Natural Language Processing Using Latent Dirichlet Model

Qiang Wang, Renlu Zhang, Minqian Yang, Lili Zhang, Jiaqi Ling, Haofu Wu

Abstract

This paper establishes an approach to extract knowledge from structured open-source data, with a focus on natural language processing and text analytics. Our analysis consists of two components: first, best practice for topic modeling, and then, interpretation of statistical results in aviation context. In order to extract knowledge from narratives in our dataset, we utilized several packages in python to break down sentences into words and identify topics which frequently shows up. After initial exploration, a list of stop words, which refers to topics that are not relevant to aviation safety, was excluded from our topics. And iterations take place a few times until we conclude a model that comes with the best possible outcomes: 20 topics that represents 20 frequent problems in aviation safety fields.

1 Introduction

Organizations collect safety reporting data across a variety of domains in transportation to help inform analysis to improve safety. Hitherto, data size of the safety report experienced a rapid growth and format of the report became more complex to be analyzed. While various tools are available now for us to deal with big data, a key challenge in utilizing these safety reports is to extract information from text narratives within the dataset. The extracted information can help safety related studies or further downstream analysis. To make the best use of these safety reports, we develop a Natural Language Processing based approach that automatically extract and tag information in safety narratives.

2 Methodologies

Our work is in Python, with the help of several packages in data preparation, text processing, and plotting. Source code is available upon request.

2.1 Latent Dirichlet Allocation Model

Generative topic models widely applied in unstructured text data include emails, social media posts, chats, support tickets, surveys, etc. One of the most efficient models is Latent Dirichlet Allocation developed by Blei, Ng and Jordan (2003). LDA model is the hierarchical Bayesian version of pLSI with multinomial distributions sampled from Dirichlet distributions, which is the distribution of distribution. One of the explanations is via Gibbs sampling, if all conditional distributions $p(x_i|x_{i-1})$ are known (x_{i-1} is the state of all the x except $x_i, x_{i-1} = (x_1, x_2, ...x_{i-1}, x_{i+1}, ...x_n)$. In LDA the goal is to estimate the distribution p(Topic Assignment|Document already know). For Gibbs sampling, one has to calculate: p(Topic assignment for document i|Document already know except the document i). After a sufficient number of iterations, we arrive at a topic assignment sample z.

Like the fact that Beta distribution is the conjugate of binomial distribution, Dirichlet distribution is the conjugate of multinomial distribution, which is the modeling distribution in text categorization. Thus, it's a good fit for prior distribution in Bayesian equation.

$$p(\theta|\alpha) = \frac{\Gamma(\sum_{i=1}^{k} \alpha_i)}{\prod_{i=1}^{k} \Gamma(\alpha_i)} \theta^{\alpha-1} \theta^{\alpha_0} ... \theta^{\alpha_{k-1}}, \tag{1}$$

where the parameter α is a k-vector with components $\alpha_i > 0$, and $\Gamma(x)$ is the Gamma function.

The plot summarizes the key idea behind Latent Dirichlet Allocation model. Parameter α determines the Dirichlet distribution Dir (α) (The dimension is the number of topics) from which the topic multinomial distribution v: $p(z_n|\theta)$ of each document is sampled from. And for the inner layer, the topic assigned for each word is sampled from the multinomial topic distribution v. Given the topic z for the word, the word is sampled from the multinomial distribution φ: $p(w_n|z_n\beta)$. But for this multinomial distribution φ:$p(w_n|z_n\beta)$, it's sampled from the Dir(β) (The dimension is the number of words). So there are three hierarchies in the sampling model. The parameters for the two Dirichlet parameters α and β are fixed for all documents. For each

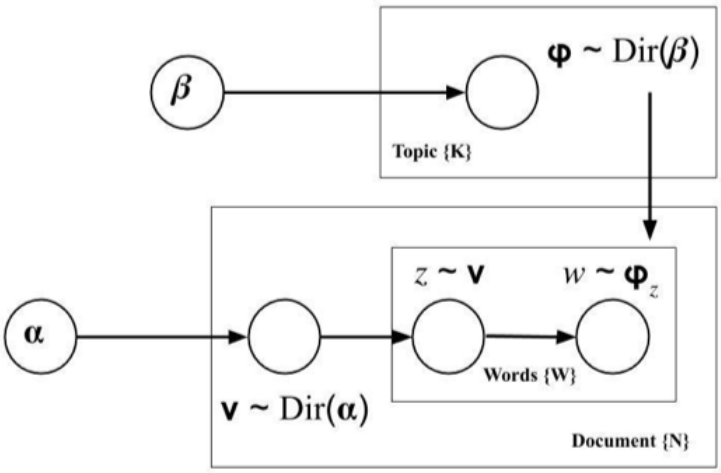

Figure 1: LDA: Bayesian Version

document, the vector parameters θ for the multinomial distribution v of topic distribution given each document is sampled from the Dirichlet distribution whose dimension is the total number of topics. For each word, the topic assigned for it is sampled from multinomial distribution v and the word likelihood is sampled from the Dirichlet parameters whose dimension is the total number of words and parameter is β. Given the parameters α and β, taking the product of the marginal probabilities of single documents, the probability of a corpus:

$$p(D|\alpha,\beta) = \prod_{d=1}^{M} \int_{\theta}^{n} p(\theta_d|\alpha)(\prod_{n=1}^{N_d} \sum_{z_{dn}}^{n} p(z_{dn}|\theta_d)p(w_{dn}|z_{dn},\beta)d\theta, \tag{2}$$

where $\{N\}$ is a set of N words w, varied for different documents, $\{D\}$ is a set of D documents and θ is the distribution of a topic mixture.

Overall, with Dirichlet and multinomial distribution the LDA model relaxes in other models like each word is generated from a single topic in Probabilistic Latent Semantic Analysis (PLSA), and is best applied to documents in which each document deals with multiple topics. But the exchangeability neglect the impact of varied position for each word, so it's not suitable

for text with many paragraphs, but in our dataset, most are short reports with one paragraph. Also, Unlike Correlated Topic Model (CTM), LDA is incapable to model relations among topics.

2.2 Data Preparation for Analysis

The 10 SDR files downloaded from FAA official website was separated by year. After checking the data frame, there is no error noticed during the import and data was properly transferred. Since the goal is to detect anomalies in the aviation industry, we are only going to focus on commercial fight records, which is indicated by c260-submitter code. Whole dataset is filtered by this column and 553,987 records are left for analysis. The text narratives were separated into 5 different columns, so we concatenated there columns to get the entire report of each row, and there were no null value for all 553,987 rows of report records.

Other processes include:

- Records are sorted by the date of occurrence, which is the value in c20 column, and generate two more column ‘year’ and ‘month’ indicating occurrence year and month correspondingly.

- The column type c18 is messy, including both numeric and string. This column is segment code, so we convert all of them to numeric type.

- Combine the column c120: Aircraft manufacturer’s name, c130: FAA assigned code to identify aircraft group as manufacture model, like ‘BOEING’ in c120, ‘737’ in c130, thus the record in‘manu model’ is ‘BOEING 737’.

