Exploring Sentiment and topics in Media Texts:
A Case Study of 'Deepseek'

1 Introduction

A highly discussed topic in the media recently is "DeepSeek," and I am particularly interested in how different media outlets portray it. My research focuses on two key questions: first, what is the media's attitude toward DeepSeek? Second, what aspects of DeepSeek do media outlets emphasize (e.g., economic, technological, political, or social influences)?

For the first question, a subproblem involves conducting a sentiment analysis to determine "What words do media use?" "How do media describe DeepSeek?" and "Is the sentiment of an article positive or negative?" Tools like Python's TextBlob can assist with this analysis. For the second question, topic modeling methods can help identify themes, and we can track changes over time, comparing different attitudes and topics among various media outlets.

<table>
<tr><th>Research questions</th><th>Subquestions or hypothesis</th><th>method</th></tr>
<tr><td rowspan="3">Attitude of media toward deepseek?</td><td>Sentiment (positive or negative)of the outlets</td><td rowspan="5">Sentiment analysis (python TextBlob)
word frequency analysis
Topic modeling(LDA)
AntConc
Voyant
AI tool</td></tr>
<tr><td>Is there difference between medias?</td></tr>
<tr><td>Tracking sentiment changes over time</td></tr>
<tr><td rowspan="2">What aspects do people care about deepseek?</td><td>Identifying topics</td></tr>
<tr><td>Tracking topics changes over time</td></tr>
</table>

2 Material and Preprocess

2.1 Corpus setup

The selection of media sources plays a crucial role in determining the representativeness of the data. For this project, I focused on analyzing articles from three prominent news outlets: BBC, CNN, and The New York Times. These publications were chosen for their global reach, diversity in editorial stance, and prominence in the media landscape. While the articles

analyzed span from January 27 to February 14, the main criterion for inclusion was that the term "DeepSeek" appeared in the title. This resulted in a total of 40 articles.

Although this corpus may not be large enough to make broad generalizations, it serves as a valuable starting point for exploring how different media outlets cover "DeepSeek." The choice of focusing on articles with "DeepSeek" in the title also ensures that the analysis is centered on media directly discussing this topic, which provides a clearer focus for sentiment and topic modeling analysis. Furthermore, these articles represent a snapshot of how the media portrays "DeepSeek" during a specific time frame, allowing us to explore the patterns and trends emerging from these discussions.

2.2 Preprocessing

Preprocessing is a vital step in text mining, as it prepares raw data for meaningful analysis. We extract the most relevant features for further processing, such as sentiment analysis or topic modeling. It helps eliminate irrelevant noise and standardizes the text for better machine learning performance.

For this project, I used the spaCy library, a powerful natural language processing tool, to perform the preprocessing steps. First, I imported the spaCy library along with its English language model to ensure accurate text processing. Key steps in the preprocessing included:

1) Tokenization: Breaking down the text into individual words or tokens, allowing for the analysis of each word's frequency and sentiment.
2) Removing stop words: Common words like "the," "is," and "in" provide little value in text analysis, so they were removed to focus on more meaningful content.
3) Lemmatization: Converting words to their base or dictionary form (e.g., "running" to "run"), which ensures that variations of a word are treated as the same.
4) Lowercasing: Converting all text to lowercase to avoid duplicating words with different case (e.g., "DeepSeek" and "deepseek" would be treated as separate tokens without this step).
5) Removing punctuation and special characters: Non-alphabetic characters like commas, periods, and other punctuation marks were removed to focus on the words themselves.

By performing these steps, the text was refined, making it more suitable for subsequent analysis, such as sentiment analysis, word frequency analysis, and topic modeling.

3 Analysis and Visualization

3.1. Sentiment Analysis combined with Word Frequency

3.1.1 Creating a word frequency table

I want to analyze word frequency and have already pre-processed the texts, using lemmatized text directly. In my practical process, I begin by examining word frequencies with tools like Voyant and AntConc. When analyzing the data in the Voyant Tool, I noticed that some words were highlighted in green and red. （https://voyant-tools.org/?corpus=299547e5da465ecabbd4153ecaf5cb9e&view=DocumentTerms） I discovered that these colors indicate sentiment—green representing positive sentiment and red indicating negative sentiment. This discovery inspired me to explore the positive and negative words used by the media to describe "Deepseek." Additionally, the paper "History Text" led me to consider computing a sentiment score for each word. This would allow me to visualize both sentiment and frequency simultaneously, providing a clearer understanding of how sentiment and word frequency are connected.

3.1.2 Creating words table with sentiment score

Here, we use TextBlob: This is a Python library for Natural Language Processing (NLP), and it's specifically used for tasks like sentiment analysis, part-of-speech tagging, noun phrase extraction, translation, etc. For sentiment analysis, TextBlob assigns polarity (positive or negative) and subjectivity (how subjective or objective the text is) to the text. We apply sentiment analysis to each word, polarity score ranges from -1 to 1, and remove newtral words(score=0). And we can visualize to see the result clearly.

