# Augmenting Topic Finding in the NASA Aviation Safety Reporting System using Topic Modeling

Carlos Paradis[*] and Rick Kazman[†]
*University of Hawaii at Manoa, Honolulu, Hawaii, 96822, USA*

Misty D. Davies[‡] and Becky L. Hooey[§]
*NASA Ames Research Center, Moffett Field, CA, 94035, USA*

**Context: The NASA Aviation Safety Reporting System (ASRS) is a voluntary confidential aviation safety reporting system. The ASRS receives reports from pilots, air traffic controllers, flight attendants and others involved in aviation operations. The reports are de-identified and coded by ASRS expert safety analysts and a short descriptive synopsis is written to describe the safety issue. The de-identified reports are then disseminated to the aviation community in a number of ways including entry into an online database, Safety Alert Bulletins, For Your Information Notices, and the CALLBACK newsletter. In this paper we consider whether we can improve the dissemination of safety concerns through the use of topic modeling. Topic modeling may improve the dissemination of safety concerns by grouping and summarizing large collections of reports simultaneously. However, the groupings must be both meaningful and useful.**

**Aim: We investigate whether existing topic modeling techniques are suitable to ease some of the manual effort, and to enhance it with additional visual cues regarding the process of grouping, sense making and labeling reports.**

**Method: We evaluate the applicability of WarpLDA topic modeling results combined with three visualization tools, the first two of which have been extended by us in this work for ASRS: Termite, TopicFlow, and LDAVis. Based on the identified limitations in these tools, we propose a methodology for improving them, and evaluate their outputs using ASRS as our test dataset.**

**Results: The user interfaces of Termite, Topicflow and LDAVis were found insufficient for sense-making of the ASRS narratives. Our evaluation of the method shows a large spread in performance ranging from perfect to equivalent to a random model. We also identified certain report set topics to be easier to group than others.**

**Conclusion: While many tools for topic modeling and visualization have been proposed, more work is necessary before they can be applied in practical situations to improve existing manual workflows. The methodology presented and applied in this work contributes towards this effort. Our empirical results suggest there is potential in using topic modeling to separate some report set topics, such as maintenance and flight attendant reports.**

## I. Introduction

THE timely identification and communication of safety alerts has been a primary objective of ASRS since its inception. [1]:

> On December 1, 1974, TWA Flight 514 was inbound through cloudy and turbulent skies to Dulles Airport in Washington, D.C. The flight crew misunderstood an ATC clearance and descended to 1,800 feet before reaching the approach segment to which that minimum altitude applied. The aircraft collided with a Virginia mountaintop, killing all aboard.
>
> A disturbing finding emerged from the ensuing NTSB accident investigation. Six weeks prior to the TWA accident, a United Airlines flight crew had experienced an identical clearance misunderstanding and narrowly missed hitting the same Virginia mountain during a nighttime approach. The United crew discovered their close call after landing and reported the incident to their company. A cautionary notice was issued to all United pilots.

[*] Graduate Student, Department of Information & Computer Sciences.

[†] Professor, Shidler College of Business.

[‡] Research Computer Engineer, Intelligent Systems Division, Mail Stop 269-1, AIAA Associate Fellow.

[§] Director, NASA Aviation Safety Report System, MS 262-4.

> Tragically, at the time there existed no method of sharing the United pilots' knowledge with TWA and other airline operators. Following the TWA accident, it was determined that future safety information must be shared with the entire aviation community. Thus was born the idea of a national aviation incident reporting program that would be non-punitive, voluntary, and confidential.*.

Less than 2 years later, a quarterly report of ASRS noted that although its designers were acutely aware of the potential value of data derived from individual occurrences in highlighting deficiencies and discrepancies in the national aviation system, they also believe that other, perhaps more valuable, insights into system problems can be gained only by study of a large body of occurrence data [2]. However, as of the last program briefing reporting intake metrics on July 2019, ASRS has processed over 1.6 million reports in its 42 year history [3, p.13], making the analysis of a large body of occurrence data a daunting task.

Towards addressing this goal, in this work we evaluate the applicability of unsupervised learning, specifically topic modeling of ASRS data. Topic modeling is not the only possible approach. Our choice differs from more recent literature in mining safety incident reports, which relies on supervised learning [4], [5], or older literature which emphasized qualitative studies on metadata of interest (e.g. Anomalies, Human Factors, etc.) [6], [7], [8]. We share the same motivation as [9] in the choice of unsupervised learning instead of supervised learning as this allows us to identify not only existing topics, but also emerging trends, in an automated fashion. We chose topic modeling using classic approaches instead of deep learning because they are widely used in software engineering studies [10]. In this context, we pose the following research questions:

RQ1: Can existing topic modeling tools help analyze evolving safety threats in large collections of text?

We pose this research question from a pragmatic and exploratory standpoint to motivate the other two research questions, rather than seeking to falsify a statistical hypothesis. If an existing tool already suffices to facilitate the identification of relevant safety threats in ASRS, then we can focus our efforts in enhancing said tool instead of reinventing the wheel.

RQ2: Can topic modeling discover ASRS report sets from a collection of mixed reports by their topic?

To be useful, topic modeling must be able to classify documents in a meaningful way. Standard practice in papers that use topic modeling is to present a table showing the top n terms [10], which does not account for the actual separation of the documents. Here we leverage a novel model setup to evaluate the classification of documents using ASRS data. Being capable to semi-automate or automate the identification of relevant themes in aviation safety incidents reports would be useful both to ASRS analysts in issuing alerts, and to the aviation community at large to perform their own explorations of the ASRS corpus.

RQ3: Are there specific report sets that are easier to group than others?

If topic modeling does not present a fully satisfactory solution, it may be possible that it can still identify certain topics well enough for practical use.

Our findings are consistent with suggestions made in a recent review of the topic modeling literature [11] on the need for better visualization and user interfaces for sense-making. Careful evaluation of topic modeling optimization criteria is also suggested, as the most popular optimization method—held out accuracy—may lead to poor topics [12]. In addition, we agree with other authors that stability measures are needed for assessment, due to the random initialization of topic modeling algorithms [13].

Our contributions in this work are as follows:

- We refactored two published topic modeling visualization tools [14], [15] so that they can be used in other datasets, and made the changes and code publicly available†.

* https://asrs.arc.nasa.gov/publications/callback/cb_317.htm

† See: https://github.com/sailuh/termite and https://github.com/sailuh/topicflow

- We discuss practical limitations of creating a data pipeline using WarpLDA [16] for 3 different published topic visualization tools [14, 15, 17] for ASRS. Our findings can be used by future work to further refine visualization tools.
- We address the identified limitations of topic performance and stability [18] evaluation in unsupervised learning leveraging publicly available ASRS Report Sets[‡], which can also be used to improve the evaluation of new methods.

The remainder of this paper is as follows: In the Dataset Section II we introduce ASRS, and the dataset we use to evaluate our models. In the Method Section III, we introduce topic modeling, and the experimental protocol used to evaluate its applicability. We also present the topic modeling tools to answer RQ1, and the modifications performed so they could be reused. We then present our findings in the Results Section IV and how our method differs from other work in the Related Work Section V. Finally, we present the threats to validity in Section VI and our conclusions in Section VII.

## II. Data Model

As stated in our introduction, ASRS began as a reporting program for sharing aviation safety knowledge and confidentially reporting irregularities to help prevent accidents. Figure 1 presents a diagram from 1976 with the ASRS workflow for a new report. The fundamental process has not changed substantially since then. Briefly, starting from the top, a reporter fills in a form with details of the incident. Depending on the reporter role, various forms may be used. Common to all reports are multiple-choice selections of some metadata (e.g. type of weather), but also a free text box, *the narrative*, where the details are provided.

After a screener assesses the report to determine if other federal agencies should be made aware of it due to it being criminal or accident related, an ASRS analyst may follow up with the reporter for any necessary clarification, deidentifies the report, and applies additional metadata codes. In addition, the analyst will also include a *synopsis*, which summarizes the narrative and highlights safety concerns and, if evident, contributing factors.

Using the report database, ASRS also generates *data products* to distribute to the aviation community, leading to a sustainable cycle. Specifically, ASRS provides a report query capability, issues safety alerts to organizations, performs thematic searches, and publishes newsletters, [3]. Relevant to our method is the *report sets* data product. These report sets are in essence a thematic grouping of existing reports in the database. In the Method Section 2, we will explain how we leverage them in our experimental protocol to evaluate topic models.

## III. Method

The motivation of our work relies on leveraging existing ASRS data as a gold standard to understand the practical usefulness of the model in improving the existing ASRS workflow. Figure 2 provides two experimental protocols. The *Real Dataset setup* reflects the more commonly used approach, which we use to answer RQ1. The *Toy Dataset Setup* reflects the protocol we implemented for RQ2 and RQ3. This method was instantiated as an R package, Kaona[§] and is publicly available.

