# The sentiment analysis of online news articles before and after a controversial event

Matthew Simpson - MSc in Big Data

UNIVERSITY of STIRLING

*The story so far…*

## Background information

The word “Sentiment” can be traced back to medieval Latin “Sentimentum”. Meaning – “What one feels about something”.

An analysis is more commonly associated with social media posts or reviews about products and services. Understanding your customer sentiment is vital, bad reviews can put off prospective buyers and this is obviously undesirable for both small and large company alike.

This project goes a different route by understanding the sentiment of news articles posted online by cable news networks, smaller new age media sites and traditional newspapers.

Specifically, before and after longstanding controversial events where new information is brought to the table. **Does the perceived article sentiment change in relation to a public figure or person(s) noticeably?**

## Technology used and pre-processing

The data is collected and specific news outlets filtered by implementing the ‘newsplease’ crawler package with CommonCrawl.org WARC file repositories. Each article is stored as a single JSON file.

Merging and transformation into a Pandas dataframe object allows flat table conversion for ease of use and visual clarity at the analysis stage. Dropping of metadata columns deemed unuseful for the analysis.

Filtering out of non-related articles related to topic via specific keywords. Subsequent cleaning of special characters and less than useful common short words for standardising textual data.

## Planned Techniques

***Rule-based Approach***

Use pre-existing lexicon dictionaries available online to gauge polarity of articles – Postive/Negative

Pros: No training data required
Cons: Not as accurate

***Automatic Approach (Machine Learning)***

Modelled as a classification problem, an ML algorithm seeks to learn from labelled training data and apply classifications on sentiment polarity. Specific algorithms used TBC.

Pros: Can be scaled
More accurate
Cons: Manual labelling for specific story streams

***Current project stage: Data cleaning & filtering***



For further information contact: Matthew Simpson

mas00257@students.stir.ac.uk