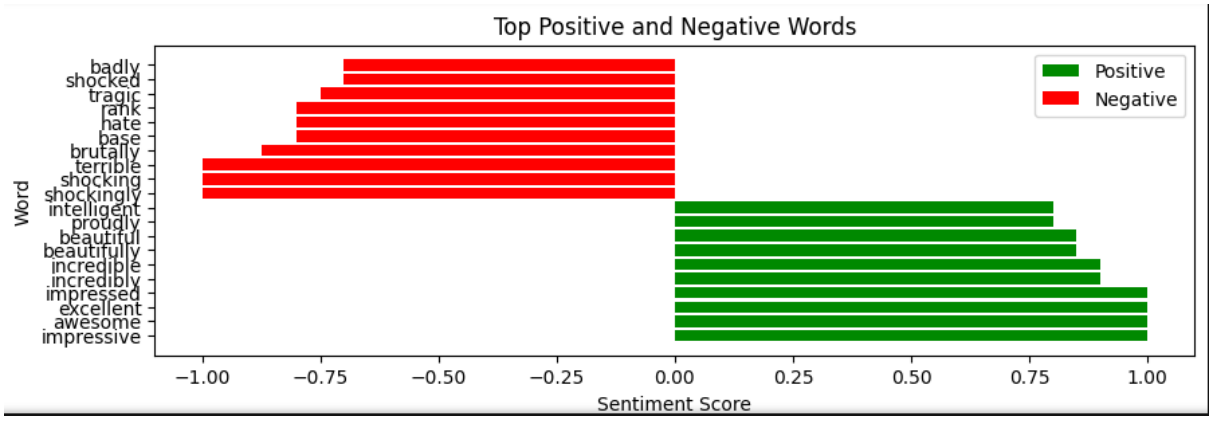


Figure 1. “Top Positive and Negative Words” generated by TextBlob

In evaluating the results of the sentiment analysis, we must consider that the machine automatically filters and assigns scores to a set of words. However, when reviewing the word frequency table, it becomes clear that some important words may be missing. For instance, words like "innovation," "create," and "ban" were not included in the original analysis. This can result in valuable terms being overlooked or misrepresented.

A specific example of this issue can be seen with the words “shocking” and “shockingly.” The machine assigned a sentiment score of -1.0 to both words, which indicates an extremely negative sentiment. However, in the context of these articles, it is not so clear that these words are negative. In some cases, such as with the term “shocking,” the word may express a highly positive emotion, especially when used to describe a breakthrough or innovation. In the Chinese context, for example, "shocking" may convey excitement and admiration for a significant development, which would make it a positive sentiment rather than a negative one.

This example demonstrates that we cannot fully rely on the automated tool’s sentiment analysis. As an interpretation, I believe that "shocking" reflects a breakthrough, so I manually assigned it a positive sentiment score of +1.0. Additionally, I combined the frequencies of both “shocking” and “shockingly” to reflect the full context in which these terms appear. Similarly, I included other important words such as “innovation” (which appeared 47 times) and assigned it a positive sentiment score of +1.0, as innovation is typically viewed favorably in the context of technological advancement.

Other words with significant frequency, such as “ban” (frequency 40), were assigned a negative sentiment score of -0.8, as bans often suggest restrictions or limitations. Words like “startup” (47 occurrences) were given a positive sentiment score of +0.3, “lead” (77 occurrences) was assigned +0.2, and “create” (42 occurrences) was given a score of +0.7, reflecting their generally positive connotations related to progress and growth.

These adjustments were made manually based on the context and frequency of the words, and the modified sentiment table was saved.

3.1.3 Words selection - machine selection and manual modification

To reduce manual effort, I decided to filter out words with low frequency and low sentiment scores using Python. After visualizing the entire dataset, I noticed that words with a frequency above 20 had a significant impact on the analysis, while words with a frequency below 20 seemed less influential. Therefore, I set a threshold of 20 for frequency to retain the

most impactful words. But then I realized some words with low frequency but high sentiment scores, so it became a challenge to decide which words to visualize.

To refine the data further, I manually adjusted the sentiment scores of certain words. Specifically, I modified the score of some adjectives from 0.5 to 0.6 to increase their weight in the analysis. I also adjusted the parameters to reflect this higher score threshold (0.6). Additionally, I combined words with similar meanings or common roots (e.g., "innovative" and "innovation," "develop" and "development," "success" and "successful") to further reduce redundancy. While this step required considerable time and effort, the goal was to highlight the most valuable data, though I acknowledge that the process was influenced by my own subjective interpretation of what is valuable.

In this process, I also added words with higher frequencies and more relevant meanings, such as "breakthrough" (frequency 30), "power" (frequency 79), "raise" (frequency 27), and "cheap" (frequency 28), to the positive sentiment word list.

At the same time, I removed some words that had high sentiment scores but were irrelevant or meaningless in the context of my analysis. For example, "honest," which only appeared once with a score of 0.6, was deleted as it didn't contribute to the topic at hand. I also deleted the word "welcome," which had a score of 0.8 but was not useful for the sentiment analysis. Once all modifications were completed, I saved the updated dataset to ensure that future runs of the code would not overwrite the results.

3.1.4 Visualization Results

To begin, we consider both frequency and sentiment score as the two key factors. Our goal is to display the words in a two-dimensional space, with frequency and sentiment score as the axes. Initially, we focus on words that have a frequency greater than 20 and a sentiment score higher than 0.6.

Next, we manually review the results and remove any meaningless words. During this process, we notice that many words have the same sentiment score, which causes them to cluster together and become difficult to distinguish. To address this, we make slight adjustments to the sentiment scores manually.

Inspired by the figure in the Lucy et al. paper(2020) , I aim to use a similar visualization. The authors employ a log-odds-ratio plot to compare how frequently certain words are used in reference to Black people or women compared to their counterparts (p.11).

In my research, I aim to find words related to "Deepseek" and explore the sentiment or topics surrounding it, considering both frequency and sentiment polarity. By plotting words in a two-dimensional space, I hope to uncover meaningful insights.

To ensure better visualization (as some words may overlap and become invisible), I will reduce the number of words displayed and make minor adjustments to the sentiment scores, ensuring that the points are spread out and not concentrated in one area.

