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
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