Text Mining the Contributors to Rail Accidents

Rail accidents represent an important safety concern for the transportation industry in many countries. In the 11 years from 2001 to 2012, the U.S. had more than 40,000 rail accidents that cost more than $45 million. While most of the accidents during this period had very little cost, about 5,200 had damages in excess of $141,500. To better understand the contributors to these extreme accidents, the Federal Railroad Administration has required the railroads involved in accidents to submit reports that contain both fixed field entries and narratives that describe the characteristics of the accident. While a number of studies have looked at the fixed fields, none have done an extensive analysis of the narratives. This paper describes the use of text mining with a combination of techniques to automatically discover accident characteristics that can inform a better understanding of the contributors to the accidents. The study evaluates the efficacy of text mining of accident narratives by assessing predictive performance for the costs of extreme accidents. The results show that predictive accuracy for accident costs significantly improves through the use of features found by text mining and predictive accuracy further improves through the use of modern ensemble methods. Importantly, this study also shows through case examples how the findings from text mining of the narratives can improve understanding of the contributors to rail accidents in ways not possible through only fixed field analysis of the accident reports.