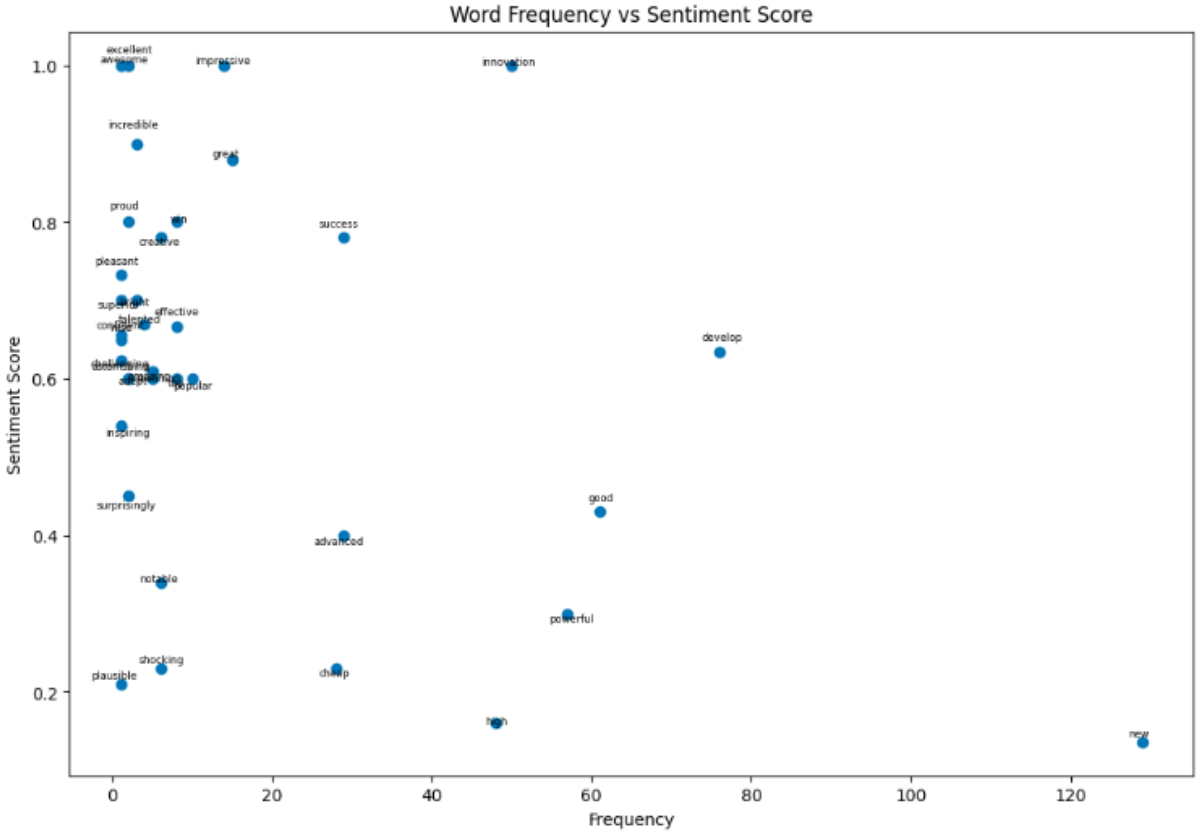


Figure 2 Positive Words talking about “deepseek”

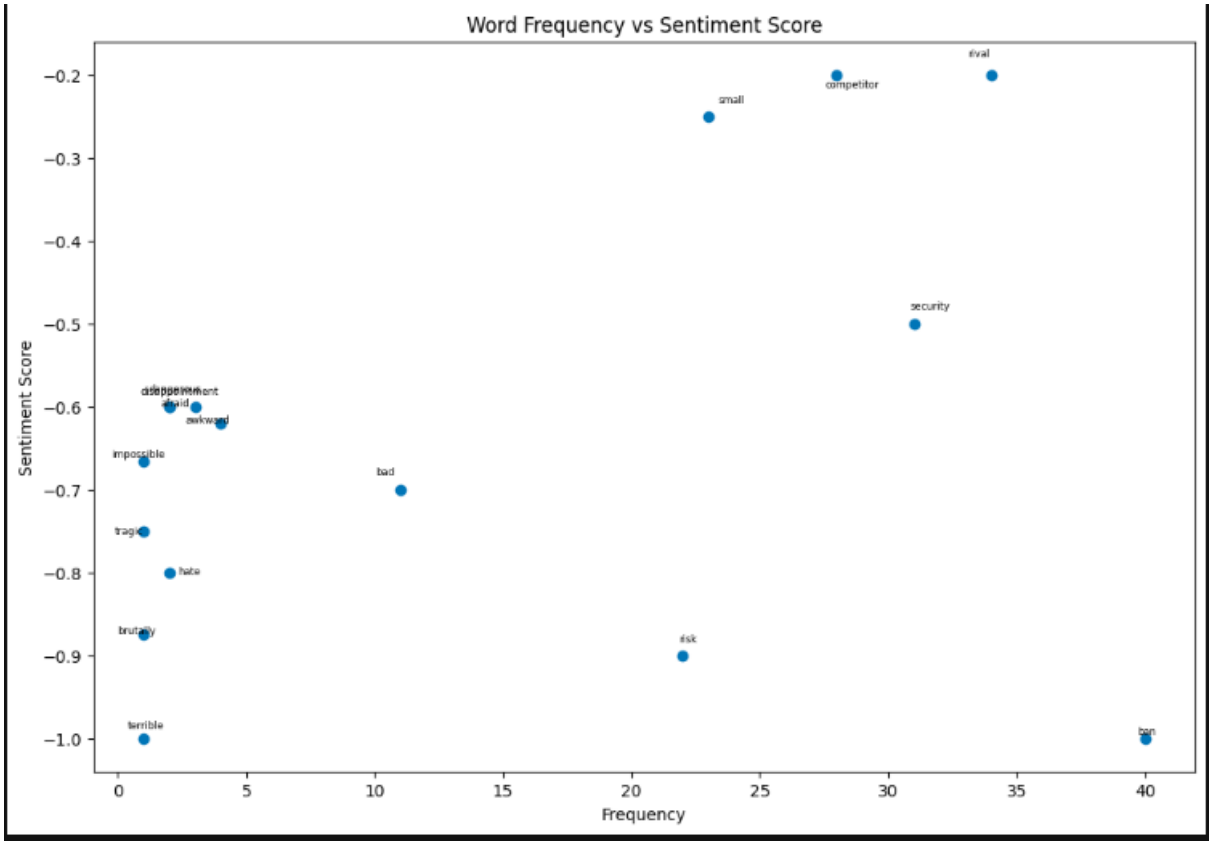


Figure 3 Negative words talking about “deepseek”