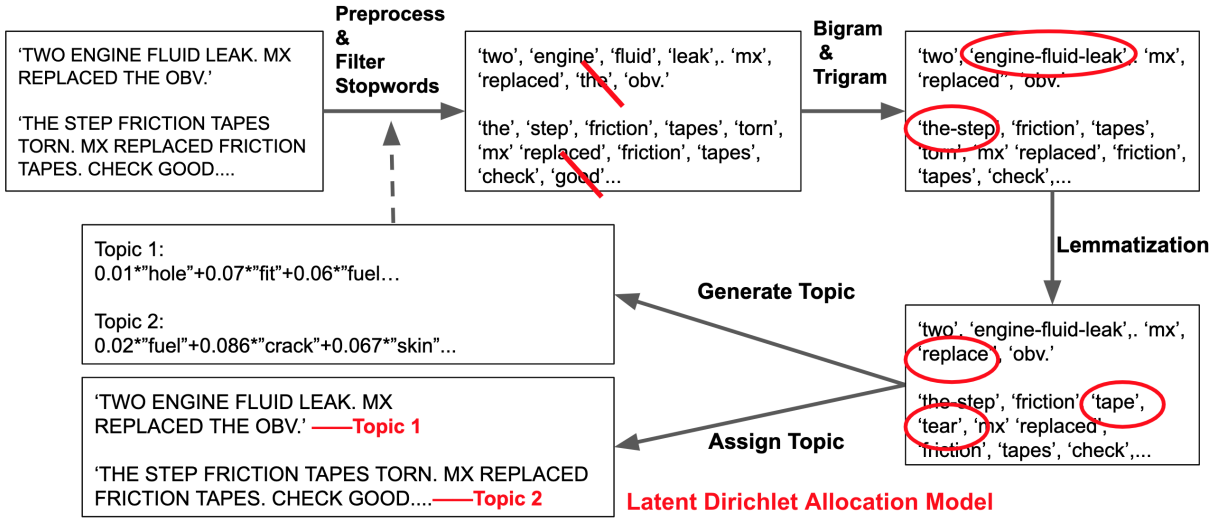

Figure 2: Topic Modeling Flowchart

2.3 Topic Modeling

2.3.1 Text Preprocess and Filter Stop Words

First step is using simple preprocess function in Gensim to convert all documents into lists of lowercase tokens and ignore tokens that are too short or too long. The next step is filtering out stop words based on the word lists from NLTK package. Besides, we summarized the term frequency score of each word across all documents and selected the top 50 most frequent words, then read each word one by one to determine if the word is important in describing the issue, by doing so, we obtained additional stop-words includes: srm, limit, normal, iaw, amm, op, replace, repair, check, find, remove, install, service, maintenance, perform, aft, acft, good, ea, note, right, left, fail, inspection, classify, area, part and operational. After building the stop-word list, we used it to filter through all lists of words.

2.3.2 Create Bigram and Trigram

The third step is to combine words into bigrams and trigrams, this step can help us to aggregate separate words into phrases that are common in the file and also, we believe these phrases are usually related to specific aviation terms that can be analyzed later in the project. For example, in the diagram, 'top-step' and 'engine-fluid-leak' are detected to be in the list of bigram and trigram. Overall, we trained Phrases Model built inside Gensim with the entire lists of words and applied the model back to our data, eventually we got 4,436 unique phrases from our data.

2.3.3 Lemmatization

The fourth step is lemmatization, we employed the lemmatization function in Spacy package, this step will convert each word to its original word, for example, 'replaced' to 'replace', 'tapes' to 'tape'. This way we can minimize the number of words in our corpus and improve the accuracy of the model by minimizing training data noise.

2.3.4 Latent Dirichlet Allocation Model

The fifth step is to convert the preprocessed data into corpus and dictionary that can be used in building topic analysis model. Dictionary includes all unique words used in the data and there will be ID code assigned for each word. Corpus is lists of codes of words and its number of occurrences and each list corresponds to one document, they can be generated by using the Dictionary and doc2bow function in Gensim package.

To build our model, we selected Latent Dirichlet Allocation model1. With Dirichlet and Multinomial Distribution, the LDA model relaxes assumptions in other models such as each word is generated from a single topic in Probabilistic Latent Semantic Analysis (PLSA), so it is best applied to documents in which each document is associated with multiple topics. But the exchangeability in this model neglects the impact of varied position for each word, so it's

not suitable for text with many paragraphs, but in our dataset, most are short reports with one paragraph. Also, Unlike Correlated Topic Model (CTM), LDA is incapable to model relations among topics.

By using grid search, we determined the number of topics that will separate the documents in the best manner is 20 and adjusted all other parameters according to the size of all training data. The result of the model indicates a rather convincing performance with a coherence score of 0.55. We read through the topics and the clustered documents to check if they are mutually exclusive and meaningful for domain expert.

2.3.5 Topic Interpretation

After having each document assigned into a single topic and the lists of words that represents a topic, we labeled documents under each topic and interpreted them also based on other attributes. For each topic, we selected at least 1% of the documents equally across the 10-year period as interpretation sample and focus on the topic words, and after first round of interpretation, we swapped topics among each group members to do 2nd round interpretation. This way we can minimize the human bias.

Apart from the text content records under each topic, each topic is also summarized based on their attributes. The attributes searched are c90: Descriptive name of part, c120: Aircraft manufacturer's name, c130: FAA assigned code to identify aircraft group, 'manu_model': combined columns of c120+c130, c240: Location on aircraft of the defective or malfunctioning part, c250: Text reflecting condition of failed part.

2.3.6 Phrase Analysis

After applying bigram and trigram models into our data, we are able to mark the phrase in format like xxx-xxx. So, we used regular expressions to extract all these phrases from the lists of words and summarized the occurrence of these phrases by year to create a word frequency

rank among the entire word lists and only looked into those phrases that have occurred in more than 100 documents. After reading through all the phrases we only selected those phrases meaningful in describing issues and mapped the trend of occurrence across the 10-year period.

2.3.7 Abbreviation Analysis

The abbreviations are another area that we believe is useful to extract from the data. Since in aviation reports, abbreviations are usually important terms representing the issue related. And the method we used to get those abbreviations was by web scraping the entire abbreviations list published by FAA from https://www.faa.gov/jobs/abbreviations/. And used all the abbreviations listed to create a filter and then count the occurrence of these abbreviations by year. By reading through the meanings of each abbreviation, we selected abbreviations that appeared in more than 100 files. Then we can do trend analysis on some of the important abbreviation terms by year.

3 Dataset

In this project, our group pick the Service Difficulty Reports (SDR) data to demonstrate the analytics in the aviation domain. This dataset contains information related to aviation incidents collected by Federal Aviation Administration (FAA), an organization operated by U.S. Department of Transportation. Records in SDR come mainly from certificate holders and certificated repair stations. FAA collects this data for the purpose of planning, directing, controlling and evaluating certain assigned safety-related programs in order to improve aviation safety.

Service Difficulty Reports is ideal for us to analyze topic trend of the report over the years to help FAA to achieve global harmonization of aviation systems for five reasons. First of all, the data source is reliable. Records were submitted and collected by authorized group which we mentioned before. Second, yearly distribution of the records is stable. We would like to

know main incident topics happened in recent years, so we focus on data from 2010 to 2019. The records in 2010 to 2018 distributed between 50,000 to 80,000, while the data in 2019 is still collecting. Third, this dataset offers enough records for us to analyze. It contains more than 500,000 rows and 80 columns. Furthermore, it contains abundant features like aircraft model, defective parts and incident reports, so that we can analyze it from various aspects. Last but not least, we would like to present some business advice to the stakeholders in the aviation industry, so the ideal dataset should not only contain military or unscheduled plane. While this dataset records data related to commercial aircraft which makes our intention possible.

By identifying trends of topics in aviation safety reports over the years, our research could promote the changes in the aviation industry. Based on the changes of topics, our goal was to provide a further analysis of effective improvements and potential risk.

4 Exploratory Data Analysis

Our dataset is a tabular one with 86 columns and 553,987 rows. Data range starts from 2010/01 to 2019/06. There is one caveat to use the dataset properly. When our group downloaded open source data, the first column 'C5' value is missing while its column name remain, which results in the problem that the dataset is not aligned correctly. By removing the first column name 'C5', the value matches with the column name again.

- The number of reported occurrence remains steady from 4 to 6 thousand from 2010/01∼2019/04 and drops drastically after 2019/05.
- The top three Air Transport Association that was reported are 3350, 5320, 5330, while by looking at the percentage of the total over the year, 5320 is the only one experienced growth after 2017.
- The top three Manufacturer that was reported are Boeing, Airbus, CNAIR; Boeing takes

up almost 50% of the total and the percentage of Airbus increased greatly in 2017 and remained to be the second after Boeing. In particular, Boeing 737 appears to be the most frequent model in Service Difficulty Report and Canadair cl 600 shows a decreasing trend in recent years.