To stay consistent with the related literature in topic modeling terminology we refer to each *report's narrative* in ASRS as a *document*. Furthermore, we use a notation similar to Blei et al. [19] to describe the equations in this section, except by distinguishing words, documents, and corpora by distinct letters and adding an explicit definition for topics:

- A word w is an item from a vocabulary indexed by $1, \ldots, V$. The $v$th word in the vocabulary is represented by a V-vector $w$ such that $w^v = 1$ and $w^u = 0$ for $u \neq v$.
- A document is a sequence of N words denoted by $d = (w_1, w_2, \ldots, w_N)$, where $w_n$ is the $n$th word in the sequence.
- A corpus is a collection of M documents denoted by $\mathbf{c} = d_1, d_2, \ldots, d_M$.
- A topic $z$ is a distribution over documents.

### A. Experimental Protocol

An important step in using any automated method is to evaluate if the results of the automated step are meaningful and useful. Topic modeling is often evaluated subjectively, by inspecting the the ‘top-terms’ in the topic-term matrix resulting from topic modeling. This, however, does not serve to evaluate if the groupings are correct. A contribution of

[‡] https://asrs.arc.nasa.gov/search/reportsets.html

[§] http://github.com/sailuh/kaona

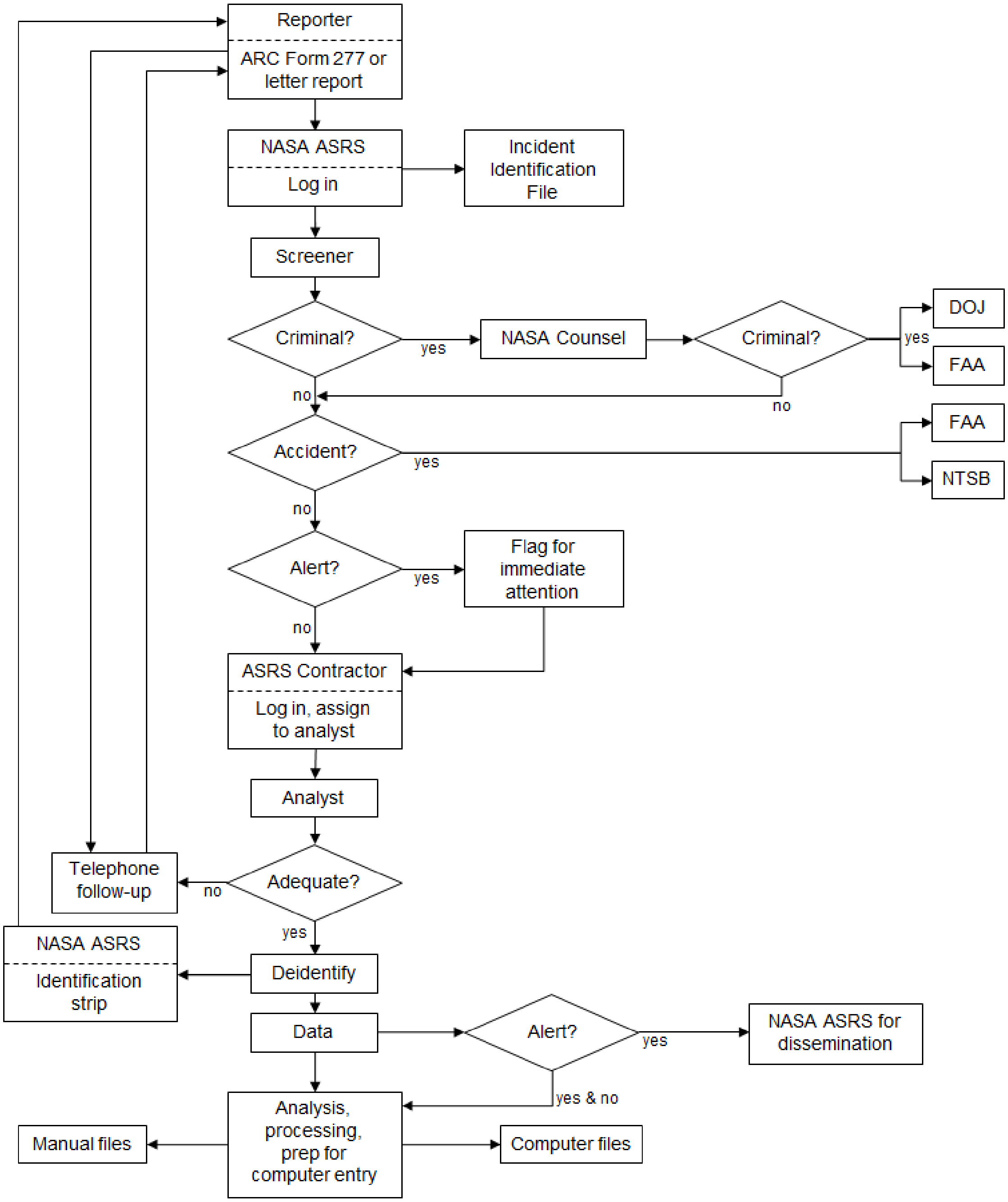


**Fig. 1 How Reports are Generated: Handling of NASA Aviation Safety Reports (ASRS) [2].**

this work is the proposed Toy Dataset Setup, where a quantitative evaluation is possible. To do so, we must have a-priori annotated data, in this case grouping labels, to verify if the automated step was or was not correct.

We begin explaining in detail how the ‘top-terms’ evaluation is currently done, using the Real Dataset Setup. Then, we use it to motivate the Toy Dataset Setup, which provides us with grouping labels, and which we used for RQ2 and RQ3.

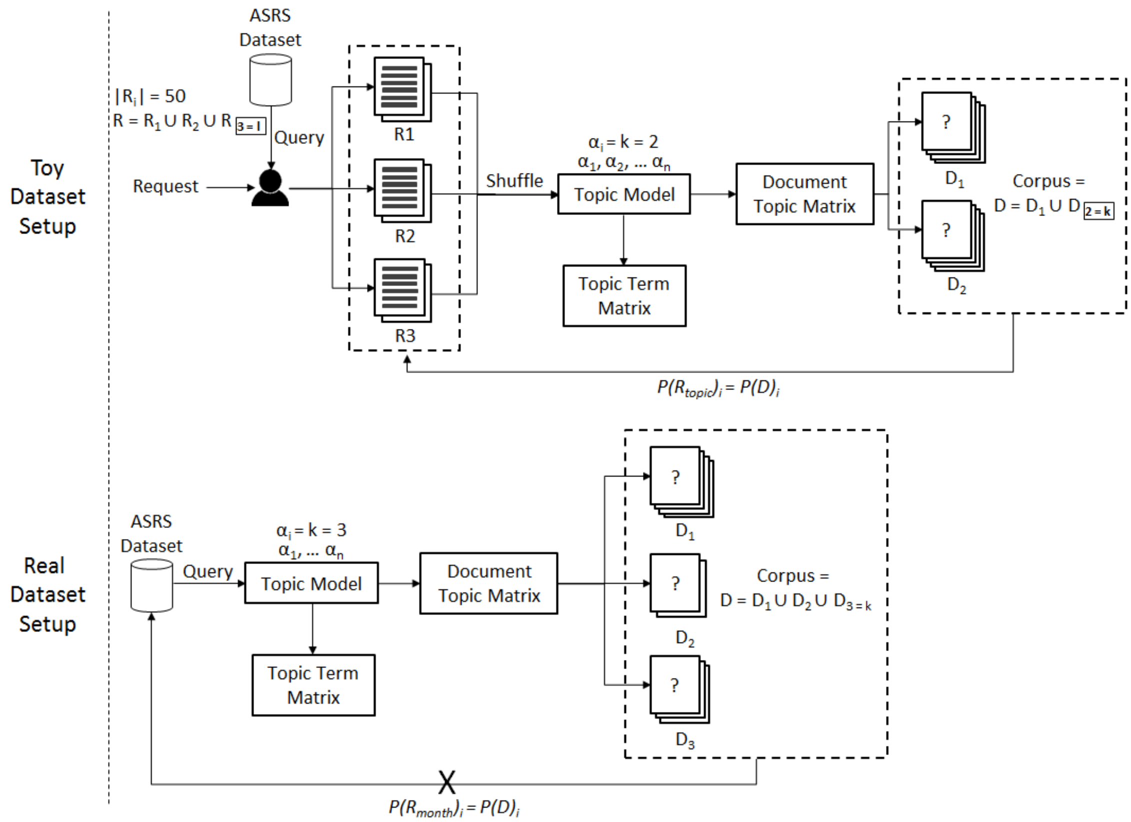


**Fig. 2 Report Set (Toy Dataset) vs Real Dataset (ASRS Database) Setup.**

*1. Real Dataset Setup*

As noted in the Data Model Section II, ASRS receives several reports a day. These reports are indexed by year and month, and so we are able to *query* the order in which they are received to a month granularity. In Figure 2 (bottom), we showcase a single month query, $R_{month}$, being performed, which contains several individual reports. While these reports could at this point be manually inspected to identify groups of reports based on their topic, in the Figure, these reports are instead input to a *topic model*. We choose a-priori the number of groupings, as required by the algorithm, in the example to be $k = 3$ topics. The algorithm then creates a partition $P(D)_i = \bigcup_{i=1}^{k=3} D_i$, assigning each report of the query one of the 3 groupings, where $P(D)_i$ is defined as one of the possible partitions of the set of reports $D$.