We try different visualization, since some words have low frequency but still are meaningful, with high sentiment meaning.see figure below:

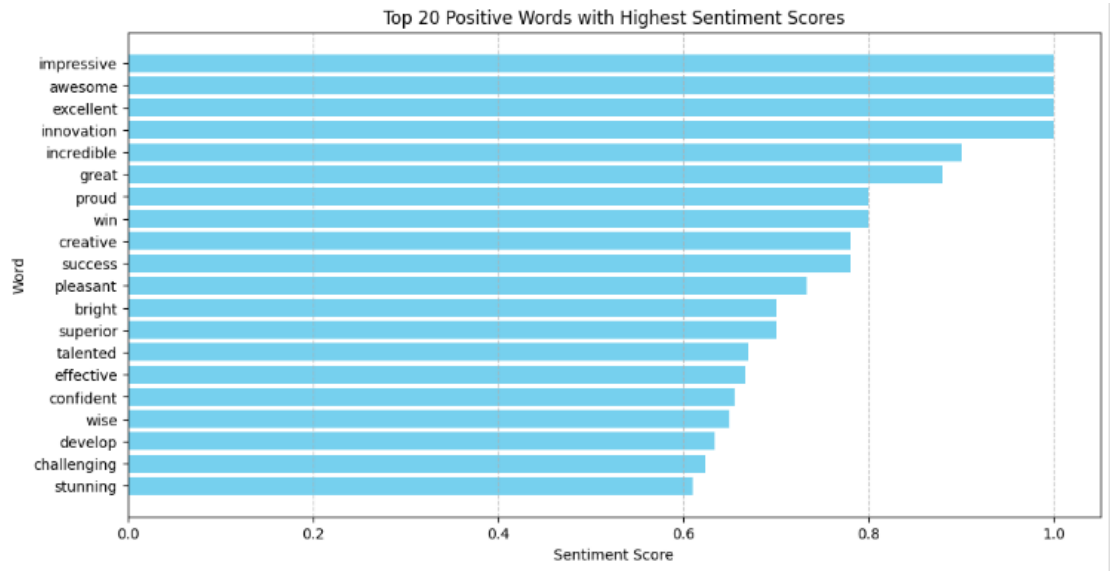


Figure 4 Top 20 positive words about “deepseek”

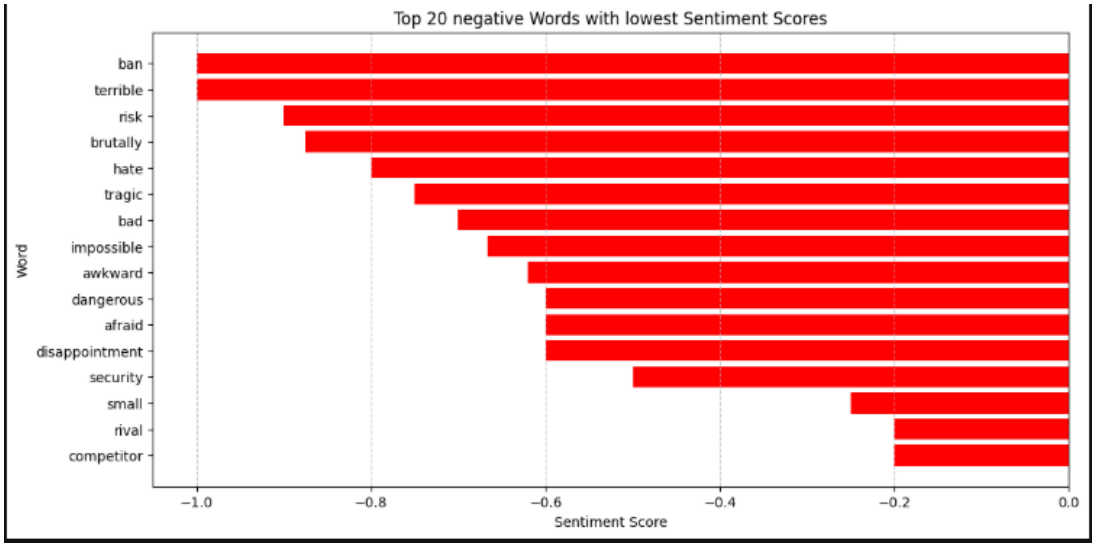


Figure 5 Top 20 negative words about “deepseek”