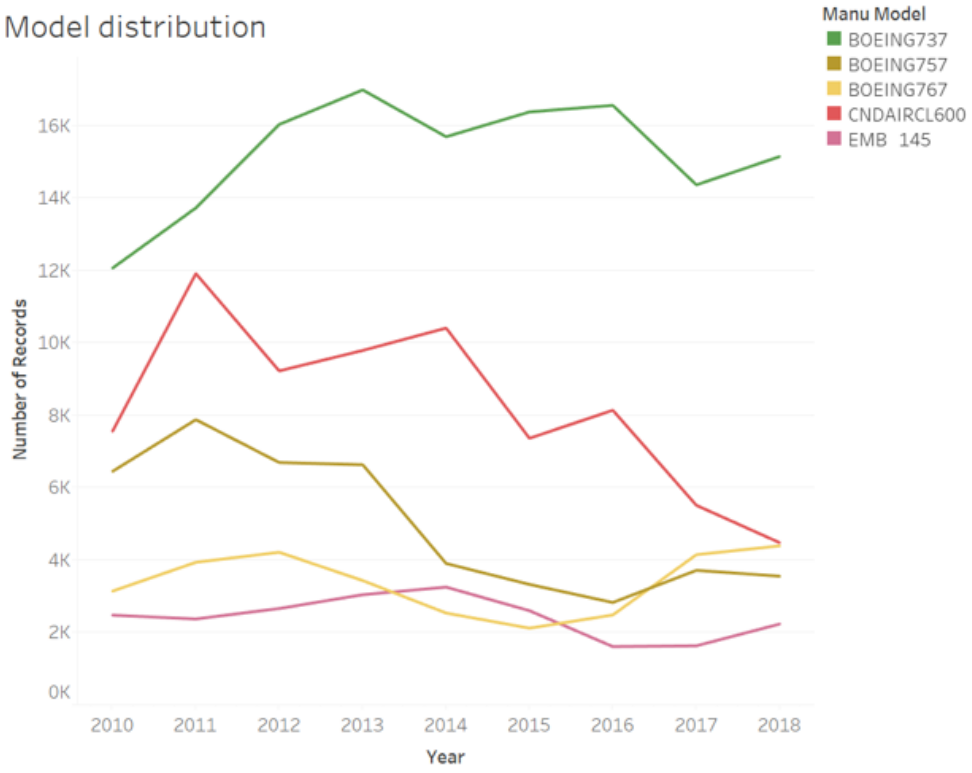

Figure 3: Top Five Model Distribution over Year

- The top three Region responsible for aircraft that was reported are NM, EA, EU; especially the percentage of NM experienced a steady growth and took up over 70% in 2019.
- The top three Descriptive name of part that was reported are skin, light, seat track; The percentage increased part contains seat track and floor panel.

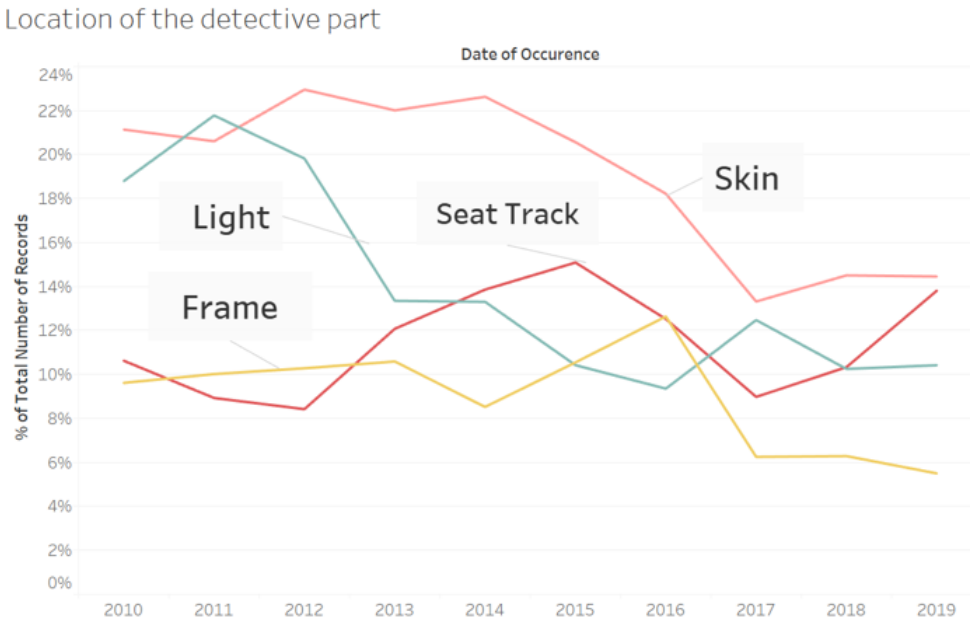

Figure 4: Top Four Location of the Detective Part

- The top three Location on aircraft of the defective or malfunctioning part that was reported are fuselage, cabin, emergency light; The percentage increased location was cockpit; The percentage of fuselage experienced a sharp dip in 2017, while the percentage of cabin and emergency light increased greatly in 2017).

- The top three Text reflecting condition of failed part Top condition that was reported are corroded, cracked, inoperative; The percentage of corroded are increasing over the year.

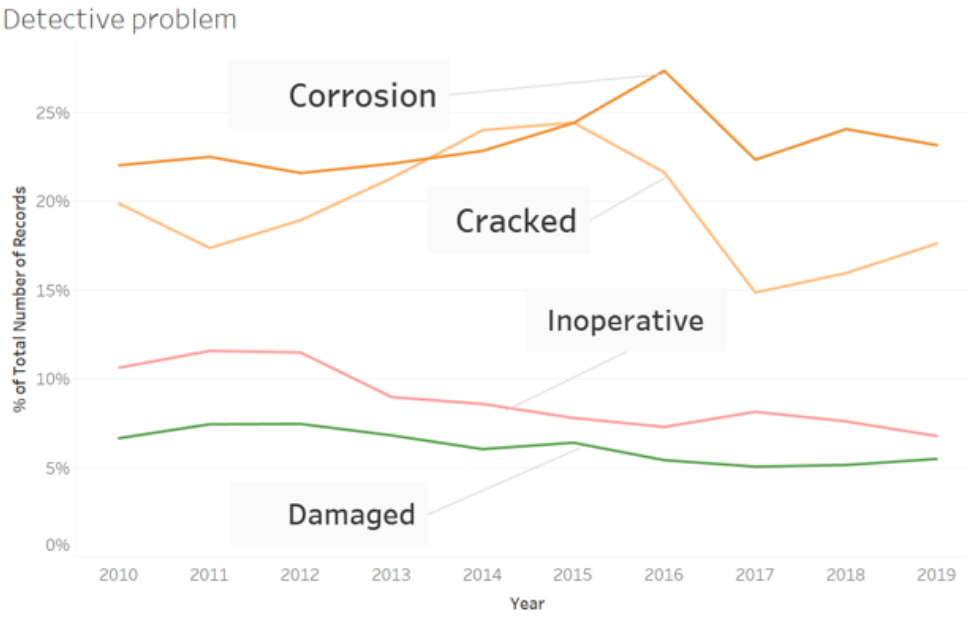

Figure 5: Top Four Detective Problem

- The top three District office receiving report were 29, 21, 27; The percentage of District office 15 increased over years.

5 Topic Overview

Topic ID	Topic Summary	Top 3 Aircraft Model
Topic0	Skin Enhancement Structure Crack	DOUG DC8;BOMBDR CL600; DOUG MD10
Topic1	Fuselage Frame Damage & Crack	DOUG DC6;DOUG DC8; BOEING 727
Topic2	Emergency Exit Light Lamp Inoperative	DHAV DHC8; EMB 120; BOEING 747
Topic3	Error Warnings and Messages	BOEING 787; BEECH 1900; BOEING 767
Topic4	Problem of Wing	DOUG MD11; EMB 135; BOEING 787
Topic5	Corrosion on Plane Fixed Accessories	LKHEED 382; BOEING 717; BOEING 727
Topic6	Main Cabin Floorboard Crack/ Main Deck	AIRBUS 300; DOUG MD11;

	Floor damaged	BOEING 747
Topic7	Seat Track (Flange) Corrosion	AIRBUS 330; BOMBDR CL600; BOEING 727
Topic8	Floor Support Corrosion/ Delamination On Floor Panel	AIRBUS 300; AIRBUS 310; BOEING 717
Topic9	Fluid/ Fuel/ Oil leakage + Pressure Bulkhead Damage	DOUG DC10; LKHEED 382; BOEING 767
Topic10	Smoke or Burning Smell, Flame and Fume	BOEING 787; EMB 145; SAAB SF340
Topic11	Emergency lights and Signs Inoperative due to Charger or Battery System	DOUG MD 80; BOEING 787; BOEING 777
Topic12	Landing System Malfunction	SAAB SF 340; DOUG MD 88; BOMDR DHC8
Topic13	Emergency Exit Sign Lens/Cover Crack	BOEING 737; BOEING 777; AIRBUS 320
Topic14	Crack on Skin or Skin Attached Part	BEECH 1900; DOUG MD11; LKHEED 382
Topic15	Engine Malfunction or Valve Failure	SAAB SF340; BOEING 767; DHAV DHC8
Topic16	Cargo Area or Frame Corrosion	AEROSP ATR42; BOMBDR DHC8; AEROSP ATR72
Topic17	Door Emergency Evacuation Slide out of Position	BOEING 737; BOEING 787; BOEING 717
Topic18	Door Difficult to Open or Part Damage	BEECH 1900; AIRBUS 320; EMB ERJ170
Topic19	Intercostal Area Part Crack or Escape Path Light Inoperative	BOEING 777; DOUG DC9; BOEING 737

Table 1: Topic Overview

As we can see from the bar chart below, the legend shows the first most relevant word in the word lists. For example, topic 1 is about damage and dent issue with the fuselage, skin of the aircraft, and the legend of topic 1 demonstrates "'crack' + 'skin' + 'frame' + 'fuselage'". It stands out as the most frequent topic, since there are 54,021 counts over 10 years.

