TEXT MINING AND SENTIMENT ANALYSIS OF NEWS ARTICLE: THE STAR MALAYSIA

Phiraphong A/L A Watt, 288584[1]

[1]*Universiti Utara Malaysia, Malaysia, phiraphong_a_watt@soc.uum.edu.my*

ABSTRACT. This study explores the application of text mining and sentiment analysis on a dataset of news articles from The Star Malaysia. The purpose of this study is to uncover key themes and sentiments across various news sections, including Business, Tech, Sport, and AseanPlus, through the extraction and visualization of top words, bigrams, and trigrams. Additionally, it aims to evaluate the performance of three machine learning models in predicting sentiments within the articles. The problem addressed is the imbalance in sentiment distribution, predominantly skewed towards positive sentiments, which affects model performance. The approach involves preprocessing the text data, generating n-grams, and applying sentiment classification models, followed by an analysis of confusion matrices and classification reports. The results reveal that while the models perform well in predicting positive sentiments, they struggle with negative and neutral sentiments due to the class imbalance. This study highlights the potential of text mining in extracting valuable insights from large textual datasets and underscores the importance of addressing class imbalance in sentiment analysis.

Keywords: text mining, sentiment analysis, news articles

INTRODUCTION

In today's digital age, the media plays a pivotal role in shaping public opinion and influencing societal norms. News articles serve as a primary source of information for the public, providing insights into current events, political developments, and social issues. As the volume of news content increases, so does the need for effective methods to analyze and interpret this information. Understanding how news narratives are constructed and perceived is essential for both researchers and consumers of media, particularly in diverse societies where multiple perspectives coexist.

Text mining is a powerful analytical technique that involves extracting meaningful information from unstructured textual data. By employing various computational methods, text mining enables researchers to identify patterns, trends, and relationships within large datasets. In the context of news articles, text mining can be used to categorize content, track topic evolution over time, and uncover hidden insights within the text. This study will utilize text mining techniques to analyze a selection of articles from The Star Malaysia, focusing on the themes and topics prevalent in the Malaysian media landscape.

Sentiment analysis, a subfield of text mining, specifically focuses on determining the emotional tone behind a body of text. It involves the use of natural language processing (NLP) techniques to classify sentiments expressed in written content as positive, negative, or neutral.

In the realm of news articles, sentiment analysis can reveal how the media frames issues and influences public perception. By analyzing the sentiment of articles from The Star, this study aims to provide insights into how news narratives reflect and shape public opinion on critical national issues, highlighting the broader implications of media sentiment in contemporary society.

RELATED WORK

Numerous studies have examined sentiment analysis within the context of news articles. For instance, a study on China-related news in The Star revealed that sentiment analysis could effectively capture public sentiment and its implications for diplomatic relations (Wu et al., 2022). This research highlighted the complexities involved in analyzing news sentiment compared to other domains such as product reviews, due to the multifaceted nature of news content.

Another relevant study conducted a comprehensive analysis of newspaper headlines using text mining techniques, demonstrating how sentiment analysis can be applied to gauge public sentiment over time (Hossain et al., 2021). The findings indicated that the emotional tone of headlines could significantly influence readers' perceptions and reactions.

The existing literature underscores the importance of text mining and sentiment analysis in media studies, particularly in understanding how news narratives shape public opinion and societal values. This study will build on these foundations by applying similar methodologies to analyze sentiment in articles from The Star Malaysia, contributing to the broader discourse on media influence in Malaysia.

DATASET DESCRIPTION

The Star is a well-known newspaper and news website in Malaysia. It covers a wide range of issues, such as business, politics, sports, entertainment, lifestyle, and national news. The Star, an established English-language daily in Malaysia, has widespread circulation and maintains a substantial online presence via its website.

The dataset utilized in this study is version 22 contains news items from The Star Malaysia (Azrai Mahadan, 2024). Its articles range from brief updates to in-depth news pieces, covering a wide range of themes including crime, technology, and the country. The dataset consists of 13 columns of data. The columns in the dataset are as follows:

- content_id: A unique identifier assigned to each news article in the dataset. This ID helps in tracking and referencing specific articles.
- title: The headline or title of the news article. It provides a brief overview of the article's content and is often designed to attract readers' attention.
- text: The main body of the news article, containing the full content. This includes all the information, narratives, and details presented in the article.
- section: This attribute indicates the specific section of the news outlet where the article is published, such as News, Sports, Lifestyle, etc. It helps categorize the article within the broader context of the publication.
- category: Similar to the section, this field categorizes the article into broader themes or genres, which may include business, health, technology, etc. It aids in organizing articles based on their subject matter.

- content_tier: This attribute may denote the level of accessibility or exclusivity of the content, indicating whether the article is freely available or part of a subscription service.
- content_length: This field categorized the content in the article into three category which are short, medium and long.
- authors: The names of the individuals who wrote the article. This field reflecting the collaborative nature of journalism.
- published_date: The date when the article was published. This is crucial for understanding the timeliness of the content and its relevance to current events.
- keywords: A list of significant terms or tags associated with the article. Keywords help in search optimization and categorizing content based on key themes.
- summary: A brief overview or abstract of the article, highlighting the main points and findings. This allows readers to quickly grasp the essence of the article without reading the entire text.
- url: The web address where the article can be accessed online. This provides a direct link to the source for further reading.
- top_image: This field contains the URL of the main image associated with the article, often used to enhance the visual appeal of the article, and attract readers.

Data Distribution

This section will explain about the information in the dataset. This study focuses on articles published from January 2024 to June 2024. The timeline plot of articles as show in Figure 1 from January to June 2024 shows an initial low and steady publication rate until mid-February, followed by a significant increase and periodic peaks indicative of a regular, likely weekly, publishing schedule. A dramatic spike in article counts at the end of June suggests an unusual surge in activity, which warrants further investigation to understand the underlying cause. This pattern highlights a structured and escalating publishing trend, punctuated by a notable peak in late June.

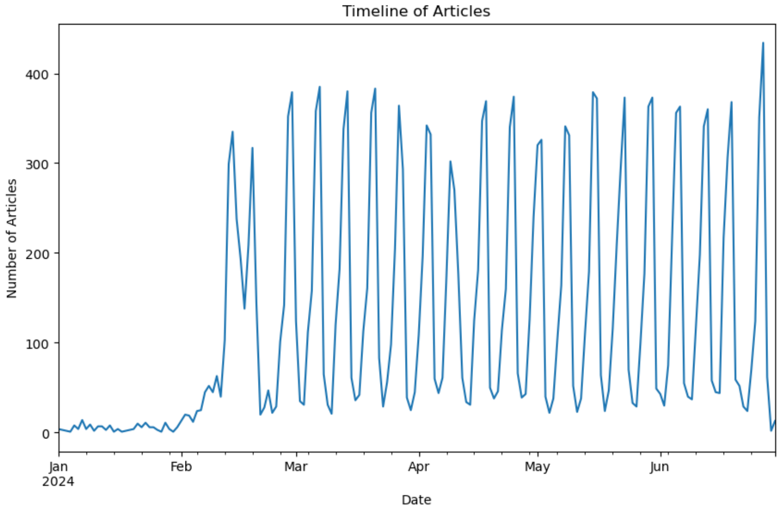


Figure 1. Timeline of Article.

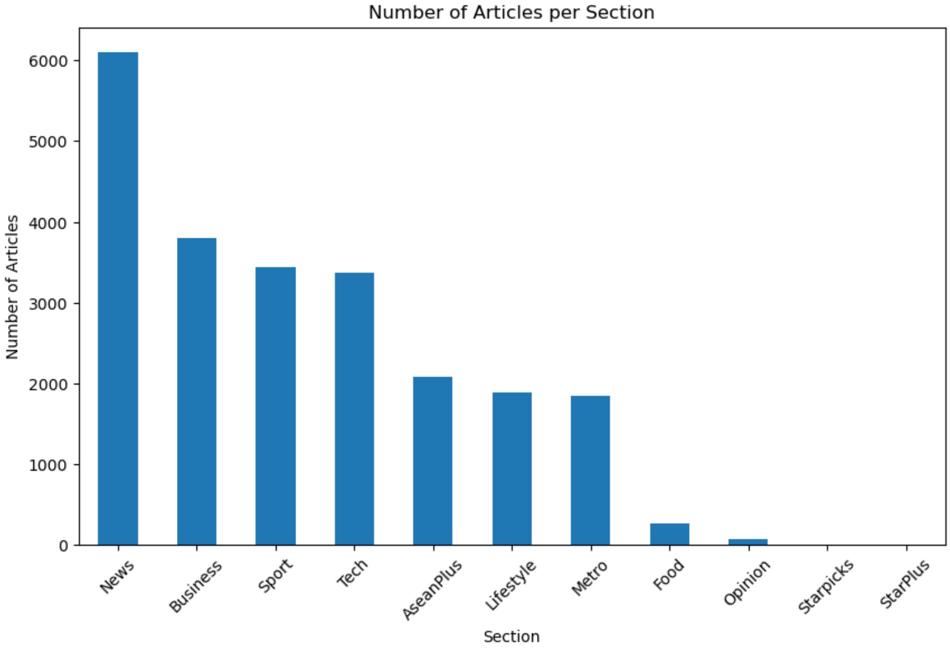


Figure 2. Number of Articles per Section.

