Analysing Sentiment Trends in News Headlines

1. Introduction and Motivation

Advancements in digital communication have transformed the way news headlines are shared and consumed. Notably, socioeconomic events, such as the COVID-19 pandemic, have severely affected the emotional tone of these headlines. News headlines influence public perception significantly and reflect prevailing societal sentiments, shaping individual worldviews. They offer concise depictions of significant events, enabling an examination of societal attitudes and thus highlighting the importance of exploring evolving sentiment trends and their wider consequences. This report aims to inform readers on how sentiment analysis was used to examine the trends of news headlines from 2007 to 2021 in several countries. Analysing these results would provide insight into the consequences of sentiment shifts, illuminating broader societal impacts and influencing our interpretation of modern events and how media and communication strategies are developed and employed. The results are expected to reveal a discernible trend of heightened negative sentiment over time due to increased adverse global events.

2. Methods and Dataset

2.1 Data Collection

For this project, a total of four datasets from Kaggle (see links in references) were used, all consisting of news headlines from 2007 to 2021 and by prominent new publishers, with the countries including Australia (ABC News), Ireland (The Irish Times), India (Times of India), England (Daily Mail), and the United States (The New York Times). The datasets were in the form of CSV files, consisting of two string attributes: the publish date and the headline text. Their respective authors originally created the data by scraping the different news publisher's websites. Three datasets consisted of news headlines from a single country and one from multiple countries and news outlets, thus requiring some pre-processing before being analysed.

2.2 Data Pre-processing

Before conducting sentiment analysis, data pre-processing was performed to ensure the dataset's quality and consistency. This involved creating Python code and using the Pandas library to filter the CSV file with multiple news outlets into separate files for England and the United States. Additionally, the dataset's format was standardised, and the headlines were limited to 128 characters to suit the input of the sentiment analysis models. Specifically for the Australian dataset, the word "abc" had to be omitted because some values were categories rather than valid headlines.

2.3 Analysis

The method employed for the analysis encompassed a systematic approach consistent across sentiment, emotion, and keyword analyses. Multiple text classification models on Hugging Face were used to process the respective datasets. These models allowed for the extraction of valuable insights from the text data. The code analysing the news headlines was run within the Google Collab environment to use the cloud processing power for reduced model inference times. Subsequently, the results of each analysis were saved to CSV files and grouped in three-month intervals for further examination and creation of appropriate visualisations using Matplotlib and the WordCloud library. This standardised approach ensured consistency and facilitated the comprehensive exploration of sentiment, emotion, and keyword trends within the news headlines dataset.

3. Experimental Setup and Results

3.1 Sentiment Analysis Setup

Sentiment analysis was performed on the collected dataset of news headlines first. The sentiment analysis model "siebert/sentiment-roberta-large-english" from Hugging Face was used. The headlines were analysed individually, and their sentiment labels were recorded as either "positive" or "negative". A line graph was then produced from the results.

3.2 Emotion Analysis Setup

Emotion analysis was conducted on the same dataset, employing the "finiteautomata/bertweet-base-emotion-analysis" model from Hugging Face. A batch size of 64 was employed for faster processing. Emotion labels such as 'joy,' 'surprise,' 'sadness,' 'fear,' 'anger,' 'disgust,' and 'neutral' were recorded for each headline. A horizontally stacked bar graph with a slider for the time to animate the graph was then made from the result.

3.3 Keyword Analysis Setup

Keyword analysis involved the extraction of significant keywords from the headlines. The "yanekyuk/bert-keyword-extractor" model was used from Hugging Face. Keywords were filtered to limit keywords to at least three characters and to exclude hashtags generated from the model. Furthermore, the sentiment of the keywords was then analysed using the "cardiffnlp/twitter-roberta-base-sentiment-latest" model, consistent with the sentiment analysis process applied to headlines previously. This sentiment analysis approach was conducted for each keyword, generating sentiment labels for individual keywords. Five separate word clouds were then made using the WordCloud library for each country, again with a slider for the time to animate the visualisation.

