Transportation Research Part C 87 (2018) 105–122

Contents lists available at ScienceDirect

Transportation Research Part C

journal homepage: www.elsevier.com/locate/trc

Using structural topic modeling to identify latent topics and trends in aviation incident reports

Kenneth D. Kuhn

Operations Researcher, RAND Corporation, 1776 Main St, Santa Monica, CA 90407, United States

ARTICLE INFO

Keywords:
Aviation
Aviation safety
Natural language processing
Topic modeling
Text mining
Aviation safety reporting system

ABSTRACT

The Aviation Safety Reporting System includes over a million confidential reports describing aviation safety incidents. Natural language processing techniques allow for relatively rapid and largely automated analysis of large collections of text data. Interpretation of the results and further investigations by subject matter experts can produce meaningful results. This explains the many commercial and academic applications of natural language processing to aviation safety reports. Relatively few published articles have, however, employed topic modeling, an approach that can identify latent structure within a corpus of documents. Topic modeling is more flexible and relies less on subject matter experts than alternative document categorization and clustering methods. It can, for example, uncover any number of topics hidden in a set of incident reports that have been, or would be, assigned to the same category when using labels and methods applied in earlier research. This article describes the application of structural topic modeling to Aviation Safety Reporting System data. The application identifies known issues. The method also reveals previously unreported connections. Sample results reported here highlight fuel pump, tank, and landing gear issues and the relative insignificance of smoke and fire issues for private aircraft. The results also reveal the prominence of the Quiet Bridge Visual and Tip Toe Visual approach paths at San Francisco International Airport in safety incident reports. These results would, ideally, be verified by subject matter experts before being used to set priorities when planning future safety studies.

1. Introduction

This article describes and applies methods that allow an analyst to identify topics and trends in aviation incident reports. The article shows how the specific methods used here, which have not been applied in the aviation systems domain before, and other similar methods offer promise for setting priorities when planning more detailed aviation safety research.

Public and private agencies in Europe, the United States, and elsewhere are implementing many substantive changes to air transportation operations. These changes seek to address challenges posed by increasing air traffic, increasing diversity of air traffic (for example due to the emergence of small Unmanned Aerial Systems), aging infrastructure, and ongoing efforts to make air transportation safer and more efficient. For example, the Federal Aviation Administration (FAA) is in the midst of implementing "Wake Recategorization" procedures at busy airports, "reducing separation criteria for multiple runway operations."[1] Wake Recategorization is a part of the NextGen initiative to modernize air traffic control.

Researchers have developed models and simulations for assessing the safety implications of possible operational changes. For

E-mail address: kkuhn@rand.org.

[1] NextGen Priorities Joint Implementation Plan, 2017–2019. available online at: https://www.faa.gov/nextgen/media/ng_priorities.pdf.

https://doi.org/10.1016/j.trc.2017.12.018
Received 12 September 2017; Received in revised form 28 December 2017; Accepted 30 December 2017
Available online 05 January 2018
0968-090X/

example, Netjasov (2012) introduce a mathematical risk assessment model for evaluating airspace (re)designs in terms of potential conflicts at intersections and along airways. Zhao et al. (2013) report on simulation experiments using a novel Computational Red Teaming framework to identify vulnerabilities and bottlenecks in a network representing an airport terminal area. Barnett et al. (2015) compare historical and simulation data to evaluate alternative configurations of the North Airfield at Los Angeles International Airport. Chittaro (2017) evaluate different methods for delivering aviation safety briefings using surveys of university students.

This effort is, instead, based on exploratory analysis of recently reported text data describing actual aviation incidents. There is a wealth of such data available to analysts; in fact the scale and scope of the data have proven problematic in the past. Techniques developed in the fields of natural language processing (NLP) and machine learning (ML) are used to highlight areas of concern, for example the latent *topics* that are at the heart of recent reports.

Many articles published during the last fifteen years highlight the promise of applying NLP methods to aviation safety reports. For example, Péladeau and Stovall (2005) claim that JetBlue airlines representatives found such methods "show great potential for providing airline managers with clear, relevant insights."

Relevant academic research has largely focused on developing models and algorithms for categorizing incident reports (Posse et al., 2005; Oza et al., 2009; Switzer et al., 2011; Wolfe, 2007; Persing and Ng, 2009). As Wolfe (2007) notes, the challenge of gaining useful insights from the large volume of (diverse) incident reports has been partially "mitigated by the use of classification schemes." Wolfe (2007) and the other authors cited rely on an initial set of labels and a training data set that includes incident reports which have been labeled by subject matter experts. The goal is to develop an algorithm that can stand in for a domain expert.

This article, in contrast, introduces a method to automatically identify topics (not categories) within incident reports based only on the specific reports being analyzed. Topics represent latent structure in the corpus being analyzed. Relatively many topics will be present, to varying degrees, in any one report and many reports will include material about any one topic. The overall methodology is quite flexible. An analyst could, for example, use the methods introduced here to find any number of topics within a set of incident reports that have been, or would be, assigned to the same category (or combination of categories) when using methods introduced in earlier research. The methods introduced here do not rely on previously defined categories or training data and can be applied relatively rapidly and easily by a single analyst (although discussion among, or subsequent studies by, subject matter experts would be helpful).

Topic modeling has previously been applied in the transportation domain. Pereira et al. (2013) previously used topic modeling to study (ground) traffic incidents. The authors show that they can predict the duration of incidents using text data available at traffic management centers with up to 28% less error than alternative methods that do not use the text data. Zhang et al. (2018) detect traffic accidents using social media data, and show how one method based on 'deep learning' outperforms an alternate method based on a form of supervised topic modeling. Das et al. (2016) use topic modeling and other NLP methods (including listing the most frequently used terms) to explore papers published by the Transportation Research Board (TRB). The authors are particularly interested in topic prevalence over time. Sun and Yin (2017) use topic modeling to identify 50 key topics within transportation research before pointing out which topics are prevalent in specific academic journals and in papers written by authors from specific countries. Sun and Rahwan (2017) go on to develop a network model of collaborations among transportation researchers. Although not an application of topic modeling, this work is interesting as it illustrates a way to use document metadata to find hidden structure and because it provides metrics that the authors claim can be used "to evaluate researchers' performance and impact" (Sun and Rahwan, 2017).

Das et al. (2017) extend some of the works cited in the preceding paragraph, finding topics within 15,357 TRB compendium papers using a modeling approach that explicitly makes use of document metadata including publication year, responsible review committee (e.g., ADC80: Alternative Transportation Fuels and Technologies), author name, and author affiliation. The authors used structural topic modeling (STM), a particular form of topic modeling that has gained prominence relatively recently. Das et al. (2017) remark that "STM provides fast, transparent, replicable analyses." STM, unlike other forms of topic modeling, directly estimates the impacts of metadata on topic prevalence. The results can reveal trends in the frequency with which topics appear over time as well as relationships between covariates and topic prevalence or word use within a topic. More specifically, STM is a generalization of more commonly used latent Dirichlet allocation (LDA) and correlated topic models. STM is unique in that it includes document-level random variables whose distribution functions can depend on covariate data. STM allows analysts to directly model the trends and relationships mentioned previously. They would otherwise need to perform post hoc, heuristic analyses of the results of, for example, an LDA model. More details on LDA and STM models are provided in later sections of this article.

STM is here applied to a large corpus of aviation incident reports. One goal of this study was to evaluate the usefulness of this novel method for identifying safety issues. A more ambitious goal was to begin using this method to find previously unreported connections or themes in incident reports. It was hoped that STM would highlight intuitively interesting results that would be otherwise difficult to uncover such as: trends in aviation incident reports over time, seasonal patterns in reports, or relationships between report metadata and topic prevalence. STM is ideal for analyzing the relevant data set, described later in this article, because the data set includes free text narratives of aviation safety incidents accompanied by much metadata that is, intuitively and empirically, linked to topics within the narratives.

2. Aviation safety reporting system data

2.1. Data description and initial exploration

The Aviation Safety Reporting System (ASRS) lets pilots, air traffic controllers, airline dispatchers, and others submit confidential

JFK Tower cleared us for takeoff on 31R; Kennedy 1 Departure (Breezy Point Climb). After our takeoff roll the Tower cleared a heavy aircraft into position on 22R Intersection YA and hold at idle thrust at an intersecting runway. At around 100 knots we received a pretty good jolt from his thrust buffet. Quick left rudder and left aileron was used to counteract the thrust buffet. We notified ATC Departure to relay the message to JFK Tower about the event. ATC should not position and hold someone on 22R YA after clearing someone for takeoff on 31R. The risk is too high and is only dependent on the other aircraft not using more than idle thrust or turning his aircraft. Of course the best way is to not have an aircrafts thrust directly in the path of another aircraft taking off. However; it . . .

Fig. 1. Example of a narrative from an ASRS record.

reports of safety incidents. The FAA and the National Aeronautics and Space Administration (NASA) developed and manage the ASRS, in part to "provide data for planning and improvements to the future National Airspace System."[2] In 2015, the ASRS database included over 1.3 million records and roughly 7500 additional reports were being added each month (see Footnote 2). Similar systems exist elsewhere, including CHIRP in the United Kingdom and REPCON in Australia.

Analysts anonymize submitted reports and code the results into a database that is available to the public. Database records include several structured fields but also lengthy blocks of free text, notably narratives describing incidents. Fig. 1 shows the beginning of a narrative portion of an ASRS record.

When this research project began there were 25,706 ASRS records available to the public based on incidents that took place between January 2011 and December 2015. These records are explored in this article. The studied data set is large and diverse enough to illustrate some of the difficulties encountered when analyzing ASRS data. The data set also contains relatively recently reported and thus still relevant data.