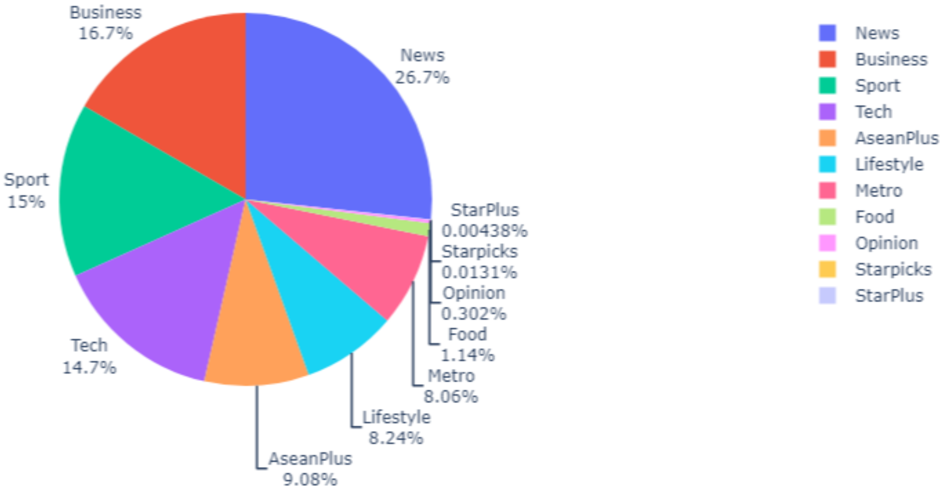


Figure 3. Distribution of Section.

Figure 2 and Figure 3 both illustrate the distribution of articles across sections. The News section has the highest number of articles, accounting for 26.7% of the total, with the total above 6000 articles. Business follows with 16.7%, and Sport with 15%, both having number below 4000 articles. The Tech section, representing 14.7%, has slightly fewer articles than Sport, with a count above 3000. AseanPlus accounts for 9.08%, Lifestyle for 8.24%, and Metro for 8.06%. The Food section, with 1.14%, and Opinion, Starpicks and StarPlus have less than 0.5%. These visualizations highlight a significant concentration of articles in the News section, indicating high activity or interest, while sections like Food, Opinion, and StarPlus have fewer contributions, suggesting less focus in these areas.

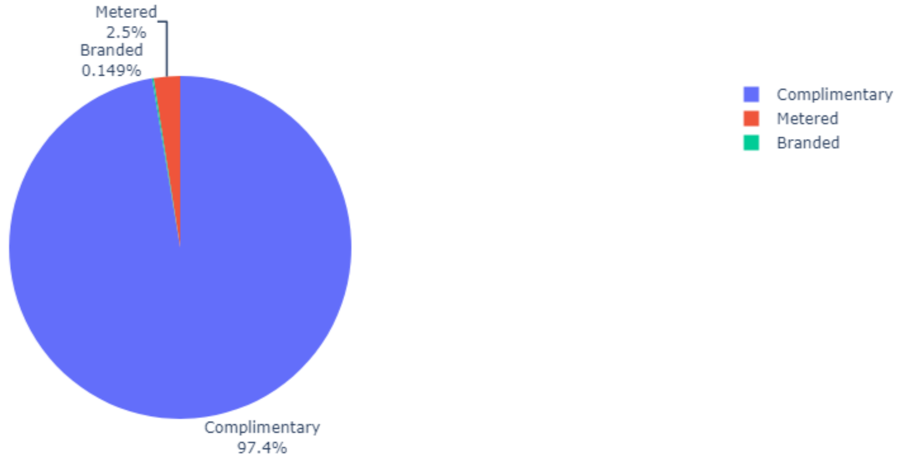


Figure 4. Distribution of Content Tier.

Figure 4 represents the distribution of content tiers in the dataset. The largest segment, labeled "Complimentary," accounts for 97.4% of the total. The "Metered" segment, which makes up 2.5%. The smallest segment, "Branded," at 0.14%, signifies a minimal portion of the dataset.

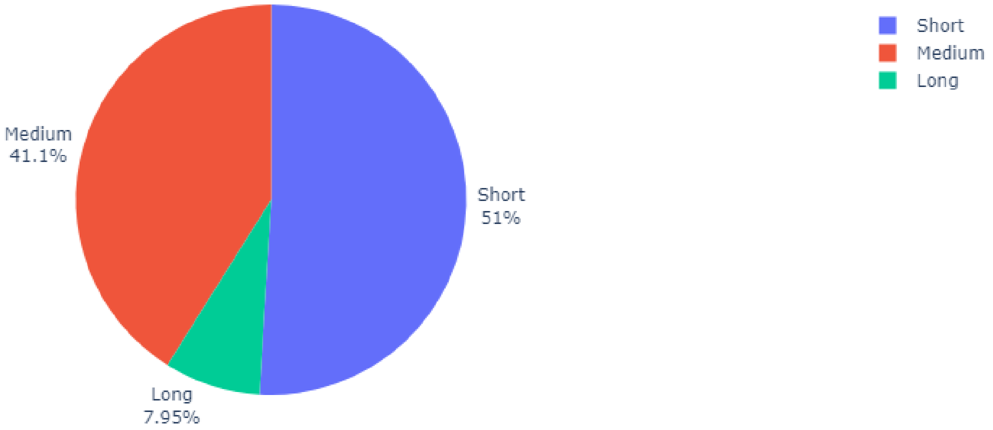


Figure 5. Distribution of Content Length.

Figure 5 represents the distribution of content length in a dataset. The Short category constitutes the majority, making up 51% of the total content. The Medium category follows, accounting for 41.1%, while the Long category is the smallest segment, representing 7.95%. This distribution indicates that over half of the content is short, with a significant portion being medium-length, and a smaller fraction being long.

The analysis of the dataset will focus on the top five sections identified in the article distribution, namely News, Business, Sport, Tech, and AseanPlus. Given that the News section comprises the largest share of articles, it will be particularly interesting to examine the sentiment expressed in this category, as it likely reflects current public opinion on significant events. Additionally, the Business and Sport sections, which also have substantial representation, will provide insights into sentiments surrounding economic and athletic developments, respectively. The Tech section's focus on technology-related news will allow for an exploration of public sentiment towards innovation and digital advancements, while the AseanPlus section can shed light on regional perspectives and international relations. By conducting sentiment analysis across all articles in these top sections, the study aims to uncover underlying emotional tones and trends, contributing to a deeper understanding of how media narratives shape public perception in various domains. This comprehensive analysis will not only highlight the sentiments prevalent in each section but also reveal potential correlations between article themes and public sentiment.

METHODOLOGY

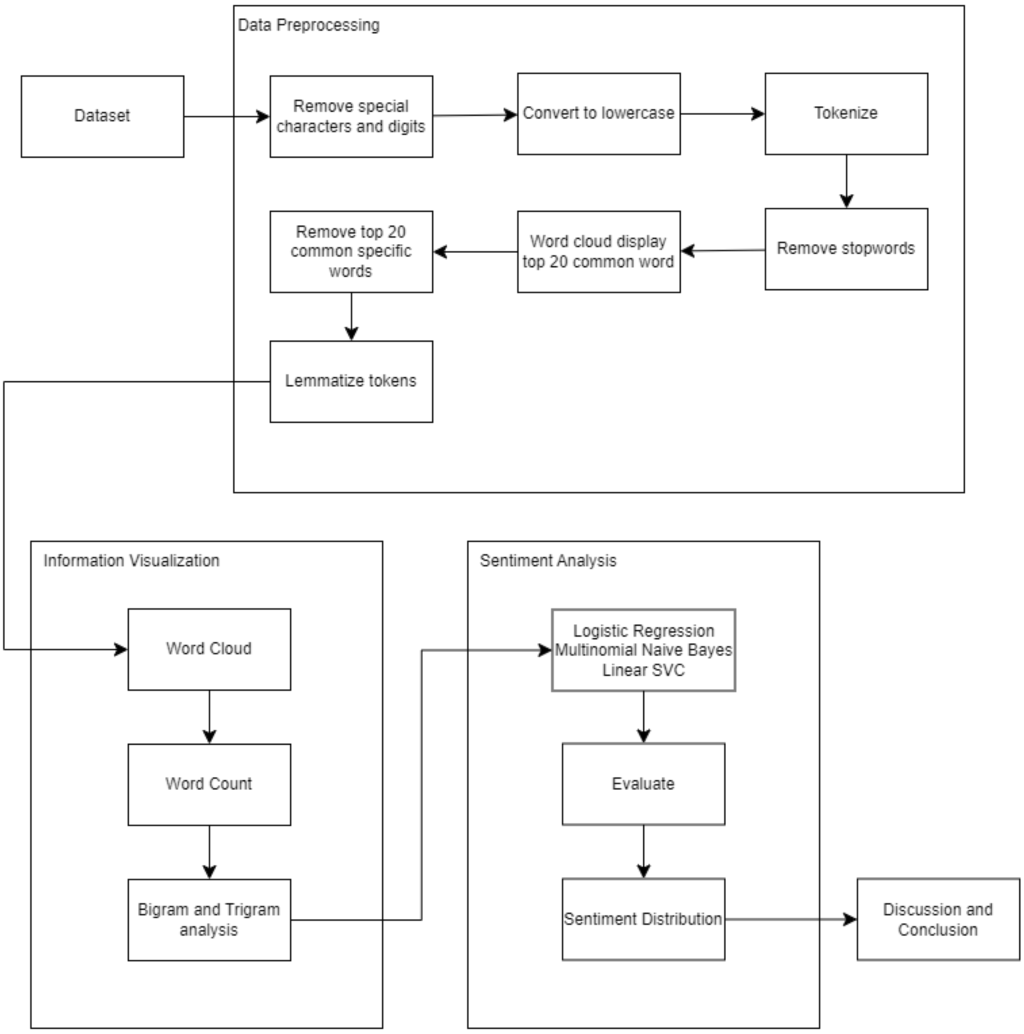


Figure 6. Research Flow Work Diagram.