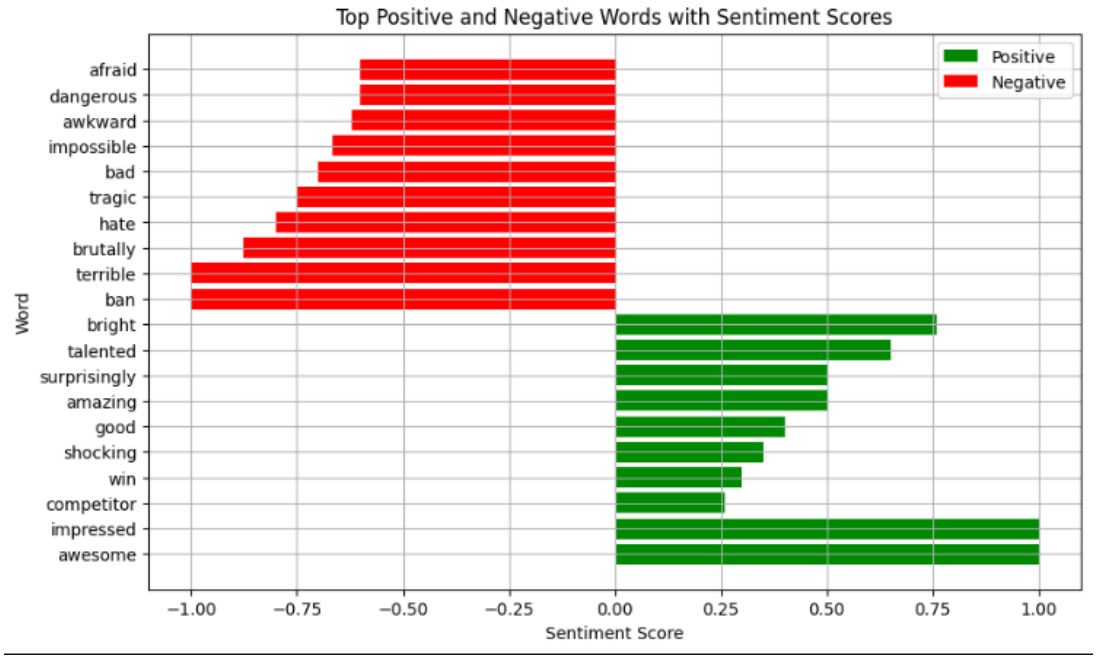


Figure 6 Top Positive and Negative Words about “deepseek”

And at last ,we can compare the figure6 with figure1 to see if the words we are showing changed.

When analyzing the top negative words in the text, combined with word frequency, it becomes evident that negative aspects revolve around political and security-related concerns. Words like “security,” “risk,” and “ban” frequently appear, signaling a focus on perceived threats to national interests. These terms highlight the tension around issues such as national security, government regulations, and international conflicts. For example, "risk" often carries negative connotations of potential dangers, particularly in the context of geopolitical

discussions, while words like “ban” and “security” invoke a sense of limitation or precaution, further emphasizing negative implications.

On the other hand, the positive sentiment in the text is strongly associated with terms that represent innovation and technological progress. Words such as “innovation,” “creative,” and “breakthrough” dominate the positive discourse, underlining a sense of optimism about future advancements and novel solutions. These words suggest that, within the media coverage, there is a celebration of new developments, particularly in the fields of science, technology, and industry. This distinction between the political/competitive fears (negative sentiment) and the focus on progress and creativity (positive sentiment) reflects the contrasting ways in which technological and political narratives are framed.

Moreover, when these sentiment-based word categories are analyzed alongside word frequency, it becomes clear that the media's portrayal of topics like national security and international competition often utilizes emotionally charged language. The frequent use of negative terms is likely intended to evoke concern or urgency in the audience, while the positive terms related to innovation seem to be used to inspire hope and excitement about technological advancement. Together, these word choices provide insight into how the media uses language to shape public perception, framing issues as either threats or opportunities, depending on the sentiment they wish to convey.

3.2 Perform sentiment analysis on the corpus scale

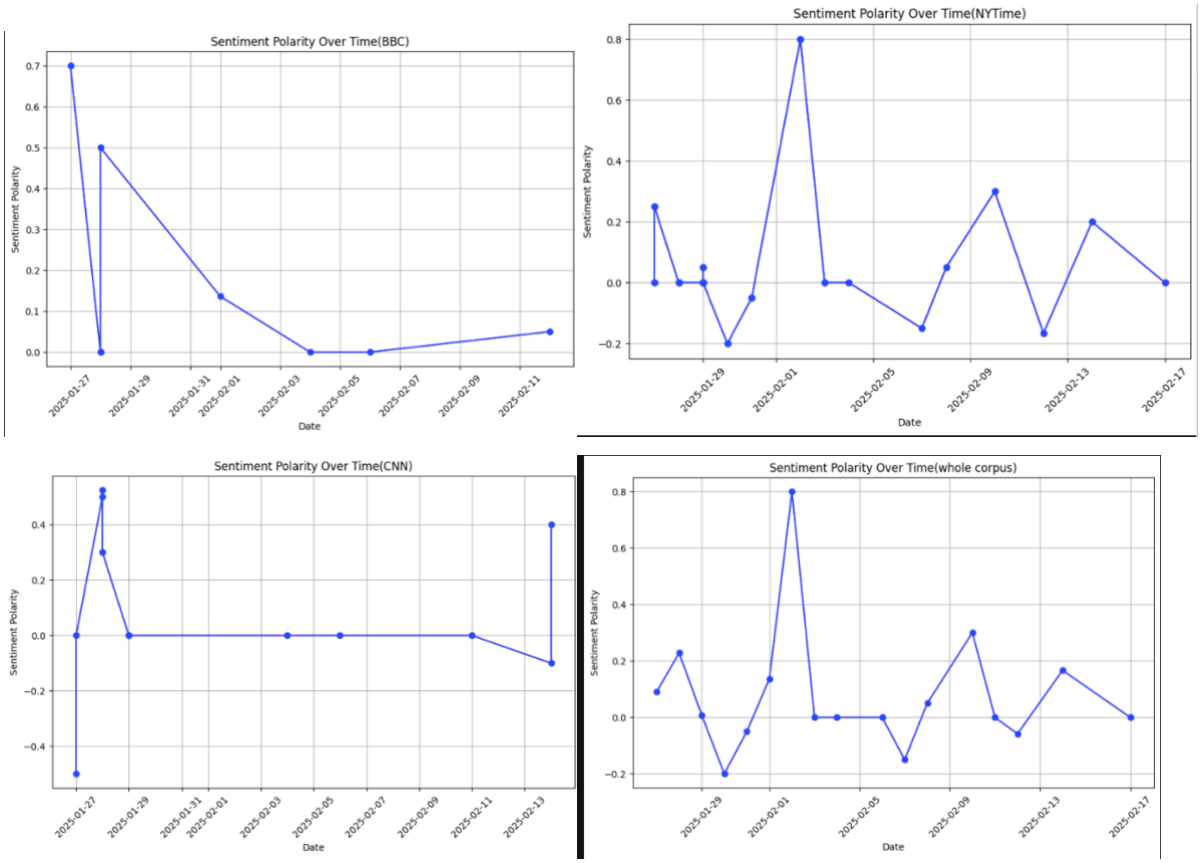
In this part, we also utilize spaCyTextBlob for sentiment analysis. Our main objectives are: 1) to determine the sentiment of each article, 2) to compare the sentiment across three different media outlets to identify any potential gaps, and 3) to examine whether there is any noticeable trend in sentiment over time. If such a trend exists, we hope to uncover shifts in the media's attitude towards "Deepseek." While we anticipate observing some trends, we acknowledge that the corpus is relatively small and the time span is limited, which means any identified changes may not be particularly pronounced.

