Analysing Tweets from news organisations regarding health

Sentiment Analysis and Topic Modelling

Project Scope

Analysis of health-related tweets from major news organisations, utilising Azure Machine Learning Studio and performing Natural Language Processing (NLP). Firstly, topic modelling will be applied to the dataset, in the hope of highlighting the major areas of focus from news organisations when reporting on health. Secondly, sentiment analysis will provide a more thorough understanding of the language and sentiment used in these tweets.

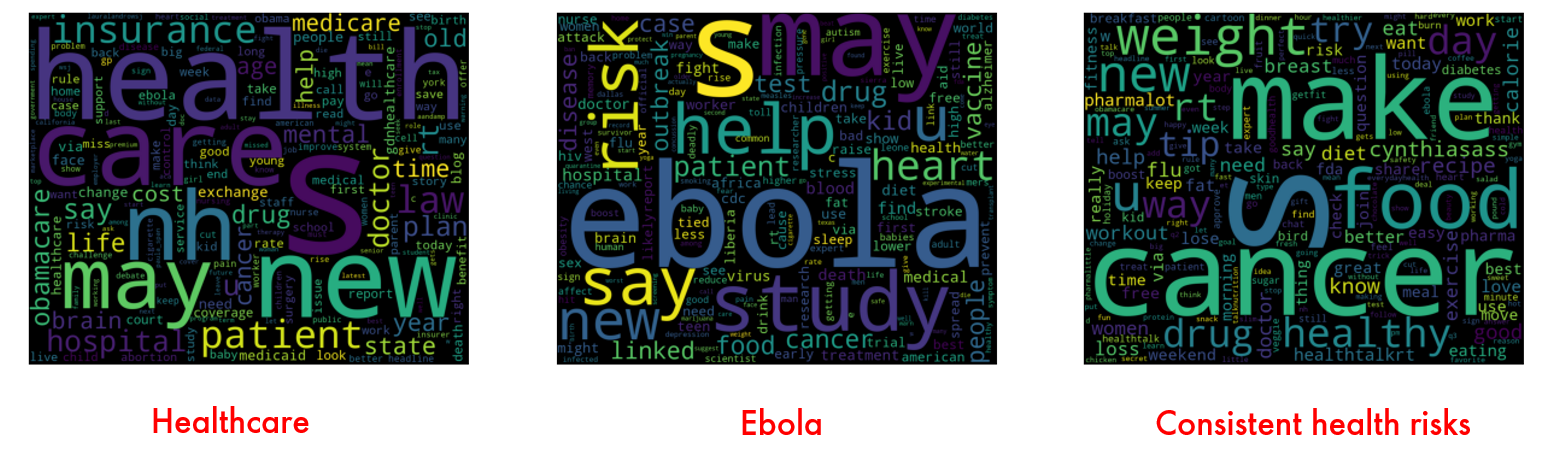
Data Cleaning & Pre-processing

Compiled all raw data files into one dataframe, removing 0.3% of the dataset during cleaning procedure. Removing stop-words, normalising case and removing URLs were among some of the pre-processing features performed on the dataset via the ‘Pre process Text’ module in Azure ML Designer.

**Topic Modelling**

The Latent Dirichlet Allocation (LDA) module in Azure designer can perform topic modelling on the tweets within the dataset. It will assign a probability score, giving the likelihood of each datapoint being assigned to a particular project. The highest probability indicates the topic the tweet was assigned to.

To choose the optimum number of topics to run the LDA model, the average of all the highest probabilities were taken for each topic number. This was then plotted, and the difference was taken with random chance probability to find the optimal number of topics, which was 3. A word cloud was then produced (as shown below), with suitable labels added for each topic.



**Sentiment Analysis**

Two methodologies two perform sentiment analysis. The first was to perform a sentiment wordcount with reference to the Loughran McDonald dictionary. The other was to use the VADER sentiment analysis package which analysis both the polarity and intensity of sentiments in text. The visualisations produced allowed a better insight into the reporting style of these organisations.