Data Preprocessing

The dataset underwent a rigorous preprocessing phase. Initially, special characters and digits were removed from the text data to ensure uniformity. The text was then converted to lowercase to eliminate case sensitivity issues. Following this, the text was tokenized, breaking it down into individual words. Stop words, which are common words that do not contribute significant meaning, were removed to reduce noise in the data.

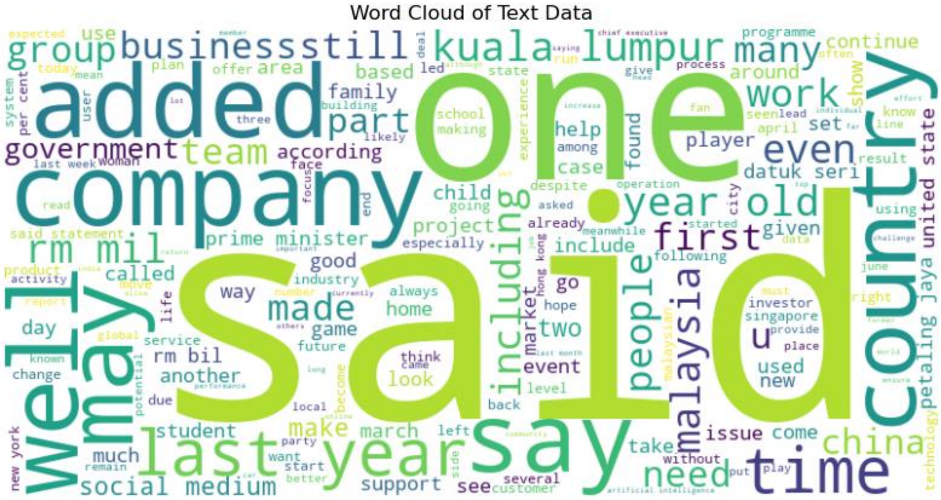
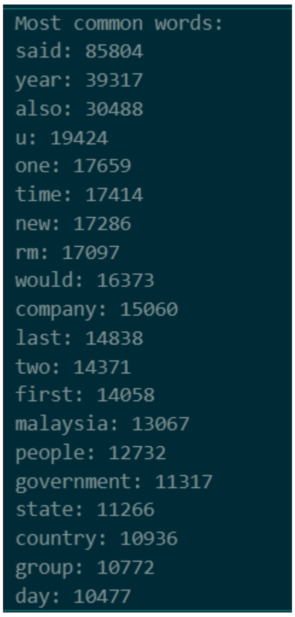


Figure 7. Word Cloud of Text Data.



```
Most common words:
said: 85804
year: 39317
also: 30488
u: 19424
one: 17659
time: 17414
new: 17286
rm: 17097
would: 16373
company: 15060
last: 14838
two: 14371
first: 14058
malaysia: 13067
people: 12732
government: 11317
state: 11266
country: 10936
group: 10772
day: 10477
```

Figure 8. List of Top 20 common word.

A word cloud was generated to display the top 20 common words in the dataset, aiding in the identification of specific words that needed removal as show in Figure 7. These specific words were subsequently removed from the dataset, and the tokens were lemmatized to reduce them to their base forms. These words were removed to enhance the quality of the text data for subsequent analysis.

Information Visualization

Post preprocessing, the data was visualized through various methods. A word cloud provided a visual representation of the most frequent words in the dataset. Word count analysis was performed to determine the frequency of each word, which further informed the preprocessing step. Additionally, bigram and trigram analyses were conducted to identify the most common pairs and triplets of words, offering deeper insights into common phrases within the text.

Sentiment Analysis

For sentiment analysis, the preprocessed text data was analyzed with TextBlob using three different classifiers: Logistic Regression, Multinomial Naive Bayes, and Linear Support Vector Classification (SVC). These classifiers were trained to identify the sentiment of each article, categorizing them as positive, negative, or neutral. The models' performance was evaluated, and sentiment distributions were plotted to visualize the results.

RESULTS & DISCUSSION

The word clouds, word count analyses, and bigram and trigram visualizations for each section provided valuable insights into the common terms and phrases used in different contexts.

News Section

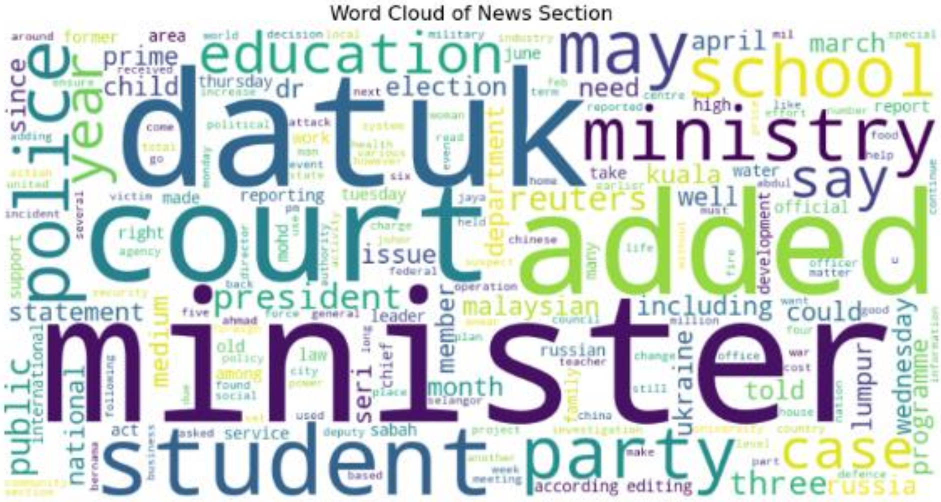


Figure 9. Word Cloud of News Section.

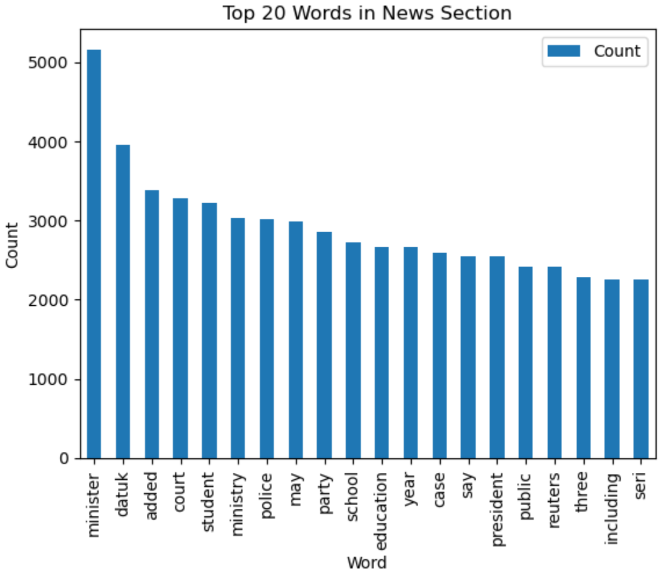


Figure 10. Top 20 Words in News Section.

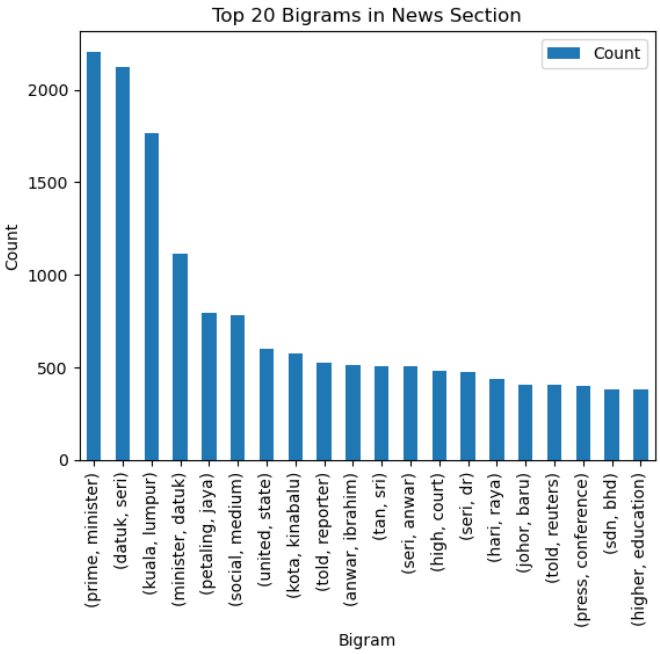


Figure 11. Top 20 Bigrams in News Section.

