Thematic analysis of TAIC's body of Knowledge Final report

James Thompson - 22nd December 2023

Introduction

This project set off to turn a proof of concept into a more polished prototype. The project consists of two main components, the engine and the viewer. The engine is the brains of the project while the viewer is how you interact with the engine (through a website). Below I have briefly discussed the work that has been done, the current state of the project and further work. For a deeper dive into how the program actually works I recommend visiting the repository which can be found publicly on GitHub (https://github.com/1jamesthompson1/TAIC-report-summary). Equally some instructions on how to use this program and webapp can be found on the GitHub repository and elsewhere.

Work completed

The proof of concept was funded by Robert Thompson and was completed as my final semester university project, which can be found here. The proof of concept had basic theme generation and weightings attribution. However it lacked three important things; a sophisticated viewer, multi transport mode support and explanation of outputs.

The first task to do was add in multi transport support. This allowed the engine to complete the analysis of reports across aviation, marine and rail reports. Due to the flexible nature of the engine this was not too difficult. However the general safety theme generation had to be tweaked to allow for safety themes that are only relevant to particular transport modes.

The engine outputs a lot of reasoning that is similar to a human. The proof of concept was compared to human answers and was found to be broadly quite similar. However it would make the output more reliable if it had evidence and references to back it up. This means forcing the model to provide citations for all claims that it makes. To make sure that the citations are valid the engine checks all of the produced references with the original reports.

The sophistication of the viewer was quite important to make it usable by a wider audience. The searching of the analysis was upgraded to allow more complex search queries (i.e AND, OR etc). Filtering of the reports through the weightings, transport mode or year, allows the user to find specific groups of reports with greater ease than just search queries. Lastly a results summary generator was added to demonstrate a basic analysis of the search results.

Current state

The program created by the project works in two stages.

The first stage is the “analysis” stage. This is where the engine does the main work of creating the database of reports and summaries. The steps are:

1. Get all the reports from the website
2. Identify safety themes present in each report
3. Identify and consolidate the general safety themes present across all reports
4. Describe how much each general safety theme contributed to a particular accident analysis

After this stage you have a folder full of the analysis from each report. With each report folder having three things:

- The safety issues extracted verbatim from the report
- A list of potential safety themes generated from the report.
- A description of how each general safety theme contributed to the accident with a percentage weightings.

In addition to each report folder there are also the general safety themes which have names, descriptions and a safety theme group assigned to them.

This first stage takes about 4-8 hours to run and costs about $150. Little time has been spent on optimising the engine to reduce run time or cost, further work would be possible to bring these numbers down.

This database then gets used by the viewer for the “viewing” stage. This is where you use the webapp (found here) to search and discover the information in the database. The website allows you to do three main actions (which can all be combined together):

- Search all the reports for a query. For example “mast bumping”. There is support for google style advanced queries, for example “crash AND runway”
- Filter the reports based on its properties. For example restricting the results to be online marine and from the 2015-2018. It also allows you to filter based on how much a particular general safety theme or safety theme group contributed towards an accident.
- Getting a summary of the search results. There is an option to make the application generate a PDF summary of the search results. This document is self-contained and provides you with some basic statistical and semantic summaries of the search results.

Further work

Given that this project is still early in its life there are many different avenues that it could go down. For brevity sake I will simply list some of the interesting ideas that I believe could be the most successful.

Firstly is making the engine work on more transport agencies accidents reports. This would increase the number of reports available from a couple of hundred to a few thousand. In doing this it would bring this tool to be a world class accident report analyzer. Furthermore the increase of data would allow the creation of a purpose trained Large Language Model on the reports. This would allow a direct conversation between a person and all of the reports. It

would also increase the scope and usefulness of any search as you would be looking through a database that is over 10 times larger.

Secondly is upgrading the viewer and the second stage of analysis. As the viewer is the user's main interaction there are many things that could be upgraded to help the experience of using the viewer. This includes making the search and filtering more intuitive as well as adding in smarter helpers to interpret the results. The viewer and subsequent search results also gives an opportunity for another layer of analysis to take place. There is an example summariser of the search results built however this is quite basic and could be taken a lot further.

Lastly, deepening and developing what the engine does on each of the reports. Currently the engine generates the safety themes and determines the contribution of general safety themes. Firstly when it reads the report it only reads the analysis and finding section. This is due to the Large Language Model being restricted in how much text it can deal with at one time. With advances in technology it is quite possible to have the whole report read which may strengthen the output. Secondly as mentioned above each report could actually be used to train a unique LLM opening up the possibility of having an intelligent chatbot.

Conclusion

With a goal of having a tool that can “‘unlock’ information” trapped inside the published accident investigation reports. This prototype is in the position to give utility to users of the software today. The prototype is also in a prime position to expand out and start working on accident reports from across the globe.

The web application is of a level that with the addition of some introductory instruction they would be able to navigate and have most features that a searcher would normally have. With the addition of the reasoning and the references the output of the engine should be trustable and tested enough to pass the standards that one would expect of a colleague. Importantly with the addition of all modes of transport into the database has meant the size is large enough for meaningful analysis and searches.

As outlined in the further work this is still very much a prototype meaning that more work is required to bring it to a production standard and equally more work would be fruitful. It should be noted that although the engine and its output is tested and validated in multiple ways there is no guarantee that everything in the output is correct. Furthermore one should apply the same level of scrutiny as they would of anything a colleague or fellow human tells them.