# Analytics Specializations & Applications

# Coursework

# Academic year 2023-2024

# ‘AI Law & Regulations’

## Executive Summary

In this report, we would like to analyze the information extracted from reddit posts and news articles to know more about topics related to AI Law and Regulation. Once we have crawled the dataset from online sources, the dataset is then pre-processed to be prepared for topic modeling and sentiment analysis.

For topic modelling, 10 significant topics were identified in both news and reddit dataset. These topic names were defined by us based on the top 20 words from each themes found. Next, we conducted a sentiment analysis on the reddit and news website dataset using Textblob, and further compared the sentiment analysis using LIWC.

Our analysis found that news articles had a generally more positive and neutral sentiment score towards AI Law and Regulations. In contrast, reddit has more variation in sentiment score with a higher negative sentiment score. Through our LIWC analysis, we concluded that there is a relationship between the sentiment score and the dictionaries available in LIWC. Furthermore, our study shows that there is a negative correlation between publics opinion on Reddit and the news articles being published.

For future research, business can extend it into a longitudinal study to understand the sentiment trend over time. They can also do further comparative analysis of sentiment scores with industry trends and policy announcements on AI Law and Regulation to provide further context to the analysis and its sentiments.

## Methodology

The web scraping process for collecting data on AI Law and Regulations involves identifying relevant sources, analyzing website structure, selecting scraping tools, sending HTTP requests, parsing HTML content, extracting desired data, cleaning and transforming data, and storing it in a suitable format. Additional techniques, such as handling pagination and dynamic content, may be required, and it is crucial to regularly monitor and maintain the scraping process. We can efficiently and effectively extract accurate and high-quality data for further analysis and use by these steps.

We identified 4 news websites that have news related to AI law and regulations and crawled them with respective API and tools together with a specified social media, Reddit. Before this, we ensure the relevant news is actively discussed as this guarantees the quality and quantity of data, for which the volume of users and contents enhances the interpretability of analysis. Then we extracted the post titles, authors, postdates and their respective articles to further analyze them. For Reddit, we first installed and registered a developer account with client id and client secret, then initialized Praw with python to access the subreddits and eventually fetched all data through iterations on information stated above as well as the votes and comments that specifically exists in Reddit.

Before proceeding with the textual analysis, we performed a basic preprocessing to clean the crawled data. In the cleaning stage, non-alphanumeric words, and symbols, stop word and some self-identified insignificant words were removed to make the analysis focused on the more relevant words. Both the news and reddit dataset was also normalized to ASCII, checked for missing and duplicating values and tokenized for text analysis.

Next, topic modelling methods such as Latent Dirichlet Allocation (LDA) and Non-Negative Matrix Factorization (NMF) were used to find the prevalent top 10 topics using 20 words for each theme found. Based on our understanding, we named the 10 topics found by looking at words for each topic/theme identified. Furthermore, Term Frequency-Inverse Document Frequency (TF-IDF) vectorization was applied to the dataset to transform the dataset into quantifiable entity. This process allows us to understand the importance of each term within the document based on their frequency.

After naming 10 different themes of our reddit and news website data, we categorized all our articles and comments with topic named. Then, we conducted a sentiment analysis using TextBlob to understand the polarity and subjectivity of each datapoint. We further categorized sentiments of each datapoint into negative, neutral and positive based on the polarity score to indicate the sentiment score underlying each datapoint.

Lastly, once we get the sentiment score of each datapoint, we conducted a secondary sentiment analysis using Linguistic Inquiry and Word Count (LIWC) to perform a correlation analysis between the result we generated with LIWC's result to reassure that our result is legitimate and further research on any potential topics related to our analysis. This comprehensive approach allowed us to explore and conduct an in-depth analysis of our dataset.

## Data Description

### News Websites

| **News Websites** | ZDNet | The Verge | CBC | The Guardian | **Total** |
|---|---|---|---|---|---|
| **Number of Unique Articles** | 1,294 | 621 | 302 | 1,200 | **3417** |
| **Data Items Used** | Title; Author; Publication Dates; Content | | | | |
| **Number of Unique Authors** | 47 | 64 | 105 | 600 | **816** |
| **Data Range** | January 1, 2023 - December 31, 2023 | | | | |
| **Geographic Area** | • English-focused<br>• Global coverage | | | | |

Table 1: News Websites Table

A comprehensive data extraction and analysis project was undertaken to sieve through the vast amounts of published content across four prominent news platforms: CBC, The Verge, ZDNet and TheGuardian. The focus was explicitly on articles related to "AI and Law Regulations". Leveraging advanced web scraping technologies, the project adeptly filtered keywords within the 2023 publication year's corpus. This meticulous process was not merely limited to retrieving articles; it involved a sophisticated data de-duplication strategy and relevance comparison to ensure the extracted information's uniqueness and pertinence. Specialized algorithms parsed through the content, identifying and cataloging entries based on their publication date, while meticulously documenting the contributing authors. This procedure ensured the extraction of not just quantitative data, but also qualitative insights, highlighting the most prolific contributors within the sphere. The endeavor underscored the significance of leveraging automation and analytical tools in distilling and synthesizing vast datasets into coherent, actionable intelligence.

Upon the completion of the data collection and refinement process, the findings presented a fascinating narrative across a trio of websites for the year 2023. CBC emerged with 326 articles, with contributions from January 2 to December 7, attributed to 111 unique writers. Elizabeth Thompson emerged as the most prolific, contributing to five articles, showcasing a broad spectrum of insights into AI legalities. Meanwhile, TheVerge charted a path with 655 articles, propagated by 67 individual authors, spanning from January 3 to December 31. Emilia David stood out remarkably with a staggering 88 articles, epitomizing an in-depth exploration into AI and law regulations. www.zdnet.com eclipsed the other platforms with an awe-inspiring assembly of 1,294 articles contributed by 47 writers, from January 1 to December 29, with Sabrina Ortiz leading the charge with 284 contributions, mirroring an intense focus on the nexus between AI and legal frameworks.