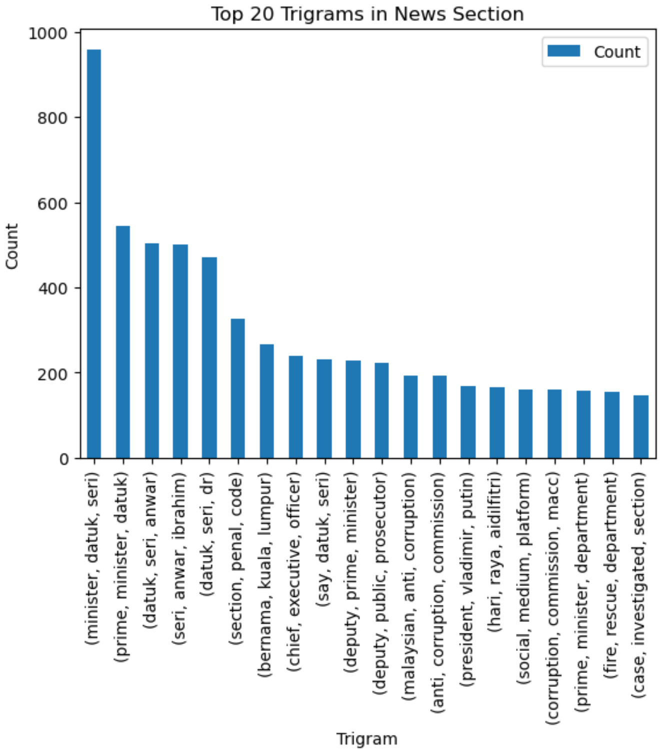


Figure 12. Top 20 Trigrams in News Section.

The word cloud and top words, bigrams, and trigrams in the news section provide a comprehensive overview of the prominent topics and entities covered. Key terms such as "minister," "datuk," "court," "student," "ministry," and "police" dominate the word cloud and word frequency diagrams. This indicates that a significant portion of the news articles focuses on political figures, legal matters, education, and law enforcement.

The bigrams and trigrams reinforce this observation. Common bigrams like "prime minister," "datuk seri," "kuala lumpur," and "minister datuk" suggest frequent reporting on high-ranking officials and events taking place in major cities. Trigrams such as "minister datuk seri," "prime minister datuk," and "chief executive officer" further emphasize the focus on specific titles and positions, indicating detailed coverage of actions and statements from notable figures.

Recent news stories that could relate to these findings include updates on government policies, high-profile court cases, educational reforms, and significant police actions. The emphasis on terms like "education," "student," and "school" suggests ongoing discussions about the education system, potentially covering topics such as policy changes, student achievements, or challenges faced by educational institutions.

The analysis of the news section highlights the media's focus on political and legal matters, education, and law enforcement, reflecting the importance of these topics in public discourse and their impact on society.

Business Section

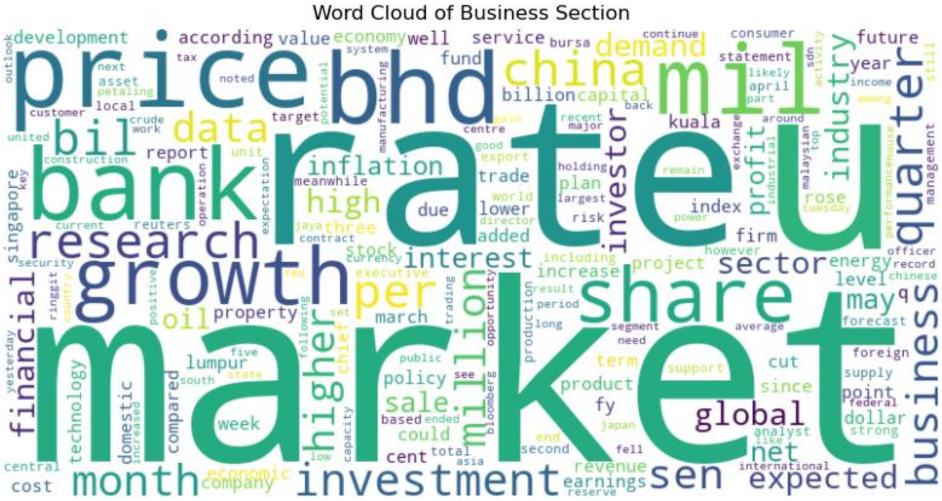


Figure 13. Word Cloud of Business Section.

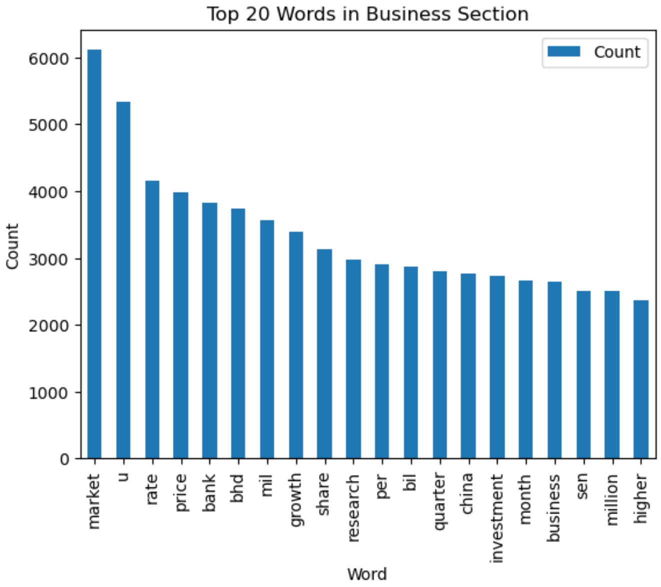


Figure 14. Top 20 Words in Business Section.

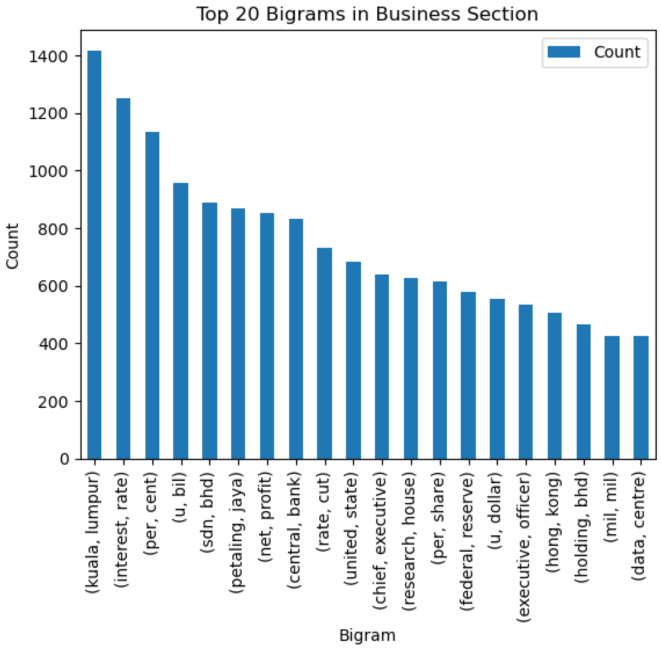


Figure 15. Top 20 Bigrams in Business Section.

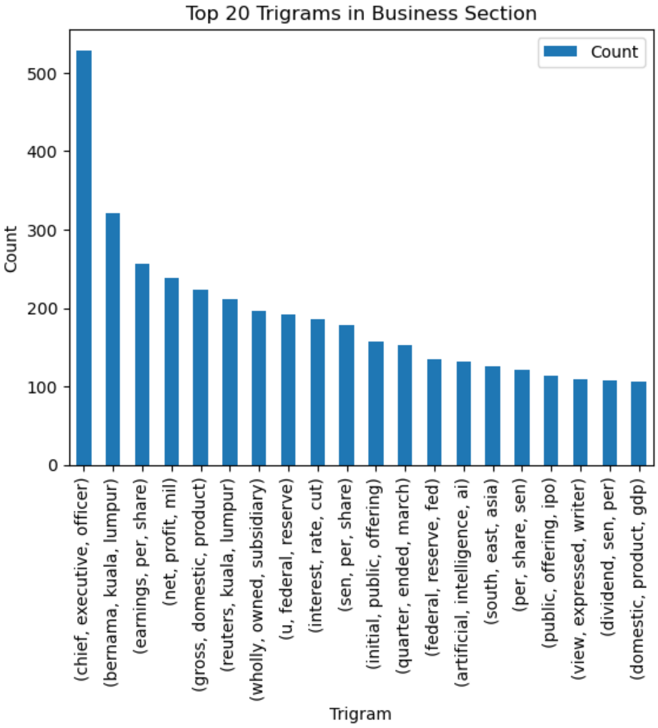


Figure 16. Top 20 Trigrams in Business Section.

The word cloud and frequency analysis of the business section highlight significant trends and themes in business reporting. Prominent words such as "market," "rate," "price," "bank," "growth," and "investment" indicate a strong focus on financial markets, economic indicators, and investment activities. These terms suggest that the business section frequently covers topics related to market performance, interest rates, pricing strategies, banking activities, and economic growth.

The bigrams and trigrams provide further insight into the specific aspects of business news. Common bigrams like "interest rate," "u.s. bil," "central bank," and "federal reserve" suggest detailed coverage of monetary policies, central banking actions, and significant financial data from the United States. Trigrams such as "chief executive officer," "interest rate cut," and "gross domestic product" reinforce the focus on executive actions, monetary policy decisions, and key economic metrics.