In the incident reports in this data set for which the flight "Mission" of the primary aircraft is listed, it is described as "Passenger" 68% of the time. 14% of the relevant incident reports focused on a "Personal" flight, 6% on a "Cargo/Freight" flight, and the remaining 12% on some other category of flight such as "Skydiving." The reporting organization was classified as an "Air Carrier" for 58% of reports, as "Government" for 16%, as "Personal" for 12%, and as some other category (including "Military" and "Corporate") of organization for the remaining 13% of reports. Other structured data fields within ASRS records include some containing information on the month when the incident was reported, the locale (e.g., LGA.Airport, TUL.TRACON), meteorological conditions, and phase of flight (e.g., Climb, Descent).

There have been prior research efforts exploring ASRS data, dating back more than 30 years, searching for issues that could be resolved to prevent or reduce the frequency of specific types of incidents. For example, Billings and Reynard (1984) describe a project that lasted for seven years, involving manual classification of incident reports and examination of relationships in derived data. An example of the authors' conclusions is that "The most common controller errors involve failure to coordinate traffic with other elements of the air traffic control system" (Billings and Reynard, 1984). Bliss et al. (1999) report on the role of cockpit alarm systems in incidents.

There has been recent progress in the fields of machine learning, computational linguistics, and natural language processing. There are established theories and tools for performing largely automated and relatively rapid analysis of a corpus. As an example, there are fast and easy-to-implement algorithms that identify the most frequently observed *n-grams*. An n-gram is a specific sequence of n items, here words, that appears in a document, here a text document. For example, in the text "JFK Tower cleared us for takeoff" there are three 4-grams including "JFK Tower cleared us," "Tower cleared us for," and "cleared us for takeoff." n-grams are used in many applications, including those that: compress files, check spelling or grammar within documents, match files, or identify the language of text or speech. A list of the most frequently observed n-grams within a corpus can be useful to see the language or jargon used within a corpus. It may also highlight the importance of specific people, objects, or actions. For example, if many incident reports referred to "Kennedy 1 Departure (Breezy Point Climb)" then this might indicate a safety issue on this particular route. (This particular n-gram was taken from the narrative shown in Fig. 1 but was not a particularly frequently observed n-gram.)

Table 1 shows the most frequently observed n-grams in the data analyzed here. Note that some of the results reflect industry jargon or how ASRS analysts code submitted reports. For example, FO is an acronym that refers to the First Officer of a flight. The raw text narratives are full of jargon, including jargon used by air traffic controllers and jargon used by pilots. Narratives also vary tremendously in terms of the level of detail provided when describing incidents. This presents an issue here. For example, the prominence of the phrase "cleared visual approach runway" in the results shown in Table 1 owes much to its universal use in the aviation domain.

Table 1 reveals the importance of the first officer and pilot. The application of other, more sophisticated NLP techniques holds the promise of more interesting results. This explains the recent proliferation of papers describing applications of NLP to ASRS data.

2.2. Prior natural language processing of incident reports

Suomi and Sjöblom (2009) provide a general introduction to the application of data mining techniques to aviation incident

[2] NASA ASRS Program Briefing. Available online at: https://asrs.arc.nasa.gov/docs/ASRS_ProgramBriefing2015.pdf.

Table 1
Frequently observed n-grams in ASRS narratives.

Phrase	Observation count
5-grams	
First officer FO pilot flying	38
Cleared visual approach runway R	33
Climb via SID except maintain	30
Declared emergency returned departure airport	29
We cleared visual approach runway	28
4-grams	
First officer pilot flying	286
In future I will	213
I asked first officer	206
Cleared visual approach runway	160
Aircraft maintenance manual AMM	149
3-grams	
Air carrier X	1096
First officer I	892
At point I	872
At time I	697
Landed without incident	614

reports and describe "test runs for four data mining products, for possible use in the Finnish civil aviation authority." The authors conclude that such products are unique and useful but warn that skilled users are needed to interpret results. Sjöblom (2014) goes on to test another three methods, this time focusing on using "clustering to find similarities between reports," but again drawing general conclusions that highlight the promise and difficulty of applying NLP methods to aviation incident reports. Tanguy et al. (2016) take a similarly broad view of the field, while describing topic modeling, classification model fitting, and other methods at a relatively high level. The authors note that they have developed tools that "are currently in test or in use both at the national and international levels, by airline companies as well as by regulation authorities" (Tanguy et al., 2016). One interesting conclusion is that "It appears that topic modelling is very suitable for [incident report] data" and that identified topics highlight "relevant aspects of [these] documents, as can be seen through an expert's interpretation" (Tanguy et al., 2016). The authors highlight the importance of interactive analysis, where human experts explore results identified by algorithms. Specific findings directly related to aviation safety are not provided.

Many of the most relevant, most detailed prior publications describe models that categorize reports using a set of incident "shaping factors" originally proposed by Posse et al. (2005). Persing and Ng (2009) call this the "cause identification" problem. Solutions to this problem enable analysts to automate a helpful but tedious task. Abedin et al. (2010) describe two different NLP methods for cause identification. Ahmed et al. (2010) similarly introduce a labeling technique the authors call Semi-supervised Impurity based Subspace Clustering - Multi Label (SISC-ML) and apply this technique to ASRS (and other) data. Switzer et al. (2011) also "demonstrate a system of automatic classification of the ASRS shaping factors" and claim "an improvement in performance" due to the use of boosting. All of the works cited in this paragraph make use of the previously established set of shaping factors and training data provided by subject matter experts. It would take substantial effort to generate a new set of labels and training data. There are, however, many reasons why analysts may want to categorize incident reports using a set of labels other than these shaping factors. Analysts may want to divide reports into more or less homogeneous sets of reports than would be possible using the shaping factors. Analysts may be studying a subset of incident reports from the ASRS database and may want to generate labels that are meaningful for, and can be used to appropriately summarize, just the specific reports of interest.

Ananyan and Goodfellow (2004) also categorize incident reports but use categories defined by the International Air Transport Association (IATA). As the authors note, "Historically, categorization of safety events to WinBASIS system of categories utilized by IATA was carried out by representatives of participating airlines based on manual evaluation of report narratives" (Ananyan and Goodfellow, 2004). The cited study also contrasts with those cited in the preceding paragraph because it promotes a specific product, the PolyAnalyst. Similar studies describe other products for parsing and analyzing aviation safety reports for commercial air carriers. Nazeri (2003) describes the Aviation Safety Data Mining Workbench, used at American Airlines. The software in question was able to identify areas of concern, for example by looking at the distributions of structured variables. Nazeri (2003) notes that "78% of {departure = FLORIDA} coincide with {event = ALT_DEVIATION}," in other words 78% of incidents involving flights departing from Florida involved an altitude deviation event. Nazeri (2003) mentions a tool called FindSimilar which enables users to identify "reports that are similar to a report selected by the user (as the target)" where similarity is defined in terms of both structured variables and insights gleamed from unstructured narrative data. Nazeri (2003) claims that American Airlines was "thrilled to have the speed, flexibility, and accuracy of these tools - especially for application on [their] coming field rich data base of self-reported crew errors and safety concerns." Péladeau and Stovall (2005) describe a "technology demonstration" of the WordStat system at JetBlue airlines. WordStat enables users to visualize text data and explore relationships among incident reports. Of special relevance here, Péladeau and Stovall (2005) note that "tools for automatic document classification in WordStat were still under development. We believe such kind of tools could potentially be valuable for the JetBlue aviation safety team to support the manual categorization

of events, to allow retrospective categorization of uncoded reports."

Several of the quotes cited in this section hint or plainly indicate that topic modeling, which aims to identify "the main themes that pervade a large and otherwise unstructured collection of documents" (Blei, 2012), could be useful. Topic modeling has gained prominence recently as analysts search for ways to organize the large volume of text data available on the internet. It does not require previously established labels or training data. It can be used to cluster or categorize incident reports, although this is but one use case. It is worth noting that topic modeling focuses on topics. As was noted previously, many topics will be present, to varying degrees, in any one report and many reports will include material about any one topic.

There are few published reports of applications of topic modeling to aviation incident reports, likely due to the relatively recent emergence of this type of natural language processing. El Ghaoui et al. (2013) describe a suite of methods, called sparse machine learning, for topic modeling and other tasks, testing the methods on ASRS data. The authors "reveal causal and contributing factors in runway incursions" and "automatically discover four main tasks that pilots perform during flight" (El Ghaoui et al., 2013). The four tasks that pilots perform are labeled aviate, navigate, communicate, and manage systems (El Ghaoui et al., 2013). The automated analysis of runway incursions reveals specific runway/taxiway intersections that are frequently mentioned in incident reports. As was mentioned previously, Tanguy et al. (2016) describe at a high level, the potential promise of topic modeling for analysis of incident reports.

3. Topic modeling

3.1. Latent Dirichlet allocation

This article describes applications of structural topic modeling to ASRS data. STM is a form of topic modeling, a probabilistic way to describe documents in terms of topics. A single document is often linked to many topics. At the same time, topics are present in many documents. As described in the preceding section, topic modeling allows analysts to organize and study a corpus. The most common form of topic modeling is latent Dirichlet allocation. A brief introduction to LDA is provided here.

LDA assumes documents and the words within them are derived from a "generative probabilistic model" (Blei et al., 2003). Each document, indexed by d which goes from 1 to D, is, in theory, generated via the following model.