The comparative analysis of four news platforms highlighted a stark divergence in content volume and author engagement. ZDNet's concentrated editorial focus on AI's legal ramifications catered to technically inclined readership, while The Verge's prolific output indicated a robust internal focus or a more engaged freelance cohort on AI-related legal issues. CBC's moderate output reflected a balanced, investigative approach to coverage. These platforms' varying editorial stances and capacities illustrated the complexity within the domain of AI and legal regulations and contributed to the evolving discourse on society's integration of AI into legal frameworks. The Guardian's data crawled yielded 1,200 articles from 600 unique journalists covering updated news about changing AI regulations across countries. The laws aimed to regulate AI across industries, like healthcare, politics, military, social media, and education. Tech companies were actively modifying regulations to optimize the use of AI.

**Reddit Social Media Platform**

| **Keyword** | AI Law | AI Ethics | Legislation on Artificial Intelligence | Regulatory Challenges in AI | AI Compliance | Data Privacy and AI | **Total** |
|---|---|---|---|---|---|---|---|
| **Number of Unique Articles** | 570 | 583 | 459 | 606 | 108 | 689 | **3,015** |
| **Data Items Used** | Keyword; Subreddit; Post Title; Post Author; Post Upvotes; Number of Comments on Post; Comment Body; Comment Date; Comment Upvotes | | | | | | |
| **Number of Unique Users** | 466 | 507 | 391 | 420 | 94 | 524 | **2,402** |
| **Data Range** | January 1, 2023 – December 31, 2023 | | | | | | |
| **Geographic Area** | • English-focused<br>• Global coverage except subreddit Legal Advice UK (tailored to United Kingdom) | | | | | | |

Table 2: Reddit Table

For Reddit, the data is crawled within 9 subreddits including Legal Advice UK, Singularity, Futurology, Technology, Law, Artificial, Artificial Intelligence, ChatGPT and OpenAI. They also have active users ranging from 0.28 million to 20 million, allowing us to collect diverse perspectives of input from these globally welcomed communities. As shown in table 2, there are 3015 unique articles, for which it neglects the multiple appearances of same articles in different keywords. The 2402 unique users also indicate the appearance of repeated users within these communities who discussed these keywords related to AI law and regulations. As far as the data is dated within the year 2023, we have extracted the top 5 relevant comments for our analysis.

Generally, users were all raising ethical concerns of using AI and embracing the capability of future AI. Many discussions were surrounded with whether AI should be heavily regulated and decelerated for various cases in the community. The positive feedback reflects that people are concerned about these issues and worried about the potential threats brought by AI.

## Data Analysis

### Topic Modelling

After performing LDA and NMF topic modeling process for both the news and reddit dataset respectively, we named the topics based on top 20 words found in each theme. Next, we went through the AI and Law articles, we saw that there is a common theme for some of the keywords found. Then, we assigned a topic name based on the top 20 keywords for each theme based on our own understanding and context. For our news dataset, NMF was chosen because the topics generated by LDA such as 'AI in Music Industry Trend', 'Social Dynamics' and 'Daily News' were not relevant to our 'AI Law and Regulation' context. Similarly, LDA was chosen for Reddit data rather than NMF because the topics such as 'AI Chatbot interactions', 'News & Blogging' and 'Forum Discussion' were not relevant to our focus of this analysis.

Table below are the finalized topic names for our News and Reddit dataset used for this analysis:

| | **News** | **Reddit** |
|---|---|---|
| **Topic 1** | **Advanced AI Technology** | **AI Ethics and Regulation** |
| **Topic 2** | **Legal and Government Regulation** | **Public Opinion on AI** |

| | | |
|---|---|---|
| **Topic 3** | **Content Management & Digital Privacy** | **Generative AI** |
| **Topic 4** | **Public Opinion** | **AI Bot Interaction** |
| **Topic 5** | **Generative AI** | **Forum Regulation** |
| **Topic 6** | **Cyber Security** | **AI Guideline** |
| **Topic 7** | **AI & Search Engines** | **AI Personification** |
| **Topic 8** | **AI in Creative Arts** | **AI Innovation** |
| **Topic 9** | **SEO & SEM** | **AI Chatbots** |
| **Topic 10** | **AI in Education** | **AI Company** |

Table 3 – Finalized Topic of News & Reddit.

### Sentiment Analysis

After we have named the topic name, we computed the cosine similarity and assigned a topic to each news article and reddit comment respectively based on the highest similarity. This allows us to better conduct sentiment analysis on our dataset and further analyze the sentiment differences between topics and media platform. Once we have gotten the polarity and subjectivity from TextBlob, we categorized that data with below – 0.1 polarity as 'Negative' sentiment, above 0.1 polarity as 'Positive' sentiment and the rest as 'Neutral'.

### *Comparative Analysis*

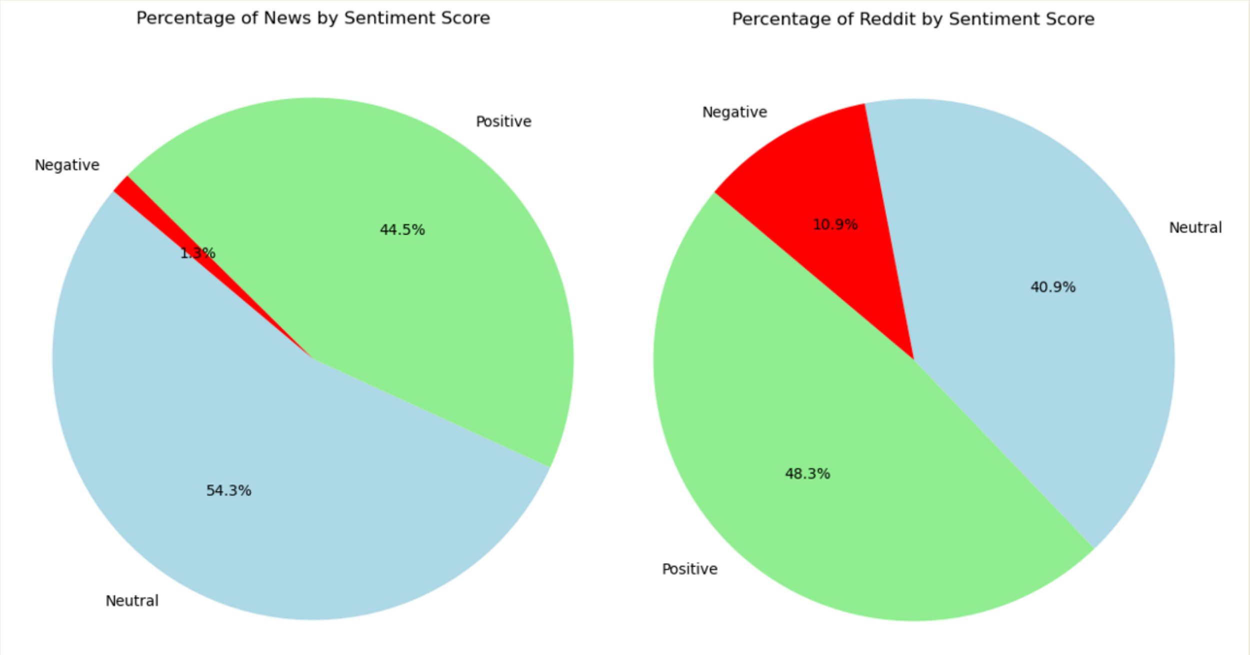

Diagram 1: Percentage of news by sentiment score (Left) and percentage of reddit by sentiment score (Right)