Recent news stories that could relate to these findings include updates on central bank interest rate changes, stock market performance, major corporate earnings reports, and significant mergers and acquisitions. The emphasis on terms like "investment," "growth," and "share" suggests ongoing discussions about investment opportunities, economic expansion, and shareholder value.

The analysis of the business section reveals a strong emphasis on financial and economic topics, reflecting the importance of these issues in the business world and their impact on the broader economy. This focus is crucial for keeping readers informed about the latest developments in markets, economic policies, and corporate actions.

Sport Section

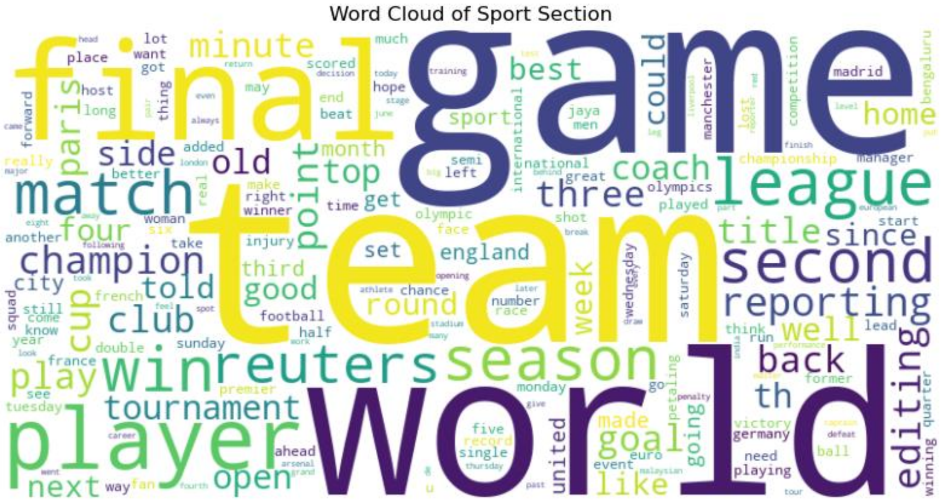


Figure 17. Word Cloud of Sport Section.

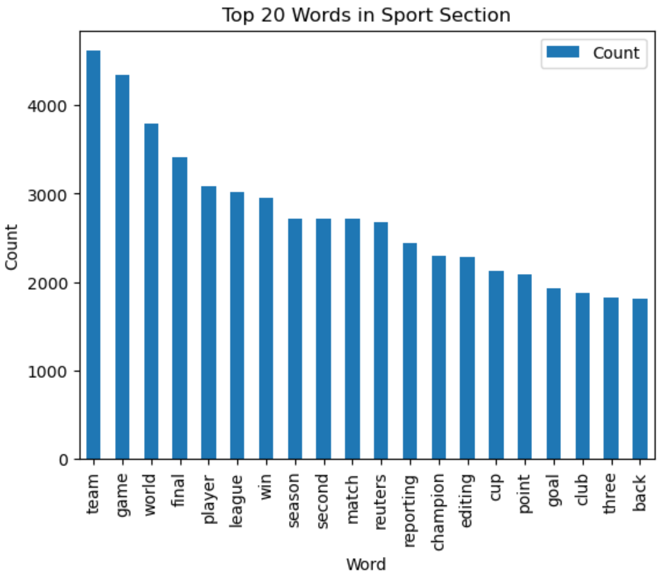


Figure 18. Top 20 Words in Sport Section.

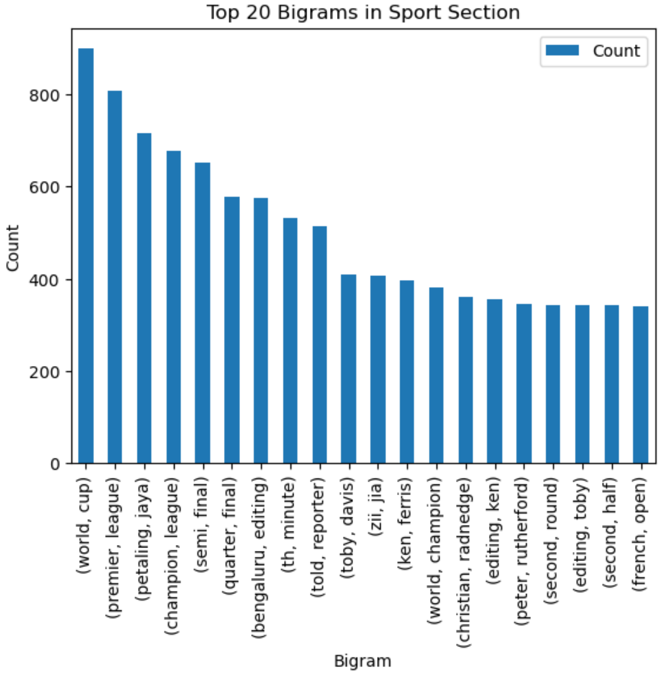


Figure 19. Top 20 Bigrams in Sport Section.

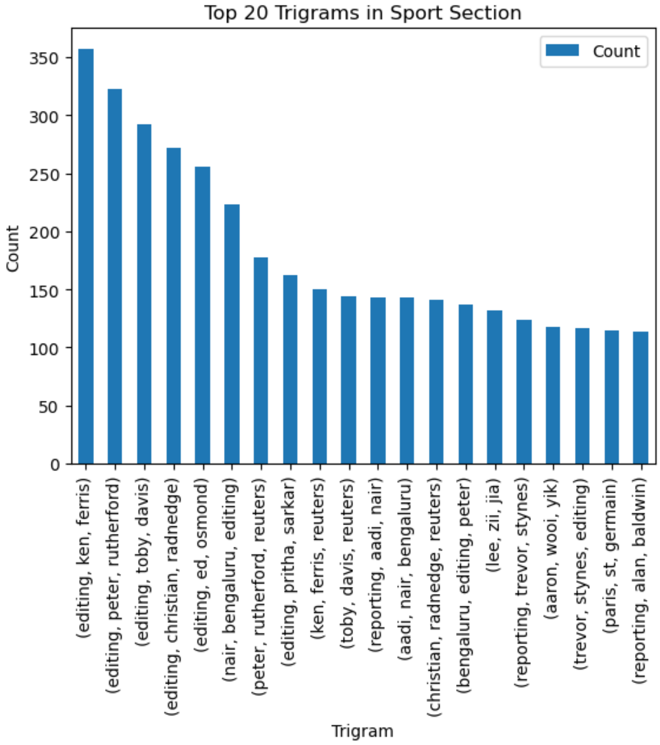


Figure 20. Top 20 Trigrams in Sport Section.

The word cloud and frequency analysis of the sports section highlight key themes and trends in sports reporting. Prominent words such as "team," "game," "world," "final," "player," "league," "win," "season," and "match" indicate a strong focus on major sporting events, team performance, individual athletes, and international competitions. These terms suggest that the sports section frequently covers final matches, game highlights, team standings, player achievements, and global sports tournaments.

The bigrams and trigrams provide further insight into the specific aspects of sports news. Common bigrams like "world cup," "premier league," "champion league," "team player," and "final match" suggest detailed coverage of prestigious tournaments, league championships, team dynamics, and crucial matches. Trigrams such as "editing ken ferris," "editing peter rutherford," and "editing tony lawry" show the emphasis on journalistic contributions, while phrases like "world cup final" and "champion league final" indicate focus on high-stakes games and international competitions.

Recent news stories that could relate to these findings include updates on major sports leagues like the Premier League and NBA, coverage of international tournaments like the World Cup and the Olympics, and reports on significant player transfers, injuries, and achievements. The emphasis on terms like "team," "game," and "world" suggests ongoing discussions about the outcomes of key matches, team strategies, and the global impact of sports events.

The analysis of the sports section reveals a strong emphasis on major sporting events, team and player performances, and international competitions, reflecting the excitement and dynamism of the sports world. This focus is crucial for keeping readers informed about the latest developments, results, and highlights in the realm of sports.

Tech Section

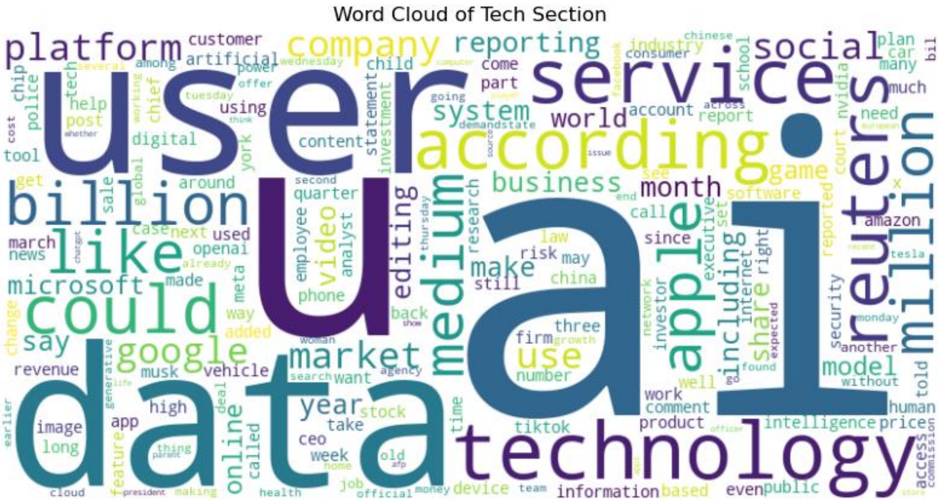


Figure 21. Word Cloud of Tech Section.