- The number of words N_d is a random variable drawn from a Poisson distribution.
- The model parameter θ_d, representing topic proportions within the document, is a random variable drawn from a Dirichlet(α) distribution.
- For each word in the document (words being indexed by n), the topic of the word is $Z_{d,n}$, a random variable drawn from a multinomial (θ_d) distribution.
- For each topic (topics being indexed by k), the model parameter β_k represents word proportions within the topic.
- The word themselves, $W_{d,n}$ terms, are random variables drawn from another multinomial distribution defining $p(W_{d,n}|Z_{d,n},\beta_k)$ terms.

Fig. 2, by Blei (2012), shows the plate notation representation of the LDA model commonly used in the relevant technical field and serves as a visualization or alternate explanation of the model. There is one 'plate' and an associated parameter θ_d for each of the D documents in a corpus. There is also another, inner plate that is replicated for each of the N words in the document. Z and W, topics and words, appear in this inner plate. There is a topic linked to each specific place where a word appears in each specific document. There is also a plate replicated for each of the K topics in a corpus that includes the β_k parameters set via the topic parameter η. According to the theory, the topic determines the distribution used to generate the word. Different topics can generate the same English language word.

Estimates of the parameters of the model described previously provide researchers with data on topic representation within each document and within the corpus. These data also reveal the words most associated with each topic, allowing analysts to ascribe intuitive meanings to topics. In its most general form, LDA can also be applied to non-text data and has proven useful in image recognition.

Applying a Bayesian approach, the "key inferential problem" is to compute the posterior distribution of the latent variables given the text from a document (Blei et al., 2003). This can be expressed as evaluating the following application of Bayes Law (Blei et al., 2003).

$$p(\theta,z|w,\alpha,\beta) = p(\theta,z,w|\alpha,\beta)/p(w|\alpha,\beta) \tag{1}$$

Although it is not typically feasible to directly evaluate Eq. (1), there are many ways to find approximate solutions using

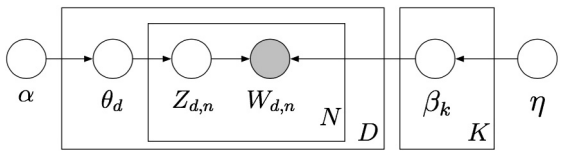

Fig. 2. Latent Dirichlet allocation model, in plate notation (by Blei, 2012).

expectation-maximization (EM) algorithms.

Note that within an LDA model, the probability of observing a particular word at a particular location within a document is a function only of the relevant topic and the model parameters. In particular, LDA does not allow us to model changes in the representation of topics and words within documents over time or as a function of (other) covariate data.

3.2. Structural topic modeling

Structural topic modeling extends the LDA framework. STM allows for correlations among topics. Covariate data including document metadata influence topic prevalence within documents. STM also uses (document-specific) covariate data to define distributions for word use within a topic.

In the preceding subsection describing LDA, there was an assumption that there was a model parameter θ_d for each document d that represented topic proportions within the document. This model parameter was assumed to be a random variable drawn from a Dirichlet(α) distribution. Note, in particular, that this distribution was common across all documents. In an application of STM, this parameter is a random variable drawn from a Log-normal distribution that is based on document-level data (Roberts et al., 2014b).

In the preceding subsection, there was also an assumption that there were β_k terms, model parameters that represented word proportions within a topic. These also were common across the corpus. In an application of STM, a multinomial logit model is used for word distributions where a word's prevalence is based on topic, document covariate data, and topic-covariate interactions (Roberts et al., 2014b). A relevant equation, provided by Roberts et al. (2013), is shown below. Here individual words in the relevant vocabulary of possible words are indexed by v. m_v is the baseline log frequency for word v. The κ terms in the equation below are used to capture adjustments based on the relevant topic and covariate data.

$$\beta_{d,k,v} \propto \exp(m_v + \kappa_v^{\cdot,k} + \kappa_v^{y,\cdot} + \kappa_v^{y,k}) \tag{2}$$

Fig. 3, also by Roberts et al. (2013), shows the plate notation representation of the structural topic model and can be compared to Fig. 2. Roberts et al. (2013) and Roberts et al. (2014a) provide further technical details on structural topic modeling.

Farrell (2016) recently applied STM to scientific texts on climate change, revealing links between corporate funding and the framing of scientific studies. STM has also been applied to social media data in a variety of ways. Mishler et al. (2015) show that "STM can be used to detect significant events such as the downing of Malaysia Air Flight 17" when applied to twitter data. Reich et al. (2015) show how STM can be used to explore relatively large data sets including course evaluations and discussion forum posts from a Massive Open Online Course. As mentioned in the Introduction, Das et al. (2017) use STM to identify topics in transportation research articles. The authors conclude that STM allowed them to develop "a unique tool to explore topical prevalence and content and to identify more relevant trends in the expensive fields of transportation research." Das et al. (2017) point out a potential

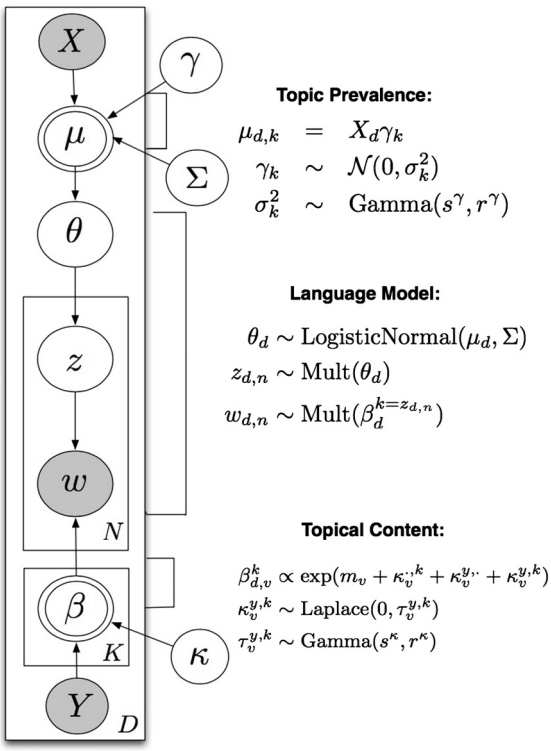

Fig. 3. Structural topic modeling, in plate notation (by Roberts et al., 2013).

practical application as well involving “the assignment of papers to the appropriate committees” within the Transportation Research Board.

STM is well suited to the analysis of ASRS data where free text narratives of aviation safety incidents are accompanied by substantial document metadata. This metadata includes several variables mentioned earlier: flight mission, the type of organization reporting the incident, meteorological conditions, and phase of flight. Tanguy et al. (2016) noted that topic modeling has proven useful for airlines and regulatory agencies because topics highlight “relevant aspects” of incident reports to human experts. STM can go a step further, highlighting the links between these relevant aspects and conditions. For example, STM might reveal that issues related to the use of landing gear are particularly prevalent for a certain type of aircraft. LDA would only show us that there is a landing gear topic in incident reports. An analyst would have to know to run post hoc analysis involving type of aircraft to ever uncover the relationship. STM works well in this context because it can take advantage of the effort that has already gone into developing the ASRS database, populating it with both a very large number of free text incident reports and with much useful metadata.

3.3. On validation

It is important to recognize that topic modeling is a form of unsupervised learning. There is no ground truth data, making it challenging to validate the output of a topic model.

Many authors frame their topic modeling efforts as exploratory in nature and do not attempt to explicitly validate their models but do allow readers to inspect the results and ascertain the validity of the results for themselves. Another common strategy is to validate a model by comparing it to other, often related models. This can be done by holding certain words or documents out when developing models and to later estimate the likelihood of the held-out data using the various models. Wallach et al. (2009) note that direct calculation of likelihood is intractable but offer reasonable approximations. Blei et al. (2003) instead use perplexity, which is “equivalent to the inverse of the geometric mean per-word likelihood.” Perplexity is commonly used in machine learning. Results will depend upon the size of the vocabulary in a corpus and how “predictable” words are within a corpus, making it difficult to say what a good or bad score is, in general. More recently, Chang et al. (2009) note that the utility of topic models ultimately depends upon how coherent and relevant (human) analysts find their results. The authors present and apply a quantitative method for validating a topic model involving tests where people were asked to identify spurious words linked to a topic and spurious topics linked to a document. The authors surprisingly find that their measures are negatively correlated with more established measures and conclude that “topic models which perform better on held-out likelihood may infer less semantically meaningful topics” (Chang et al., 2009).

4. Results: All recent reports

Structural topic modeling was applied to the ASRS data covering incidents that occurred between January 2010 and April 2015. The stm package developed by Roberts et al. (2014a) was used. Punctuation, whitespace, and stop words were first removed from the corpus. The flight mission, the phase of flight, and the time at which the incident was reported were selected as covariate data to be studied.

4.1. Selecting the number of topics

A natural first problem when applying STM involves identifying the number of topics. There is no single correct way to address this issue, but one possibility involves studying the trade-off between semantic coherence and exclusivity. These measures were also used by Das et al. (2017) to examine the ‘distinctness’ of topics.

Semantic coherence is based on measures of how frequently individual words occur and pairs of distinct words co-occur. Such measures can help analysts avoid defining topics that are problematic for one of several specific reasons. For example, words may be linked in a chain. The word corn might be linked to the word dog which is also linked to the word cat but the words corn and cat should not be assigned to the same topic in a topic model. This issue and its remedy via the semantic coherence measure were introduced by Mimno et al. (2011). As the number of topics in a model increases, the semantic coherence will decrease, generally speaking.