According to Diagram 1, we can see that in news websites, for any articles related to our topic, most of the published article is in a neutral stance with the majority of 54.3% being neutral while the negative sentiment is almost negligible with only 1.3%. However, it shows a huge difference in the reddit platform where they have a 10.9% of negative sentiment score which is almost 10 times more than the news articles. Furthermore, people in reddit generally speak more positive opinion towards AI Law and Regulation and having a lesser people stand a neutral party. This could indicate that news websites usually provide articles that have a less biased opinion towards AI Law and Regulation. Nevertheless, reddit is a social media platform where people can have freedom of speech toward their own opinions, thus we will observe more distinct negative and positive sentiment score compared to news website.

### ***News***

*Sentiment Score by Website*

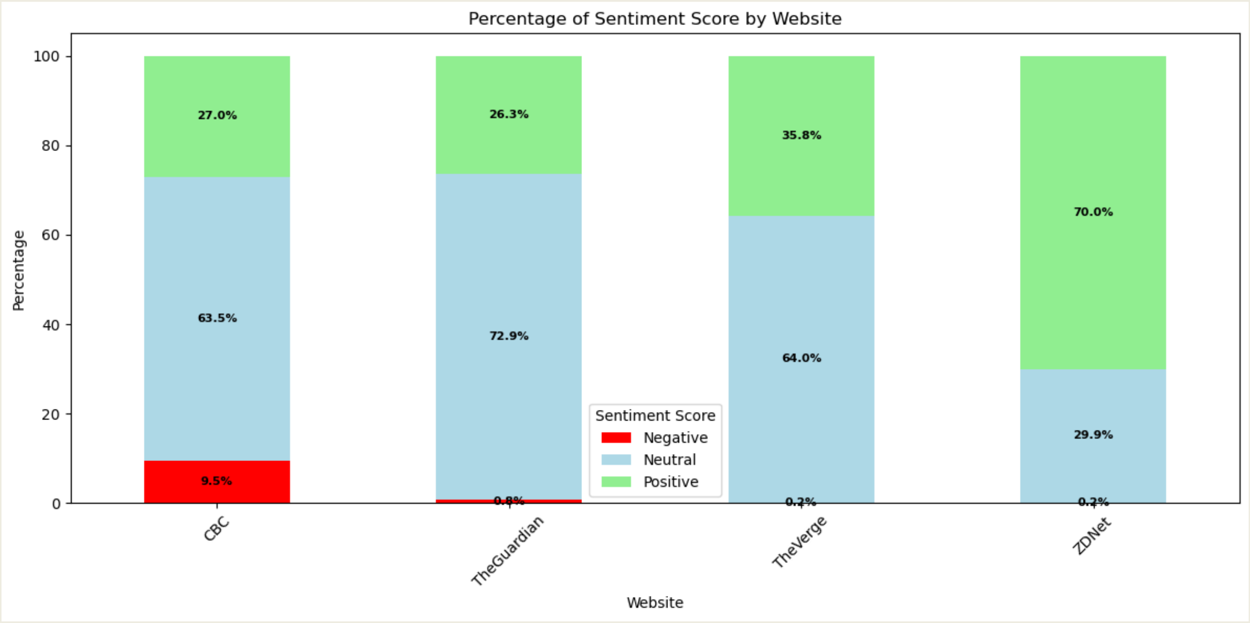

Diagram 2: Sentiment Score of News Website

In Diagram 2, ZDNet shows the highest positive sentiment score of 70% being positive towards the 'AI Law & Regulation' related articles and followed by TheVerge, CBC and TheGuardian respectively, with all 3 of the news websites having halves of the sentiment score of ZDNet. Overall, this shows that ZDNet published articles that are speaking positively towards AI Law and Regulation. Meanwhile, there's another notable point which is that CBC and TheGuardian – news that considered as general news reporting websites contain negative sentiments towards the AI regulations topic. Whereas, CBC has the most negative sentiment score out of 4 websites mentioned previously. In contrast, TheVerge and ZDNet, who are mainly reporting technology news, rarely contain any negative sentiments towards the topic.

