Identifying Emerging Safety Threats Through Topic Modeling in the Aviation Safety Reporting System: A Covid-19 Study

Carlos Paradis, Rick Kazman, Misty Davies, Becky Hooey
University of Hawaii at Manoa (Information and Computer Sciences), cvas@hawaii.edu;
University of Hawaii at Manoa (Shidler College of Business), kazman@hawaii.edu;
NASA Ames Research Center (Intelligent Systems Division), misty.d.davies@nasa.gov;
NASA Ames Research Center (NASA Aviation Safety Reporting System),
becky.l.hooey@nasa.gov

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 many ways, including via an online database, Safety Alert Bulletins, For Your Information Notices, and the CALLBACK newsletter.

In this work, we consider whether we can improve the grouping, linking, and understanding of safety concerns through topic modeling. Specifically, we use topic modeling as a building block to identify emerging safety threats over time. This unsupervised approach, we argue, offers the flexibility to identify new emerging themes in this large dataset by constructing different timelines based on the content similarity of ASRS report narratives. This method's unsupervised nature improves upon related research, which is limited to pre-defined labels and therefore can not fully capture emerging safety threats.

We apply our method to all ASRS reports in 2020 to assess if the generated timelines can highlight COVID-19 as it is emerging as a safety threat in incoming ASRS reports. We perform both a quantitative and qualitative evaluation of the automatically constructed timelines. The qualitative evaluation is performed by describing the evolution of top terms in the timelines, generated by our method, which we found explicitly convey the themes of COVID-19. Separately, we use a set of 1,213 COVID-19 reports from 2020 that were manually identified by ASRS analysts to quantitatively evaluate the COVID-19 reports distribution across the timelines.

Our results have shown that COVID-19 emergence can be identified using the top terms that were generated by topic modeling. The top terms in topic modeling therefore can serve as a summary alternative to manually inspecting reports. Moreover, leveraging the manually identified COVID-19 reports, we found the manually identified timelines accounted for over 70% of the COVID-19 reports curated by the ASRS analysts, which demonstrates the potential of this approach for facilitating the understanding of safety concerns as they emerge and evolve. This method shows great potential to understand aerospace safety threats and other narrative-driven incident report databases.