In general, we can conclude that the 20 topics were distributed evenly without any significant outliers, implying that setting parameter to 20 after grid search is a good practice in this case.

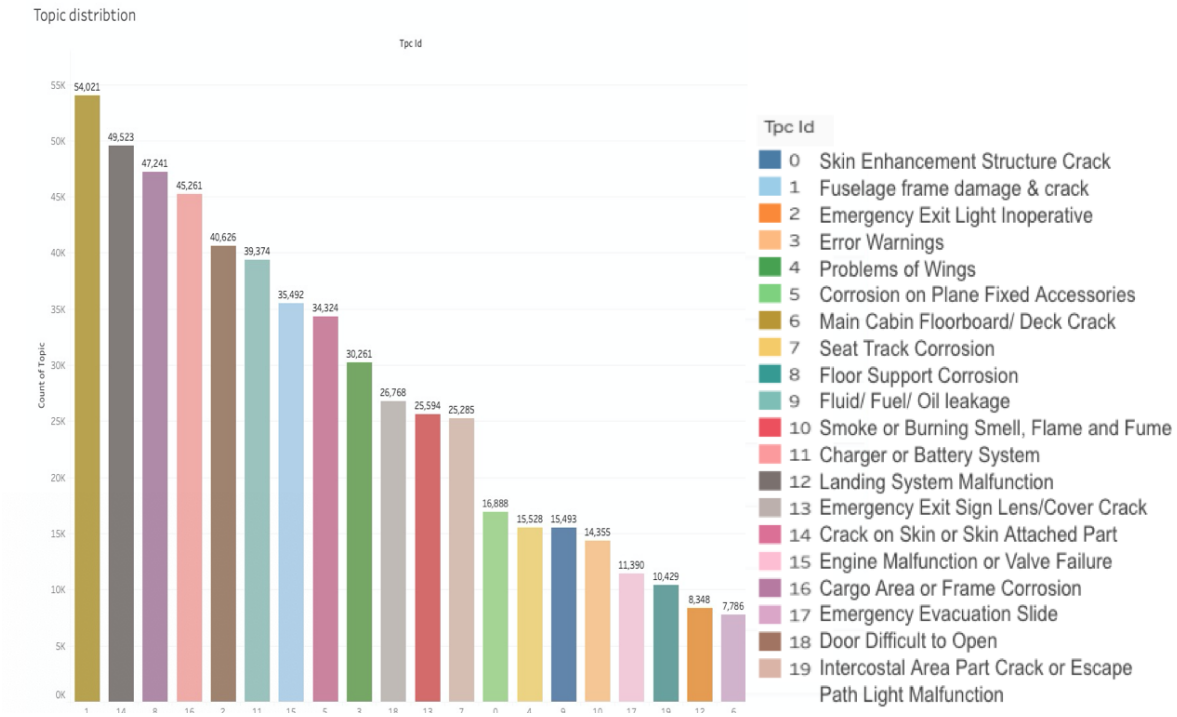

Figure 6: 20 Topic Distribution

And the percentage of each topic has been changing from year to year. As the chart below demonstrates, while some topic remained roughly constant, some others experienced dramatic change.

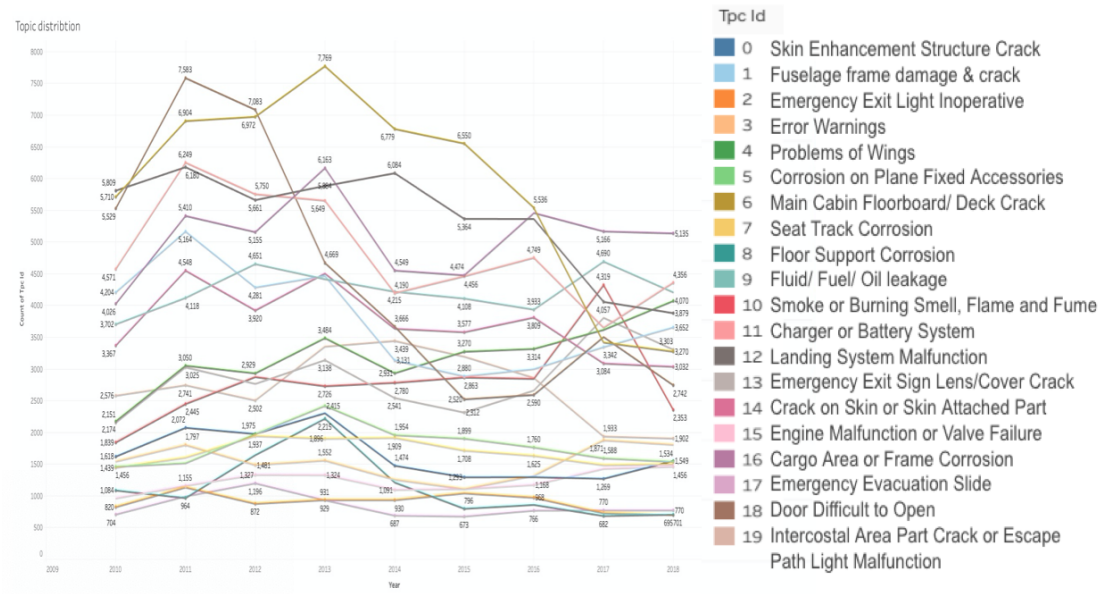

Figure 7: 20 Topic Trend by Year

6 Recommendation

6.1 Emerging Risks

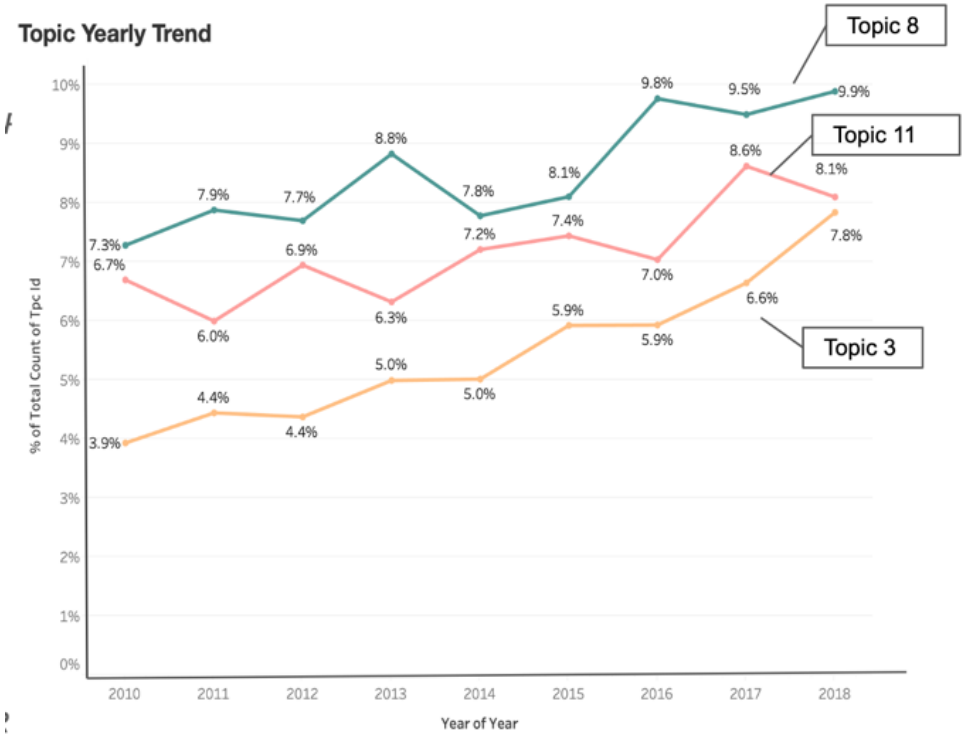

Figure 8: Yearly Trend of 3 Emerging Topics

When exploring the percentage of all the extracted topics over the year, our group identified 3 topics that have risen significantly: Topic 3, 8 and 11. Our team took a deeper look into these three topics and found some insights that might improve aviation safety.