3.4 Sentiment Analysis Results

The sentiment analysis results have unveiled noteworthy trends in the emotional tone of news headlines during the specified timeframe. These trends were scrutinised on a quarterly basis, allowing for the discernment of subtle shifts in sentiment over time, as shown in Figure 1. India emerged as the most positive sentiment contributor, while England displayed the highest negative sentiment, with all the countries becoming slightly more positive over time. This may be due to the differences in the sentiment and emotion of the topics for the headlines of India and England. The headlines for India mainly consisted of entertainment and gossip, whereas those for England involved world events and politics. These findings accentuate the divergent emotional landscapes in news headlines from different regions. Moreover, the observed shifts in sentiment over time strongly correlated with major global events such as the global financial crisis (2007 - 2009) and the COVID-19 pandemic (2020 - 2021), underscoring the profound influence of significant occurrences on public sentiment as reflected in headlines.

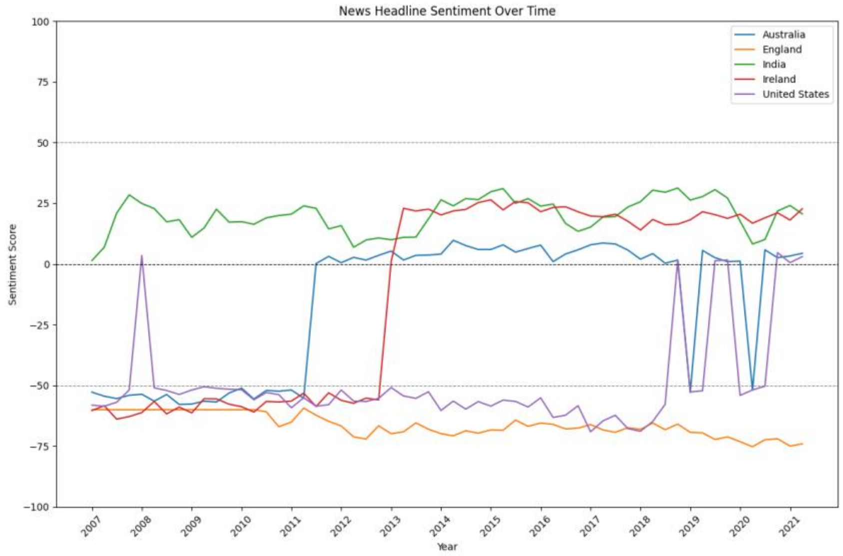

Figure 1: Sentiment analysis results

3.5 Emotion Analysis Results

The results from emotion analysis have unveiled the prevailing emotional undercurrents within news headlines, offering valuable insights into the emotional landscape surrounding significant news events, as shown in Figure 2. The analysis of emotions revealed a prevalence of negative emotions, such as disgust, sadness and fear. The results are also aligned with the sentiment analysis results, as there were more negative emotions during the global financial crisis and COVID-19, although the headlines generally contained more positive emotions, such as surprise and joy, over time. Additionally, India consistently had the greatest number of positive emotions, and England had the least. This suggests a notable correlation between sentiment and the emotional tone of headlines. These findings provided a nuanced understanding of the emotional spectrum within the news landscape.

