture as to the global economic damage that the virus would eventually cause. However, moving from February to March, we see that the proportion of negative sentiments has increased from 35% to 44%. On 11 March 2020, the World Health Organisation officially labelled COVID-19 as a global pandemic (World Health Organisation, 2020), as it became obvious that the virus had spread across continents at a rapid pace. March was also the month which had the greatest number of countries imposing their first national or state-wide lockdowns, such as United States of America, United Kingdom, Italy, and France etc. This directly translated to disaster for the aviation industry as international bans on non-essential travel were imposed and airlines began to ground most of their fleet.

Moving from March to April, the downtick in negative sentiments from 44% to 39% could be explained by the emergence of state rescue packages for airlines across the globe. For example, the US Treasury Department agreed a ~$25bn rescue package for 10 of the country's biggest airlines to support their payrolls [24]. Air France-KLM also secured €7bn in French government aid to prop up the airline [25]. The slew of bailouts for airlines helped to dampen the negative sentiments from the previous month.

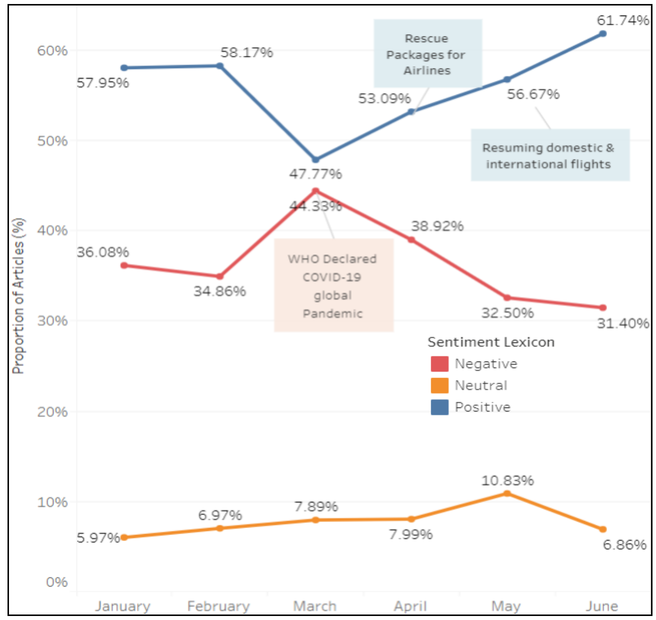


Fig. 10. Sentiment distribution by month and related key world events

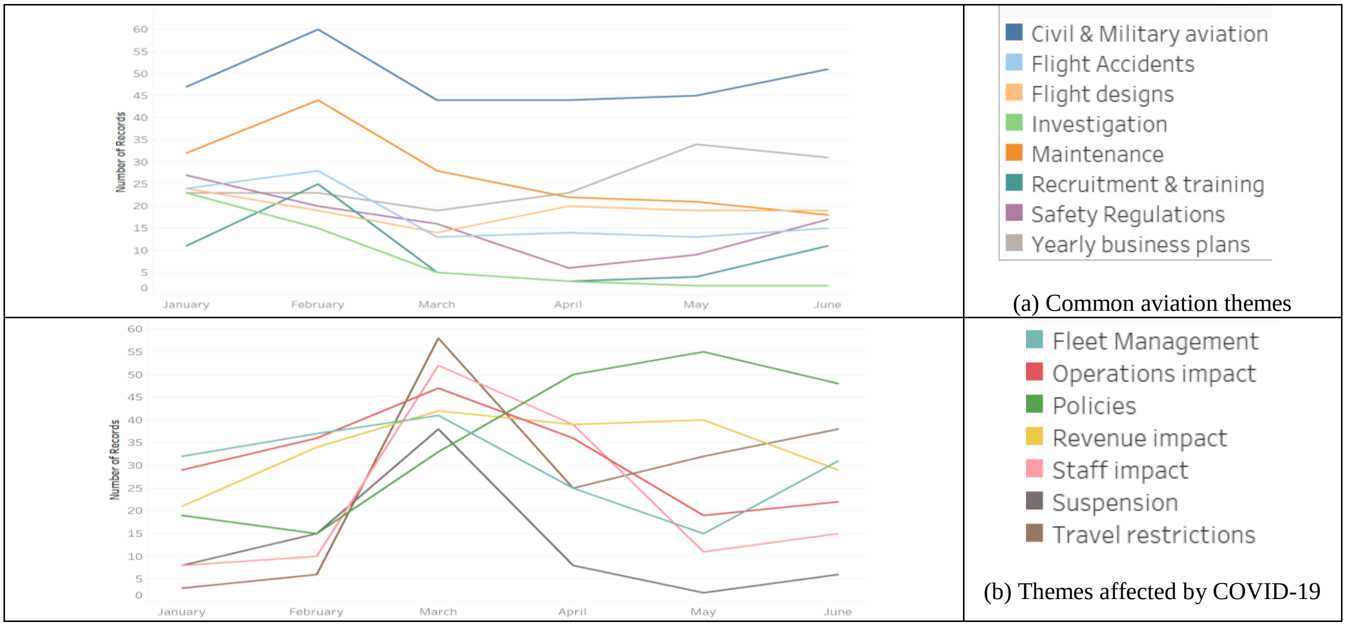
The even bigger dip in Negative sentiments from 39% in April to 32% in May and 31% in June was largely due to countries beginning to lift their lockdowns gradually. As May went by, many countries began to announce their plans to begin reopening their borders, and intra-regional travel in Europe was expected to start in June [26]. This likely helped to lift the gloomy outlook for air travel.