A topic is considered to be exclusive if the words that have a high probability of appearing conditional on that topic have low probabilities conditional on other topics. For example, there may be a topic present within ASRS data that refers to problems with landing gear. If the word gear frequently occurs when this topic comes up but rarely occurs otherwise, then this finding would be evidence of the topic’s exclusivity. As the number of topics in a model increases, the exclusivity of the model as a whole will typically increase.

Fig. 4 graphs the values of semantic coherence and exclusivity when applying STM to select between 6 and 100 topics in ASRS data. Each data point is based on a distinct analysis and a distinct set of topics. The label shows how many topics were found. So the location of the data point labeled 97 reflects the semantic coherence and exclusivity of a model that identified 97 separate topics in the ASRS data. The fact that the observed values of exclusivity (semantic coherence) are between 9 and 10 (-100 and -50) reflects details about the frequency of word occurrence in the ASRS data that are unimportant here. Attention should instead focus on comparisons among the data points.

Fig. 4 shows the expected trends for, and trade-off between, semantic coherence and exclusivity. There is no clear correct number of topics in the data. A few observations do stand out, including the cases where 9, 10, 11, 17, 20, and 38 topics were found. Arguably

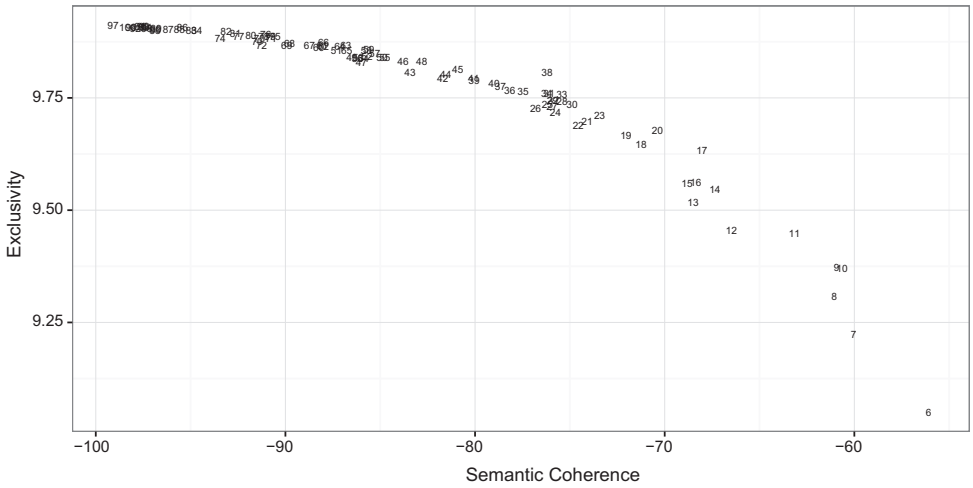

Fig. 4. Statistics used to select the number of topics.

the biggest outlier is the case where 17 topics were found. The following sub-sections focus on this case.

4.2. Identified topics and intuitive meanings

After applying STM, specific words are linked to specific topics. Explicit definitions, labels, or intuitive meanings of topics are not automatically generated. It is up to the analyst or the consumer of the analysis to determine what the topics represent based on the frequency with which different words are used within the topics. This requires considering each topic in turn and examining, for each, a small collection of words that are associated with the topic. If, for example, the words rain, precipitation, snow, and visibility are linked to the topic, then we may say that the topic represents weather.

The preceding paragraph is purposefully vague when it comes to the determination of the words associated with a topic. The most straightforward approach would be to select the words that have the highest probability of occurring conditional on the topic within an STM model. The problem here is that certain words such as aircraft and airport will show up as high probability words for many topics. This can include words such as have or get that are universal but add relatively little value when stripped of context. Statisticians have developed other metrics to use when and where these issues are evident. One commonly used metric is known as lift and refers to the probability of word occurrence conditional on topic divided by the probability of word occurrence across the corpus. This metric will highlight words that are much more common within a topic than they are across a corpus. The problem with this metric is that words which appear infrequently are likely to score well. It would be difficult and arguably unwise to assign an intuitive meaning to a large topic based on the outsized importance of the word waterspout within that topic, given how rare the word waterspout is and how rare issues related to waterspouts are. Bischof and Airoldi (2012) suggest using the FREX statistic, defined as the ratio of word frequency conditional on a topic to word-topic exclusivity (described informally in the preceding section of this article). This metric has less intuitive meaning than the other metrics described in this paragraph but also avoids referring to only the most common words. In their work (cited earlier), Das et al. (2017) provide readers with the highest ranked words based on the probability of occurrence conditional on topic and FREX metrics alongside labels that the authors have selected. These authors studied a corpus that is likely more diverse than the one studied here, as it covers all modes of transportation and concerns other than safety.

Table 2 shows, for each topic found in the ASRS database, intuitive topic label(s) identified using expert judgement, the prevalence of the topic within the corpus, and the five highest ranked words when ordering by probability of occurrence conditional on topic (Prob), by lift (Lift), and by the FREX statistic (FREX). All three of the metrics described in the preceding paragraph are used, as there is no single correct statistic to use. The topic labels were selected based on the other data shown in Table 2. Some topics appear to cover distinct but related issues or systems and were therefore assigned more than one label. For example, one topic is linked to words describing passengers and other words describing aircraft cargo. Since the goal of the topic labels is to succinctly describe the topics, however, no topic is assigned more than three labels.

Table 2 orders the topics in terms of estimates of how frequently each topic appears in the ASRS data. The human factors topic is the topic found to be the best represented in ASRS records. The next two most common topics relate to air traffic control and airspace management issues. The individual topics that describe mechanical issues, including, among others, those ascribed the engine, oil, pressure, and landing gear, fuel labels are each relatively rare. This analysis has not revealed any single mechanical issue that frequently appears in aviation safety incident reports. The topic assigned the label surface is more common than the approach topic, which is more common than the climb topic.

Analysts with limited resources could choose to study the role of human factors in aviation safety given its prominence here. These analysts could then focus exclusively on the narratives that contain many words linked to the human factors topic. They could even use the frequency with which the most relevant words appear in different narratives to rank the narratives, which could allow

Table 2
Topics identified and labeled.

Topic Label	Topic Proportion	Criteria	Word 1	Word 2	Word 3	Word 4	Word 5
human factors	0.107	*Prob*	get	time	just	said	need
		Lift	deep	stupid	leadership	pride	imagine
		FREX	thing	something	think	know	realize
airspace	0.082	*Prob*	aircraft	control	traffic	sector	airspace
		Lift	apreq	datablock	dside	jurisdic	loa
		FREX	sector	carrier	train	dside	separate
ATC	0.078	*Prob*	arriv	clearance	atc	departure	cross
		Lift	domno	fms	mistook	sefr	trup
		FREX	fms	restrict	rnav	sid	waypoint
surface, routing	0.074	*Prob*	runway	tower	aircraft	taxi	clear
		Lift	backtaxi	ogg	quebec	foxtrot	papa
		FREX	taxiway	runway	hold	short	taxi
approach	0.071	*Prob*	approach	visual	final	land	runway
		Lift	phanom	mateo	glidepath	stable	loc
		FREX	approach	visual	tcas	sight	terrain
smoke, fire	0.064	*Prob*	land	emerg	airport	fire	declar
		Lift	smoke	midcabin	fire	fum	tailpipe
		FREX	fire	smoke	declare	emerg	divert
low-altitude traffic	0.061	*Prob*	airport	pilot	radio	flight	traffic
		Lift	civilian	foreflight	laser	tfr	tfrs
		FREX	helicopter	ctaf	class	tfr	pattern
fatigue	0.057	*Prob*	flight	plan	dispatch	crew	hour
		Lift	circadian	polar	nighter	fdp	awake
		FREX	fatigue	schedule	sleep	hour	duty
thrust, flaps	0.056	*Prob*	captain	flap	takeoff	first	officer
		Lift	dual	rto	asymmetric	thrust	autothrust
		FREX	flap	trim	thrust	autothrottle	lever
climb	0.052	*Prob*	altitude	climb	feet	level	atc
		Lift	barometric	altimeter	gyro	rime	compass
		FREX	climb	altitude	cloud	altimeter	feet
tug, brake	0.049	*Prob*	aircraft	brake	left	right	ramp
		Lift	tug	wand	rope	traction	towbar
		FREX	tug	wheel	brake	deice	snow
maintenance, fault	0.047	*Prob*	maintain	aircraft	system	control	mel
		Lift	veil	dmi	nef	mel	elac
		FREX	mel	inop	fault	maintain	breaker
passengers, cargo	0.047	*Prob*	flight	passenger	door	attend	captain
		Lift	clinic	csr	lightheaded	mail	monoxide
		FREX	agent	door	galley	cargo	bag
weather	0.041	*Prob*	aircraft	speed	wind	weather	turbulence
		Lift	pub	recat	vortex	chop	turbulence
		FREX	turbulence	wake	wind	encounter	moderate
landing gear, fuel	0.041	*Prob*	fuel	gear	land	engine	tank
		Lift	sputter	desert	enrich	pump	imbalance
		FREX	tank	gear	pump	fuel	hydraulic
mechanic	0.040	*Prob*	aircraft	mechanic	inspect	install	remove
		Lift	jobcard	rii	washer	bolt	bundle
		FREX	install	card	cable	repair	bolt
engine, oil, pressure	0.034	*Prob*	engine	pressure	cabin	start	oil
		Lift	buy	outflow	psi	bleed	pressure
		FREX	oil	bleed	pressure	mask	temperature