6.1.1 Topic 3 False Activation of Warning System/Light/Messages was Increasing over the Year

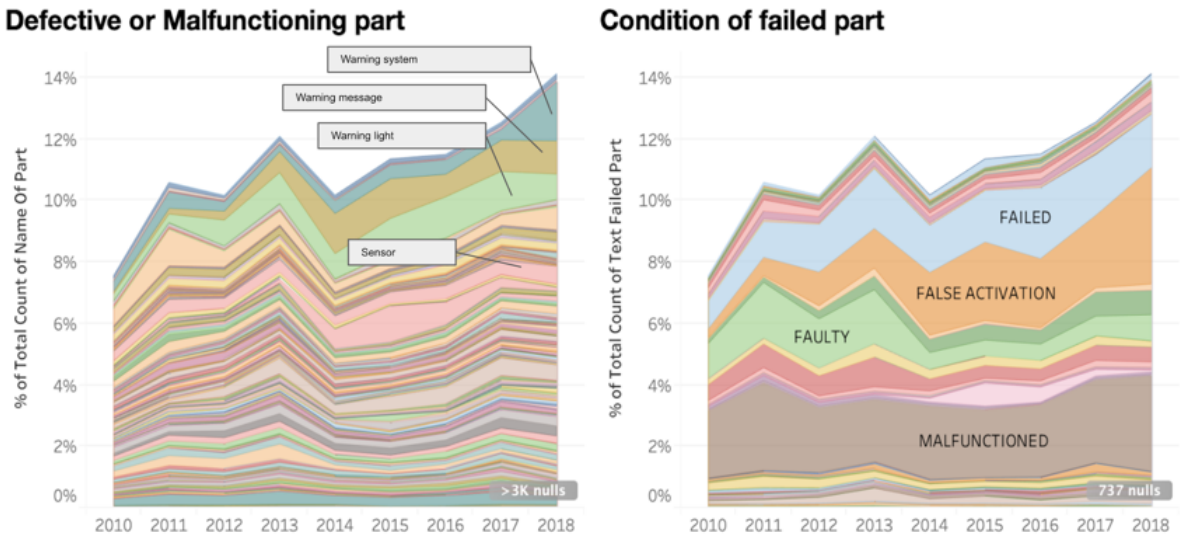

Figure 9: Overall Attributes of Topic 3

The defective part of topic 3 concentrated in warning system/ warning light and warning message/ sensors. Besides, the top 3 conditions of the failed part are malfunctioned, failed and false activation. Though 'malfunctioned', 'failed' remains steady over the years, the rise of 'false activation' issue is considered the drive for the topic surge.

When we typically looked at what kind of manufacturers and models have the highest 'false activation' issue. We found that the percentage of BOEING 767 increased in spite of the fact that the percentage of BOEING 767 didn't increase in the total reports.

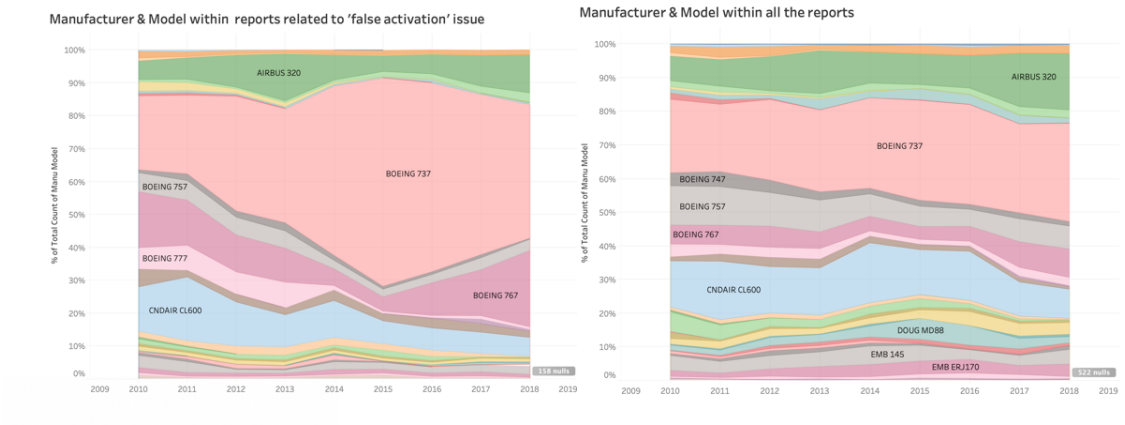

Figure 10: Overall Attributes of Topic 3

According to FAA, more than 98 percent of the 8,898 alerts from aviation 406-MHz emergency locator transmitters (ELTs) in 2017 in the U.S. were false alarms.[10] Only 112 alerts were authentic distress situations, according to figures from the NOAA Search and Rescue Satellite Aided Tracking (SARSAT) team. Every false alert has the potential to put rescuers in harm's way and waste valuable resources. The problem of ELT false alerts should raise more attention to aircraft operators and pilots since it shows the sign of continual growth.

False activation due to sensor can also be a potential threat. False aviation not only exits in ELT, it also appears to MCAS function. The pilot of the Ethiopian plane, which crashed just six minutes after takeoff from Addis Ababa, said that false activation of the MCAS function, as happened in the Ethiopian crash, can add to what is already a high workload environment. This bad data feeding into flight system could potentially cause severe problem.

It is worth mentioning here more work needs to be done to evident these observations. However, with the frequency and fatal consequences, more attention needs to be paid to this issue.

6.1.2 Topic 8Floor Support/ Floorboard/ Floor Panel Damage and Corrosion Requires Maintenance

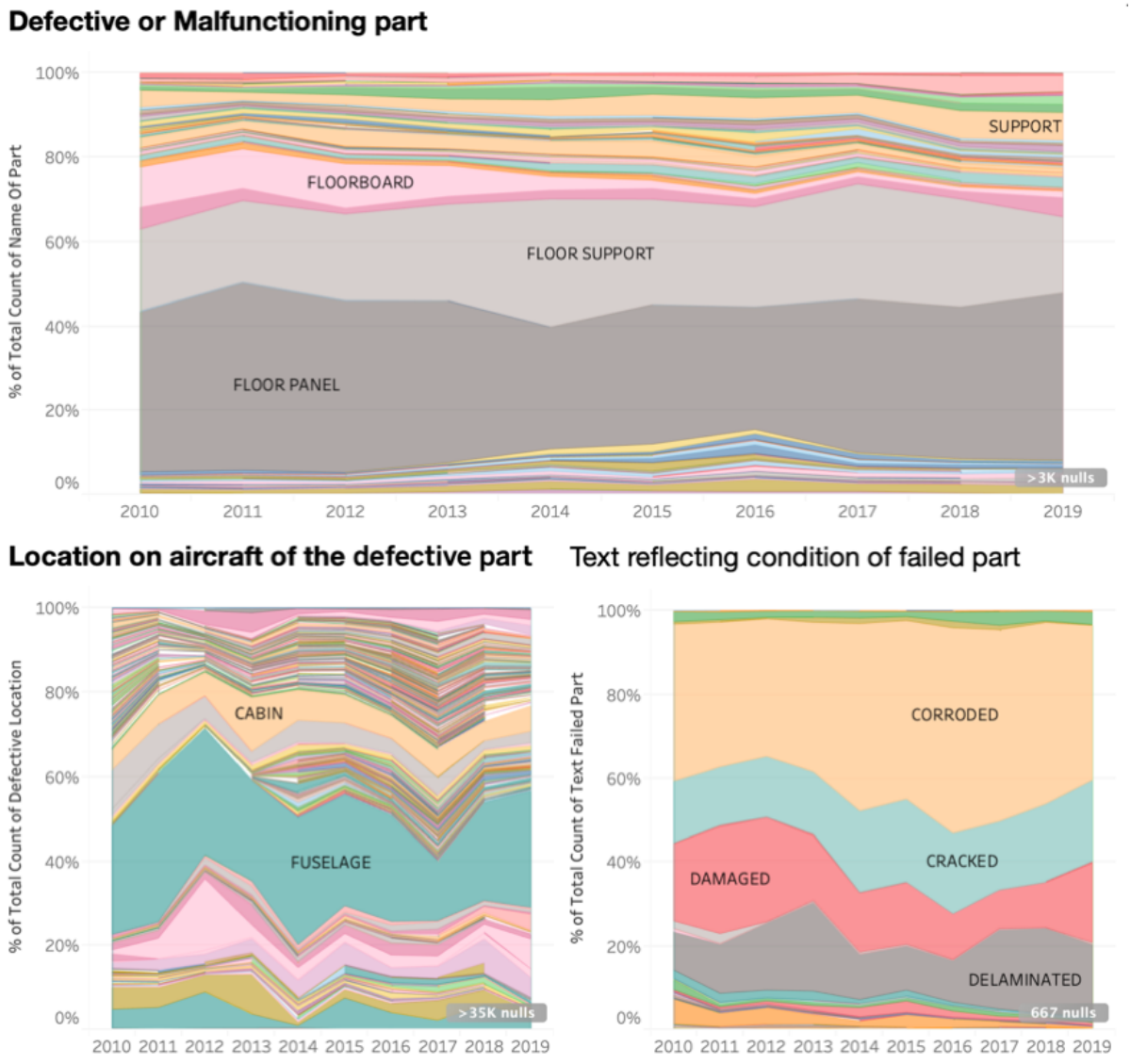

Figure 11: Overall Attributes of Topic 8

In the 8th extracted topic, a large percentage of defective part were related to floor in the cabin or fuselage. Their main issue was corrosion, crack, damage and delamination, which suggests more efforts on the maintenance of the aircraft should caught our attention.