In summary, the shift in the proportion of negative sentiments against positive sentiments over the months as COVID-19 panned out can indeed be explained by the development of real-world events. Thus, article sentiment analysis is a potentially useful quick tool for industry participants to assess the nature of developments in the industry.

B. Article Theme Results Analysis

Extending the results further, we reviewed the theme clusters generated by the topic modelling implementation. We analysed the document topic probabilities for each article in the entire corpus to determine the number of dominant topics. The analysis showed that the majority of the articles are dominated by one topic (i.e. only one topic taking up ≥28.6% of the probability). As such, each article was represented as a single topic for subsequent analysis of topic modelling results.

Table I shows the topics we identified in both the corpus. From Table I, we observe that some themes are common for both the corpora whereas some themes are unique for each corpus. From this table, we infer the topics but not understand how the topic evolved. Italics indicate unique corpus and bold indicates the themes significantly affected.



(a) Common aviation themes

(b) Themes affected by COVID-19

Fig. 11. : Theme trends over six months.

Table 1: Topics from both corpora

Topic label	Top Words
Civil & military aviation	force, helicopter, system, capability, mission test, force, helicopter, fire, system, service
Policies	government, state, share, agreement, financial government, uk, carrier, business, state
Flight designs	*test, design, technology, system, land*
Investigation	*crew, runway, inquiry, land, approach*
Flight accidents	*international, authority, civil, state, investigation*
Maintenance	production, engine, work, part, facility engine, maintenance, fuel, takeoff, result
Operations impact	carrier, service, international, route, operate industry, crisis, demand, coronavirus, global
Fleet management	fleet, order, jet, delivery, customer, service airport, fleet, service, data, carrier, operation
Recruitment & training	*market, chief, executive, industry, make*
Revenue impact	expect, cost, demand, end, reduce passenger, loss, operating, revenue, traffic
Safety Regulations	pilot, work, report, safety, change pilot, faa, issue, safety, include
Staff impact	march, employee, april, support, industry executive, chief, cut, add, situation
Suspension	*carrier, march, international, outbreak, suspend*
Travel restrictions	passenger, coronavirus, travel, airport, cargo travel, government, country, coronavirus, measure
Yearly business plans	plan, business, operation, support, include order, business, jet, delivery, production

We then performed the analysis of trends in the topics over six months. Fig. 11 shows the comparisons of the normal and COVID-19 affected topics. We observe that the common topics have fewer fluctuations and all these topics drop during March and re-trend back slowly.

Fig. 11 (a) shows trends of the aviation articles which represent the day to day business and management strategies irrespective of the pandemic. The impact can be seen but not very significant. Whereas, Fig. 11 (b) shows the themes affected by COVID-19. The fluctuations are very high and all the topics show a sudden peak during March. This is the period where most of the countries imposed lockdown measures and the Government imposed policies and regulations on travel. As discussed in the previous analysis, the impact on the aviation industry is very significant.

C. *Analysis by Themes and Sentiment*

In this sub-section, we analyse the themes and the sentiments to gain deeper insights. Fig. 12 shows the themes, sentiments and percentage of articles in each row across.