Topic: human factors Someone needs to look into our Air Carrier deeply before someone gets hurt. This is a serious safety issue. . . .
Topic: airspace I handed off Aircraft X to HTO sector. I still had communications with Aircraft X. I had radar and communications on Aircraft Y. I climbed a departure off of JFK to FL190. . .
Topic: ATC There is a discrepancy in altitudes between the charted arrival and the FMC database for the TELLR 1 STAR into DEN. The altitude at CREDE Intersection on the charted arrival is. . .
Topic: surface, routing After landing and departing Runway 2L; I taxied onto Taxiway Alpha; an active taxiway; prior to receiving clearance from Ground Control to do so. I saw no clear. . .
Topic: approach We were cleared for visual approach to Runway 6 from (approximately) a 6 mile right downwind. [We] flew base over antenna tower. We had excellent visibility and clearly had. . .
Topic: smoke, fire Avionics smoke ECAM Came on as we climbed toward FL250. We handled the ECAM and determined that there were no other indications of smoke; fire or instrument. . .
Topic: low-altitude traffic I was flying VFR from RIV and my iPad overheated and my panel mount GPS lost signal and I might have entered Class Bravo airspace while transitioning from PDZ. . .
Topic: fatigue While preplanning the release for Flight ABC ZZZ-ZZZ1 arrival time XA45Z; [Flight Planning Software] advised the Dispatcher with the following message: NO ALTERNATE. . .
Topic: thrust, flaps We received erroneous takeoff configuration warning when thrust levers advanced for takeoff. Immediately discontinued takeoff at approximately 30 KIAS and taxied clear of. . .
Topic: climb During descent; after leveling at ATC assigned 12;000 FT; ATC queried about our assigned altitude. Pilot not flying indicated 12;000 FT assigned and asked what ATC was showing and asked. . .
Topic: tug, brake I was pushing aircraft to park at hangar. Once chocked was going to pull the push tractor back and in to set it up for a crewmate to pull the pin out. As I thought I had my foot. . .
Topic: maintenance, fault We arrived at the airplane and noticed Fire Detection Loop A for the APU was deferred. I tested the squibs and noticed no squibs messages (1 and 2) for the APU. . .
Topic: passengers, cargo I noticed a strong metallic smell and at the same time I could hear the bell going off for the interphone. I picked up the interphone and I think it was Flight Attendant B. . .
Topic: weather During arrival into ATL we were descending thru approximately 10000 FT when we encountered wake turbulence. Over an approximate 5 minute period the wake induced three. . .
Topic: landing gear, fuel On May 2013 at approximately XA:00 hours; in good VFR weather; on base leg to the airport; the engine lost power. I switched the fuel selector to the opposite tank; but the. . .
Topic: mechanic I; as an Inspector; and Mechanic X were involved with the Left-hand Elevator Assembly that was removed for damage and a replacement Serviceable [Elevator]; according to the. . .
Topic: engine, oil, pressure Received 'ENG 2 OIL TEMP HI' in descent. Temperature started in the yellow then jumped to the red range. Temperature top [at] 225 degrees. Temperature was erratic. . .

Fig. 5. Example narratives for each topic.

the analysts to review the most relevant narratives first. As was noted during discussion of the most frequently observed phrases, the use of jargon, and likely of different types of jargon, presents an issue here. Pereira et al. (2013), in their study of (ground) traffic incidents, find that "topic modeling is particularly well suited for the domain of incident messages since these have a relatively homogeneous lexicon and each message aims to objectively convey clear and synthetic information to help emergency response." Unfortunately, a similar claim cannot be made when analyzing records from the ASRS database. Still, topic modeling can be used to relatively easily and rapidly gain insight into aviation safety incident reports. We are beginning to see some of the hidden structure behind the collection of incident reports within the ASRS database.

Fig. 5 shows the start of the ASRS narratives that are most strongly linked to each topic. This figure further illustrates the form and substance of incident reports. Fig. 5 also provides some intuitive evidence that topic modeling and topic label assignment were successful insofar as the narrative portions provided do appear, generally speaking, to match their topic labels but not other topic labels. The narrative linked to the climb topic appears to be an exception, although later in that narrative the author notes "Reset altimeters to correct 29.06 and made immediate climb to 12,000 FT assigned altitude."

Fig. 6 presents a visualization of the correlations among the topics listed in Table 2. Note that use of the smoke and fire topic is correlated with use of the engine, oil, and pressure topic. The smoke and fire topic has a substantial negative correlation with the human factors topic. A report that focuses extensively on one of these topics will likely barely mention the other topic, if it mentions

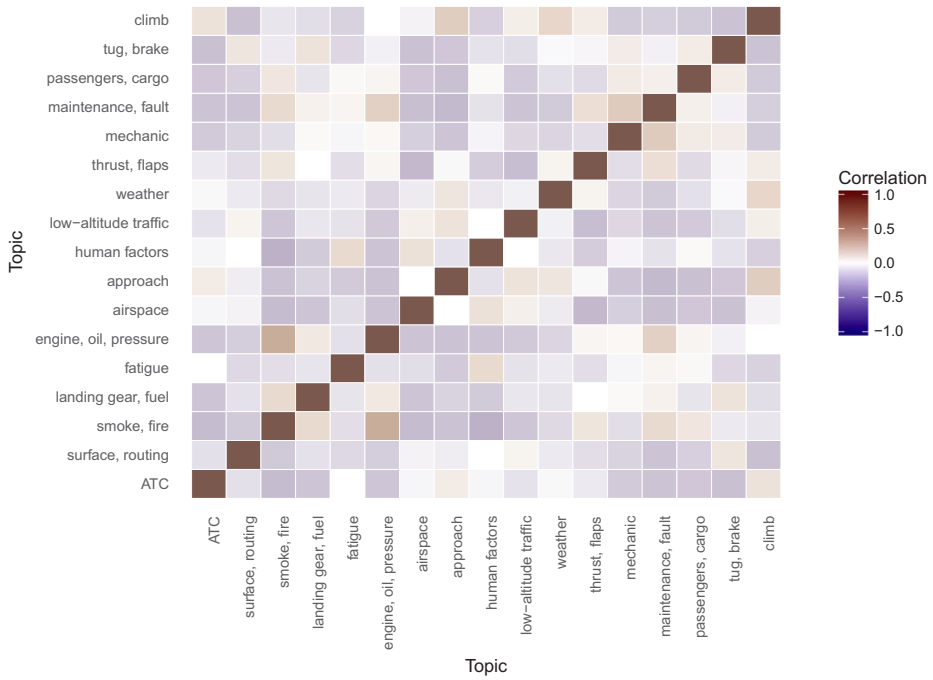

Fig. 6. Correlations among topics.

the other topic at all. The engine, oil, and pressure topic is also correlated with the maintenance and fault topic. These results all make intuitive sense.

As mentioned in Section 3.3, it is difficult to validate the topics found during a topic modeling exercise. The value of the topics described in Table 2 and Fig. 5 and the model behind them depends upon how convincing and valuable subject matter experts find these results (Chang et al., 2009). It is, however, possible to compare these results to alternate results found by another model as a sort of sanity check. Towards that end, a 17-topic LDA model was also developed based on the same corpus of incident reports used to fit the STM model. In this case, the topicmodels package developed by Grün and Hornik (2011) was used.

The 10 ASRS narratives that are most strongly linked to each STM topic were recorded. The LDA topic which is mostly strongly linked to each of these narratives was then also noted. If a STM topic and a LDA topic represent similar themes, one might expect that the 10 ASRS narratives linked to the STM topic would also be linked to the LDA topic. Table 3 summarizes the results of this matching exercise. Certain cells in the table are highlighted; here all or all but one of the narratives linked to a specific STM topic are also linked to a single LDA topic. The STM and LDA topics appear to be analogues, representing similar themes in the data.

Table 3
Comparison of topics found via STM and LDA.

	LDA topic																
STM topic	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
human factors	1	0	0	0	0	0	0	0	0	0	0	0	**9**	0	0	0	0
airspace	0	0	0	0	0	8	0	0	0	0	0	0	0	0	0	0	2
ATC	0	0	0	0	0	0	**9**	0	0	0	0	0	0	0	0	0	1
surface, routing	0	0	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0
approach	0	0	0	0	0	0	0	2	0	0	0	8	0	0	0	0	0
smoke, fire	0	7	0	0	0	0	0	0	0	0	1	0	0	0	2	0	0
low-altitude traffic	1	0	0	1	0	0	0	5	0	0	0	0	0	0	0	3	0
fatigue	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
thrust, flaps	0	1	0	0	0	0	0	0	0	0	8	1	0	0	0	0	0
climb	0	2	0	0	0	0	1	0	0	2	0	4	0	0	0	0	1
tub, brake	0	0	0	0	6	0	0	0	0	0	1	0	0	0	3	0	0
maintenance, fault	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**	0	0	0
passengers, cargo	0	0	0	0	0	0	0	0	0	0	0	0	0	0	**10**	0	0
weather	0	0	0	0	0	0	0	0	0	**9**	0	1	0	0	0	0	0
landing gear, fuel	0	1	0	0	0	0	0	0	0	0	3	0	0	6	0	0	0
mechanic	0	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	0
engine, oil, pressure	0	**10**	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

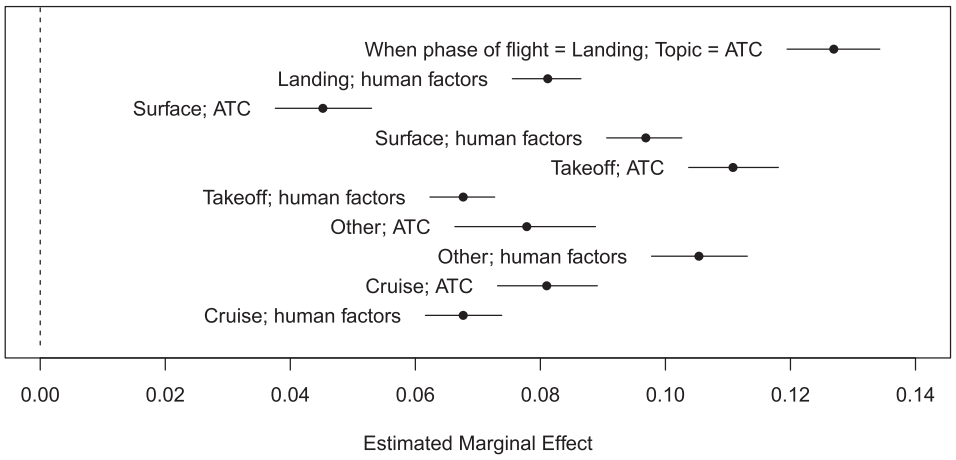

Fig. 7. Topic and phase of flight.