The choice of $k = 3$ thus encodes the number of groupings in the partition that the algorithm is to find given a set of reports. A limitation of this setup commonly reported in the literature, is that we do not know the correct partition $P(R_{month})_i$, i.e. to what grouping each report should be associated nor the number of report sets $k$. Therefore, we can't assess if $P(R_{month})_i = P(D)_i$, as $P(R_{month})_i$. The use of, and inspection of, the 'top-terms' is thus commonly used as a subjective replacement to assess the quality of the groupings.

To be precise, we do not claim here when we refer to a correct partition that there is a "true grouping" $P(R_{month})_i$, or universal ground truth. Success here is a grouping that is useful for pilots to identify a theme of interest. Because the related literature assumes the Real Dataset Setup, we used it to answer RQ1 despite not being able to evaluate if the groupings in the Real Dataset Setup are or not useful. To evaluate if topic modeling in general can be useful to ASRS we thus propose the creation of an alternative way to prepare the data for evaluation, i.e. the *Toy Dataset Setup* by leveraging existing ASRS *report sets*, as shown in the upper portion of Figure 2, to answer RQ2 and RQ3. We now explain how we leverage the report sets to evaluate the groupings.

### *2. Toy Dataset Setup*

Because our goal is to evaluate the ability of topic modeling to group reports, we prepared the Toy Dataset Setup to conduct rigorous experimental trials. A single trial is showcased on the upper portion of Figure 2. In the Figure, we assume three groups, $l = 3$, of documents exist, unknown to the topic modeling algorithm. We then 'shuffle' these three groups, and input them to topic modeling. We must then choose a number of topics we expect to exist; in the Figure we show $k = 2$ (two groups) into which the shuffled reports must be grouped. We then evaluate 'how close' the algorithm comes, per trial, of uncovering the original groupings.

Each report set $R_{topic}$. is already made available publicly on ASRS website and prepared by ASRS analysts, we just leverage them here. Each report set contains 50 reports each about a given topic (e.g. bird strikes, flight deviation). This means in a single trial we assess if 100 reports were correctly assigned to their group.

While at first it may sound counter-intuitive that we did not choose $k = 3$ in the Figure III, remember this is a realistic scenario in the operational setting. As the value of $l$ is unknown, $k$ may be chosen either to be higher or lower than $l$. The merit of the Toy Dataset Setup is that knowing $l$ we can assess the performance of the algorithm in the potential different settings that may happen in operation. The error measure we chose, discussed in Section III.G, is also viable for $k \neq l$.

For each trial, therefore, we can then obtain a measure of performance for using topic modeling. However, a single trial is insufficient to draw conclusions of the topic modeling performance (similarly to how one coin toss would not afford us discerning if a coin is biased). Since we must choose a rationale for the value for $l$ and $k$, we decided to present the results of $l = k = 2$, and defer $l = 3, k = 2$, as shown in Figure 2 to future work.

Because there are a total of 30 report sets available in ASRS, we can construct 30 choose 2 (hereafter referred as 30C2) = 435 trials, where $l = 2$, and $k = 2$, i.e. we shuffle a pair of report sets $l$, and evaluate if the topic model, knowing a-priori $k = 2$ topics exist, can correctly separate the reports based on their narrative content.

We highlight here that our Toy Dataset Setup is synthetic (artificially chosen), and does not exactly reflect the groupings one would find in $P(R_{month})_i$ (hence why we refer to it as $R_{topic}$). The main motivation for using synthetic (artificially chosen, such as the report sets) datasets is that they are easier to generate and the independent generative factors can be easily controlled [20]. When compared to the *Real Dataset Setup*, the number of report sets, $l$, and the size of each report set $|R_i|$ is *synthetic*, however, the corpus itself is not artificially generated, although a trained topic modeling algorithm could also generate report sets. Due to this distinction, we chose to name our setup *Toy Dataset Setup*, as the reports are not artificially generated, just their groupings.

To create an experimental trial to evaluate topic modeling in the *Toy Dataset Setup*, we combine different report sets together (in Figure 2 we exemplify $l = 3$ report sets) for each trial, and evaluate if the topic model is capable to recover the set to which each report was originally assigned, given the choice of $k$. **While we exemplify in the diagram $l = 3$, and $k = 2$, in our experiment we tested for $l = 2$, $k = 2$.** More specifically, to answer RQ2 and RQ3 we evaluated in the 30C2 = 435 possible combinations of report sets, if the topic model was able to uncover the original groupings of report sets. Our rationale to use $l = k = 2$ instead of $l = 3$ and $k = 2$ as in the Figure III, is because we would like to assess how well the topic model performs on an easier task first. That is, we assume the learning task to separate a pair of report sets on each trial to be easier than a large number of topics.

## B. Topic modeling

We chose Warp Latent Dirichlet Allocation (WarpLDA) [16], an implementation of topic modeling [19], to be applied to each pair of documents (report set narratives) due to its efficient implementation and given the growing number of reports in ASRS. LDA is a generative probabilistic model of a corpus. In LDA, documents are represented as random mixtures over latent topics, and each topic is characterized by a distribution over words.

For each of the 435 possible pair combinations of documents, i.e. each of the 435 trials, we obtain 435 document-frequency matrices as input to LDA which provides (for the chosen number of topics, here $k = 2$): the distribution of the documents over topics (the document-topic matrix), and the distribution of topics over words (terms), commonly known as the topic-term matrix. As the name implies, a document-topic matrix lists the documents as rows, and topics as columns, where the number of columns is specified by the a priori choice of the number of topics, and cells indicate the proportion each document is about a given topic. A topic-term-matrix, in turn, lists topics as rows and terms as columns, and cells can be interpreted as the relevance or contribution of the term in representing the topic, expressed as a probability. It is very common in the literature that the identified topics meaning are interpreted from the terms of highest probability value for each topic (e.g. "top 10 terms").

### C. Model Tuning

LDA requires the number of topics $k$ to be specified *a priori*, which can be obtained through measurements such as perplexity [19], where a lower perplexity score indicates better generalization performance. Several other optimization methods also exist such as differential evolution [21, 22] and approaches which rely on the elbow method [¶]. Due to our experimental setup explained in the previous section, we do not need to learn the number of topics $k$, as $l$, the number of topics in the corpus, is known due to the shuffle of report sets.

### D. Deterministic Mapping

Topic modeling provides a probabilistic mapping from documents to each topic in the document-topic matrix. To obtain the actual groupings shown in Figure 2, so we can assess if $R_topic = D$ for each of the 30C2 pairs, we must establish a deterministic mapping.

Similar to our prior work [23], we define a deterministic map from documents to topics in the document-topic matrix for each of the 435 pairs of report sets tested, using the highest probability of the topic given the document in Eq. 1.

$$argmax_{z_i} p(z_i|d) \tag{1}$$

Thus we claim a document is *assigned* to a topic if that topic has the highest probability in the document-topic matrix. Since each row is a document, we can interpret the document-topic matrix as labeling the topic to which a document belongs. In addition, because we know for each document to what report set they were originally associated before the shuffling (although the algorithm does not), we can evaluate empirically if $R = D$ (see Figure 2).

### E. Tools

To answer RQ1, we surveyed the literature for existing tools which enhanced the presentation of topic modeling in ASRS. This led us to Termite [14], LDAVis [17], and TopicFlow [15].

Termite and LDAVis take as input and augment just the topic-term matrix explained in the the Topic Model Section III.B. This means that, for a given set of reports, users can explore topics using the visualization. Both Termite and LDAVis also introduce term ranking heuristics for each topic, which are intended to facilitate comprehension of the topics by identifying "top-terms". TopicFlow augments both the topic-term matrix and document-topic matrix. Moreover, TopicFlow provides an additional capability for topic modeling, by creating a notion of time evolution. This is done by assessing the similarity of topics between consecutive pairs of months.