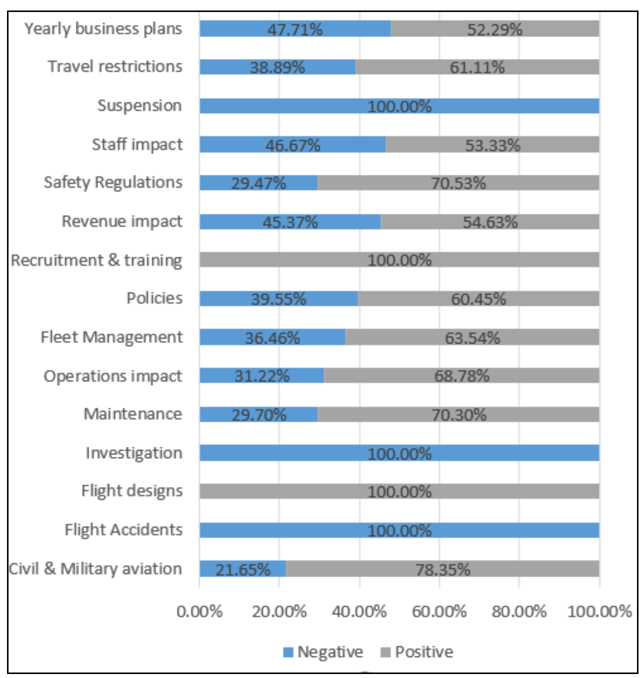


Fig. 12. Themes, sentiments & percentage of articles

Topics such as “suspension” and “flight accidents” are negative sentiment only whereas topics such as “recruitment & training” and “flight designs are only of positive sentiment. This observation is logical as these articles state the facts with subjectivity to create an impact on the readers. The highest negative impact is on “Policies”, “Staff”, “Business plans”, “Revenue” and “Suspension”. These are the articles which indicated the impact on the industry and the action items that the company is taking in response to the impact. For industry experts, the topic and sentiment-based articles would provide deeper analysis.

VI. LIMITATIONS OF OUR WORK

Traditional sentiment analysis stops short at polarity classification. This project, however, went further by generating the sentiment scores for each news articles and broke down the evolution of sentiments across months. We successfully evaluated and applied the most suitable sentiment analysis model that could produce a sentiment scoring that made the most logical sense. Thereafter, we trained the most suitable topic model that produces topic clusters of satisfactory quality and applied it on the positive and negative sentiment corpora to derive topics in each corpus. The results derived were easy to be interpreted and provided us with more insights into the corpus. Our domain knowledge in the airline industry has also provided advantages in the pre-processing and analysis of the news articles.

Nonetheless, there are some limitations to our solution approach and techniques. Sentiment analysis lexicon model has limitations in identifying the negative and positive terms related to the specific aviation industry. This affects the performance of the model in assigning the accurate sentiment to the articles. Further, news articles contain statements which are both subjective and objective. The sentiment of the article should be detected only after understanding the subjectivity of the sentences. Hence the model can be improved by first analysing the subjectivity status in the article and then apply polarity calculations. During the topic modelling process, we observed that there is a lack of standardization across articles in the use of terms with different variations (e.g. COVID-19, Coronavirus, Corona / United States, US, USA). Such terms are manually standardized, however, there may be terms which are not flagged out due to the large vocabulary of words.

For future work, some possible extension of this project will be to explore readily available pre-labelled training data for sentiment analysis on airline-related news, which could potentially increase the classification accuracy. In addition, we propose to expand the corpus by crawling airline news articles from other platforms which aid in identifying more granular aviation themes. We may also extend the work to study the impact on other related fields such as insurance.

VII. RELATED WORK

Analysing comments or news articles for gathering industry changes or impact on the domain themes is useful for the stakeholders to gather insights and make informed decisions [28, 32]. Obthong et al. surveyed research on text analysis on news articles and concluded that machine learning algorithms are effective in generating the analytics on news articles. [27] Sentiment analysis of health-related news articles aid in understanding the concerns in specific medical areas [28]. Applying classification models and with an automated computational framework on news articles related to companies listed in Stock exchanges, the authors predicted the changes in stock prices along the day [3]. Applying lexicon-based sentiment models, Xiaodong et al. extracted, and analyzed the effects of news sentiments on stocks in the pharmaceutical sector [4]. Karamibekr el al. focused on the sentiment analysis of social issues in news articles. The conducted a statistical investigation on the differences between sentiment analysis of products and social issues [6]. Junque et al. applied opinion mining to automatically detect the sentiment of each article, thereby allowing to visualize how the sentiment of reporting evolved throughout the year, on a party, politician and newspaper level [5]. In our work, we applied sentiment analysis and thematic analysis similar above research on the aviation industry news articles to extract sentiment, themes, and evolutions over time.

