# Assignment 3

## Document 1: Overview

**MA5851 Data Science Master Class 1**

**James Cook University**

**Sukhchain Singh Bhathal**

**13863320**

**Issue:**

- **The Issue:**
  The newspapers that follow celebrities or famous people and make up stories on them in order to attract readers' attraction this trend is called **news sensationalism**. According to new research… (Why Humans Value Sensational News: An Evolutionary Perspective, 2003) categorization of the front-page stories and content of such stories is socially constructed by many individuals. Recently, after analyzing the content of 14 television systems and 29 TV stations that found that dependency on commercial revenue restores the use of vivid storytelling. News sensationalism is the main point behind this project to find the public interest of individuals on such stories.
- **How the issue relates to Natural Language, meta-data, and or other data sources:**
  I choose this topic for this project as this topic news sensationalism is becoming a grim issue nowadays. This project will research vivid storytelling by fetching data from different news websites to investigate their story's content and see how it is present in current days.

  News sensationalism is also related to meta-data as every news website contain some additional information for each of the story published on the news website. However, over time, it has evolved into a race to find the most sensationalist, stirring, and spectacular stories to maintain profits as well as a top rating.

- **Where the Issue is present on the world wide web:**
  This issue is not a centralized location that can be centralized. This is an issue is related to every news website and almost every news website have the

**WebCrawler Alignment:**

- **Relevant domain(s) on the world wide web:**
  https://www.bbc.com/news
  https://www.bbc.com/news/world/asia
  https://www.abc.net.au/news/

- **Alignment of the chosen domains to the issue with linkages to how the chosen domains could be expanded:**

The chosen domains are related to the issue of news sensationalism as chosen domains are the news-related domains and by investigating the news stories from these domains we can identify the correctness of the news and how authentic they are.

- **Identification and discussion of the NL variables from the domain(s) to the issue:** From the chosen domains, I have selected the news title, summary as on the above domains there is not must information related to the news story.
- **The reasoning for the use of a web crawler with considerations to scalability:** In order to fetch the news storying from the above-mentioned domains, so that their content and titles can be checked against other websites via Natural Language tasks such as **Sentiment Analysis, Topic Modelling** by implementing Natural language techniques such as **Tokenization, Stemming/Lemmatization.**

**NLP Tasks Alignment:**

- **Purpose of the two NLP tasks:** The two NLP tasks which I have used in this project are used to break down the human language into computer-readable form and identify news **sensationalism** by reading news titles and summaries to identify the public option on such stories.
- **Alignment of the NLP tasks and the issue, with linkages between the NLP tasks:** As the sentimental analysis is the analysis that is widely used to monitor social media sites to see the public interest on certain topics. That's why I choose this analysis to identify the public option on news stories. Secondly, I choose,
- **Alignment of the data harvested from the web crawler(s), with linkages to the NLP tasks:** The web crawler will get the URL of a website and then crawl the website via the python **selenium package** which will fetch the news title and summary of each website and then follow the **pagination** of that website to get the content of the international pages as well. In the end, data will be written down into a .csv file for further processing.
- **A concise summary of sequencing of NLP methodologies in each NLP task:** In this project, I have used two NLP methodologies which are Sentiment Analysis and Summarization both the methods used the same sequencing in order to activate the desired results.

  Firstly, I have to fetch the data from news websites by fetching their story title and summary. Once the content is fetched. Secondly, I apply the Exploratory data analysis by applying data cleaning and data manipulation tasks. Then, I apply the Sentiment Analysis for task 1 and Summarisation for task 2 by implementing NL techniques Tokenization and Steeming respectively.

- **A concise summary of the ML used in the NLP task:**
  In this project, I have used two ML techniques to check the text from the above-mentioned domains. **Sentiment analysis:** studies the subjective information in an expression, that is, the opinions, appraisals, emotions, or attitudes towards a topic, person, or entity. Expressions can be classified as positive, negative, or neutral.

**Word Count: 620**

**References:**

1. Why humans value sensational news: An evolutionary perspective. (2003, May 1). ScienceDirect. https://www.sciencedirect.com/science/article/abs/pii/S1090513803000126?via%3Dihub
2. *SAGE Journals: Your gateway to world-class research journals*. (n.d.). SAGE Journals. Retrieved December 9, 2021, from https://journals.sagepub.com/action/cookieAbsent
3. Williams, A. (2005, December 6). *Celebrating the everyday heroes*. Townhall. https://townhall.com/columnists/armstrongwilliams/2005/12/06/celebrating-the-everyday-heroes-n1121673