Termite's authors [14] argue that ranking topic terms from higher to low probability, as derived by the topic modeling, is arbitrary. The authors propose a *seriation method*, which modifies topic term probabilities to account for co-occurrence and collocation likelihood between all pairs of words [24]. For example, when reading each topic terms, it is more likely that users will read "social networks" instead of "networks social". An example of Termite output is shown in Figure 3. The original Termite implementation contained many source code dependencies which made reusing the tool difficult. As a contribution of this work, we created a forked implementation which can be applied to any domain[‖].

Similar to Termite, LDAVis [17] also provides a ranking heuristic, however it prioritizes distinguishing the terms between topics to facilitate topic comparison. Specifically, they define a new metric, *relevance* (Equation 2), where $\lambda$ determines the weight given to the probability of the topic's word relative to its lift (the ratio of a term's probability within a topic to its marginal probability across the corpus).

$$rel(w_i, z_j|\lambda) = \lambda \log\big(p(w_i|z_j)\big) + (1-\lambda)\log\left(\frac{p(w_i|z_j)}{p(w_i)}\right) \tag{2}$$

Setting $\lambda = 1$ results in a ranking of terms in decreasing order of their topic-specific probability, and setting $\lambda = 0$ ranks terms solely by their lift. In addition to the term ranking heuristic, LDAVis provide an additional visualization of

[¶] https://cran.r-project.org/web/packages/ldatuning

[‖] https://github.com/sailuh/termite

topic similarity in a 2D plane. This is done through dimensionality reduction by principal coordinate analysis [25] over the terms each topic has. LDAVis does not provide a data pipeline, but it provides a simple interface to input a topic-term matrix and the additional metadata it requires. An example of LDAVis applied to the same dataset as Termite in ASRS is shown in Figure 5.

TopicFlow [15] does not provide any ranking heuristics, but augments the presentation of topics to present topic evolution. The original implementation of TopicFlow could only be applied to existing datasets [**], which had its own data format. We created a tool to transform any topic modeling output into the format required by TopicFlow, and the necessary code changes so any other dataset can be used[††], as its published code did not provide an interface to other datasets. An example of TopicFlow applied to the same dataset as the previous tools in ASRS is shown in Figure 4.

### F. Preprocessing

For the tools in RQ1, we did not perform any preprocessing as they each had different heuristics in their displays, which were presented in the prior section. For RQ2 and RQ3 we removed stop words[‡‡] and performed stemming[§§].

### G. Error Measures

There is a disconnect between how topic models are evaluated and why we expect topic models to be useful, with the most common evaluation being holdout [11]. However, topic models are often used to organize, summarize, and help users explore large corpora, and there is no technical reason to suppose that held out accuracy corresponds to better organization or easier interpretation [11, 26].

Holdout is a type of internal criterion, but there are also external criteria methods. External criteria use direct evaluation in the application of interest, such as a gold standard (a surrogate for user judgments for the groupings) [26]. The gold standard is ideally produced by human judges with a good level of inter-judge agreement [26]. In this work, our gold standard is defined by ASRS *report sets*.

An intuitive way to evaluate clusters using an external criterion is using *set matching* [18], specifically classification error rate. In Figure 2 on the *Toy Dataset Setup*, we could compare the identified groups of documents $D_1$ and $D_2$ to the report sets $R_1, R_2, R_3$. However, this approach has faced criticism due to disregarding part of the unmatched groups in the evaluation metric [18]. An improvement over this metric is pair counting, such as the Adjusted Rand Index (ARI). Intuitively, we consider every possible pair of reports within each report set, and evaluate if a) they were originally or not part of the same report set and b) if topic modeling assigned them or not to the same report set. Thus, we are able to generate a confusion matrix, and calculate the ARI, a real value where 0 indicates the performance equivalent to a random model, -1 worse than random, and 1 better than random.

## IV. Results

We now present our findings to the three posed research questions.

RQ1: Can existing topic modeling tools help analyze evolving safety threats in large collections of text?

**Topic Sense Making**. First, we observed the representation of top-10 terms, extensively employed in recent research studies, including the original LDA work [11], and also used by the tools of Figures 3, 4 and 5 proved insufficient for sense-making. While tools such as LDAVis are visually appealing and interactive, if the presentation of the documents, groupings and top terms is not relevant and meaningful, these tools and methods can not be used in practice.

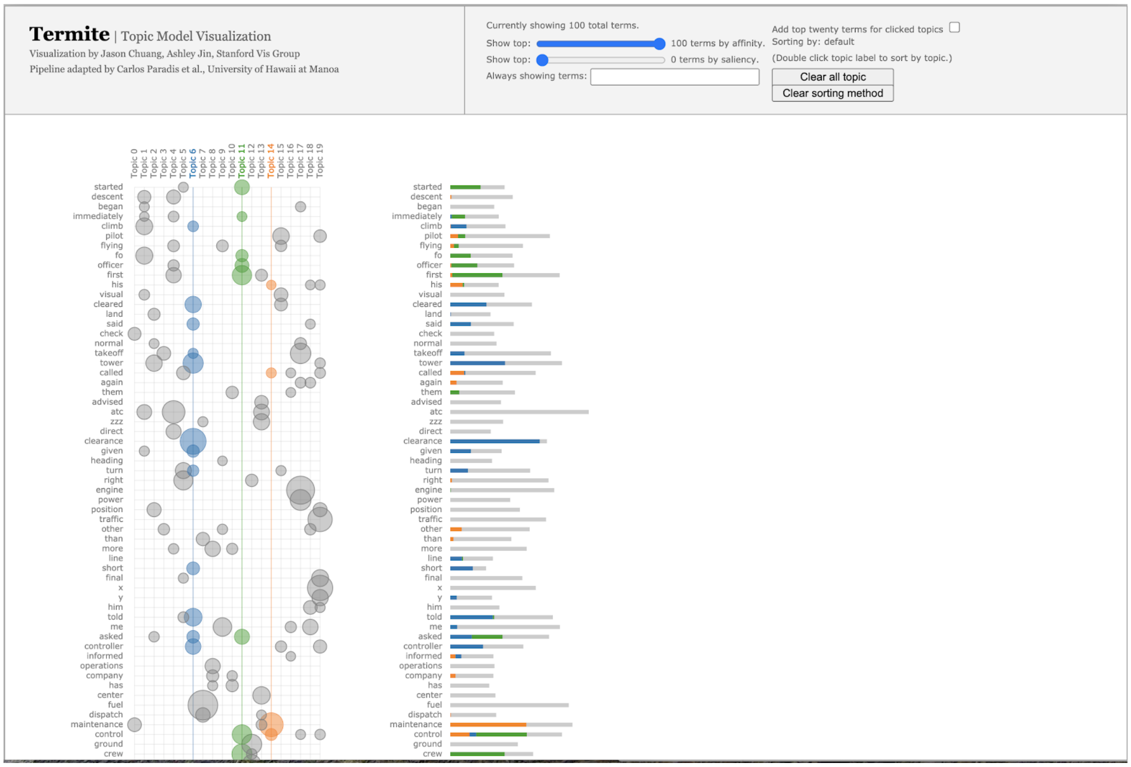
**Document-Topic Assignment**. Second, the number of documents presented by existing tools was too great. Consider TopicFlow for example, as shown in Figure 4. The green rectangle highlighting Topic 1_7 groups 501 documents (right panel in the figure), where each document is a narrative from ASRS's 2018 corpus. While a quick exploration of the documents suggests narratives associated with weather conditions and fuel, the large number of documents included in this topic made proper evaluation difficult. As we stated in the introduction, ASRS has processed over 1.6 million reports in its 42 year history, and the number of reports per month trends to increase. Therefore, a more systematic approach to evaluate document-topic assignment is needed.

[**] https://github.com/alisonmsmith/topicflow/tree/master/data

[††] https://github.com/sailuh/topicflow

[‡‡] https://algs4.cs.princeton.edu/35applications/stopwords.txt

[§§] https://www.rdocumentation.org/packages/tokenizers/versions/0.2.1/topics/tokenize_word_stems



**Termite** | Topic Model Visualization

Visualization by Jason Chuang, Ashley Jin, Stanford Vis Group

Pipeline adapted by Carlos Paradis et al., University of Hawaii at Manoa

Currently showing 100 total terms.

Show top: 100 terms by affinity.

Show top: 0 terms by saliency.