Recently, research on COVID-19 in the area of airlines industry is taking the toll due to the demand in the quick response and action planning required by the stakeholders. Some studied common health-related themes while others studied airlines domain-specific aspects of the aviation industry. For example, Quilty et al. studied the effectiveness of thermal screening [29]. Travel control measures and the impact of airlines role on the global spread of the COVID-19 outbreak shows the risk increased in January 2020 [30]. Sobieralski analyzed the effects of uncertainty shocks on airline employment [31]. The hardest-hit employees are ones related to passenger handling and flight operations. Suspension of airlines and the effect of such cancellations in delaying the estimated arrival time for all other countries is studied by collecting data [32]. The above works used statistical models for studying the impact of COVID-19 on the airlines industry. In our study, we analysed the aviation related news articles to gain insights on the impact of COVID-19 on airlines industry in terms of aviation themes. We applied text mining and NLP models to achieve the goal and presented the analysis in terms of visualizations by themes, sentiments and time.

VIII. CONCLUSION

In this project, we analysed news articles collected from FlightGlobal during the first half of 2020 to gain insights into the themes and sentiments in the aviation industry across six months of the COVID-19 pandemic. We observed that variation in the theme and sentiment trends aligned with the real-world developments of the pandemic and the article sentiment analysis could potentially offer industry players a quick sense of nature of developments in the industry. Article theme analysis adds further value by summarizing the common key topics within the positive and negative corpora of articles, allowing stakeholders to gain more insights on areas of concerns or aspects that are gradually worsening or recovering from the impact of the pandemic. Our results show that Operations, Fleet management, Recruitment & training, Revenue and Staff are the themes which are affected significantly due to suspension and travel restrictions. We also observed that topics such as civil & military aviation, flight designs, maintenance, safety regulations and yearly business plans are affected moderately. To understand the impact we also presented the article sentiment together with time analysis.

ACKNOWLEDGEMENT

This research is supported by the National Research Foundation, Singapore under its International Research Centres in Singapore Funding Initiative. Any opinions, findings and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of National Research Foundation, Singapore. We thank Selene Chi Lian Choong, Yejin Son and Jia Wei Wong for the assistance in this research study.

REFERENCES

[1] Countries where COVID-19 has spread (August 27, 2020). Retrieved August 27, 2020, from https://www.worldometers.info/coronavirus/countries-where-coronavirus-has-spread/

[2] Covid-19: Implications for the commercial aviation industry (June, 2020). Retrieved August 27, 2020 from https://altonaviation.com/wp-content/uploads/2020/06/Alton-2020-Covid-19-Implications-for-the-Commercial-Aviation-Industry.pdf

[3] Beckmann, Marcelo. (2017). Stock Price Change Prediction Using News Text Mining. Thesis.

[4] Li, Xiaodong & Xie, Haoran & Chen, Li & Wang, Jianping & Deng, Xiaotie. (2014). News Impact on Stock Price Return via Sentiment Analysis. Knowledge-Based Systems. 69. 10.1016/j.knosys.2014.04.022.

[5] Junque de Fortuny, Enric & De Smedt, Tom & Martens, David & Daelemans, Walter. (2012). Media coverage in times of political crisis: A text mining approach. Expert Systems with Applications. 39. 10.1016/j.eswa.2012.04.013.

[6] Karamibekr, Mostafa & Ghorbani, Ali. (2012). Sentiment Analysis of Social Issues. Proceedings of the 2012 ASE International Conference on Social Informatics, SocialInformatics 2012. 215-221. 10.1109/SocialInformatics.2012.49.

[7] Neri, Federico & Aliprandi, Carlo & Capeci, Federico & Cuadros, Montse & By, Tomas. (2012). Sentiment Analysis on Social Media. 10.1109/ASONAM.2012.164.

[8] Blei, David & Ng, Andrew & Jordan, Michael. (2001). Latent Dirichlet Allocation. The Journal of Machine Learning Research. 3. 601-608.

[9] Flight Global. Retrieved August 27,2020 from https://www.flightglobal.com/

[10] Westerlund, M., Leminen, S., Rajahonka M., (2018) A Topic Modelling Analysis of Living Labs Research Technology Innovation Management Review, 8(7), 40-51.

[11] Yang, S., Zhang H., (2018) Text Mining of Twitter Data using a Latent Dirichlet Allocation Topic Model and Sentiment Analysis. International Journal of Computer and Information Engineering 12(7) 525-529.

[12] L, M., Ch'ng, E., Chong, A., See, S., (2016) Twitter Sentiment Analysis of the 2016 US Presidential Election using an Emoji Training Heuristic Applied Informatics and Technology Innovation Conference (AITIC).

[13] Angiani, G., Ferrari, L., Fontanini, T., Fornacciari, P., Iotti, E., Magliani, F., & Manicardi, S. (2016). A Comparison between Preprocessing Techniques for Sentiment Analysis in Twitter. KDWeb.

