RTI Technical Exercise - Investigating the Relationship Between Fatalities and Narrative Topic In the NTSB Aviation Dataset

Prudvi Gaddam

Investigating the Relationship Between Fatalities and Narrative Topic in the NTSB Aviation Dataset

Overview

The accident narratives within the National Transportation Safety Board (NTSB) dataset were analyzed using natural language programming (NLP) to identify topic clusters. Five topic clusters were identified:

1. Helicopter accidents
2. Mechanical accidents
3. Student pilot accidents
4. Unknown topic category
5. Weather-related accidents

These topic clusters were then compared against accident fatality to understand if any trends within the data could be identified. Non-fatal student pilot accidents were the most common narrative, making up 33.8% of all observations. Weather-related accidents were the most common category of fatal accidents making up 44.6% of all fatal accidents.

Methodology & Analysis

The structured data in the dataset was analyzed using descriptive statistics to understand the various continuous and categorical variables in the model. Counts were used for categorical variables and histograms were used for continuous variables. The incident was classified as being binarily fatal or non-fatal by feature engineering the *Injury Severity* variable. The unstructured data was explored by analyzing word use trends across various decades, the most common words used, and histograms of word count across the observations.

Prior to building the model, the unstructured narrative text data for each observation was first stripped of sentences associated with NTSB actions after the accident had occurred, as these sentences do not add information about the accident itself. The text data was then pre-processed using standard NLP techniques and subsequently, the topics were determined using a Latent Dirichlet Allocation (LDA) model, an unsupervised topic clustering algorithm. The optimal *number of topics* was identified as 5. Each narrative was then fit to the final LDA model, and the topic with the highest score was assigned to the narrative. The data was then subset by topic and fatality class and proportions were analyzed.

Results and Discussion

Figure 1 shows the proportional breakdown of topics by fatality class. Non-fatal and fatal accidents made up 79.1% and 20.9% of the observations, respectively. Four discernable topics were identified from the text data – helicopter accidents, mechanical accidents, student pilot accidents, and weather-related accidents. The fifth category labeled as unknown was difficult to decipher without adequate domain knowledge and has been left as the “Unknown” category.

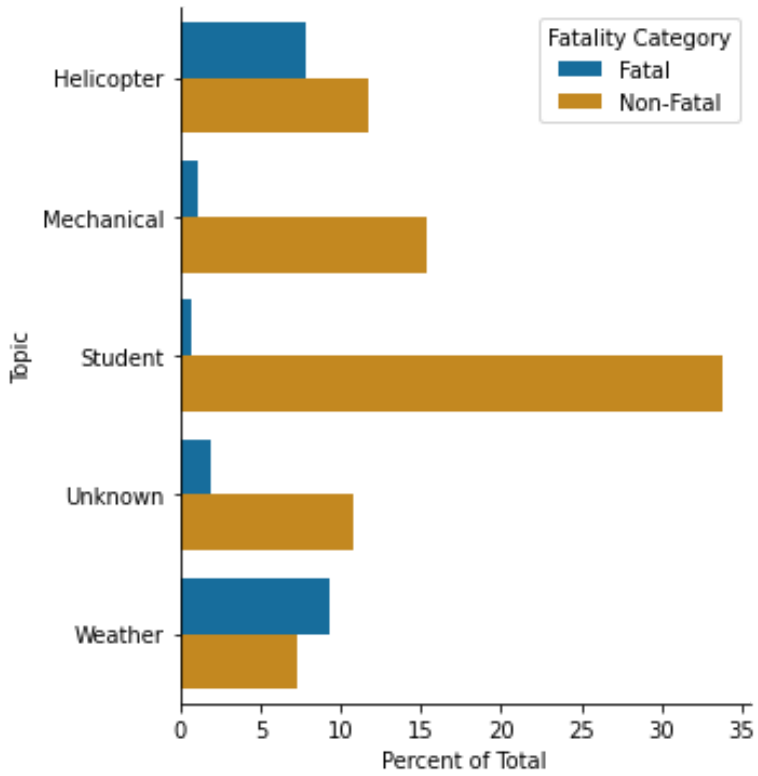

Figure 1: Bar chart showing the proportional breakdown of narrative topics within fatal and non-fatal accident categories

The most common accident theme was non-fatal student pilot accidents which make up 33.8% of the narratives present within the dataset. In contrast, Student pilot accidents were the least common type of fatal accident making up only 0.7% of the dataset. Within the fatal category, weather-related accidents were the most common type of accident making up 9.3% of the data overall and accounting for 44.6% of all fatal narratives. Additionally, there are 1.3 times more fatal weather-related accidents than non-fatal weather-related accidents, and this is the only category where fatal accidents outnumber non-fatal accidents.

Helicopter and mechanical-failure accidents are primarily non-fatal, with the non-fatal accidents making up 11.8% and 15.4% of the data respectively. The fatal versions of these accidents make up 7.8% and 1.2% of the data respectively.

This analysis indicates that special care must be afforded to minimize the risks associated with weather-related accidents as well as helicopter accidents, given that they make up the bulk of the fatal accidents. Further analysis would also investigate the impact of category intersectionality on fatalities. Currently, each narrative is categorized as a single topic however one could have student-weather-related accidents or helicopter accidents due to mechanical failures. Additionally, the fatalities category can be expanded to be continuous and the relationship between narrative topics and the number of fatalities can be analyzed.