*Sentiment Score by Topic*

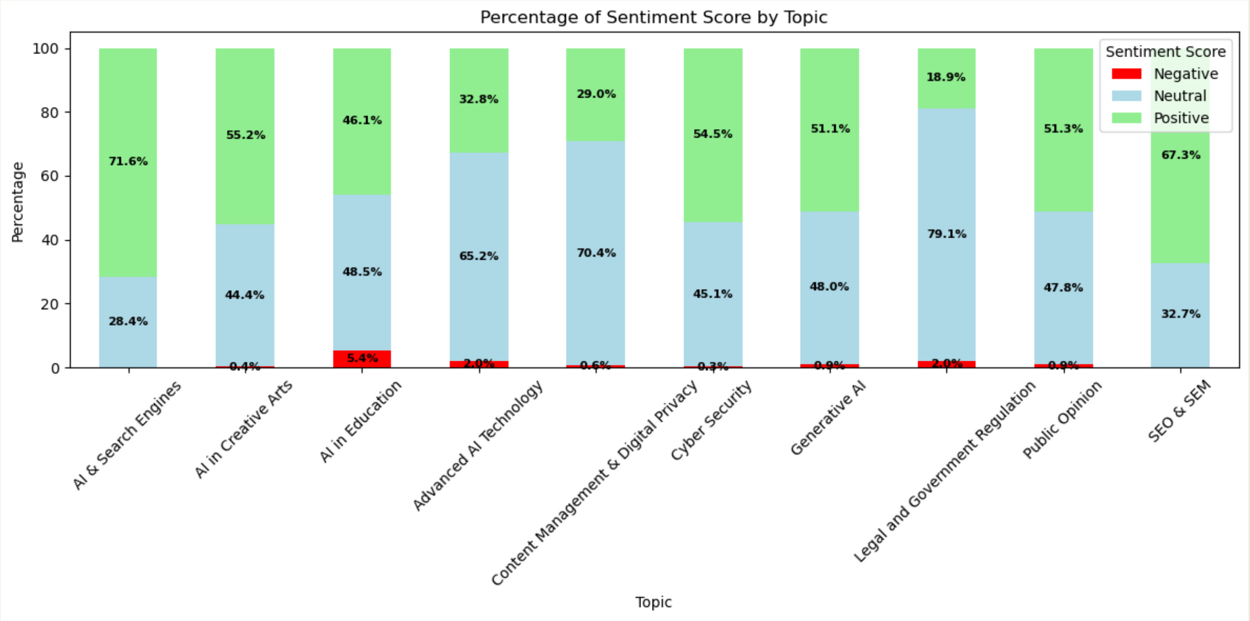

Diagram 3: Sentiment Score of News Website Topic

As shown in Diagram 3, we can see that the topic regarding AI & Search Engines and SEO & SEM have a highly positive sentiment score and there is no negative sentiment in such topics at all. Thus, showing that news websites are generating mostly positive perspective when it comes to articles related to search engines. In contrast, when it comes to topics related to AI in Education, Advanced AI Technology and Legal and Government Regulation, it contains a higher negative sentiment than other topics. This could be caused by multiple factors such as AI in Education being misused by the public and there are many concerns surrounding AI technology and legal restrictions. However, most articles generated by news websites are in neutral and positive stance. Only a few of the topics contain a negligible number of negative sentiments.

*Sentiment Score by Month in News Articles*

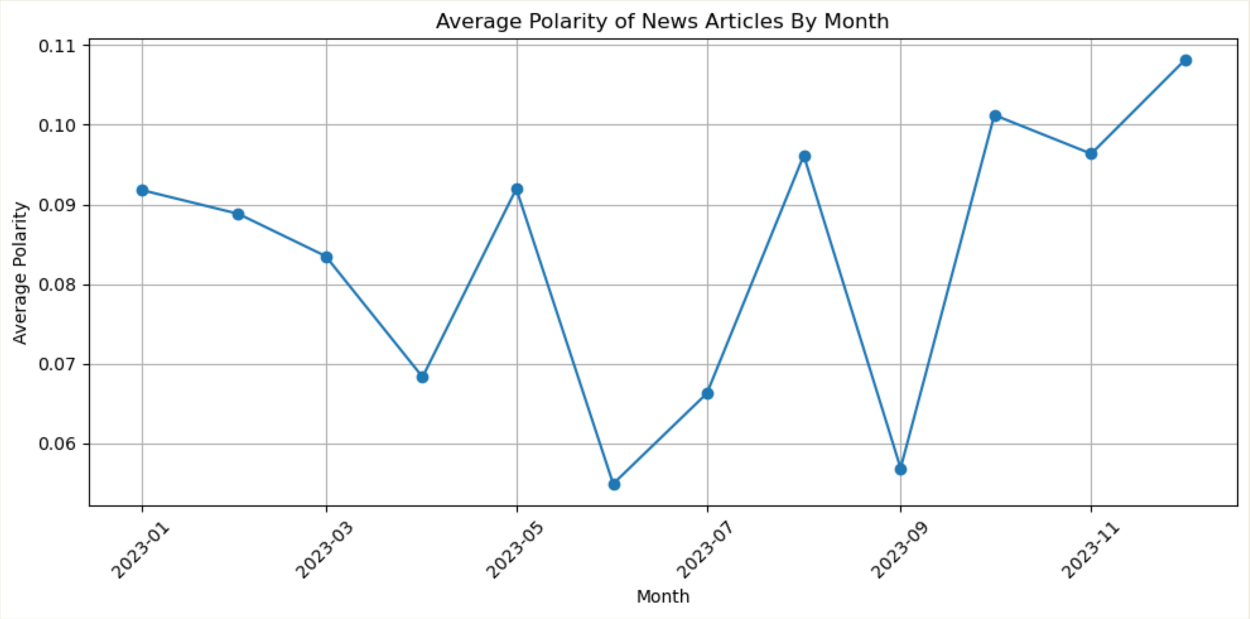

Diagram 4: Average Polarity of News Articles by Month

According to Diagram 4, it is shown that the average sentiment score of articles written by all 4 news websites generally shows a neutral sentiment score with polarity between −0.1 to 0.1. However, the polarity fluctuates throughout the whole year 2023, and it is showing an increasing trend during the last quarter. This shows that news articles are starting to write more positive opinions towards the topic AI Law and Regulation. Thus, indicating to us that news websites are capable of manipulating the articles published to increase or decrease the sentiment score towards any specific topic/industry.