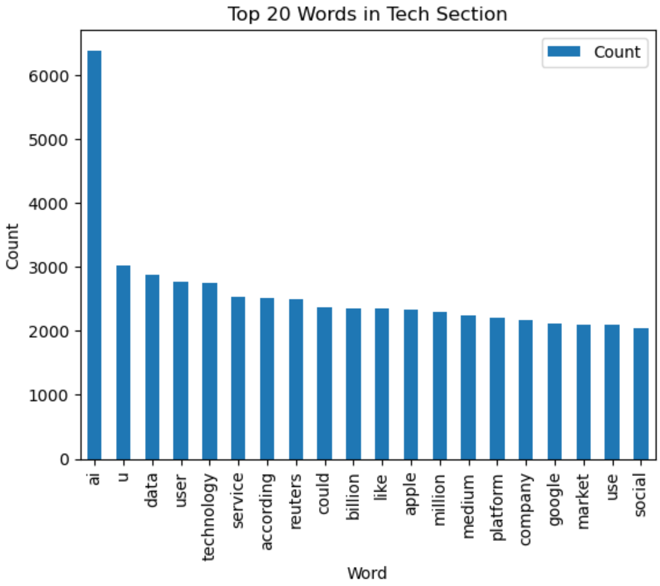


Figure 22. Top 20 Words in Tech Section.

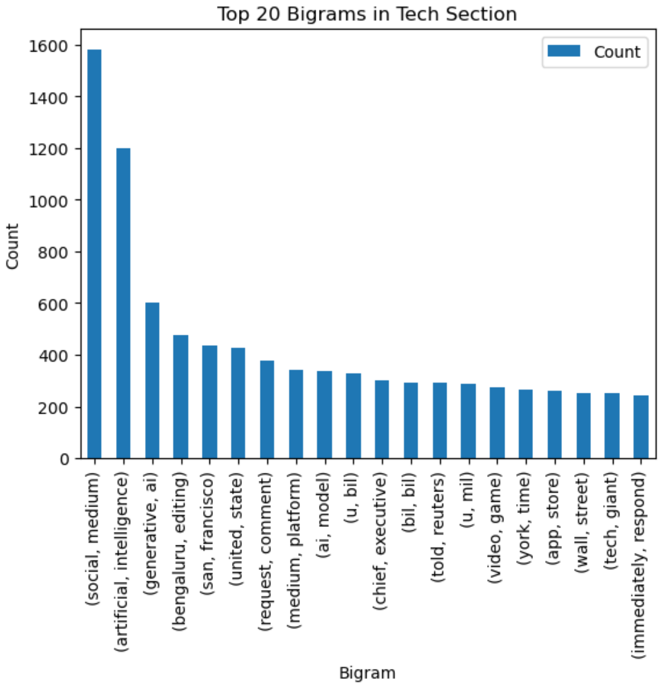


Figure 23. Top 20 Bigrams in Tech Section.

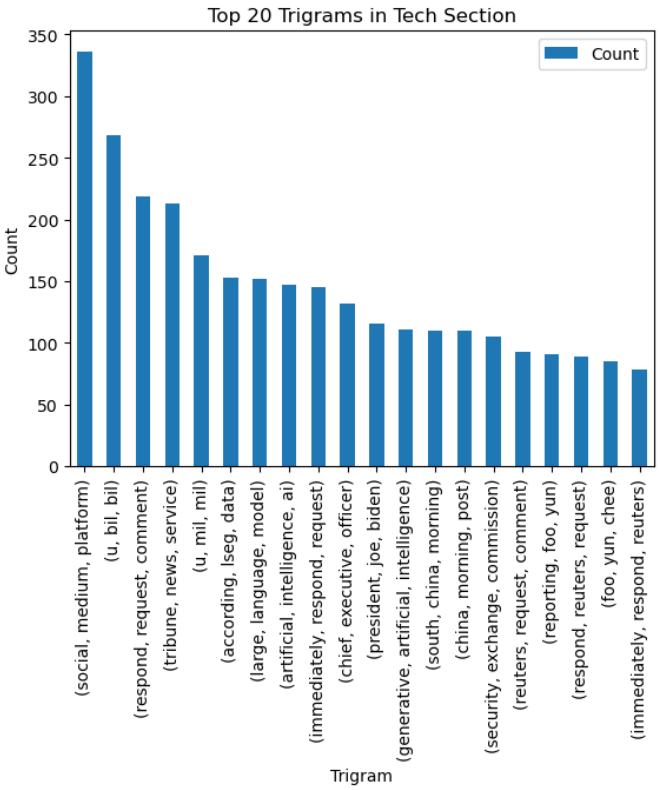


Figure 24. Top 20 Trigrams in Tech Section.

The word cloud and frequency analysis of the tech section reveal several key trends and themes within technology reporting. Prominent words such as "AI," "data," "user," "technology," "service," "platform," "company," and "market" suggest a strong emphasis on artificial intelligence, data management, user experience, technological advancements, and market dynamics. This indicates that the tech section frequently covers innovations in AI, the importance of data, user-centric services, and the performance of tech companies in the market.

The bigrams and trigrams provide further context. Common bigrams like "artificial intelligence," "social medium," "chief executive," and "video game" indicate detailed coverage of AI developments, social media platforms, executive decisions, and the gaming industry. Trigrams such as "social medium platform," "artificial intelligence ai," and "chief executive officer" reinforce the focus on social media, AI technology, and corporate leadership within the tech industry.

Recent news stories that relate to these findings include advancements in AI technology, updates from major tech companies like Google, Apple, and Microsoft, developments in social media platforms, and significant trends in the gaming industry. The prominence of terms like "AI," "data," and "user" suggests ongoing discussions about AI innovations, data privacy and security, and enhancing user experience in various tech products and services.

The analysis of the tech section reveals a strong focus on AI, data, user-centric technologies, and the performance of tech companies, reflecting the dynamic and rapidly evolving nature of the technology industry. This focus is crucial for keeping readers informed about the latest technological advancements, market trends, and the strategic decisions of leading tech companies.

AseanPlus Section

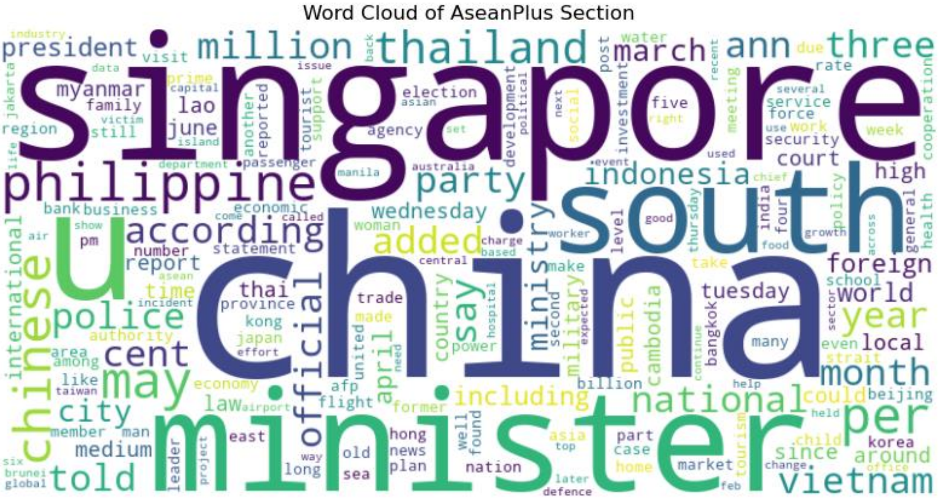


Figure 25. Word Cloud of AseanPlus Section.

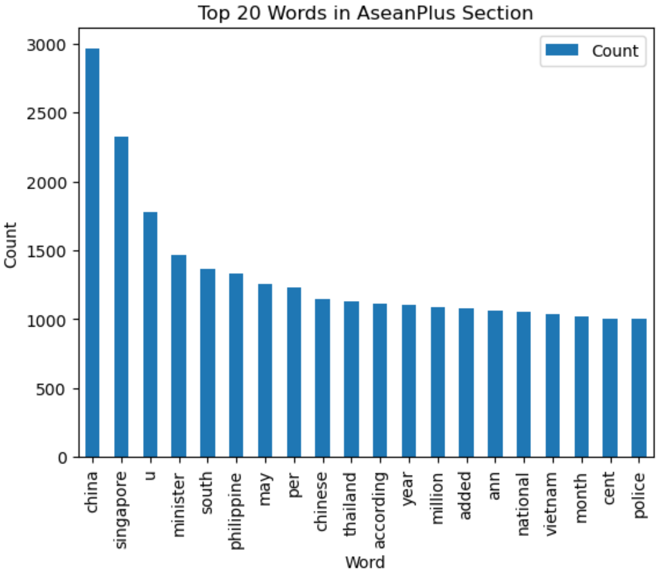


Figure 26. Top 20 Words in AseanPlus Section.