Many STM topics do appear to have an analogue LDA topic. For example, 9 of the 10 incidents assigned to the weather topic by STM are assigned to LDA topic #13. No other incidents considered during this experiment were assigned to LDA topic #13. Topic #13 appears to be the weather topic in the LDA model. For all but two of the STM topics, the majority of the representative narratives are assigned to a single LDA topic. The two exceptions are the low-altitude traffic and climb topics. These topics are apparently less readibly identifiable within the corpus.

The results presented in Table 3 give weak evidence in support of the topics found via STM. These topics at least appear to match the topics found when applying an alternate methodology. The benefit of using STM is that the impacts of covariates including document metadata are explicitly accounted for within the model.

4.3. Impact of covariates on topic prominence

Table 2 lists expected topic proportions across all documents. We can also look at the effects of covariate data on topic proportions. Roberts et al. (2014a) provide further technical details. The statistical software package developed by the authors also generates statistics on the uncertainty surrounding the coefficients of the regression models mentioned previously based on an approximation of the relevant covariance matrix (Roberts et al., 2014a). The result is that we can estimate topic proportion as a function of covariate data and, further, produce confidence intervals around estimated topic proportions.

Fig. 7 is an example, contrasting the estimated proportions of the air traffic control (ATC) and human factors topics as a function of the phase of flight of the primary aircraft listed in the incident report. The dots on the chart depict expected topic proportions while the horizontal bars reflect uncertainty in these estimates.

When the aircraft is reported to be in the takeoff, cruise, or landing phases of flight, the ATC topic is more prominent than the human factors topic. In all other phases of flight, the human factors topic is more prominent than the ATC topic. This is particularly true when aircraft are reported to be on the surface of an airport.

The results presented in Fig. 7 also show that the most prominent (topic, phase of flight) pairs involve the ATC topic and the takeoff and landing phases of flight. The least prominent pair involves ATC issues in reports issued involving an aircraft on the airport surface. This makes some intuitive sense. On the airport surface, tower controllers are responsible for clearing ground support vehicles and aircraft to use runways and taxiways. They also must ensure ATC instructions and established policies and procedures are adhered to. They provide taxi instructions to pilots, working to ensure that the aircraft are properly sequenced for efficient runway and gate use. But pilots are largely responsible for maintaining separation during taxi operations. In contrast, ATC must take a more active role in ensuring separation standards are met during takeoff and landing. The result appears to be that safety issues during the takeoff and landing phases of flight are relatively strongly linked to ATC.

Fig. 8 shows the estimated proportions of the smoke, fire topic and of the fuel pump, tank, landing gear topic as a function of flight mission. Issues involving smoke and fire are more prominent for cargo and, particularly, passenger flights. Issues involving fuel pumps, tanks, and landing gear are more prominent for other flights, and particularly for private aircraft.

The data points that stand out the most in Fig. 8 reveal the significance of landing gear and fuel-related issues and the insignificance of smoke and fire issues in incident reports where the primary aircraft listed in the report is a private aircraft. A cursory inspection of a sample of the relevant reports reveals incidents where there was smoke or an unpleasant smell or an alarm from a smoke detector in the cabin. All of these issues would seem to be more likely on a larger commercial transport aircraft as opposed to a smaller private aircraft. For example, there would seem to be more opportunities for passengers or cargo to create issues on the larger commercial aircraft. Further analysis by subject matter experts would be helpful here, to come to more definitive conclusions.

Document metadata on when incidents occurred are used by the structural topic model studied here to develop (document-specific) estimates of topic prevalence. Fig. 9 shows the estimated topic proportions for the ‘ATC,’ ‘fuel pump, tank, and landing gear,’ ‘weather,’ and ‘maintenance, fault’ topics as a function of time. The chart shows a strong cyclical pattern in the prominence of the

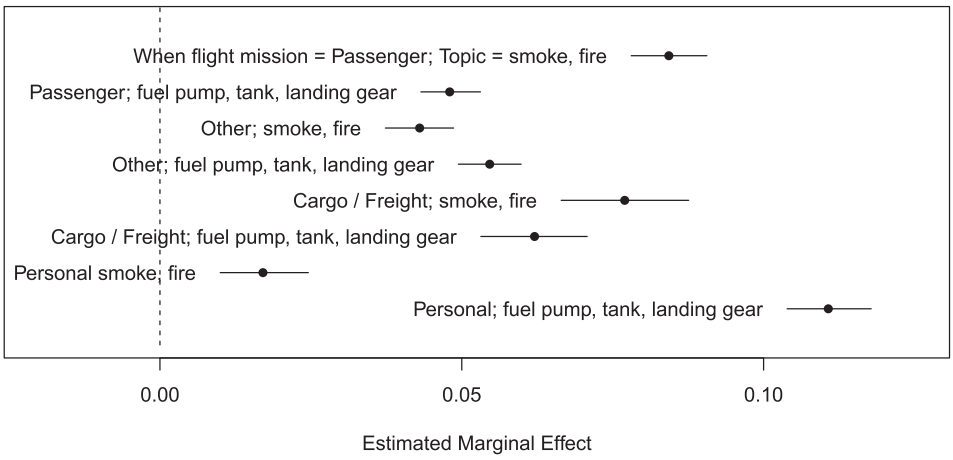

Fig. 8. Topic and flight mission.

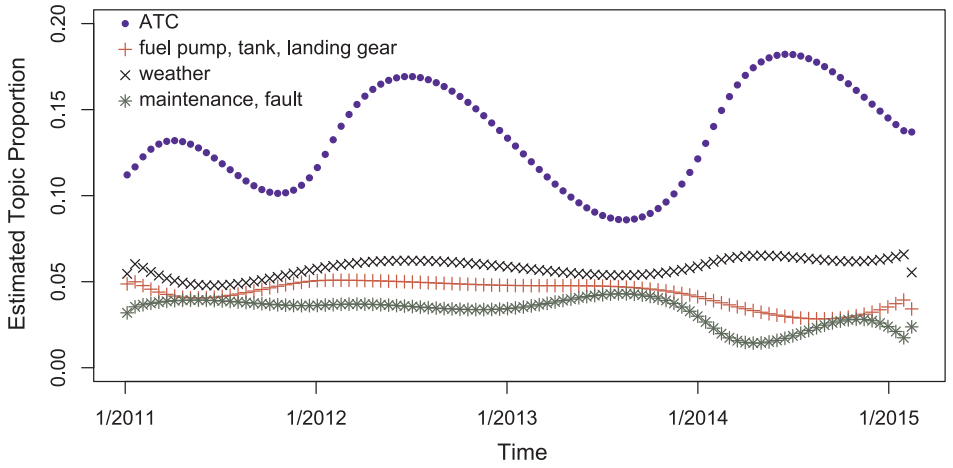

Fig. 9. Estimated topic proportion over time.

ATC topic over time. The other topics are included here to show that this pattern is unusual. It's not clear what is driving this result. It may be that ATC is a more prominent issue in the spring and summer months, generally speaking, and that the STM model is picking up on this. There is, however, no clearly correct, intuitive explanation of this result.

Further research is needed to explore and extend the results presented here. For example, a more thorough inspection of incident reports would explain if and perhaps why there are more aviation safety incidents related to air traffic control but not weather in the Spring and Summer months. Is this simply a result of misunderstanding the meaning of these two topics or indicative of something interesting regarding the prominence of air traffic control-related incidents in certain months of the year?

5. San Francisco International Airport

The preceding sections of this article are based on an analysis of all of the ASRS records reported between January 2010 and April 2015. There might be more specific, actionable insights gained from analysis of particular subsets of these data. This section applies similar methods to data where the locale is listed as SFO.Airport. San Francisco International Airport (SFO) was chosen because of the large number of records linked to it and its familiarity to the author.

Table 4 shows the most common n-grams, where n is an integer greater than or equal to three, found in ASRS narratives at SFO. Recall that details on n-gram use can provide insight into the use of language within a corpus. The popularity of specific n-grams can highlight the outsized importance of specific events or objects. Many of the n-grams shown in Table 4 reference the San Mateo Bridge, include the word visual, or reference either Quiet Bridge Visual or Tip Toe Visual. The San Mateo Bridge is a well-known landmark on the routes of many aircraft arriving at runways 28L and 28R at SFO. Visibility at the bridge is critical when Simultaneous Offset Instrument Approach (SOIA) procedures are in effect for these runways. The Quiet Bridge Visual and Tip Toe Visual are two of the approach paths into these runways that pass over the bridge. The results presented in Table 4 indicate that the San Mateo Bridge, SOIA procedures, and the Quiet Bridge Visual and Tip Toe Visual approach paths are likely to be important topics of discussions when it comes to safety at SFO. These results were obtained with minimal (human) effort and minimal reliance on subject matter expertise.