According to Continued Operational Safety – Aging Aircraft published by FAA in 2018, the FAA is responsible for overseeing the continued airworthiness of more than 150,000 type-certificated GA airplanes over 30 years old. The average age of the aircraft in use today is getting older, thus it's more important to focus on the corrosion control and preventive maintenance to reduce the likelihood of related issues.

6.1.3 Topic 11Increasing Occurrence of Aircraft Batteries, Battery Pack Emergency Light Power Supply issue

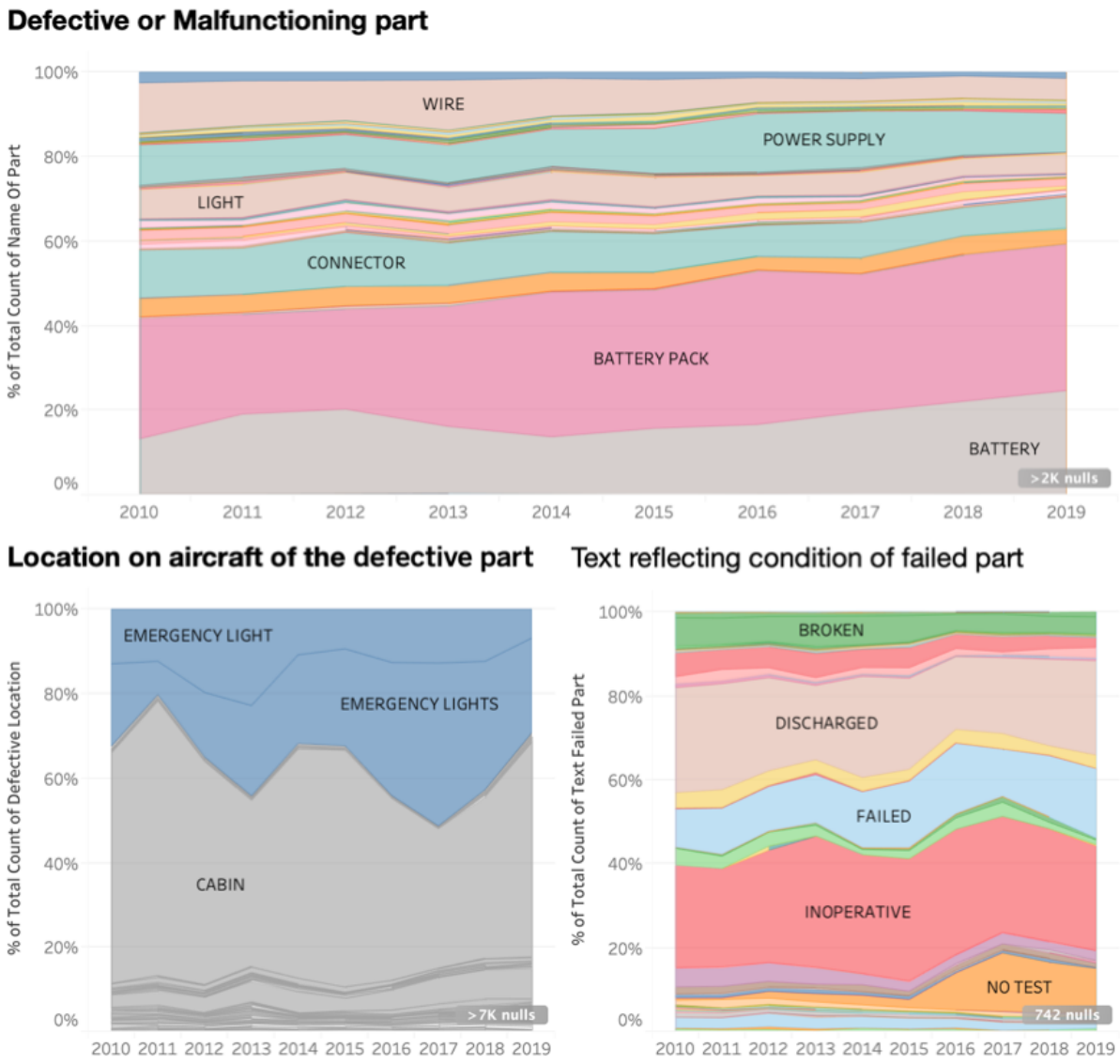

Figure 12: Overall Attributes of Topic 11

In the 11th extracted topic, a large percentage of defective part were related to battery/ battery pack/ power supply in the cabin or emergency lights. Their main issue was discharged, inoperative and failed.

Since the battery is one of the highest-maintenance components on board, for organizations with heavy flight schedules, such as charter services, batteries should have shorter maintenance cycles. Periodical check, regular check and develop the back-up battery system are suggested to reduce their occurrence.

6.2 Improving Issues

From the yearly percentage distribution, we found two of the topics have an obvious downward trend during the years, especially among the most recent years.

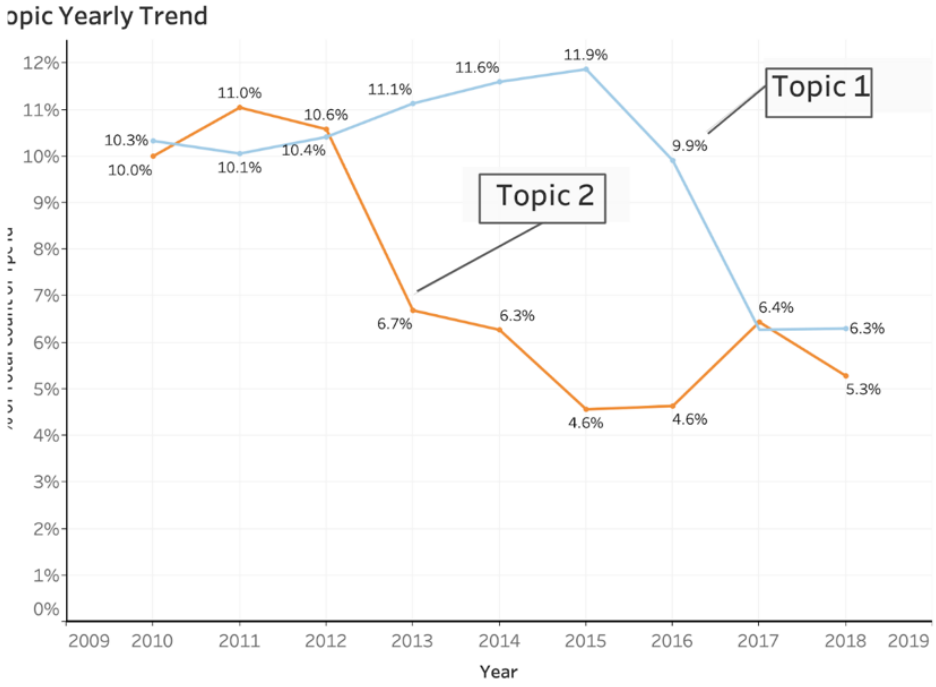

Figure 13: Yearly Trend of 2 Issues Mitigated

6.2.1 Topic 1. Fuselage Frame Damage Crack

This topic shows a sharp downward trend from Year 2015 to Year 2017. Moreover, the percentage of the topic is stable from Year 2017, keeping at the percentage around 6.3

A large portion of reports in this topic also happens in the fuselage area. it is in accordance with what the topic indicated. Also, from the reports, the largest amount of the reasons reported are cracked, dented which are also shown in the keywords of the topic.