### *Reddit*

*Sentiment Score by Subreddit*

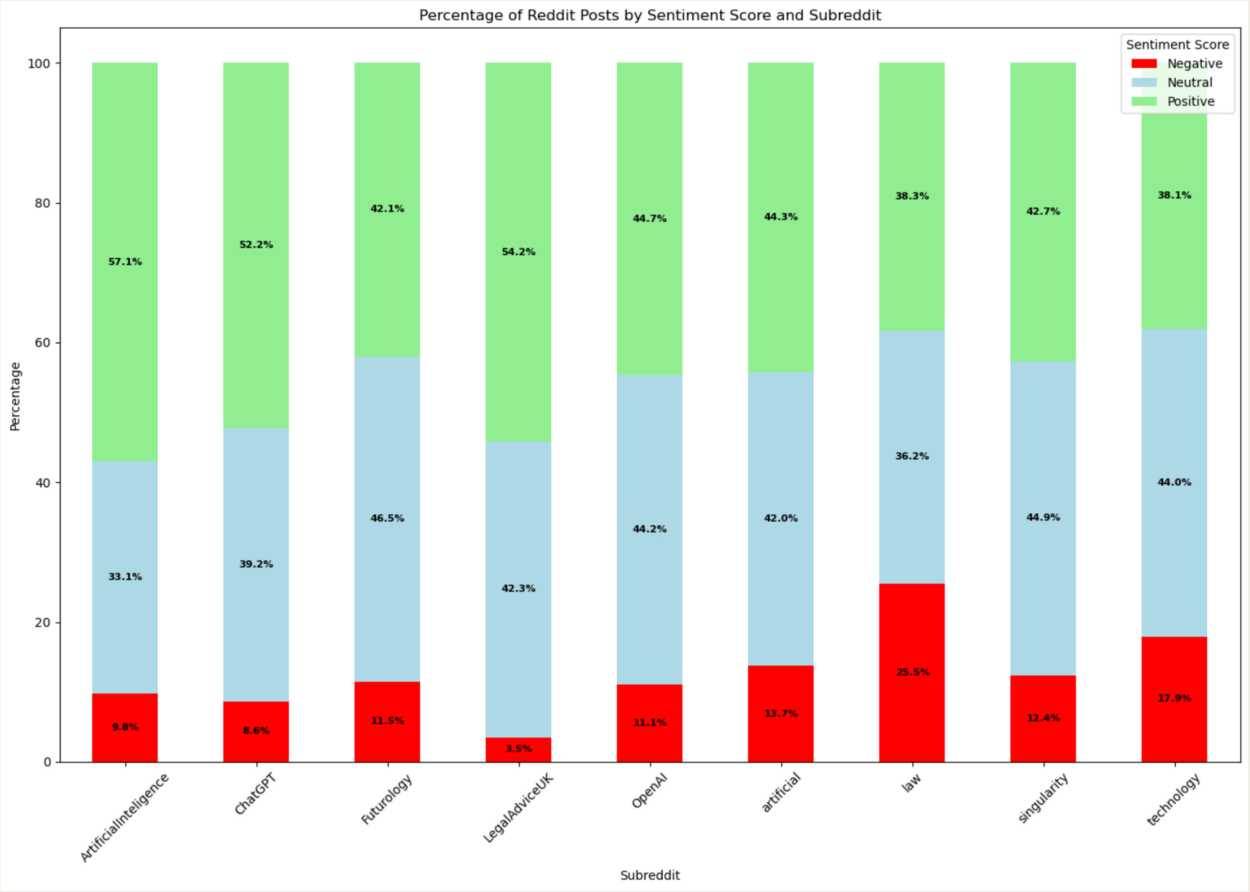

Diagram 5: Sentiment score by subreddit

Generally, we can see that there is more negative sentiment score in subreddit compared to news website in Diagram 5. When we are searching for keyword surrounding AI Law and Regulation, we can see that the subreddit 'law' have the highest negative sentiment score among other subreddit, followed by technology and artificial as the top 3 subreddit with the most negative sentiment score. In contradiction, the subreddit ArtificialInteligence, LegalAdviceUK and ChatGPT are the top 3 subreddit with most positive sentiment score. This could show that people comment more bad opinion towards AI Law & Regulation topic in law and technology related subreddit compared to other subreddit. As LegalAdviceUK mainly provides advice to European regarding legal support, they tend to receive better sentiment score and lesser negative comment. Furthermore, with ArtificialInteligence and ChatGPT subreddit is a fairly new technology that is widely accepted by the public for its convenience and efficiency, thus people might comment more of its benefits rather than leaving bad impression on it.