bbc_sentiment_analysis

article	date	polarity	subjectivity
BBC News: Is China's AI tool DeepSeek as good as it seems?	2025-01-27	0.7	0.6
BBC News: DeepSeek: How China's 'AI heroes' overcame US curbs to stun Silicon Valley	2025-01-28	0.0	0.0
BBC News: China's DeepSeek AI shakes industry and dents America's swagger	2025-01-28	0.0	0.0
BBC News: UK will not be able to resist China's tech dominance	2025-01-28	0.5	0.6
BBC News: DeepSeek: What lies under the bonnet of the new AI chatbot?	2025-02-01	0.14	0.45
BBC News: DeepSeek: The Chinese AI app that has the world talking	2025-02-04	0.0	0.0
BBC News: World leaders set to vie for AI domination at Paris summit has shaken up the world of AI	2025-02-06	0.0	0.0
BBC News: 'DeepSeek moved me to tears': How young Chinese find therapy in AI	2025-02-12	0.05	0.2

Figure 7: Sentiment Analysis on BBC corpus

Figure 8: Sentiment Polarity Over Time(below)



As shown in the figures, the highest polarity values occur in late January and early February. We have listed these articles below. After performing sentiment analysis, we can filter the articles based on their sentiment polarity using the provided code. Specifically, we apply a filter to retain only those articles with a polarity greater than 0.4.

filtered_0.4_sentiment_analysis

article	date	polarity	subjectivity
CNN:DeepSeek just blew up the AI industrys narrative that it needs more money and power	2025-01-28	0.5	0.5
CNN:DeepSeek chaos suggests America First may not always win	2025-01-28	0.53	0.37
NYTime: DeepSeek Is a Win for China in the A.I. Race. Will the Party Stifle It?	2025-02-02	0.8	0.4
BBC News: Is China's AI tool DeepSeek as good as it seems?	2025-01-27	0.7	0.6
BBC News: UK will not be able to resist China's tech dominance	2025-01-28	0.5	0.63

Figure9: articles which sentiment polarity greater than 0.4

Regarding the negative sentiment in the articles, we observe that the reports from BBC and CNN exhibited significant fluctuations at the end of January but gradually stabilized, eventually hovering around a polarity of 0. In contrast, the New York Times' reports consistently fluctuated between positive and negative values. This aligns with the New York Times' characteristics as a more diverse news outlet, not limited to straightforward reporting like the BBC. It features multiple sections, including technology, finance, and opinion columns, where articles tend to have stronger subjective tones and incorporate more expert opinions.

In terms of the topics covered in these articles with relative sentiment polarity, we aim to identify what these articles discuss. Additionally, we focus on finding articles with negative sentiment polarity to explore their content further. To do this, we apply a filter to retain only those articles with a polarity less than -0.1 and briefly review their titles first.

To conduct the analysis, I used tools like AntConc and Python coding methods, but finding meaningful insights proved challenging. To deepen the analysis, I performed word frequency analysis on both high-sentiment and low-sentiment articles, as well as on the full corpus. The tools I used for this include:

TF-IDF (Term Frequency-Inverse Document Frequency): This technique helps identify significant words that are unique to certain articles.

Word Cloud: To visualize the most frequent words in different sentiment groups.

Topic Modeling (e.g., Latent Dirichlet Allocation, LDA): This method uncovers the main topics in subcorpus articles, helping us see if high-sentiment articles focus on specific topics compared to low-sentiment ones.

filtered_-0.1_sentiment_analysis

article	date	polarity	subjectivity
CNN:A shocking Chinese AI advancement called DeepSeek is sending US stocks plunging	2025-01-27	-0.5	0.5
CNN:The real reason behind the DeepSeek hype, according to AI experts	2025-02-14	-0.1	0.5
NYTime: Is Artificial Intelligence Really Worth the Hype?	2025-02-07	-0.15	0.55
NYTime: How Did DeepSeek Build Its A.I. With Less Money?	2025-02-12	-0.17	0.07

Figure 10: articles which sentiment polarity less than -0.1

3.3 topic modeling on extreme sentiment articles

I tried combining high and low sentiment articles (polarity > 0.4 and polarity < -0.1) to find common themes in extreme sentiment, and to compare their topics in order to understand why certain articles have more positive or negative sentiment. By analyzing only high and low sentiment articles (instead of all articles), we aim to uncover the factors that drive sentiment. For example, building on the previous analysis of word-level sentiment, we explore whether positive articles focus on innovation and success, while negative ones emphasize controversies or risks.