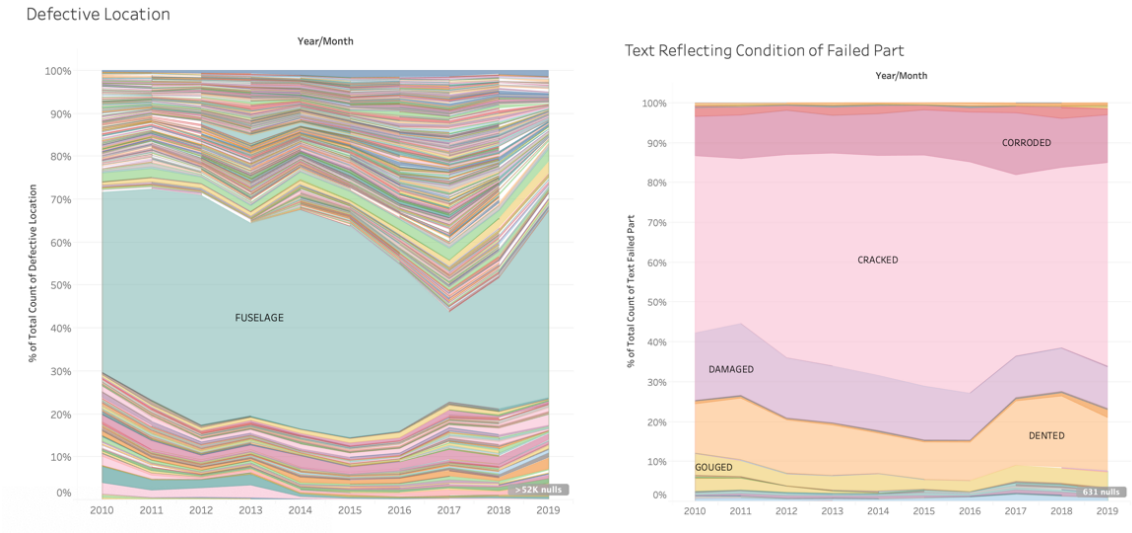

Figure 14: Overall Attributes of Topic 1

From the report and news, we infer that the downward trend may be because of the new technology. According to the BBC news from January Year 2014, the carbon fiber is mixed into the airframe to increase fuel efficiency and improve the aerodynamic performance of new aircraft is leading designers to move away from using aluminum in airframes. Furthermore, from the research of MIT in Year 2016, the newest Airbus and Boeing passenger jets flying today are made primarily from advanced composite materials such as carbon fiber reinforced plastic. MIT aerospace engineers have found a way to bond composite layers in such a way that the resulting material is substantially stronger and more resistant to damage than other advanced composites. In experiments to test the material's strength, the team found that, compared with existing composite materials, the stitched composites were 30 percent stronger, withstanding greater forces before breaking apart.

Moreover, the topic trend from year 2017 shows a plateau sign, which can possibly mean that the large quantity of aircraft in operation is using the new material composite. Therefore, we can infer that the downward trend of Fuselage frame damage crack is result from the new composite of the aircraft. And after 2017, the authority will receive around 6% incident reports compared to over 10% reports before 2015.

6.2.2 Topic 2 Emergency Exit Light/Lamp Inoperative

The issue with inoperative emergency light bulb/lamp shows a downward trend from Year 2012 to Year 2015. Then it shows a little upward trends during the year 2016 and 2017 and goes down again.

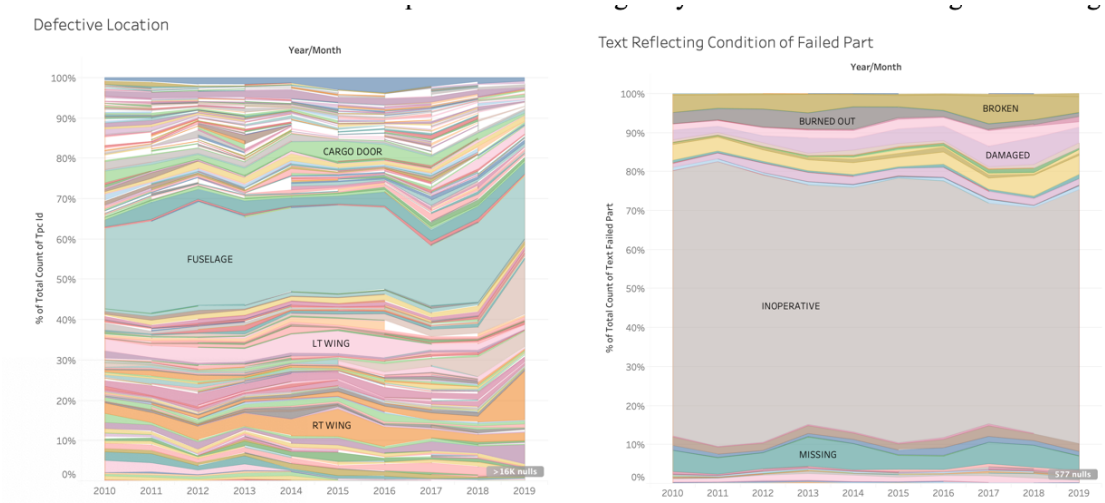

Figure 15: Overall Attributes of Topic 2

The defective location of this project is dominantly in the cabin and emergency lights part. The composition of emergency lights increases from Year 2014 and larger than cabin during Year 2017. The main reason of defective is inoperative, which is also an important keyword in the topic.

With technology accelerating in battery industry, efficient lighting systems are applied al-

most everywhere from cockpit displays to landing lights to mood lighting in the cabins. In one report published in 2011, LED and wireless systems are defined at the vanguard of new technologies whose applications in aviation are being explored. The research that day by STG Aerospace indicated that over 75% of operators would say yes to an LED lighting system. According to another aviation report published on May 1, 2013, 'The transition to light emitting diode (LED) lighting technology on aircraft is nearly complete, at least for new platforms.'. During that time, Boeing reports "more than 90 percent" of (its) backlog of more than 2,800 Next-Generation 737-800s will be delivered with the LED-lit Boeing Sky Interiors (BSI).

The trend of emerging light inoperative occurrence witnessed the plummet in 2012-2013, which can be inferred as the starting point of universal adoption in terms of efficient LED technology in aviation industry. And after 2013, the authority will receive around 4%-6% incident reports compared to over 10% reports before 2012.

6.3 Topic 18: Door Difficult to Open or Part Damage Topic 17: Door Emergency Evacuation Slide out of Position

Topic 18: Door Difficult to Open or Part Damage

Topic 17: Door Emergency Evacuation Slide out of Position

For these 2 topics, they are all related to the door of the aircraft, as we dig deeper into the locations of each issue, we noticed that there is a significant pattern.

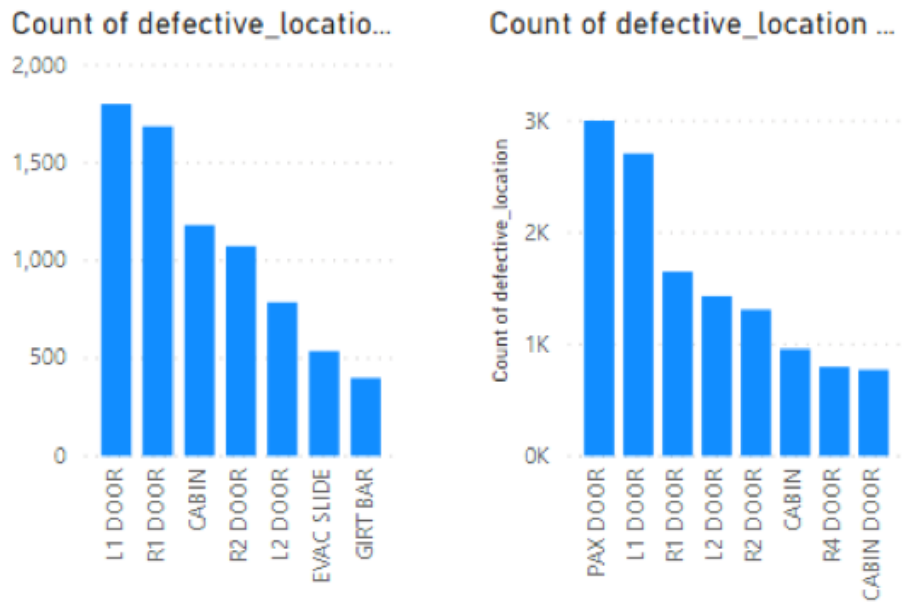

Figure 16: Overall Attributes of Topic 18

For topic 18, top 3 issue locations are Pax door, L1 door and R1 door. And L1 door has record counts of 2709, while R1 door has records count of 1651. For topic 18, top 3 issue locations are L1 door, R1 door and cabin. L1 door has record counts of 1800, while R1 door has records count of 1687, which indicates that the issues mostly appear on front doors and left doors are more common. By consulting with domain experts, we learned that L1 door are more often used for passengers getting on board. And the R1, R4/5 (Depends on model or size of the aircraft are used to load supplies. So, it seems that the most frequently used doors like L1 have more chance of part failure. So more frequent inspection and inspection could be scheduled for such doors to minimize this issue.