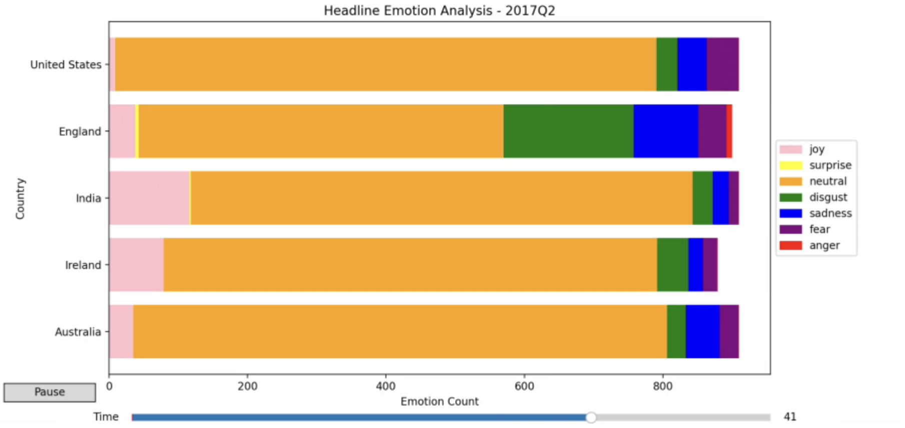

Figure 2: Emotion analysis results

3.6 Keyword Analysis Results

The keyword analysis delved into identifying pivotal keywords embedded within the headlines, illuminating recurring themes and notable subjects, as shown in Figure 3. The influence of specific keywords and topics on sentiment and emotion within headlines predominantly encompassed geographical locations, including China and Russia, and names such as Donald Trump and Joe Biden. This is clearly linked to the US elections and the political tension between the US, China and Russia. This signifies the role of places and people in shaping public sentiment. Furthermore, the analysis highlighted that Australia and Ireland exhibited the most substantial variations in keyword sentiment, with a notable trend towards increasing positivity over time. This evolution reflects changing societal attitudes and perceptions.

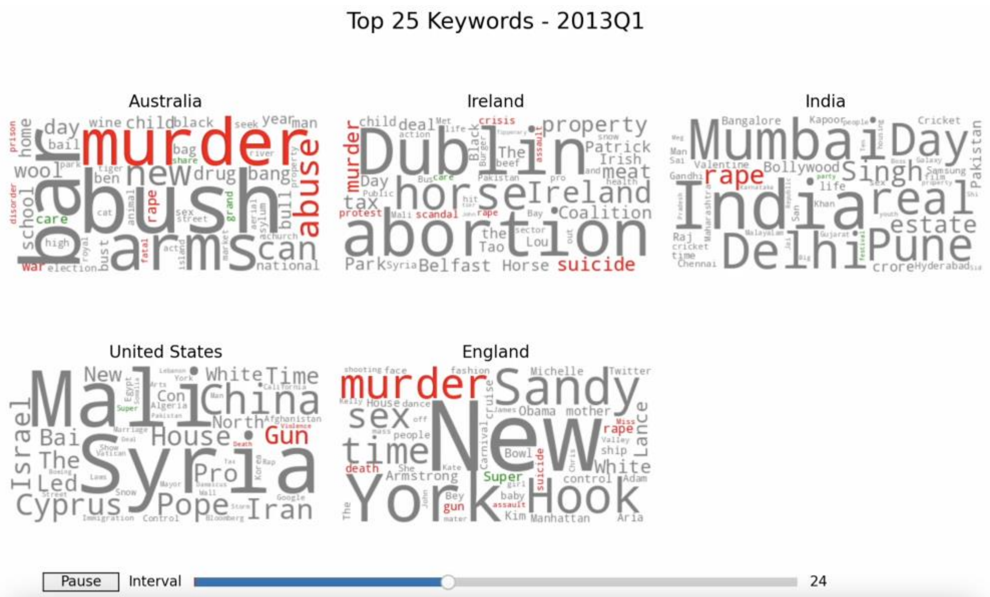

Figure 3: Keyword analysis results

4. Conclusion

The findings presented in this report offer valuable insights into the evolution of news headline sentiment and emotions over a period spanning from 2007 to 2021. Several key conclusions emerge from the analysis conducted. A notable trend is the general shift towards more positive headlines over time. This suggests an overarching pattern of increased positivity in news reporting, potentially reflecting changing societal attitudes and a preference for more uplifting news narratives, opposite the expected results. Secondly, the study reveals that India consistently exhibited the most positive sentiment in its headlines, while England consistently displayed the highest negative sentiment. These regional variations underscore the diverse emotional landscapes present in news headlines across different countries. Thirdly, it becomes evident that major global events significantly influenced sentiment and emotions within news headlines. The correlation observed between shifts in sentiment and noteworthy events highlights the dynamic and responsive nature of news reporting to the world's changing landscape.