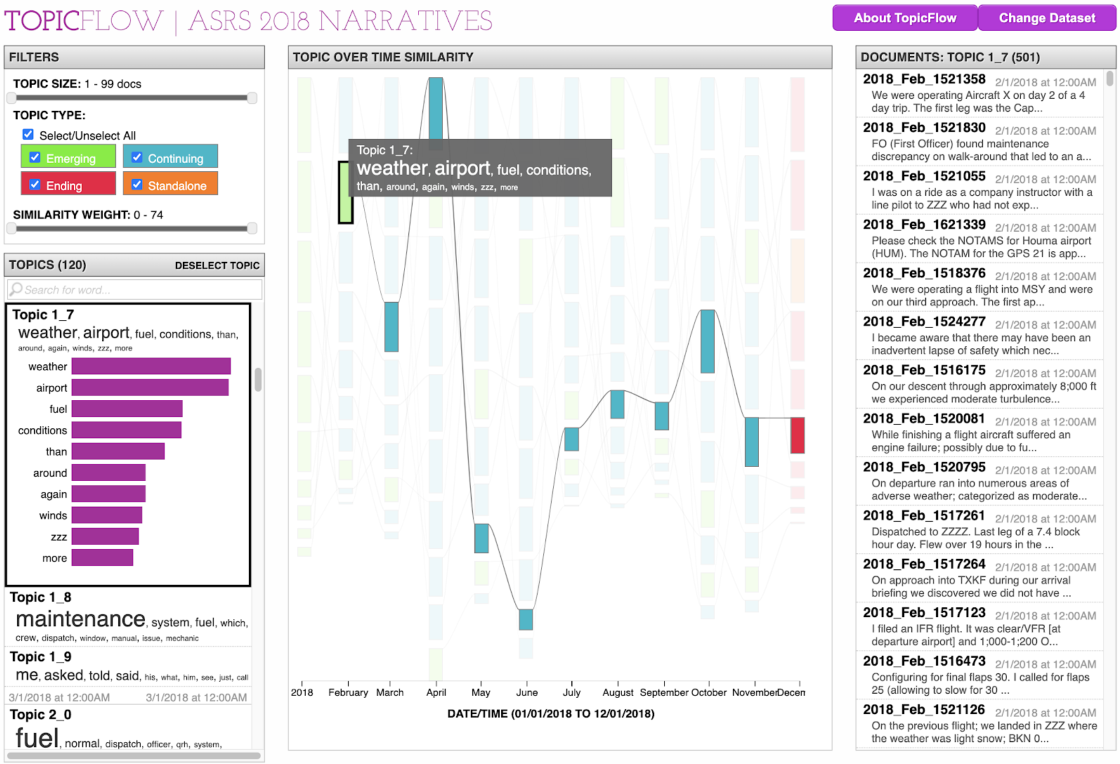
Always showing terms:

Add top twenty terms for clicked topics ☐

Sorting by: default

(Double click topic label to sort by topic.)

**Fig. 3 Termite [14] Visualization of ASRS 2018 Corpus.**



TOPICFLOW | ASRS 2018 NARRATIVES

Topic 1_7: weather, airport, fuel, conditions, than, around, again, winds, zzz, more

Topic 1_7
weather, airport, fuel, conditions, than, around, again, winds, zzz, more

Topic 1_8
maintenance, system, fuel, which, crew, dispatch, window, manual, issue, mechanic

Topic 1_9
me, asked, told, said, his, what, him, see, just, call

3/1/2018 at 12:00AM 3/1/2018 at 12:00AM

Topic 2_0
fuel, normal, dispatch, officer, qrh, system,

DOCUMENTS: TOPIC 1_7 (501)

**2018_Feb_1521358** 2/1/2018 at 12:00AM
We were operating Aircraft X on day 2 of a 4 day trip. The first leg was the Cap...

**2018_Feb_1521830** 2/1/2018 at 12:00AM
FO (First Officer) found maintenance discrepancy on walk-around that led to an a...

**2018_Feb_1521055** 2/1/2018 at 12:00AM
I was on a ride as a company instructor with a line pilot to ZZZ who had not exp...

**2018_Feb_1621339** 2/1/2018 at 12:00AM
Please check the NOTAMS for Houma airport (HUM). The NOTAM for the GPS 21 is app...

**2018_Feb_1518376** 2/1/2018 at 12:00AM
We were operating a flight into MSY and were on our third approach. The first ap...

**2018_Feb_1524277** 2/1/2018 at 12:00AM
I became aware that there may have been an inadvertent lapse of safety which nec...

**2018_Feb_1516175** 2/1/2018 at 12:00AM
On our descent through approximately 8;000 ft we experienced moderate turbulence...

**2018_Feb_1520081** 2/1/2018 at 12:00AM
While finishing a flight aircraft suffered an engine failure; possibly due to fu...

**2018_Feb_1520795** 2/1/2018 at 12:00AM
On departure ran into numerous areas of adverse weather; categorized as moderate...

**2018_Feb_1517261** 2/1/2018 at 12:00AM
Dispatched to ZZZZ. Last leg of a 7.4 block hour day. Flew over 19 hours in the ...

**2018_Feb_1517264** 2/1/2018 at 12:00AM
On approach into TXKF during our arrival briefing we discovered we did not have ...

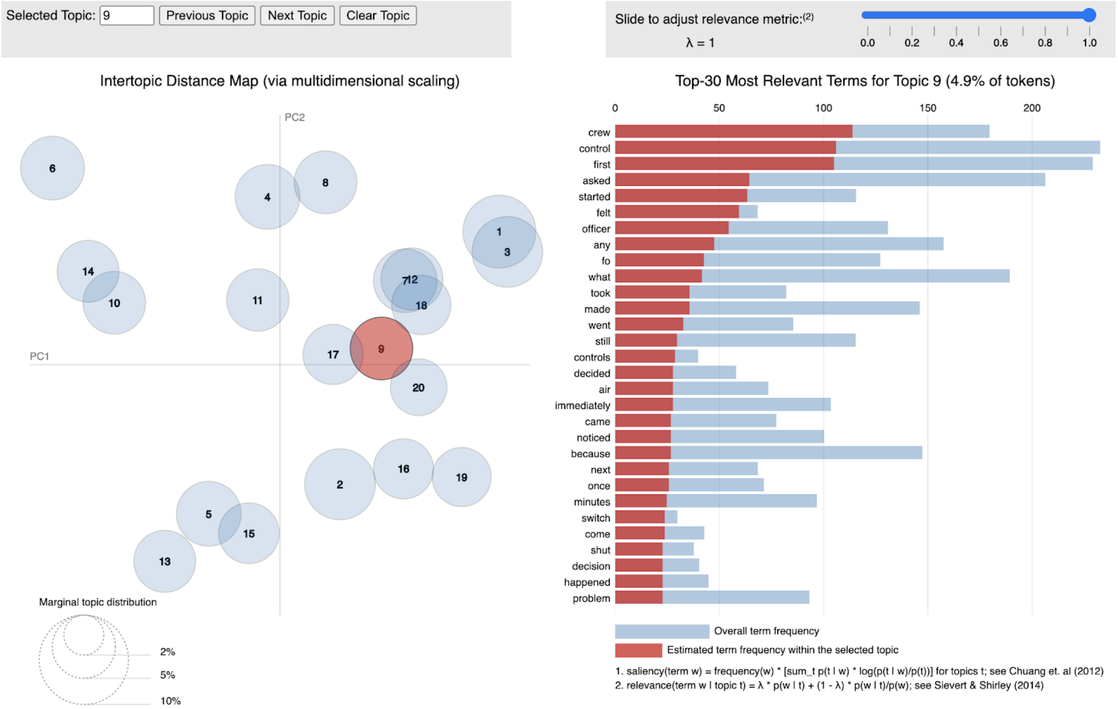
**2018_Feb_1517123** 2/1/2018 at 12:00AM
I filed an IFR flight. It was clear/VFR [at departure airport] and 1;000-1;200 O...

**2018_Feb_1516473** 2/1/2018 at 12:00AM
Configuring for final flaps 30. I called for flaps 25 (allowing to slow for 30 ...

**2018_Feb_1521126** 2/1/2018 at 12:00AM
On the previous flight; we landed in ZZZ where the weather was light snow; BKN 0...

**Fig. 4 TopicFlow [15] Visualization of ASRS 2018 Corpus.**

**Stability**. Third, because none of the existing tools provides a proper quantitative evaluation mechanism to verify documents assignment, we are unable to assess if the documents are consistently assigned to the same topic over repeated executions, due to the random nature of LDA methods. Ideally we would want results to be stable. If the same data is provided as input, and a small number of documents are reassigned to different clusters (and therefore the cluster interpretation may change) we would like this change to be minimal. Current tools do not account for, or even measure, this stability for document-topic assignment or for topic descriptions for sense making.



Selected Topic: 9

Slide to adjust relevance metric:(2)

λ = 1

1. saliency(term w) = frequency(w) * [sum_t p(t | w) * log(p(t | w)/p(t))] for topics t; see Chuang et. al (2012)

2. relevance(term w | topic t) = λ * p(w | t) + (1 - λ) * p(w | t)/p(w); see Sievert & Shirley (2014)

**Fig. 5 LDAVis [17] Visualization of ASRS 2018 Corpus.**

RQ2: Can topic modeling discover ASRS report sets from a collection of mixed reports by their topic?