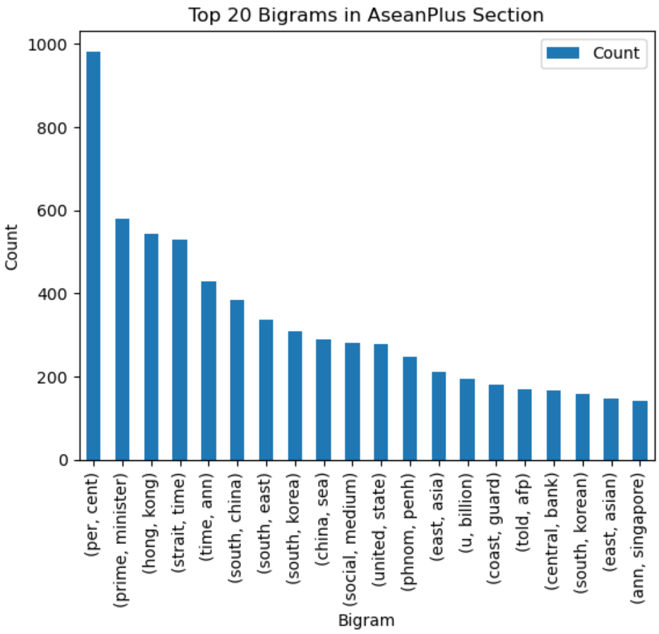


Figure 27. Top 20 Bigrams in AseanPlus Section.

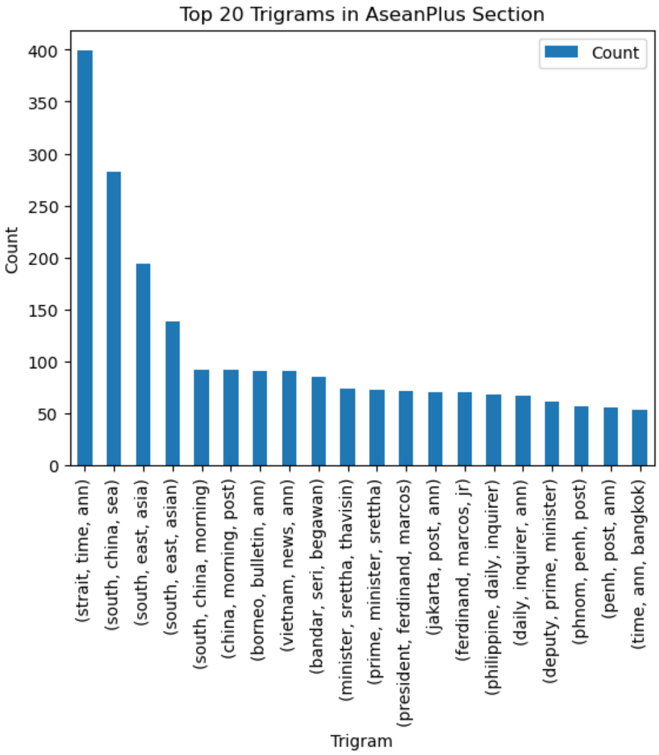


Figure 28. Top 20 Trigrams in AseanPlus Section.

The word cloud and frequency analysis of the AseanPlus section highlight key themes and trends in regional reporting. Prominent words such as "China," "Singapore," "U," "minister," "south," "Philippine," "may," "per," "Chinese," and "Thailand" indicate a strong focus on geo-political issues, national affairs, and regional cooperation. These terms suggest that the

AseanPlus section frequently covers diplomatic relations, ministerial activities, economic developments, and significant events in Southeast Asian countries.

The bigrams and trigrams provide further insight into the specific aspects of regional news. Common bigrams like "per cent," "prime minister," "hong kong," "strait time," "ann," "south east," "china sea," "social medium," and "central bank" suggest detailed coverage of economic indicators, political leaders, regional disputes, and financial institutions. Trigrams such as "strait time ann," "south china sea," "south east asia," and "central bank ann" highlight the emphasis on specific publications, regional waters, and economic governance.

Recent news stories that could relate to these findings include updates on China's Belt and Road Initiative, ASEAN summits, regional security issues, trade agreements, and bilateral relations between Southeast Asian countries and major global powers. The emphasis on terms like "China," "minister," and "south" suggests ongoing discussions about China's influence in the region, ministerial dialogues, and geopolitical dynamics in the southern parts of Asia.

The analysis of the AseanPlus section reveals a strong emphasis on geopolitical affairs, economic developments, and regional cooperation, reflecting the importance of Southeast Asia in global politics and economics. This focus is crucial for keeping readers informed about the latest developments, policies, and events that impact the region and its relations with the rest of the world.

Sentiment Analysis Model Evaluation

The confusion matrices for the sentiment analysis models (Logistic Regression, Multinomial Naive Bayes, and Linear SVC) provide a detailed comparison of their performance in classifying sentiments as negative, neutral, or positive.

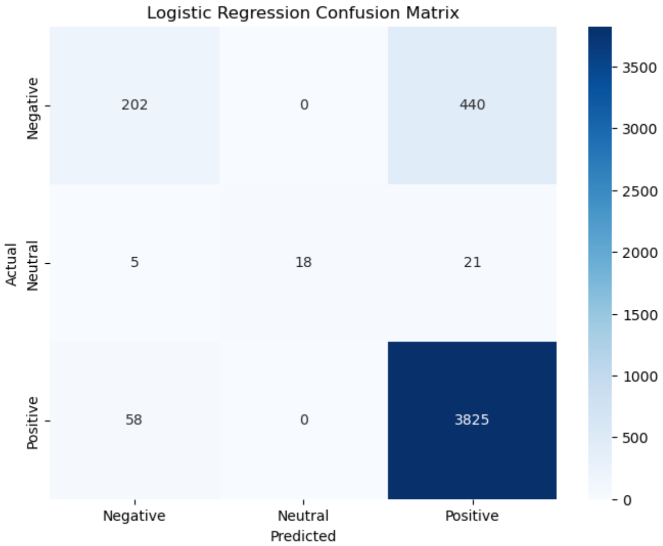


Figure 29. Logistic Regression Confusion Matrix.

The confusion matrix for the Logistic Regression model reveals its effectiveness in classifying positive sentiments but highlights significant challenges in distinguishing between negative and neutral sentiments. The model correctly classifies 202 out of 642 negative instances, resulting in a considerable number of false positives (440 instances misclassified as positive). Additionally, the model shows a weakness in identifying neutral sentiments, with only 18 out of 44 neutral instances correctly classified. However, it excels in predicting positive sentiments, correctly classifying 3825 out of 3883 instances. This imbalance suggests that while Logistic

Regression is highly reliable for positive sentiment detection, it requires improvement in handling negative and neutral sentiments.

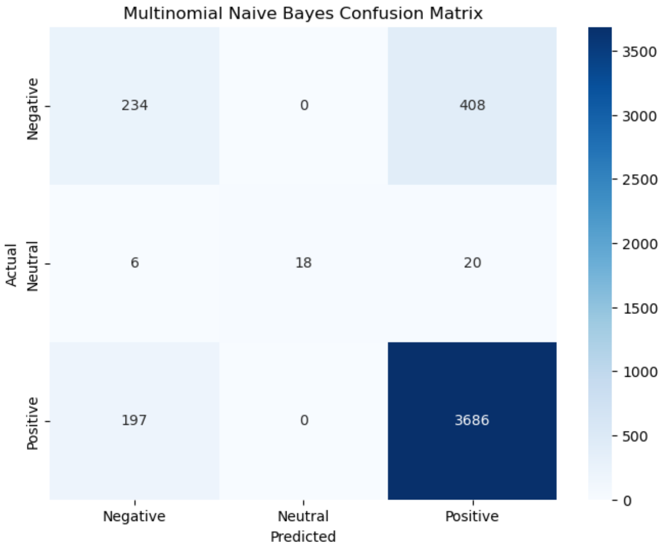


Figure 30. Multinomial Naive Bayes Confusion Matrix.

The Multinomial Naive Bayes confusion matrix demonstrates a similar trend, with strong performance in classifying positive sentiments but weaker results for negative and neutral sentiments. The model correctly identifies 234 out of 642 negative instances, leading to a high number of false positives (408 instances misclassified as positive). It also shows limited capability in detecting neutral sentiments, with only 18 out of 44 instances correctly classified. For positive sentiments, the model performs well, accurately classifying 3686 out of 3883 instances. This indicates that while Multinomial Naive Bayes is effective for positive sentiment classification, it struggles significantly with negative and neutral sentiments, suggesting a need for better handling of class imbalances.