Table 4
Frequently observed n-grams in ASRS narratives at SFO.

n-grams	Observation count
FMS bridge visual	38
bridge visual 28R	22
FMS bridge visual 28R	19
hold short line	19
air carrier X	18
maintain visual separation	16
Quiet Bridge Visual	15
visual approach runway	15
Tip Toe Visual	13
first officer FO	13
cross runway 28L	12
San Mateo Bridge	11
hold short runway	11
I pilot flying	11

Indeed, the names Quiet Bridge Visual and Tip Toe Visual were uncovered by an algorithm and not provided by an analyst.

Table 5 describes the topics identified in the narratives from SFO. (Results not shown here recommended the identification of seven topics.) The most prominent topic is assigned the labels visual and approach. Topics related to the taxiway, pushing, and holding/runways are individually less prominent but together point to (separate) issues on the surface at SFO.

Fig. 10, similar to Fig. 5 from an earlier section of this article, shows the start of the ASRS narratives that are most strongly linked to each topic found at SFO. This figure again provides limited but valuable intuitive evidence that topic modeling and topic label assignment have been successful here. The narrative portions provided do again appear, generally speaking, to match their topic labels but not other topic labels.

Das et al. (2016) and Sun and Yin (2017) provide wordclouds that visualize the topics the authors find in transportation research papers. A similar approach was employed here to generate Fig. 11. For the entire corpus and separately for each topic, this figure shows the 16 words with the highest probability of occurrence. The size of each word is proportional to its probability.

Fig. 12 shows the correlations among the topics listed in Table 5. There isn't evidence of any strong positive correlations. The approach topic is negatively correlated with the taxiway, push, and hold, runway topics. This result highlights the difference between incidents on approach versus on the surface at SFO.

Table 5
Topics in ASRS narratives at SFO.

Topic Label	Topic Proportion	Word 1	Word 2	Word 3
approach	0.22	approach	visual	aircraft
		visual	approach	sight
		tcas	sight	visual
altitude	0.17	feet	altitude	arrive
		feet	fms	altitude
		difficult	rnav	quiet
hold, runway	0.14	runway	aircraft	tower
		hold	runway	takeoff
		hold	line	across
push	0.13	flight	time	call
		push	flight	crew
		push	maintain	dispatch
separate	0.12	aircraft	carrier	departure
		carrier	air	separate
		air	carrier	nct
flap, gear	0.12	captain	flap	speed
		flap	captain	wind
		increase	gear	flap
taxiway	0.10	taxiway	taxi	ground
		taxiway	taxi	ground
		taxiway	taxi	ground

Topic: approach
We were on localizer 28L at FAF. Descending on glide slope at 1400 AGL we had TCAS alert to descend 1200 fpm. Another aircraft overshot parallel approach on 28R from above; behind and right of our aircraft. We followed TCAS guidance to 900 FT AGL and had to abandon its direction due to terrain closure. At 1200 FT First Officer visually acquired [the other aircraft] and said he was 1 plane. . .

Topic: altitude
While flying the SERFR ONE RNAV ARRIVAL; Oakland Enroute Controller cleared us direct to NRRLI intersection for the arrival Runway 28 transition and pilot's discretion descend to Flight Level 200. The First Officer (Flying Pilot) selected FL200 in the altitude alerter and selected direct to the fix in the FMS. He confirmed the selections with me and executed the selection. After a few minutes. . .

Topic: hold, runway
After taxiing across runways 1L and 1R; we were instructed to hold short of Runway 28L. There was a line-up of departure aircraft for runways 1L and 1R; and a steady stream of arrivals for runway 28L and 28R. While we waited; holding short of Runway 28L; for approximately 15 minutes; a single controller was handling all arrivals; departures; and hold short operations. She was task saturated; working. . .

Topic: push
I just did my first flight with lavatory O2 inoperative. We took a 53 minute gate delay before an acceptable solution was found to safely operate the flight. The Flight Ops Bulletin and flight attendant hot topic boards sheets address passengers stuck in the lavatory during a rapid decompression/emergency descent event (RD/ED); and basically the passengers are left without supplemental oxygen until the event. . .

Topic: separate
I was working both Sutro and Quake Departure Sectors combined. Traffic was light/minimal for Quake but slightly moderate for Sutro with a total of approximately eight aircraft on frequency. A B737 departed off of OAK on a Skyline departure. On initial contact I climbed the B737 to 5;000 FT to separate the aircraft from a light Cessna aircraft transitioning west of San Francisco northbound to. . .

Topic: flap, gear
On Briefing before departure; the Captain informed me that he had just come off of Medical Leave and had just checked out in the this aircraft last spring. He said he knew I was very experienced and asked my help in keeping him safe and giving him any helpful pointers I could provide from my experience. The pilot not flying also informed me that he had only 3 months experience in the aircraft. . .

Topic: taxiway
Transitioned to inner Taxiway [A] from outer at Taxiway [B] prior to assigned taxiway due to misidentification of the taxiway we were on. What led up to the event was poor signage of the taxiways; radio congestion; sorting out a weight and balance issue. During the event we visually cleared the taxiways adjacent ramp areas prior to our turning onto them. The outcome of the event was successful. . .

Fig. 10. Example narratives for each topic at SFO.

Fig. 13 shows the estimated topic proportions for the taxi and visual, approach topics for different types of flights. It is interesting to note that passenger flights are more likely to report problems related to the approach topic. Personal flights are (slightly) more likely to report problems related to the taxi topic. There is a relatively large amount of uncertainty when estimating topic proportions for personal or cargo flights at SFO. This is because there are relatively few recent incident reports in the ASRS database at SFO where the primary aircraft is a cargo or personal aircraft.

The application of NLP tools and techniques revealed the prominence of the Quiet Bridge Visual and Tip Toe Visual approach paths in reports of aviation safety incidents at SFO. Many traffic managers and analysts familiar with the airport will already know of issues related to the use of these approach paths. One benefit of applying NLP methods is that it allows analysts to learn of, and to quantify, the importance of specific approach paths and other specific procedures or jargon which operators will know but which analysts may not know. A clear conclusion of this work is that NLP can help focus discussion or be used to set priorities for further analysis.

6. Conclusion

The ASRS database is a useful resource for aviation systems researchers interested in safety. There are over a million incident reports in the database. Techniques developed in the fields of natural language processing and machine learning can be used for analysis of this database. This article describes applications of structural topic modeling to ASRS records from January 2010 through April 2015. STM was applied both to all the relevant records and to the subset consisting of reports from SFO.

Fig. 11. Wordcloud visualizations of corpus and topics at SFO.

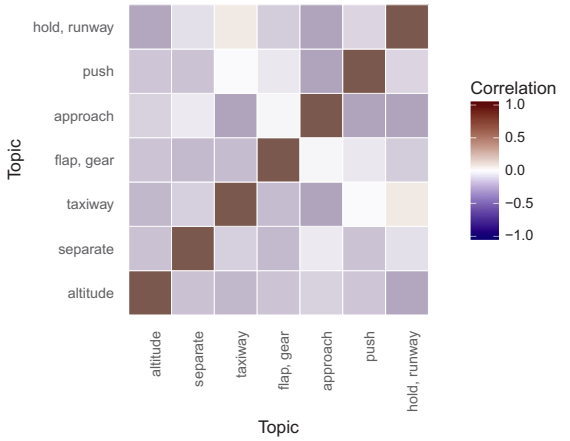

Fig. 12. Topic correlations at SFO.

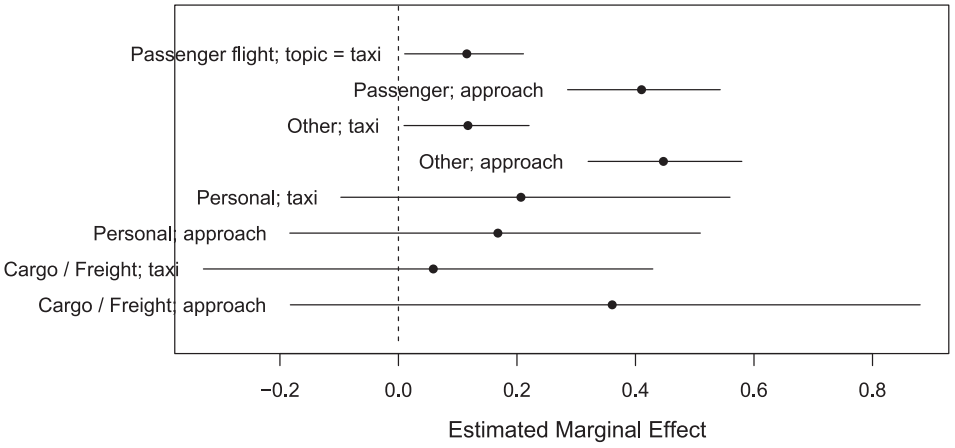

Fig. 13. Topic and flight mission at SFO.

Methods highlighted the Quiet Bridge Visual and Tip Toe Visual approach paths as particularly prominent in incident reports at SFO. Looking nationwide, results demonstrated the importance of human factors and air traffic control, with the former being more prominent in reports of incidents on the airport surface and the latter more prominent in reports of incidents during flight. The frequency of landing gear, fuel issues and the sparsity of smoke and fire issues for private aircraft were also recorded. The best way to verify these results would be to run experiments involving subject matter experts that measure the relevance and coherence of topics, for example using the verification methodology described by Chang et al. (2009).