By analyzing only extreme sentiment articles and removing neutral ones, we improve topic modeling accuracy. Focusing on extreme sentiment helps to generate clearer topic clusters. If we were to analyze all articles, many of them might be neutral, making it harder to identify what makes certain articles more emotionally charged. Using LDA (Latent Dirichlet Allocation) on all articles might result in neutral articles blending topics, leading to less distinct results. Filtering ensures that LDA identifies stronger themes in positive and negative content.

The results of the topic modeling are as follows:

Topic 1: Hype, experts, real reason, resist, UK

Topic 2: Win, America, chaos, build

Topic 3: Money, shocking, stocks, plunging, advancement, US, blew

Now, let's compare the topics generated from the entire corpus using Gensim (see topic_modeling_genism.ipynb). When setting the number of topics to 5, and each topic containing 10 words, the outcome was:

Topic 1: Model, Chinese, chip, market, government, power, OpenAI, world, data, billion

Topic 2: Deepseek, say, China, new, weekly, spend, big, start, ban

Topic 3: Company, year, cost, app, ChatGPT, Meta, race, powerful, research, good

Topic 4: Technology, build, firm, industry, giant, intelligence, need, people, source, United

Topic 5: AI, tech, like, lead, stock, American, question, call, chatbot, high

We also verifying the topics by using Voyant tool:

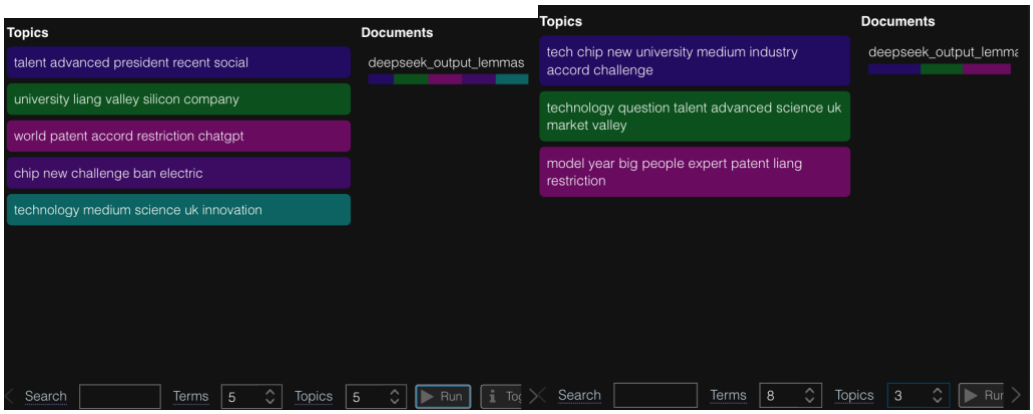


Figure: Topics by Voyant Tool(with different parameters)

When comparing this outcome with the previous one, we can see that the topics generated from the full corpus focus more on main themes like technology, AI, and industry. In contrast, the extreme sentiment analysis appears to highlight more immediate or emotional themes, such as "hype" and "chaos." This difference may point to the varying focus of articles depending on their sentiment: neutral or positive articles seem to concentrate on technological advancements and corporate success, while negative articles tend to focus more on risks, controversies, and financial instability.

4. Discussion and Conclusion

4.1 Evaluating Results

As Nina Tahmasebi & Simon Hengchen(2020) argued, "All data-intensive projects that aim to answer broad research questions, like those in the humanities, should make their model of interpretation clear and preferably evaluate it with respect to alternative models" (p.218). Transparency and evaluation of results are crucial aspects of data-intensive projects, which the authors refer to as the "model of interpretation." Evaluation should include alternative methods, different corpora, varying parameters of the method, and the use of different tools. When analyzing the results, particularly from sentiment analysis and topic modeling, one key consideration is the validity and representativeness of the outcomes. To evaluate the results thoroughly, several strategies were employed.

For topic modeling, I experimented with different parameters within the same method, such as adjusting the number of passes (e.g., passes=20). I also tested alternative methods to

improve the robustness of the findings. To validate the topic modeling results, I used Gensim to assess the coherence of the topics.

However, topic modeling is constrained by word frequency and computational limitations, as infrequent words contribute little to the overall topic distribution, while highly frequent words tend to dominate the clustering process. This can result in rare but meaningful words being excluded, leading to potential bias and incomplete topic representation (Shadrova, 2021,p.7-8). In my project, when analyzing sentiment at the word level, I found many words with extremely high sentiment polarity (either strongly positive or negative) but low frequency, often appearing just once in the corpus. While these words share similar meanings, their rarity means they contribute little to the overall topic distribution. However, removing them would reduce the diversity of language expression in the analysis. Therefore, I chose to retain and highlight these words in the visualizations to emphasize the richness of the language used in the corpus, even though they appear only once. This was a challenging decision-making process that balances the trade-off between frequency and diversity.

For sentiment analysis, if time allowed, I would have manually sampled and reviewed a few sentences to check their assigned sentiment scores. In practice, I did check some sentences but was sometimes confused by the scores assigned by the tool. However, I did not have a clear method for assigning sentiment scores to these sentences manually.

To improve transparency, I should show the original table generated by the sentiment analysis tool alongside the places where I made manual adjustments. This would make the process more transparent. Unfortunately, due to time constraints and the many tasks still remaining to improve the project, I was unable to complete this level of detail.

4.2 Explore Possibilities

One potential avenue for expanding this research is to broaden the corpus by incorporating more diverse sources and extending the research timeline. By creating a larger, more comprehensive corpus focused on the broader topic of "AI," we could compare public attitudes toward AI in general and "Deepseek" specifically. This would offer clearer insights into the factors driving public interest in "Deepseek" and allow for a more nuanced understanding of how it fits within the broader discourse on AI.

Close reading is also essential for validating the results. As Kettunen and La Mela(2022) highlight, close reading serves as a necessary step to verify the accuracy of the analysis and address issues related to classification.