In Figure 6 we present the results of all the 435 trials as a histogram, using ARI, as defined in Section III.G. The top histogram uses WarpLDA to obtain the groupings. The bottom histogram replaces the algorithm by a random assignment for each report to one of the report sets of each trial. Example top, middle and bottom pairs of report sets (with respect to their ARI scores) is shown in Table 1.

Figure 6 illustrates that WarpLDA results are overall better than random assignment for most pairs. However, the algorithm's ability to identify the original topics is not ideal. Ideally, the upper histogram results would be closer to ARI = 1 (i.e. with most values shifted to the right).

We conclude that the large standard deviation (width of the histogram in the Figure) is not promising for usage on a Real Dataset Setup.

RQ3: Are there specific topics that are easier to group than others?

A question that follows from RQ2 is whether certain report set topics are easier to be separated from *any* other report set (e.g. for problems described using a unique vocabulary). For example, does Cabin Smoke, Fire, Fumes or Odor Incidents in Table 1 consistently receive a high ARI score across all pairs that include it? To answer this we present the mean, median and standard deviation (SD) of each report set across all pairs in Table 2. We also show a histogram of the median ARI column of the table in Figure 7.

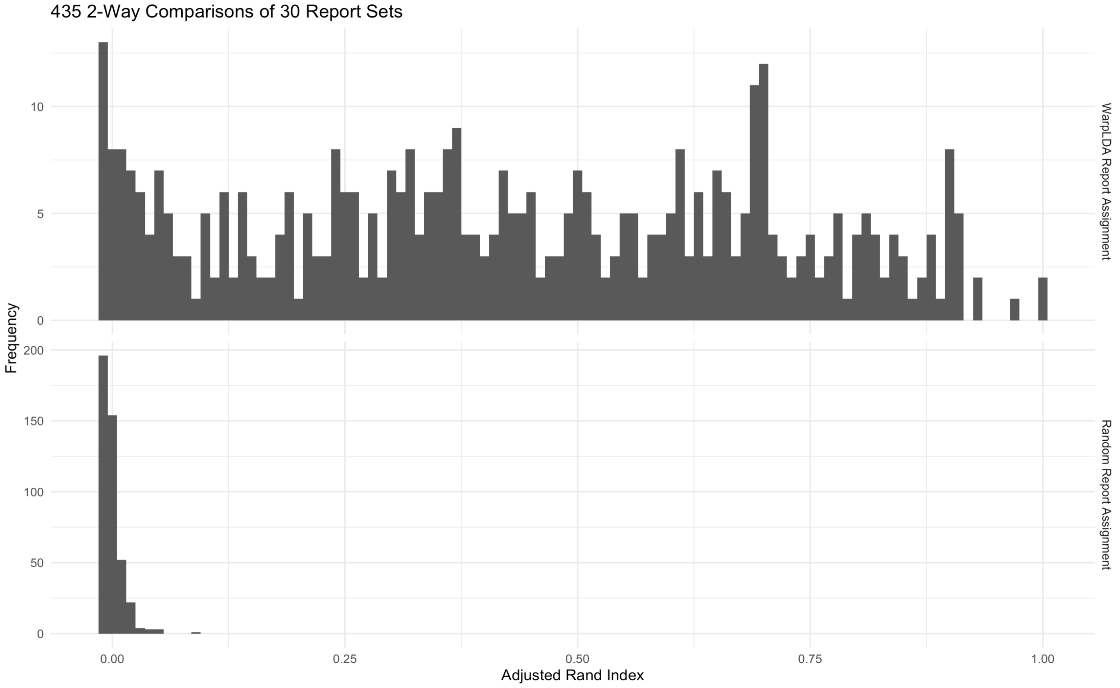


**Fig. 6 WarpLDA evaluation of distinguishing Report Sets "Cabin Fumes, Fire, Fumes, or Odor Incidents", vs "Controller Reports" when compared to manual separation of reports over multiple runs: Accuracy is defined as Adjusted Rand Index, ranging from -1 to 1.**

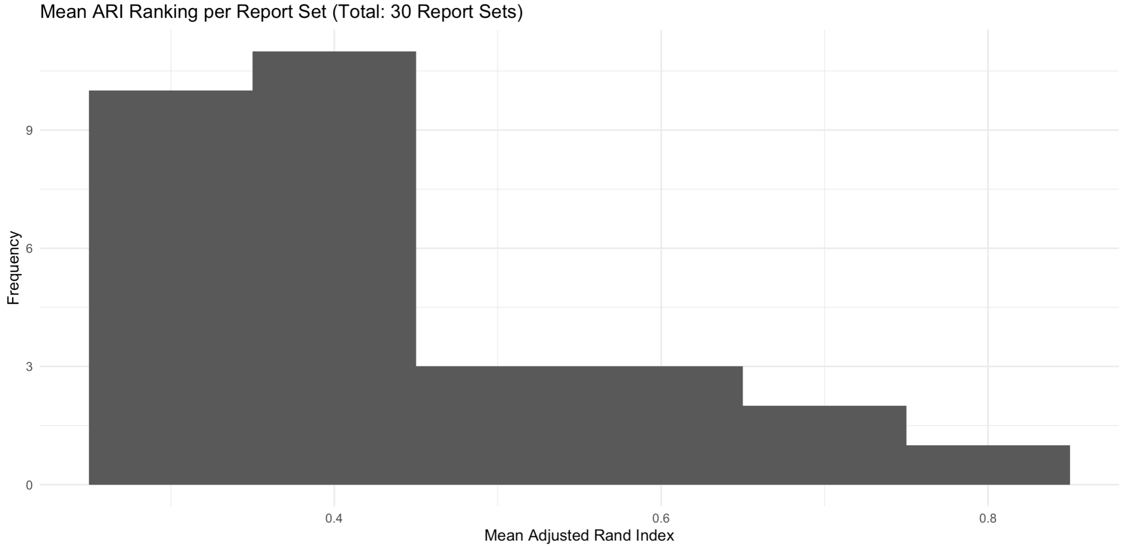


**Fig. 7 Ranking per Report Set.**

From Figure 7, we can see that there are a few report sets that are consistently separable (Mean ARI > 0.8) from other report sets. In particular, we noticed from Table 2 that the most consistently separable report set is not Cabin Smoke (which ranks 3rd) but rather Maintenance Reports.

In future work, we intend to further investigate why report sets such as Maintenance Reports are more separable than lower performance ones, such as Cockpit Resource Management. WarpLDA relies on the distribution of the narrative's words to separate report sets. We expect that the vocabulary used in Maintenance Reports to be substantially different than what is found in other report sets. It is also possible that lower performance report sets can in fact be multi-topic,

**Table 1 Top 5, Middle 5, and Bottom 5 Adjusted Rand Index (ARI) Report Set Pairs out of 30C2 = 435 combinations.**

| Report Set Title 1 | Report Set Title 2 | ARI |
|---|---|---|
| Cabin Smoke, Fire, Fumes, or Odor Incidents | Controller Reports | 1.00 |
| Controller Reports | Flight Attendant Reports | 1.00 |
| Maintenance Reports | Wake Turbulence Encounters | 0.97 |
| Flight Attendant Reports | Global Positioning System (GPS) Reports | 0.93 |
| Controlled Flight Toward Terrain | Flight Attendant Reports | 0.93 |
| Bird or Animal Strike Reports | Rotary Wing Aircraft Flight Crew Reports | 0.46 |
| Commuter and Corporate Flight Crew Fatigue Reports | Multi-Engine Turbojet Aircraft Upsets Incidents | 0.45 |
| Controlled Flight Toward Terrain | Non-Tower Airport Incidents | 0.45 |
| Pilot / Controller Communications | Inflight Weather Encounters | 0.45 |
| Checklist Incidents | General Aviation Flight Training Incidents | 0.45 |
| Rotary Wing Aircraft Flight Crew Reports | Non-Tower Airport Incidents | -0.01 |
| Controller Reports | Parachutist / Aircraft Conflicts | -0.01 |
| Unmanned Aerial Vehicle (UAV) Reports | Inflight Weather Encounters | -0.01 |
| RNAV Arrival Reports | Unmanned Aerial Vehicle (UAV) Reports | -0.01 |
| Commuter and Corporate Flight Crew Fatigue Reports | Emergency Medical Service Incidents | -0.01 |

which violates our assumption of deterministic mapping discussed in section III.D.