[14] Bird, S., Klein, E., Loper, E. (2009). Natural Language Processing with Python. O'Reilly Media Inc. http://nltk.org/book

[15] Liu, B. (2015). Sentiment Analysis: Mining Opinions, Sentiments, and Emotions. Cambridge: Cambridge University Press.

[16] Baccianella, Stefano & Esuli, Andrea & Sebastiani, Fabrizio. (2010). SentiWordNet 3.0: An Enhanced Lexical Resource for Sentiment Analysis and Opinion Mining.. Proceedings of LREC. 10.

[17] TextBlob, 2020. Textblob: Simplified Text Processing — Textblob 0.16.0 Documentation. [online] Textblob.readthedocs.io. Available at: https://textblob.readthedocs.io/en/dev/ [Accessed 27 August 2020].

[18] Hutto, C. J. & Gilbert, Eric. (2015). VADER: A Parsimonious Rule-based Model for Sentiment Analysis of Social Media Text. Proceedings of the 8th International Conference on Weblogs and Social Media, ICWSM 2014.

[19] Rehurek, R., Sojka, P. (2011) Gensim-statistical semantics in python.

[20] McCallum, Andrew Kachites. "MALLET: A Machine Learning for Language Toolkit." http://mallet.cs.umass.edu. 2002.

[21] Buitinck, L., Louppe, G., Blondel, M., Pedregosa, Fabian, Mueller, A., Grisel, O., Ga"el Varoquaux. (2013). API design for machine learning software: experiences from the scikit-learn project. In ECML PKDD Workshop: Languages for Data Mining and Machine Learning (pp. 108–122).

[22] Wolfsteller, Pilar (June 30, 2020). Norwegian cancels 97 Boeing aircraft orders. Retrieved August 27 2020, from https://www.flightglobal.com/fleets/norwegian-cancels-97-boeing-aircraft-orders/139050.article

[23] Chua, Alfred. (January 31, 2020). SIA Group slashes https://www.flightglobal.com/strategy/sia-group-slashes-capacity-to-china-over-coronavirus-outbreak/136466.article

[24] Dunford D., Dale B., Stylianou N., Lowther E., Ahmed M., Torres Arenas I. (2020, Apr 7) Coronavirus: The world in lockdown in maps and charts. BBC News. https://www.bbc.com/news/world-52103747

[25] Frost, Laurence., Deutsch, Anthony. (April 25, 2020). Air France-KLM wins aid deal, flags likely share issue. (2020, April 25). Retrieved July 24 2020, from https://www.reuters.com/article/us-health-coronavirus-airfranceklm-renau/air-france-klm-wins-aid-deal-flags-likely-share-issue-idUSKCN22631H

[26] International Air Transport Association. (2020). Economic Performance of the Airline Industry. Retrieved July 24 2020, from https://www.iata.org/en/iata-repository/publications/economic-reports/airline-industry-economic-performance-june-2020-report/

[27] Obthong, Mehtabhorn, Nongnuch Tantisantiwong, Watthanasak Jeamwatthanachai and Gary B. Wills. "A Survey on Machine Learning for Stock Price Prediction: Algorithms and Techniques." FEMIB (2020).

[28] K Todd, A P Lapointe, S P Broglio, Sentiment Analysis of Journal Articles and News Articles Pertaining to CTE, Archives of Clinical Neuropsychology, Volume 34, Issue 5, July 2019, Page 738.

[29] Quilty Billy J , Clifford Sam , CMMID nCoV working group , Flasche Stefan , Eggo Rosalind M . Effectiveness of airport screening at detecting travellers infected with novel coronavirus (2019-nCoV). Euro Surveill. 2020;25(5):pii=2000080

[30] Wells, Chad & Sah, Pratha & Moghadas, Seyed & Pandey, Abhishek & Shoukat, Affan & Wang, Yaning & Wang, Zheng & Meyers, Lauren & Singer, Burton & Galvani, Alison. (2020). Impact of international travel and border control measures on the global spread of the novel 2019 coronavirus outbreak. Proceedings of the National Academy of Sciences.

[31] Sobieralski, Joseph. (2020). COVID-19 and airline employment: Insights from historical uncertainty shocks to the industry. Transportation Research Interdisciplinary Perspectives. 5. 100123. 10.1016/j.trip.2020.100123.

[32] Gottipati, Swapna; Shankararaman, Venky; and Lin, Jeff. Latent Dirichlet Allocation for textual student feedback analysis. (2018). Proceedings of the 26th International Conference on Computers in Education ICCE 2018: Manila, Philippines, November 28-30. 220-227.

[33] Adiga, Aniruddha et al. (2020). Evaluating the impact of international airline suspensions on COVID-19 direct importation risk.