*Sentiment Score by Topic*

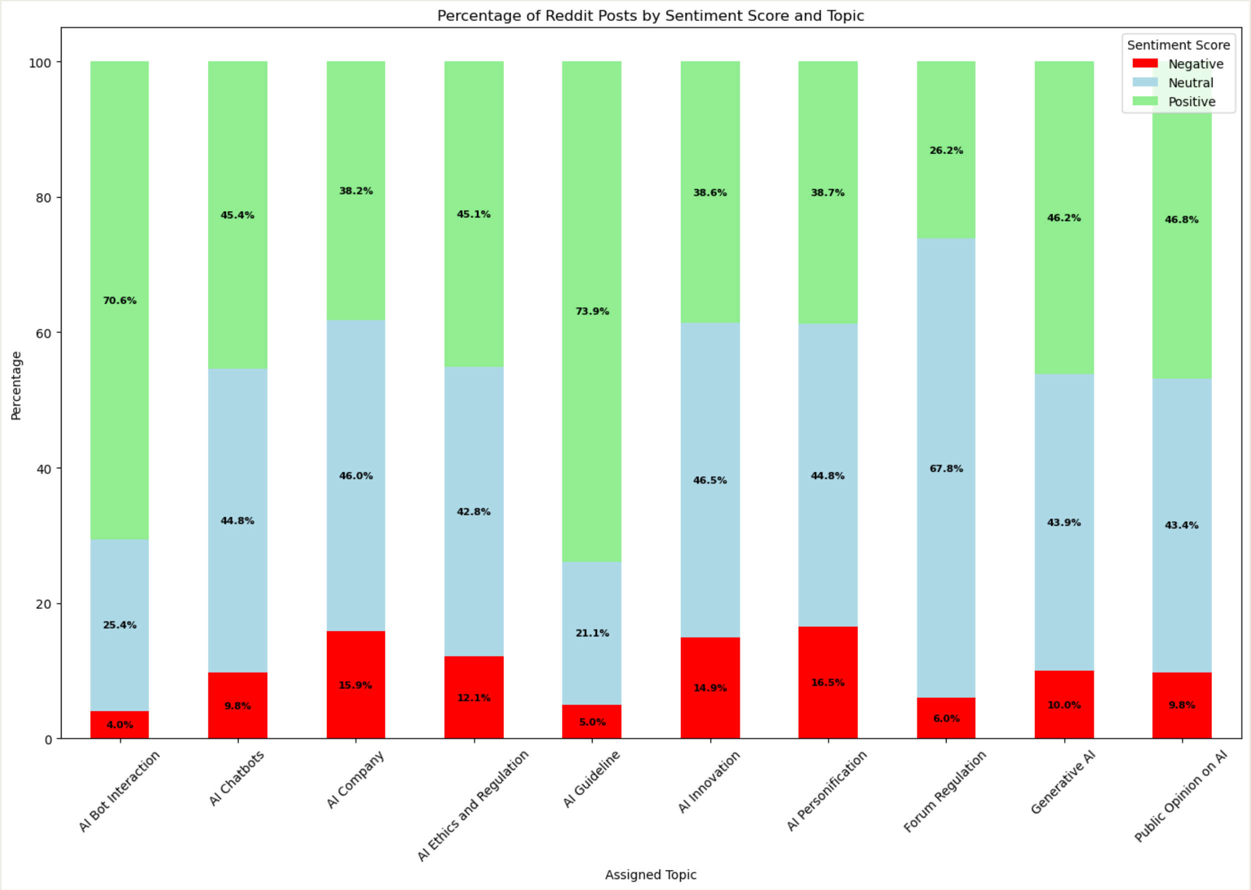

Diagram 6: Sentiment score by topic

Based on Diagram 6, we can see that topics regarding AI Bot Interaction and AI Guideline have a generally high positive sentiment score and there is a negligible amount of negative sentiment score compared to other topics. This shows that the public has a more positive sentiment towards AI Bot and the guidelines established in the industry. In contrary to this, topics that talk about AI Company, AI Innovation and AI Personification have a higher negative sentiment score. This evokes that the public are skeptical towards new AI technologies, and the companies that are developing it, especially when AI is personifying humans. At the same time, their positive sentiment scores are lower than the rest of the topics.

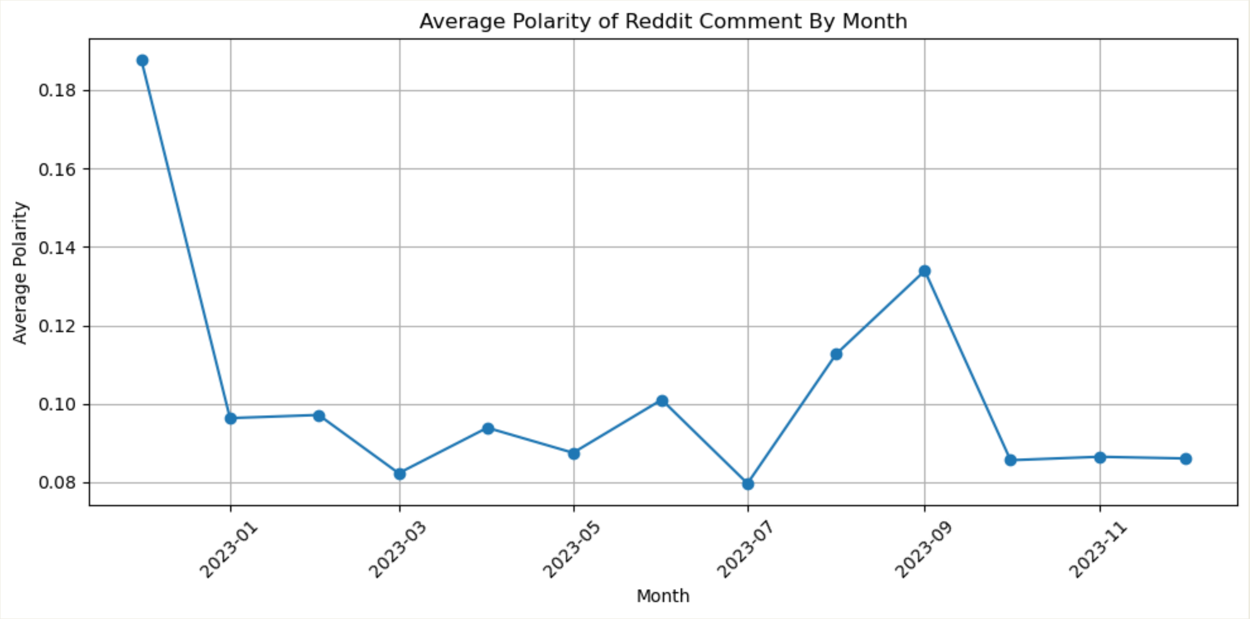

Diagram 7: Sentiment score by month in reddit.

From Diagram 7, we can see that there is a generally neutral sentiment score of articles throughout the entire year of 2023. There was sharp decrease in the sentiment score from September 2023, perhaps this was caused by the U.S. President Joe Biden signed an executive order on AI in October. This order brings up the risk from AI systems, the effects of automation in the workforce and invasion of civil rights and privacy (Henshall, 2023). Thus, proving the news article will affect the sentiment of the public.