## V. Literature Review

Our work is not the first focused on analyzing aviation safety incident narratives. In older work [6–8, 27], we observed that the ASRS-related literature focused on just a specific subset of the available taxonomy. This is not ideal to provide context for new reports nor to observe a given time window for changes in what was reported. For example, [6] provided a qualitative study of human errors in aviation. The data was obtained using Aircraft/Make Model involving passenger fights on Part 121 carriers and conventional aircraft, and analyzed in terms of observable error outcomes and underlying cognitive mechanisms. A related taxonomy component available in ASRS is Person/Human Factors, which include labels such as confusion, distraction, time pressure, etc. A qualitative study technical report was also published by NASA on memory errors in the cockpit of a random sample from 2001 records [7]. Another qualitative work by Jones et al. [8] used Person/Human Factor reports labeled with "Situational Awareness" to better understand the types of situational awareness errors in aviation. The 3-level taxonomy reflects errors due to perception of information, comprehension, and projection of actions, and was previously applied by the author on the National Transportation Safety Board (NTSB) accident investigation reports [27].

More recent work has focused extensively on building classifiers [4, 5, 28] for auto-labeling narratives with existing ASRS taxonomy metadata, which do not align with our primary interest, as our long-term goal is to identify *emerging* safety threats. Specifically, [28] evaluated the accuracy to query a new narrative, and assign it the correct label from Person/Primary Problem and Person/Contributing Factors. The model was trained using ASRS data from 01/2011-01/2013, and the test data from 01/2009-12/2009. The results are not encouraging, with Person/Primary Problem having 49% accuracy, and it is also unclear why the author decided to predict past instead of future data. Oza et al [4] built a classifier on both ASRS and the Aviation Safety Action Plan (ASAP), a private dataset similar to ASRS, to classify a subset of 22 out of the 60 categories due to data sparsity. The results showed it outperformed LDA in a classification task, and manual evaluation of 100 reports suggested that the top three categories achieved agreement of at least 70%. A similar goal to classify metadata was done in [5], however, the metadata was converted in a smaller set of labels based on their risk. The authors also used two learning models in an ensemble, where one leveraged metadata, and the other the narratives.

We believe our approach better aligns with the older literature, in that we seek to provide navigation over the processed narratives on a monthly basis to facilitate, instead of fully automate, the identification of emerging threat

**Table 2 Ranking of Mean, Median and SD Adjusted Random Index (ARI) ordered by Mean per Topic.**

| Report Set Title | Mean | Median | SD |
|---|---|---|---|
| Maintenance Reports | 0.81 | 0.83 | 0.11 |
| Flight Attendant Reports | 0.73 | 0.81 | 0.22 |
| Cabin Smoke, Fire, Fumes, or Odor Incidents | 0.66 | 0.70 | 0.18 |
| Runway Incursions | 0.61 | 0.66 | 0.19 |
| Passenger Misconduct Reports | 0.61 | 0.69 | 0.20 |
| Passenger Electronic Devices | 0.60 | 0.63 | 0.14 |
| Controller Reports | 0.53 | 0.51 | 0.28 |
| Multi-Engine Turbojet Aircraft Upsets Incidents | 0.50 | 0.51 | 0.24 |
| NMAC Incidents | 0.49 | 0.56 | 0.23 |
| Non-Tower Airport Incidents | 0.45 | 0.48 | 0.26 |
| Commuter and GA Icing Incidents | 0.43 | 0.41 | 0.25 |
| Fuel Management Issues | 0.42 | 0.42 | 0.21 |
| Bird or Animal Strike Reports | 0.41 | 0.40 | 0.16 |
| Checklist Incidents | 0.41 | 0.39 | 0.24 |
| Controlled Flight Toward Terrain | 0.40 | 0.37 | 0.29 |
| General Aviation Flight Training Incidents | 0.39 | 0.36 | 0.28 |
| Wake Turbulence Encounters | 0.39 | 0.31 | 0.25 |
| Altitude Deviations | 0.38 | 0.32 | 0.29 |
| Parachutist / Aircraft Conflicts | 0.38 | 0.37 | 0.21 |
| Penetration of Prohibited Airspace Incidents | 0.35 | 0.31 | 0.23 |
| Air Carrier (FAR 121) Flight Crew Fatigue Reports | 0.34 | 0.35 | 0.23 |
| Inflight Weather Encounters | 0.32 | 0.32 | 0.27 |
| Commuter and Corporate Flight Crew Fatigue Reports | 0.31 | 0.31 | 0.25 |
| Pilot / Controller Communications | 0.30 | 0.30 | 0.22 |
| Emergency Medical Service Incidents | 0.30 | 0.26 | 0.23 |
| Unmanned Aerial Vehicle (UAV) Reports | 0.30 | 0.27 | 0.27 |
| RNAV Arrival Reports | 0.30 | 0.21 | 0.28 |
| Rotary Wing Aircraft Flight Crew Reports | 0.28 | 0.26 | 0.25 |
| Global Positioning System (GPS) Reports | 0.27 | 0.19 | 0.28 |
| Cockpit Resource Management (CRM) Issues | 0.26 | 0.32 | 0.21 |

topics, while retaining the context of prior metadata and narratives. Specifically, when applying topic modeling to the ASRS, we identify groupings of reports, which can be understood as an additional piece of information added to each report—the grouping label—however the reports can still be read individually in their original format. The automated step thus serves to reduce the search space of the reader.

## VI. Threats to Validity

In Figure 1, we can see that an analyst must decide (diamond "Adequate?" towards the bottom) if a follow up phone call should be made before the document is deidentified. Such modifications introduce a threat to validity in the underlying distribution of the report set corpus we used in our experimental protocol (Toy Dataset Setup in Figure 2). However, given that ASRS intentionally attempts to keep the original narrative unchanged (in contrast to the *summary* field of the report which is made by an analyst entirely), we believe that this threat to validity is minimal.

The results presented in this work only use one implementation of LDA, i.e. WarpLDA [16], however other implementations could be used. Indeed, other non-generative machine learning methods which are discriminative such as various clustering algorithms could be applied to our experimental protocol which could provide worse or better results. We chose WarpLDA under the rationale that as the number of reports continue to grow, an efficient implementation of algorithm would be preferred in practice, but a more detailed investigation of the trade-off between performance versus accuracy could be performed in further studies utilizing our experimental protocol.

As noted in the Method Section III, in utilizing the *Toy Dataset Setup* instead of the *Real Dataset Setup*, we accept the risk that the Toy Dataset Setup's underlying report set word distribution may not fully reflecting that of the Real Dataset Setup's incoming monthly corpus. We did so to be able to use the report sets already produced by the ASRS as a gold standard to evaluate the automated step results. Specifically, in the experimental protocol Figure 2, the parameter $l$, the number of report sets, could be much larger than what we have tested, and $|R_i|$ may be a different size. It may also be possible that reports about a given topic are associated with different months or years, which may or not affect their underlying distribution compared to the real dataset setup. While this is a limitation of the generalization of the results, we believe our work provides a starting point for tests in future work.

## VII. Conclusion and Future Work

In this work, we began exploring how topic modeling could be used to assist in identifying relevant topics in the ASRS database. We chose to use unsupervised learning methods such as topic modeling, because they offer the potential to identify *emerging* safety threats as compared to pre-trained classifiers, which can only identify what they have been trained for. A common limitation of topic modeling is that evaluation is often done subjectively, using 'top terms' as shown in the tools. We proposed a Toy Dataset Setup experimental protocol, and executed this protocol on ASRS report sets. We concluded that although there is potential for topic modeling to be used, as a few report sets were consistently separable, results still have too large a standard deviation to be used in practice. In future work, we intend to experiment with different topic modeling models besides WarpLDA. Specifically, we want to test variants of LDA that allow for the manual input of 'seed terms' to identify topics of interest [29], or the usage of existing connections between documents [30]; for example, in ASRS co-occurring metadata could be leveraged as connections between documents. There is also potential of performance improvement by using word embedding, such as GloVe [31].

## Acknowledgments

The material is based upon work supported by NASA under award No. 80NSSC19M0202. This research was partially conducted at NASA Ames Research Center. Reference herein to any specific commercial product, process, or service by trade name, trademark, manufacturer, or otherwise, does not constitute or imply its endorsement by the United States Government.

## References

[1] ASRS, "Callback Number 317 March/April 2006," `https://asrs.arc.nasa.gov/publications/callback/cb_317.html`, 2006. [Online; accessed 18-June-2020].

[2] Billings, C., Lauber, J., Funkhouser, H., Lyman, E., and Huff, E., "NASA aviation safety reporting system," 1976.

[3] System, A. S. R., "Program Briefing," https://asrs.arc.nasa.gov/overview/summary.html, 2019. [Online; accessed 18-June-2020].

[4] Oza, N., Castle, J. P., and Stutz, J., "Classification of Aeronautics System Health and Safety Documents," *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews)*, Vol. 39, No. 6, 2009, pp. 670–680.