The results demonstrate that structural topic modeling and other methods applied here are able to identify known issues. These methods are also able to uncover some issues that have not been previously reported, but do not necessarily provide detail that could be used to produce actionable insights. Subject matter expertise is important and needed to assign intuitive meanings to topics and to otherwise interpret and extend the results of topic modeling efforts. For example, a more manual study could focus on the reports that mention the Quiet Bridge Visual and Tip Toe Visual approach paths at SFO and see if there are any common themes in the reports that suggest ways to improve safety.

Structural topic modeling allows analysts to take full advantage of both the large number of free text incident reports and the large amount of useful metadata included in ASRS data. The utility of natural language processing methods such as this to the study of aviation safety could be enhanced with further enhancement of the ASRS program. For example, those who submit incident reports could be asked to (manually) identify topics within their reports or to otherwise label or categorize incidents. Even if only a subset of submitters respond to such a request, it could provide useful training data for future categorization efforts which assign labels to incidents that were not manually labeled. It could also provide test data to validate the results of prior topic modeling and other efforts. In this case, submitters should be required to select several topics from a fixed list of previously identified topics. More generally, it makes sense to try to standardize incident reports, for example to encourage submitters to use a common vocabulary within their narratives. This will make it easier for algorithms to identify similar reports or themes within reports and, ultimately, to highlight the true underlying topics and trends in aviation safety. It would be interesting to study the links between topics identified via topic modeling and metadata not studied here such as severity of incident. Some intuitively interesting attributes of incidents, such as severity of incident, are not present in ASRS data. It would be valuable if such metadata were added to the ASRS database. There may be a way to derive such new variables from the narrative reports themselves using natural language processing techniques.

Acknowledgements

The author would like to thank Heather Arneson, Jeremy Eckhause, Amy Kim, and the anonymous reviewers of an earlier draft of this paper. This research was partially funded by the NASA NextGen - Concepts and Technology Development Project under program announcement NNH14ZEA001N-CTD1.

References

Abedin, M., Ng, V., Khan, L., 2010. Cause identification from aviation safety incident reports via weakly supervised semantic lexicon construction. J. Artif. Intell. Res. 38, 569–631.

Ahmed, M., Khan, L., Oza, N., Rajeswari, M., 2010. Multi-label ASRS dataset classification using semi-supervised subspace clustering. In: Proceedings of the Conference on Intelligent Data Understanding, pp. 285–299.

Ananyan, S., Goodfellow, M., 2004. New Capabilities of PolyAnalyst Text and Data Mining Applied to STEADES Data at the International Air Transport Association (IATA): A Technology Demonstration. Global Aviation Information Network.

Barnett, A., Ball, M., Donohue, G., Hansen, M., Odoni, A., Trani, A., 2015. Collision course? The North Airfield Safety Study at Los Angeles International Airport (LAX). Transport. Res. Part A 77, 14–34.

Billings, C., Reynard, W., 1984. Human factors in aircraft incidents: results of a 7-year study. Aviat., Space, Environ. Med. 55, 960–965.

Bischof, J., Airoldi, E., 2012. Summarizing topical content with word frequency and exclusivity. In: Proceedings of the 29th International Conference on Machine Learning, pp. 201–208.

Blei, D., 2012. Probabilistic topic models. Commun. ACM 55, 77–84.

Blei, D., Ng, A., Jordan, M., 2003. Latent Dirichlet allocation. J. Mach. Learn. Res. 3, 993–1022.

Bliss, J., Freeland, M., Millard, J., 1999. Alarm related incidents in aviation: a survey of the aviation safety reporting system database. In: Proceedings of the Human Factors and Ergonomics Society Annual Meeting, vol. 43, pp. 6–10.

Chang, J., Gerrish, S., Wang, C., Boyd-Graber, J.L., Blei, D.M., 2009. Reading tea leaves: how humans interpret topic models. In: Advances in neural information processing systems, pp. 288–296.

Chittaro, L., 2017. A comparative study of aviation safety briefing media: card, video, and video with interactive controls. Transport. Res. Part C: Emerg. Technol. 85, 415–428.

Das, S., Dixon, K., Sun, X., Dutta, A., Zupancich, M., 2017. Trends in transportation research: exploring content analysis in topics. Transport. Res. Rec.: J. Transport. Res. Board 2614, 27–38.

Das, S., Sun, X., Dutta, A., 2016. Text mining and topic modeling of compendiums of papers from transportation research board annual meetings. Transport. Res. Rec.: J. Transport. Res. Board 2552, 48–56.

El Ghaoui, L., Pham, V., Li, G., Duong, V., Srivastava, A., Bhaduri, K., 2013. Understanding large text corpora via sparse machine learning. Statist. Anal. Data Min. 6, 221–242.

Farrell, J., 2016. Corporate funding and ideological polarization about climate change. Proc. Nat. Acad. Sci. 113, 92–97.

Grün, B., Hornik, K., 2011. topicmodels: An R Package for Fitting Topic Models, R package.

Mimno, D., Wallach, H., Talley, E., Leenders, M., McCallum, A., 2011. Optimizing semantic coherence in topic models. In: Proceedings of the Conference on Empirical Methods in Natural Language Processing, pp. 262–272.

Mishler, A., Crabb, E., Paletz, S., Hefright, B., Golonka, E., 2015. Using structural topic modeling to detect events and cluster Twitter users in the Ukrainian crisis. In: International Conference on Human-Computer Interaction, pp. 639–644.

Nazeri, Z., 2003. Application of Aviation Safety Data Mining Workbench at American Airlines Proof-of-Concept Demonstration of Data and Text Mining. Global Aviation Information Network.

Netjasov, F., 2012. Framework for airspace planning and design based on conflict risk assessment: part 1: conflict risk assessment model for airspace strategic planning. Transport. Res. Part C 24, 190–212.

Oza, N., Castle, J.P., Stutz, J., 2009. Classification of aeronautics system health and safety documents. IEEE Trans. Syst., Man, Cybernet., Part C (Appl. Rev.) 39, 670–680.
Péladeau, N., Stovall, C., 2005. Application of Provalis Research Corps Statistical Content Analysis Text Mining to Airline Safety Reports. Global Aviation Information Network.
Pereira, F.C., Rodrigues, F., Ben-Akiva, M., 2013. Text analysis in incident duration prediction. Transport. Res. Part C 37, 177–192.
Persing, I., Ng, V., 2009. Semi-supervised cause identification from aviation safety reports. In: Proceedings of the 47th Annual Meeting of the ACL and the 4th IJCNLP of the AFNLP, vol. 2, pp. 843–851.
Posse, C., Matzke, B., Anderson, C., Brothers, A., Matzke, M., Ferryman, T., 2005. Extracting information from narratives: an application to aviation safety reports. In: Proceedings of the IEEE Aerospace Conference, pp. 3678–3690.
Reich, J., Tingley, D., Leder-Luis, J., Roberts, M., Stewart, B., 2015. Computer-assisted reading and discovery for student generated text in massive open online courses. J. Learn. Anal. 2, 156–184.
Roberts, M., Stewart, B., Tingley, D., 2014. stm: R Package for Structural Topic Models, R package.
Roberts, M., Stewart, B., Tingley, D., Airoldi, E., 2013. The structural topic model and applied social science. In: Advances in Neural Information Processing Systems Workshop on Topic Models: Computation, Application, and Evaluation.
Roberts, M., Stewart, B., Tingley, D., Lucas, C., Leder-Luis, J., Gadarian, S.K., Albertson, B., Rand, D., 2014b. Structural topic models for open-ended survey responses. Am. J. Polit. Sci. 58, 1064–1082.
Sjöblom, O., 2014. Data mining in promoting aviation safety management. In: International Conference on Well-Being in the Information Society.
Sun, L., Rahwan, I., 2017. Coauthorship network in transportation research. Transport. Res. Part A: Pol. Pract. 100, 135–151.
Sun, L., Yin, Y., 2017. Discovering themes and trends in transportation research using topic modeling. Transport. Res. Part C 77, 49–66.
Suomi, R., Sjöblom, O., 2009. Data mining in aviation safety data analysis. In: Social and Political Implications of Data Mining: Knowledge Management in E-Government.
Switzer, J., Khan, L., Muhaya, F.B., 2011. Subjectivity classification and analysis of the ASRS corpus. In: IEEE International Conference on Information Reuse and Integration, pp. 160–165.
Tanguy, L., Tulechki, N., Urieli, A., Hermann, E., Raynal, C., 2016. Natural language processing for aviation safety reports: from classification to interactive analysis. Comp. Indust. 78, 80–95.
Wallach, H.M., Murray, I., Salakhutdinov, R., D.Mimno, 2009. Evaluation methods for topic models. In: Proceedings of the 26th Annual International Conference on Machine Learning, pp. 1105–1112.
Wolfe, S., 2007. Wordplay: an examination of semantic approaches to classify safety reports. In: American Institute of Aeronautics and Astronautics Conference, pp. 1–13.
Zhang, Z., He, Q., Gao, J., Ni, M., 2018. A deep learning approach for detecting traffic accidents from social media data. Transport. Res. Part C: Emerg. Technol. 86, 580–596.
Zhao, W., Alam, S., Abbass, H., 2013. Evaluating ground air network vulnerabilities in an integrated terminal maneuvering area using co-evolutionary computational red teaming. Transport. Res. Part C 29, 32–54.