4.3 A Reflection on the Methodology

The research process demonstrates the interplay between human interpretation and tool results. As noted by Tahmasebi and ChenHeng (2020), there are two main fashions in the data driven research, exporatory fashion and validating fashion.(p.216) In practical experiments, however, these paths may blend, as our thinking is often nonlinear. Defining these paths as separate models helps reflect on the research process, just as defining the relationship between digital tools and human interpretation is critical for using computational methods effectively. We are not guided by the tool or the technique alone, but by our own research questions, purposes, and reasoning. In line with Ted Underwood(2017), this experimental methodology, rooted in the social sciences, should take precedence over digital computational techniques in scholarly work.

Therefore, there is a slide between subjectivity—represented by the research questions, hypothesis, and interpretation—and objectivity—represented by the data, tools, and results. In my project, the exploratory and validating approaches interacted throughout the process. Initially, I applied topic modeling to identify overarching themes (e.g., economic, technical, political, or societal focus) within the articles. However, the results were unremarkable. I found Voyant to be the most user-friendly tool, so I started with it. But when I repeated the analysis, the results were similar and difficult to interpret. I realized that the corpus was too small for effective topic modeling. Consequently, I shifted to sentiment analysis. After analyzing the entire corpus as well as separate media outlets (CNN, BBC, The New York Time), the results still proved challenging to interpret, and I struggled to answer the primary question I had posed at the outset.

Then, when I revisited Voyant, the results sparked a new direction for exploring sentiment at the word level. By combining word frequency and sentiment, I was able to examine trends in attitudes or potential topics. This led to the discovery of interesting patterns. In this exploratory phase, the new sub-questions emerged directly from the results of a preliminary experiment. Through sentiment analysis at the word level, I revisited the results from the corpus-level sentiment analysis and made new discoveries. This process also led me to focus on articles with extreme sentiment polarity, which produced clearer and more interpretable topic models. Ultimately, this allowed me to compare the results more effectively.

On the other hand, this process also shows that computational methods are inherently embedded in human subjectivity and interpretation. A key example in my project is the sentiment analysis at the word level. For instance, the tool assigns a negative sentiment score of -1 to "shocking," and "competitor" is also classified negatively. However, I find these words worth discussing. Is "shocking" inherently negative? How do people perceive a "competitor"? In my interpretation, a new app making the world (or market/industry) "shocking" could be seen as a positive success. Therefore, it’s not always appropriate to classify "shocking" as negative in this context. Context is crucial in determining whether a word expresses negative or positive emotion. Even when I manually modify the sentiment score, determining the right value is difficult. This is especially challenging in a small-scale project like mine, where nuanced judgment plays a critical role. This is akin to the example in Guldi (2023,p.240), where the term "base" could refer to either a baseball base or a military base, demonstrating the complexity of word interpretation based on context.

4.4 Conclusion

This research delved into two key inquiries regarding media coverage of "DeepSeek": the media's attitude towards it and the aspects emphasized. Through a series of analyses, several significant findings emerged.

In terms of the media's attitude, sentiment analysis, in combination with word frequency analysis, offered valuable insights. Positive sentiment was strongly associated with innovation and technological progress, as words like "innovation," "creative," and "breakthrough" were frequently used. Conversely, negative sentiment revolved around political and security - related concerns, with terms such as "security," "risk," and "ban" being prominent. When examining sentiment at the article level across different media outlets (BBC, CNN, and The New York Times), it was found that the sentiment varied. BBC and CNN's sentiment showed stable trend after initial fluctuation, while The New York Times' reports had more consistent fluctuations between positive and negative values, which can be attributed to its diverse content structure.

Regarding the aspects of DeepSeek that media outlets emphasized, topic modeling was a crucial method. Analyzing extreme sentiment articles (polarity > 0.4 and polarity < -0.1)

revealed topics like "Hype, experts, real reason, resist, UK" and "Win, America, chaos, build," which were more focused on immediate and emotional themes. In contrast, topic modeling of the entire corpus using Gensim highlighted broader themes related to technology, AI, and the industry, such as "Model, Chinese, chip, market, government, power, OpenAI, world, data, billion." This indicates that neutral or positive articles concentrated on technological advancements and corporate success, while negative articles tended to focus on risks, controversies, and financial instability.

However, the research also faced limitations. Topic modeling was constrained by word frequency and computational limitations, which might lead to bias and incomplete topic representation. For sentiment analysis, the automated tool's scores were sometimes inaccurate, and manual adjustment was challenging due to the lack of a clear method and time constraints.

To address these limitations and expand the research, future studies could broaden the corpus by including more diverse sources and extending the research timeline. Close reading of articles with different sentiment scores would also be beneficial to identify the exact linguistic features driving sentiment. In conclusion, this research has provided an initial understanding of media coverage of "DeepSeek," but there is still much room for further exploration and improvement.

References

Guldi, J. (2023). The Dangerous Art of Text Mining. Cambridge University Press, p229-271.
Kettunen, K. and La Mela, M. (2022) .'Semantic tagging and the Nordic tradition of everyman's rights', *Digital Scholarship in the Humanities*, 37(2), pp. 483-496.
Lucy, L., Demszky, D., Bromley, P. and Jurafsky, D. (2020) .'Content Analysis of Textbooks via Natural Language Processing: Findings on Gender, Race, and Ethnicity in Texas U.S. History Textbooks', *AERA Open*, 6(3).
Shadrova, A. (2021) .'Topic models do not model topics: epistemological remarks and steps towards best practices', *Journal of Data Mining and Digital Humanities*, 10, pp. 1-28.

Tahmasebi, N. and Hengchen, S. (2020) .’The strengths and pitfalls of large-scale text mining for the digital humanities’, Samlaren: *Journal for Research on Swedish and Other Nordic Literature*, 140, pp. 198–227.
Underwood, T. (2017) .A Genealogy of Distant Reading', Digital Humanities Quarterly, 11(2).