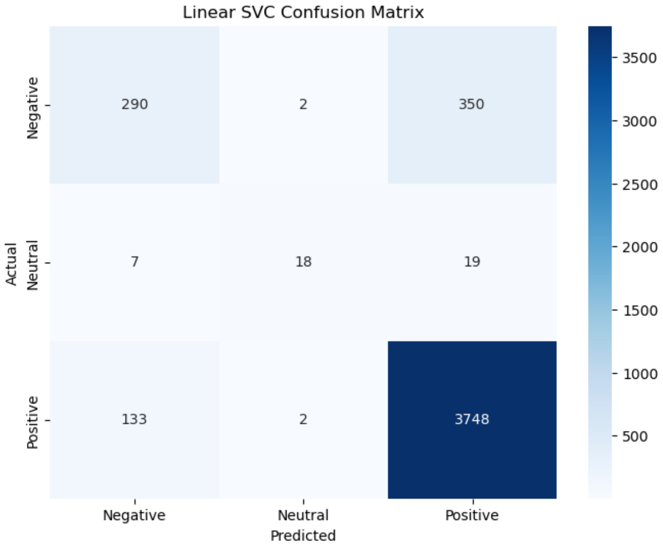
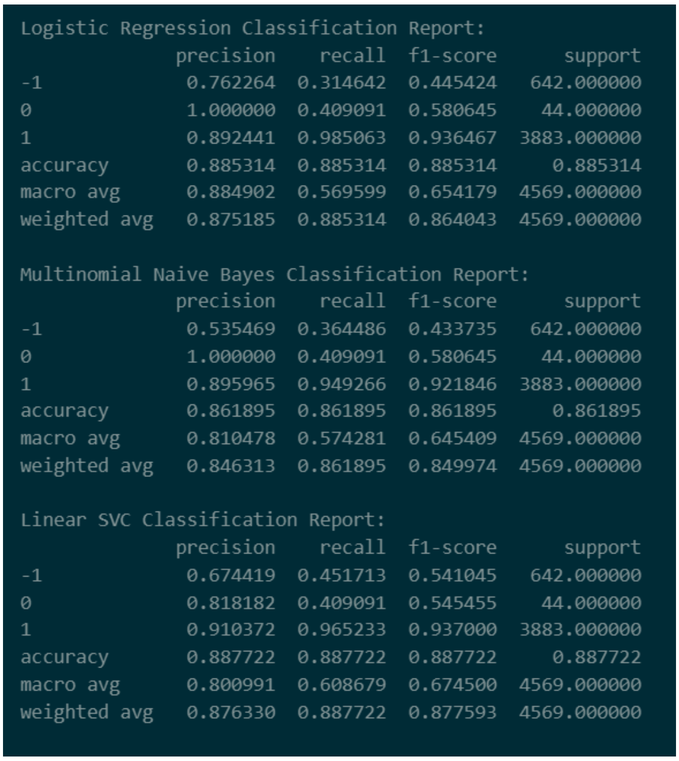


Figure 31. Multinomial Naive Bayes Confusion Matrix.

The Linear SVC model exhibits a more balanced performance across all sentiment categories, as seen in its confusion matrix. The model correctly classifies 290 out of 642 negative instances and shows moderate improvement in identifying neutral sentiments, with 18 out of 44 instances correctly classified. Notably, the model also performs well with positive sentiments, accurately predicting 3748 out of 3883 instances. Despite its relatively high number of false positives for negative sentiments (350 instances misclassified as positive), Linear SVC demonstrates the best overall balance among the three models. It shows a consistent ability to handle both positive and negative sentiments more effectively than the other models, making it the most reliable choice for balanced sentiment analysis.



```
Logistic Regression Classification Report:
              precision    recall  f1-score      support
-1             0.762264  0.314642  0.445424   642.000000
0              1.000000  0.409091  0.580645    44.000000
1              0.892441  0.985063  0.936467  3883.000000
accuracy       0.885314  0.885314  0.885314     0.885314
macro avg      0.884902  0.569599  0.654179  4569.000000
weighted avg   0.875185  0.885314  0.864043  4569.000000

Multinomial Naive Bayes Classification Report:
              precision    recall  f1-score      support
-1             0.535469  0.364486  0.433735   642.000000
0              1.000000  0.409091  0.580645    44.000000
1              0.895965  0.949266  0.921846  3883.000000
accuracy       0.861895  0.861895  0.861895     0.861895
macro avg      0.810478  0.574281  0.645409  4569.000000
weighted avg   0.846313  0.861895  0.849974  4569.000000

Linear SVC Classification Report:
              precision    recall  f1-score      support
-1             0.674419  0.451713  0.541045   642.000000
0              0.818182  0.409091  0.545455    44.000000
1              0.910372  0.965233  0.937000  3883.000000
accuracy       0.887722  0.887722  0.887722     0.887722
macro avg      0.800991  0.608679  0.674500  4569.000000
weighted avg   0.876330  0.887722  0.877593  4569.000000
```

Figure 32. Model Classification Report.

Table 1. Comparison of Accuracy.

Model	Accuracy (%)
Logistic Regression	88.53
Multinomial Naive Bayes	86.19
Linear SVC	88.77

All three models exhibit strong performance in predicting positive sentiments, with high precision, recall, and F1-scores. However, they struggle to accurately classify negative sentiments, with Logistic Regression and Multinomial Naive Bayes showing particularly low recall and F1-scores for this category. Logistic Regression demonstrates the highest precision and recall for positive sentiments but performs poorly with negative ones. Multinomial Naive Bayes provides a balanced performance but falls short overall compared to the other models. Linear SVC stands out with the highest accuracy (88.77%) and the best weighted average F1-score (0.88), indicating a balanced and robust performance across all sentiment categories, making it the most reliable model for sentiment analysis among the three.

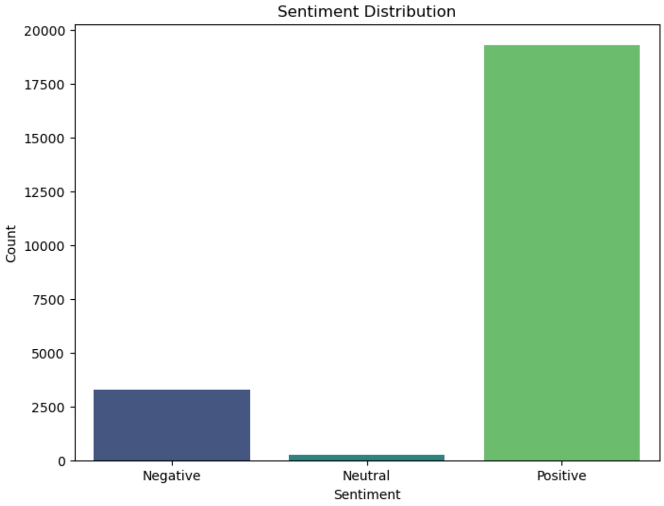


Figure 33. Sentiment Distribution.

The sentiment distribution chart provides an overview of the sentiment labels across the dataset. It reveals a significant imbalance in sentiment categories, with most of the instances being classified as positive. Specifically, positive sentiments dominate with approximately 18,000 instances, followed by negative sentiments with around 2,500 instances, and a minimal number of neutral sentiments. This imbalance indicates a strong positive bias in the data, which can influence the performance of sentiment analysis models. Models tend to perform better on the dominant class, as seen in the previous classification reports and confusion matrices, where positive sentiments were predicted with high accuracy and precision. Conversely, the less represented negative and neutral sentiments show lower classification performance, likely due to the limited data available for these categories. Addressing this imbalance could potentially improve the model's ability to predict minority classes accurately.

CONCLUSION

This text mining and sentiment analysis study on The Star Malaysia news articles reveals insightful patterns in the data. The analysis of top words, bigrams, and trigrams across different sections such as Business, Tech, Sport, and AseanPlus highlighted key themes and topics prevalent in each section. For instance, in the Tech section, terms related to AI and social media were dominant, whereas the Sport section emphasized team-related terms and major sports events. These findings provide a granular understanding of the content focus in each news section, showcasing the ability of text mining techniques to uncover underlying trends and important themes in large textual datasets.

The sentiment analysis portion of the study demonstrated the varying effectiveness of three machine learning models: Logistic Regression, Multinomial Naive Bayes, and Linear SVC. Logistic Regression and Linear SVC showed strong performance in predicting positive sentiments but struggled with negative and neutral sentiments due to the inherent class imbalance in the dataset, where positive sentiments were overwhelmingly dominant. The confusion matrices and classification reports confirmed that while the models performed well overall, the imbalance in sentiment distribution affected the recall and precision for the minority classes. Addressing this imbalance through techniques such as oversampling, undersampling, or more sophisticated algorithms could further enhance the model's performance in accurately predicting minority class sentiments. This study underscores the importance of considering class distribution in sentiment analysis and highlights the potential of text mining in extracting valuable insights from large textual datasets.

REFERENCES

Azrai Mahadan. (2024, July 20). *News article (Weekly updated)*. Kaggle. https://www.kaggle.com/datasets/azraimohamad/news-article-weekly-updated/data

Georgieva-Trifonova, T., & Dechev, M. (2021). Applying text mining methods to extracting information from news articles. IOP Conference Series Materials Science and Engineering, 1031(1), 012054. https://doi.org/10.1088/1757-899x/1031/1/012054

Hossain, A., Karimuzzaman, M., Hossain, M. M., & Rahman, A. (2021). Text mining and sentiment analysis of newspaper headlines. *Information*, *12*(10), 414. https://doi.org/10.3390/info12100414

Wu, H., Bakar, K. A., Jaludin, A., & Awal, N. M. (2022). Sentiment analysis of China-Related news in the Star Online newspaper. *GEMA Online Journal of Language Studies*, *22*(3), 155–175. https://doi.org/10.17576/gema-2022-2203-09