[5] Zhang, X., and Mahadevan, S., "Ensemble machine learning models for aviation incident risk prediction," *Decision Support Systems*, Vol. 116, 2019, pp. 48 – 63. https://doi.org/https://doi.org/10.1016/j.dss.2018.10.009, URL http://www.sciencedirect.com/science/article/pii/S0167923618301660.

[6] Sarter, N. B., and Alexander, H. M., "Error Types and Related Error Detection Mechanisms in the Aviation Domain: An Analysis of Aviation Safety Reporting System Incident Reports," *The International Journal of Aviation Psychology*, Vol. 10, No. 2, 2000, pp. 189–206. https://doi.org/10.1207/S15327108IJAP1002_5, URL https://doi.org/10.1207/S15327108IJAP1002_5.

[7] Nowinski, J. L., Holbrook, J. B., and Dismukes, R. K., "Human Memory and cockpit Operations: An ASRS Study," Tech. rep., NASA Ames Research Center, 2003. URL https://human-factors.arc.nasa.gov/flightcognition/Publications/Nowinski_etal_ISAP03.pdf.

[8] Jones, D., and Endsley, M., "Sources of situation awareness errors in aviation," *Aviation, space, and environmental medicine*, Vol. 67, 1996, pp. 507–12.

[9] Neuhaus, S., and Zimmermann, T., "Security Trend Analysis with CVE Topic Models," *2010 IEEE 21st International Symposium on Software Reliability Engineering*, 2010, pp. 111–120.

[10] Agrawal, A., Fu, W., and Menzies, T., "What is Wrong with Topic Modeling? (and How to Fix it Using Search-based SE)," *CoRR*, Vol. abs/1608.08176, 2016. URL http://arxiv.org/abs/1608.08176.

[11] Blei, D. M., "Probabilistic Topic Models," *Commun. ACM*, Vol. 55, No. 4, 2012, p. 77–84. https://doi.org/10.1145/2133806.2133826, URL https://doi.org/10.1145/2133806.2133826.

[12] Chang, J., Gerrish, S., Wang, C., Boyd-graber, J. L., and Blei, D. M., "Reading Tea Leaves: How Humans Interpret Topic Models," *Advances in Neural Information Processing Systems 22*, edited by Y. Bengio, D. Schuurmans, J. D. Lafferty, C. K. I. Williams, and A. Culotta, Curran Associates, Inc., 2009, pp. 288–296. URL http://papers.nips.cc/paper/3700-reading-tea-leaves-how-humans-interpret-topic-models.pdf.

[13] Mantyla, M. V., Claes, M., and Farooq, U., "Measuring LDA Topic Stability from Clusters of Replicated Runs," *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, Association for Computing Machinery, New York, NY, USA, 2018. https://doi.org/10.1145/3239235.3267435, URL https://doi.org/10.1145/3239235.3267435.

[14] Chuang, J., Manning, C. D., and Heer, J., "Termite: Visualization Techniques for Assessing Textual Topic Models," *Proceedings of the International Working Conference on Advanced Visual Interfaces*, Association for Computing Machinery, New York, NY, USA, 2012, p. 74–77. https://doi.org/10.1145/2254556.2254572, URL https://doi.org/10.1145/2254556.2254572.

[15] Malik, S., Smith, A., Hawes, T., Papadatos, P., Li, J., Dunne, C., and Shneiderman, B., "TopicFlow: Visualizing Topic Alignment of Twitter Data over Time," *Proceedings of the 2013 IEEE/ACM International Conference on Advances in Social Networks Analysis and Mining*, Association for Computing Machinery, New York, NY, USA, 2013, p. 720–726. https://doi.org/10.1145/2492517.2492639, URL https://doi.org/10.1145/2492517.2492639.

[16] Chen, J., Li, K., Zhu, J., and Chen, W., "WarpLDA: a Simple and Efficient O(1) Algorithm for Latent Dirichlet Allocation." *CoRR*, Vol. abs/1510.08628, 2015. URL http://dblp.uni-trier.de/db/journals/corr/corr1510.html#0001LZC15.

[17] Sievert, C., and Shirley, K., "LDAvis: A method for visualizing and interpreting topics," 2014. https://doi.org/10.13140/2.1.1394.3043.

[18] Vinh, N. X., Epps, J., and Bailey, J., "Information Theoretic Measures for Clusterings Comparison: Variants, Properties, Normalization and Correction for Chance," *J. Mach. Learn. Res.*, Vol. 11, 2010, p. 2837–2854.

[19] Blei, D. M., Ng, A. Y., Jordan, M. I., and Lafferty, J., "Latent dirichlet allocation," *Journal of Machine Learning Research*, Vol. 3, 2003.

[20] Gondal, M. W., Wuthrich, M., Miladinovic, D., Locatello, F., Breidt, M., Volchkov, V., Akpo, J., Bachem, O., Schölkopf, B., and Bauer, S., "On the Transfer of Inductive Bias from Simulation to the Real World: a New Disentanglement Dataset," *Advances in Neural Information Processing Systems 32*, edited by H. Wallach, H. Larochelle, A. Beygelzimer, F. d'Alché-Buc, E. Fox, and R. Garnett, Curran Associates, Inc., 2019, pp. 15740–15751. URL http://papers.nips.cc/paper/9704-on-the-transfer-of-inductive-bias-from-simulation-to-the-real-world-a-new-disentanglement-dataset.pdf.

[21] Storn, R., and Price, K., "Differential Evolution –A Simple and Efficient Heuristic for global Optimization over Continuous Spaces," *Journal of Global Optimization*, Vol. 11, No. 4, 1997, pp. 341–359. https://doi.org/10.1023/A:1008202821328, URL https://doi.org/10.1023/A:1008202821328.

[22] Mullen, K., Ardia, D., Gil, D., Windover, D., and Cline, J., "DEoptim: An R Package for Global Optimization by Differential Evolution," *Journal of Statistical Software, Articles*, Vol. 40, No. 6, 2011, pp. 1–26. https://doi.org/10.18637/jss.v040.i06, URL https://www.jstatsoft.org/v040/i06.

[23] Paradis, C., Kazman, R., and Wang, P., "Indexing Text Related to Software Vulnerabilities in Noisy Communities Through Topic Modelling," *2018 17th IEEE International Conference on Machine Learning and Applications (ICMLA)*, 2018, pp. 763–768. https://doi.org/10.1109/ICMLA.2018.00121.

[24] McCormick, W. T., Schweitzer, P. J., and White, T. W., "Problem Decomposition and Data Reorganization by a Clustering Technique," *Operations Research*, Vol. 20, No. 5, 1972, pp. 993–1009. https://doi.org/10.1287/opre.20.5.993, URL https://doi.org/10.1287/opre.20.5.993.

[25] Gower, J. C., "Some Distance Properties of Latent Root and Vector Methods Used in Multivariate Analysis," *Biometrika*, Vol. 53, No. 3/4, 1966, pp. 325–338. URL http://www.jstor.org/stable/2333639.

[26] Manning, C. D., Raghavan, P., and Schütze, H., *Introduction to Information Retrieval*, Cambridge University Press, USA, 2008.

[27] Endsley, M., "A taxonomy of situation awareness errors, human factors in aviation operations;," *Proceedings of the 21st Conference of the European Association for Aviation Psychology (EAAP)*, Vol. 3, 1995, pp. 287–292.

[28] Robinson, S. D., "Multi-Label Classification of Contributing Causal Factors in Self-Reported Safety Narratives," *Safety*, Vol. 4, No. 3, 2018. https://doi.org/10.3390/safety4030030, URL https://www.mdpi.com/2313-576X/4/3/30.

[29] Porteous, I., Newman, D., Ihler, A., Asuncion, A., Smyth, P., and Welling, M., "Fast Collapsed Gibbs Sampling for Latent Dirichlet Allocation," *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Association for Computing Machinery, New York, NY, USA, 2008, p. 569–577. https://doi.org/10.1145/1401890.1401960, URL https://doi.org/10.1145/1401890.1401960.

[30] Chang, J., and Blei, D., "Relational Topic Models for Document Networks," *Proceedings of the Twelth International Conference on Artificial Intelligence and Statistics*, Proceedings of Machine Learning Research, Vol. 5, edited by D. van Dyk and M. Welling, PMLR, Hilton Clearwater Beach Resort, Clearwater Beach, Florida USA, 2009, pp. 81–88. URL http://proceedings.mlr.press/v5/chang09a.html.

[31] Pennington, J., Socher, R., and Manning, C. D., "GloVe: Global Vectors for Word Representation," *Empirical Methods in Natural Language Processing (EMNLP)*, 2014, pp. 1532–1543. URL http://www.aclweb.org/anthology/D14-1162.