Rank	Model	Frequent Score
1	BOEING 737	4.042489
2	BOEING 787	3.327496
3	BOEING 717	3.129383

Also, by analyzing into the airplane models, we noticed that it seems Boeing models like

737, 787, and 717 has more frequent evacuation slide issues reported in SDR. By tracing back to some recent news, there is Boeing aircraft dropped its evacuation slides during flight on December 2nd. So some attention could be paid to such models.

7 Complementary Analysis

7.1 Abbreviation Analysis

After a brief glance at the content of the report, we find that there are a number of abbreviations [1] among the text and most of them are terminology. Thus, in order to better understand the report, while digging out potential topic trend from it, we first crawl abbreviations and their corresponding full description from Federal Aviation Administration, then match them in our dataset. Furthermore, we did exploratory data analysis by calculating ratio change of each abbreviation over years. By doing so we use frequency of the abbreviation divide total records of that year. The outcome is very small, considering our large size of the records. We also multiple results by 10000, so that we can see the trend more clearly.

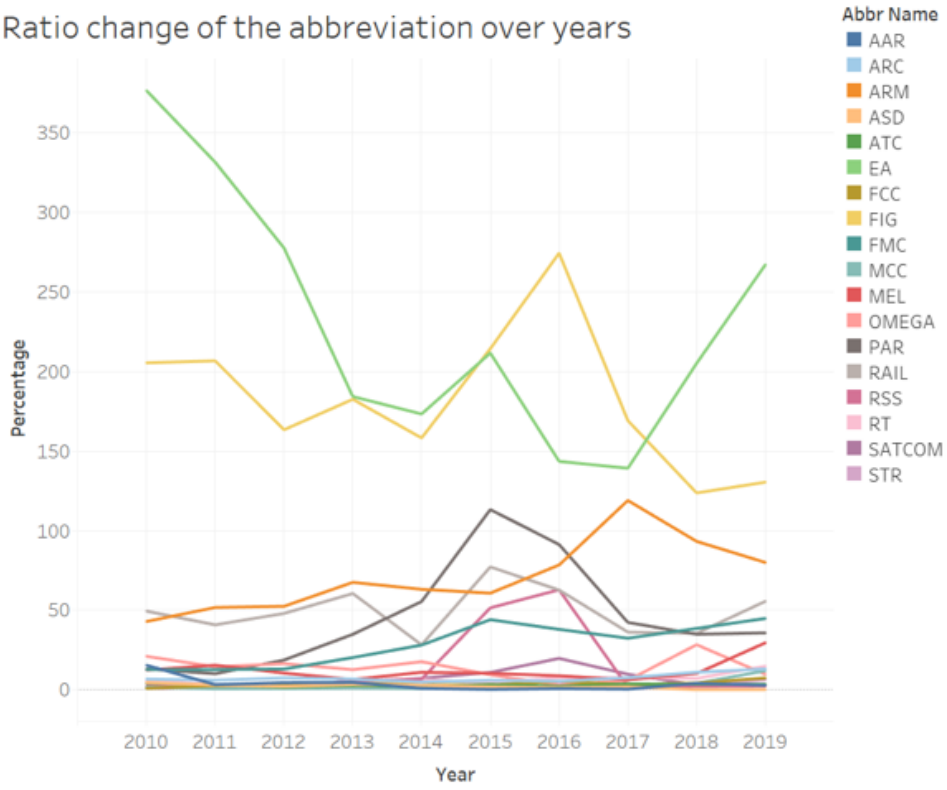

Figure 17: Ratio change of the abbreviation over years

When we compare ratio change of the abbreviation over the last ten years, we find that a) Most of the abbreviation show up more frequently in 2015 than other years; b) Compare to other abbreviations, EA (Environmental Assessment), FLG (Flight Inspection Group) appears more often than other words.

According to 2017 Preliminary Aviation Statistics provided by the National Transportation Safety Board, we could find that accidents rate for air carrier under CFR 121 is rather low in 2015 compares to other years. From our previous findings, abbreviations like EA (Environmental Assessment), FLG (Flight Inspection Group) and PAR (Precision Approach Radar) appears more frequent than other years. This could be the influence of incident happened to Malaysia Airlines Flight 370. It raised an alarm to the aviation industry to let them pay more attention

to the maintenance. However, not effective effort is a waste of time and resources. Above recommendation we provided could be one possible way to improve maintenance efficiency.

7.2 Phrase Analysis

From the model, we find 4436 Bigram and trigram in total. It is quite small compared to the total number of unique words count from the whole data set, so the trend is not the solid evidence of the emerging risk. But these phrases can give us more detailed insights with regard to aviation issues that are of low occurrence but more critical compared to the incidents in topic categories, like 'horizontal stabilizers', 'oxygen_mask'.

Initially, we filtered some phrases not emergency related. Those bigram and trigram are mainly the idioms, like the reaction after checking such as "approved_return", "without_inci--dent", "corrective_action", etc.

The plot below is the term frequency of the Top 10 phased generated by the model. From the top 10 bigram and trigram we can see most of them are related to the person who operated the aircraft such as the 'first_officer', the components of the aircraft('horizontal_stabilizer', 'tail_cone', 'circuit_breaker', 'oxygen_mask'), the condition of the aircraft('landed_safely'), as well as the safety system used by the aviation('master_caution').

8 Conclusion and Further Work

8.1 Conclusion

In this aviation contextual information extracting project, by using topic modelling with LDA, we are able to separate and summarize all the text documents into 20 distinct themes. Besides that, we also extracted some professional terms out of the text to represent some rare but more serious issues like horizontal stabilizer and oxygen mask. After that, by analyzing the trend of the occurrence of each theme and phrases, we are able to detect some emerging risk of 'False

Activation of Warning System', 'Floor support/ floorboard/ floor panel maintenance requirement', 'Aircraft Batteries, Battery Pack Emergency Light Power Supply issue' , some topics plummet because of the cutting-edge technology adoption in aviation industry and some insightful observations. With other attributes in the dataset, like aircraft model, defective location and defective reason and some external research. We can find some meaningful explanations and suggestions. We believe that entire procedures we used in this project is a rather comprehensive way to make use of text data.

Text reports are ubiquitous in this digital age. They are the essential and timely reflection of people's attitudes, malfunction records, economic foresight and can be extracted from open source platforms, indicating the cost of data collection is lower compared to well-constructed clean datasets with multiple numerical and categorical types of columns. Thus, the LDA text model and other NLP methods l;ike Bi-gram models in this project can also be applied in other industries, like detecting the heated topics in another industry or extracting the people's tweets text to detect what's their simultaneous ideas of up to date news.

In a business context, unstructured text data include emails, social media posts, chats, support tickets, surveys, etc. Sorting through all these types of information manually often results in failure. Text analysis instruments can save hours for data extraction and other resources but provide more essential and profound insights.

8.2 Future Work

For our future work, we noticed that there are still some areas that we can dig deeper and improve. Some of the topic categories are not clean enough, like topic 9 include both Fluid/ Fuel/ Oil leakage issue and pressure bulkhead damaged topic 19 is associated with intercostal area part crack or escape path light inoperative. This was due to the LDA model is rather sensitive to the number of topics. The way to deal with this messy topic is to further fine tune

the parameters or rerun the LDA in side the clustered documents to separate these topics and then detect the key issues behind.

References

Bíró, István and Jácint Szabó (2009). "Latent dirichlet allocation for automatic document categorization". In: *Joint European Conference on Machine Learning and Knowledge Discovery in Databases*. Springer, pp. 430–441.

Blei, David M, Andrew Y Ng, and Michael I Jordan (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Lee, Sangno, Jaeki Song, and Yongjin Kim (2010). "An empirical comparison of four text mining methods". In: *Journal of Computer Information Systems* 51.1, pp. 1–10.