Analysing Tweets from news organisations regarding health

Sentiment Analysis and Topic Modelling

Project Scope

Perform an analysis of tweets from major news organisations regarding any matter related to health. This will utilise NLP and so the data must first be cleaned and processed for analysis.

Firstly, topic modelling will be applied to the dataset, in the hope of highlighting the major areas of focus from the news organisations when reporting on health.

Secondly, sentiment analysis will provide a more thorough understanding of the language and sentiment employed in these tweets. The sentiment can be analysed across time and compared over different organisations to find any trends or correlations.

Data Cleaning

Gathering all the files and compiling them into one readable csv. This was done through notebook.

The files were ‘|’ delimited, however, there were tweets which contained this delimiter, causing an issue when reading these lines.211 lines were skipped which is 0.3% of the population size. As such I made the decision to skip these lines and spend time elsewhere.

Data cleaning, removing stop words and lemmatisation can be performed in Azure designer via the ‘Preprocess Text’ component. However, the lemmatization step produced an issue, where any pronouns within the text, were replaced with the following string ‘-prod-’. I first noticed this issue in the word cloud produced in Appendix A.1. Mindful of time, I skipped lemmatization, with the intention of resolving this issue after whole process has been implemented and time providing.

**Sentiment Analysis**

* EDA
* Manual wordcount
* Vader sentiment scores
* Visualisations & Conclusions

**Topic Modelling**

* EDA
* Text Pre-processing
* LDA Model
* Visualisations & Conclusions

This result also may have come from the fact that tweets are very short and this method, LDA (which works very well for longer text documents), does not work well on shorter text documents like tweets. In the bonus section to follow I suggest replacing the LDA model with an NMF model and try creating a new set of topics. In my own experiments I found that NMF generated better topics from the tweets than LDA did, even without removing ‘climate change’ and ‘global warming’ from the tweets.

Script

* Raw data
* First step
* For each tweet