However, it is essential to acknowledge certain limitations of this study. While robust, the sentiment analysis models employed are not infallible, and some inaccuracies may be present. Additionally, this study considered a relatively small selection of news headlines, and sentiment and emotions may vary based on other contextual factors not explored in this analysis. Nevertheless, the implications of these findings are substantial. The shifts in sentiment and emotional tone within news headlines hold the potential to impact various domains. Changes in sentiment can influence international relations, political dynamics, and diplomatic interactions. Furthermore, headline emotions can affect financial markets and influence investment decisions. Understanding these emotional dynamics can assist governments and organisations in crisis management, enabling more informed responses to events with widespread societal impacts.

5. References

4.5M headlines from 2007-2022 [10 largest sites] (2021) *www.kaggle.com*. Available at: https://www.kaggle.com/datasets/jordankrishnayah/45m-headlines-from-2007-2022-10-largest-sites?select=headlines.csv

A Million News Headlines (2021) *www.kaggle.com*. Available at: https://www.kaggle.com/datasets/therohk/million-headlines.

India News Headlines Dataset (2021) *www.kaggle.com*. Available at: https://www.kaggle.com/datasets/therohk/india-headlines-news-dataset

Irish Times - Waxy-Wany News (2021) *www.kaggle.com*. Available at: https://www.kaggle.com/datasets/therohk/ireland-historical-news.

6. Appendix

The feedback received after presenting my project proposal indicates several points of interest and suggestions for improvement. These comments primarily pertain to the methodology, potential sources of error, the effectiveness of sentiment analysis models, and the implications of the trends observed.

Categorisation of Feedback

Methodology: Some feedback points touch upon the methodology used for sentiment analysis, the potential risk of removing certain headlines, the effectiveness and reproducibility of AI models, and other factors impacting results.

Data Sources: There are suggestions to explore different media outlets and compare how different news sources may vary in sentiment. Additionally, there is a mention of investigating other major media companies.

Analysis and Implications: Feedback also addresses the correlation with changes in sentiment in different regions, the implications of negative headlines, and the need to investigate key influencing factors and keywords in negative news headlines.

Data Handling: Some feedback notes the difficulty of sorting through a large dataset, the need for data filtering, and the potential for misleading or inaccurately labelled headlines.

Presentation: There is a comment about the clarity of my voice during the presentation.

Addressing the Feedback

Methodology and Analysis: My final project presentation and report have effectively addressed the feedback about the methodology and analysis. The project aims to conduct sentiment analysis using specific models, and it acknowledges the potential limitations of these models, such as some degree of error. It also considers the implications of trends in headline negativity. However, some specific suggestions, considering the reproducibility of AI models, could not be implemented due to time constraints.

Data Sources: The feedback regarding exploring different media outlets and investigating other major media companies was heavily considered as I changed my dataset from just ABC News to a total of five different news outlets, all from different countries.

Analysis and Implications: Suggestions to explore key influencing factors in negative news headlines were not considered due to time constraints, and the scope of the research would have been too large.

Data Handling: The feedback about data handling, such as the difficulty in sorting through data and the need for data filtering, was addressed during the data pre-processing phase. The project aimed to present concise and clear visualisations to facilitate data interpretation.

Presentation: While the feedback mentions voice clarity, it does not impact the project's full execution as the final report is text-based. However, I practised my final presentation more to become more confident in my speaking.

In summary, the feedback received has been considered, and where applicable, it was incorporated into the project's methodology and analysis to enhance the depth and quality of the research.