### *Correlation Analysis*

*LIWC – News*

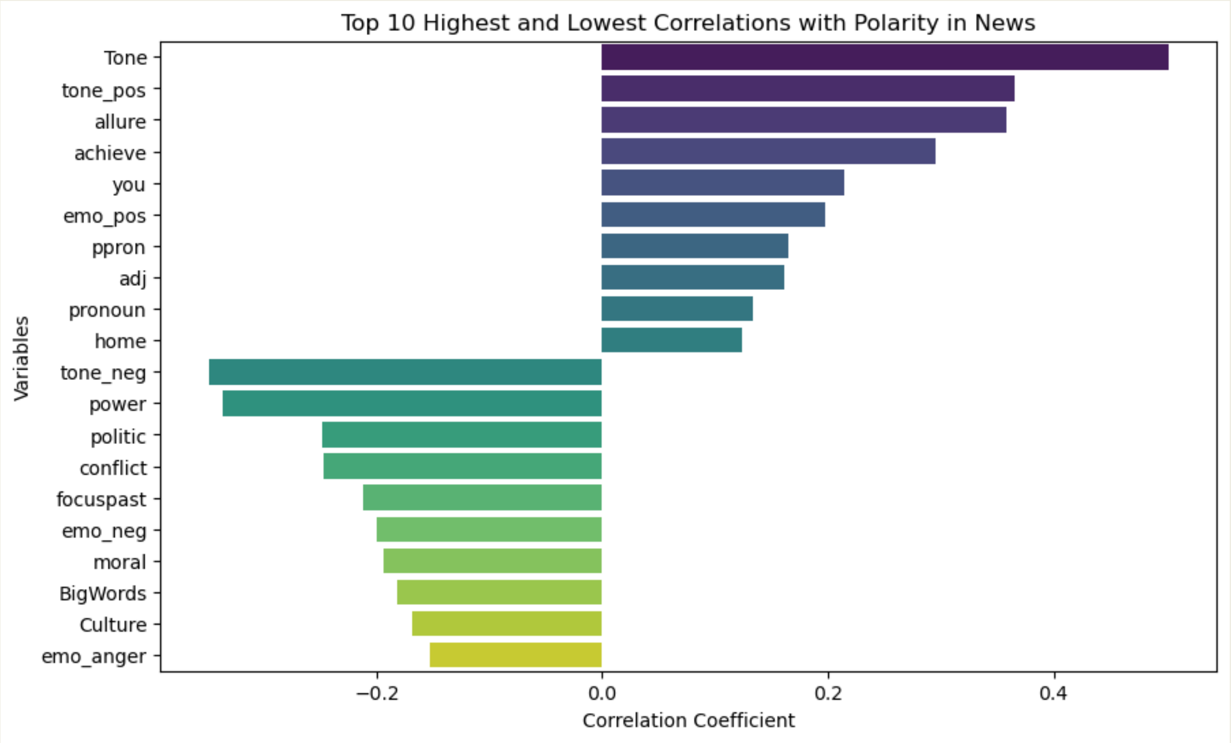

Diagram 8: Correlation of Polarity in News compared to LIWC dictionary.

Now that we have gotten the polarity in our news articles which indicates the sentiment score of each article, we further conducted a LIWC scan on our news dataset to understand how different dictionary views on the sentiment score of our dataset. Then, we performed a correlation analysis between the polarity of our articles and all dictionaries available in LIWC software. Through this analysis, we found the top 10 positive and negative correlated dictionaries against polarity. We can see that the tone, tone_pos and allure are positively correlated with the polarity which indicates a higher sentiment score will also increase the tone, positive tone and attractiveness (allure dictionary). In contrast, as the sentiment score decreases, the negative tone, power and politic element in LIWC dictionary shows an increment. This proves that when the articles are with a positive tone, it tends to get higher sentiment score but when the articles include power or politic terms, the sentiment score will be reduced.

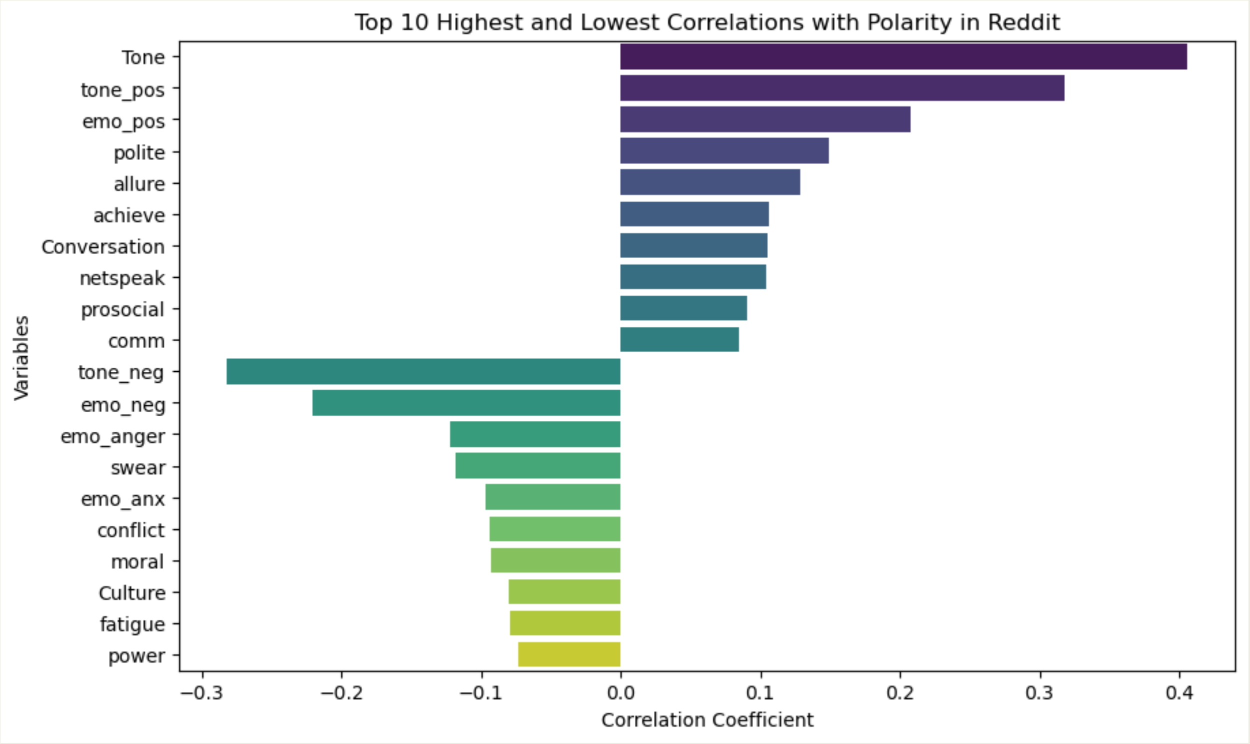

Diagram 9: Correlation of Polarity in Reddit compared to LIWC dictionary.

Similarly, we also executed a correlation analysis between the polarity of our reddit comments and the dictionaries provided by the LIWC software. In this analysis, we identified that the Tone, tone_pos, emo_pos are positively correlated with the polarity which indicates a higher sentiment score will also increase the tone, positive tone and positive emotion of the comment. On the contrary, when the sentiment score decreases, the negative tone, negative and anger emotions in LIWC dictionary increases. Hence, this suggests that when the articles are in a positive tone and emotion, it results in a higher sentiment score but when the comments detect negative and anger emotions, the sentiment score will be decreased.

*Media Platform*

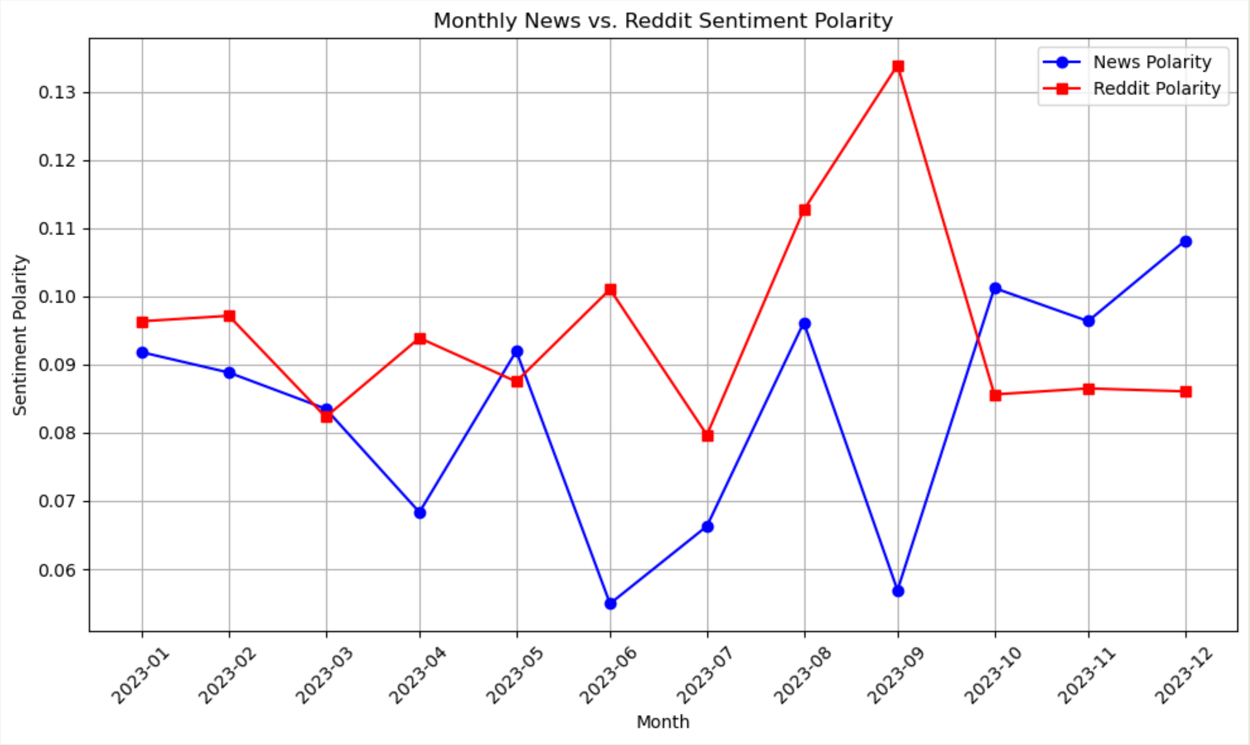

Diagram 8: Monthly News vs. Reddit Sentiment Polarity

Diagram 8 shows the polarity over time comparing between news and reddit. After we conducted a correlation analysis between news and reddit, we found that they exhibit a negative correlation of −0.42. This indicates when the news is publishing articles with positive sentiment score, the public will react negatively to Reddit. This indirectly shows that people nowadays often go against the news information when it comes to AI Law and Regulation topic. It is most likely due to the news website losing their credibility to provide accurate information and the public feels like they are being misled by the news articles (Vanderwicken 1995). Through this, we can predict that if the sentiment of news articles published is in a positive stance, we might receive a lower sentiment score in Reddit comment section.

# Recommendations

## Data Crawling

To conduct successful web scraping activities for AI laws and regulations, there are several strategies that we should prioritize. Firstly, it's essential to explore multiple sources to ensure that the dataset we collect is comprehensive and free of biases. Secondly, we should analyze the website structure to identify patterns and elements that contain relevant information. This would help optimize our scraping strategy to collect the most accurate data possible.

Thirdly, we need to incorporate user-agent and proxy rotation techniques to avoid detection and access blocked websites during the scraping process. Automating the scraping process using scripts is also a crucial practice to schedule regular runs and ensure consistent data collection while reducing manual errors. Prioritizing data cleaning and transformation is additionally important to develop robust scripts that prevent errors and inconsistencies.

Lastly, monitoring and regular maintenance of the scraping code are vital to adapt to any changes in website structure that could impact our scraping strategy. By implementing these strategies, we will be able to obtain high-quality data for our analysis of AI laws and regulations.

### Analysis

Based on the comprehensive topic modeling and sentiment analysis conducted above derived from news articles and reddit comments regarding AI Law and Regulation, there are several further analyses that company could possibly consider enhancing their analytics result.

In our report, we only highlighted the articles and reddit comments from the year 2023. A longitudinal study should be considered to track the changes in sentiment over time for a longer period to have a better insight on the changing trend of news websites and public opinion in reddit. This is due to AI is still rapidly evolving and there are numerous studies researching the potential advancement of AI (Dwivedi et.al, 2023). With a longer timeline, it allows us to capture more insight regarding the changes in sentiment score of news websites and the public.

Aside from this, we recommend performing a further comparative analysis of sentiment scores with industry trends and policy movements on AI Law and Regulation related topic to provide more context to our analysis. We can identify if there are new AI technological developments or AI related policy and orders being addressed like the one stated in our analysis previously. This could greatly affect the sentiment score and provide us with more insight into our selected topic.

## Conclusions

In summary, we can see that there are biased in different sentiment score from news websites and reddit platform. ZDNet, which primarily focuses on technology related articles, will generally show a more positive sentiment. In contrast, Reddit showed a broader range of sentiments score with a higher negative evaluation when compared to news articles. Through LIWC, we can confirm that there is relationship between the sentiment score with dictionary commonly used in the industry. On top of that, our study suggests that public perception on Reddit can be negatively correlated with the news articles published.

Based on our study, if there are any businesses who are interested to further extend this pilot study regarding AI Law and Regulation into an extensive analysis, they could further include more news websites and any new government policies introduced as it may affect the overall public sentiment score.

## References

1. Dwivedi, YK., Sharma, A., Rana, NP., Giannakis, M., Goel, P., Dutot, V (2023) ‘Evolution of artificial intelligence research in Technological Forecasting and Social Change: Research topics, trends, and future directions’, *Technological Forecasting and Social Change*, *vol. 192, doi:10.1016/j.techfore.2023.122579. (*Accessed: 2 May 2023)
2. Henshall, W (2023). The 3 Most Important AI Policy Milestones of 2023, *Times,* https://time.com/6513046/ai-policy-developments-2023/ (Accessed: 2 May 2023)
3. Vanderwicken, P (1995). Why the News Is Not the Truth, *Harvard Business Review,* https://hbr.org/1995/05/why-the-news-is-not-the-truth (Accessed: 2 May 2023)