SCHOOL OF COMPUTER SCIENCE
UNIVERSITI SAINS MALAYSIA

SEMESTER 1, ACADEMIC SESSION 2024/2025

CPC353
NATURAL LANGUAGE PROCESSING

ASSIGNMENT 1

Name	MatricNumber
KAVITASHINI A/P SELUVARAJOO	164329
TEJASHREE LAXMI A/P KANTHAN	163506
MUVENDDRAN A/L SARAVANAN	164384

LECTURER: ASSOC. PROF. DR. GAN KENG HOON

SUBMISSION DATE: 6 DECEMBER 2024

Exploring the Impact of Global Issues and Macroeconomic Factors on Stock Market Trends Using N-Gram and Part of Speech Methods

Name: KAVITASHINI A/P SELUVARAJOO
Matric Number: 164329
Email: kavita6103@student.usm.my

Name: TEJASHREE LAXMI A/P KANTHAN
Matric Number: 163506
Email: tejashree@student.usm.my

Name: MUVENDDRAN A/L SARAVANAN
Matric Number: 164384
Email: muvenddran@student.usm.my

***Abstract*—The global economy and economic factors have a big impact on stock market performance, creating complex relationships that require careful analysis. This study uses quantitative methods, specifically POS (Part-of-Speech) tagging and N-gram analysis, to examine how international economic indicators, political events, and stock market trends are connected. Important economic factors like GDP growth, inflation, interest rates, and currency changes are analyzed using these techniques to understand how they can predict market behavior and affect volatility. The research leverages advanced statistical methods to uncover patterns and gain insights for investors, policymakers, and financial analysts navigating the intricate global economic landscape.**

Keywords— Macroeconomic Analysis, Stock Market Trends, Economic Indicators, Financial Modeling, POS, n-gram Analysis

I. Introduction

Natural Language Processing (NLP) has revolutionized the way we analyze data for decision-making, enabling a deeper understanding of large volumes of textual information. One significant application is in the stock market, where NLP analyzes news, reports, and public sentiment about stocks to assist investors and analysts in making more informed decisions.

This study explores the integration of NLP with stock market data, employing techniques such as N-grams and Part-of-Speech (POS) analysis. These methods help identify patterns and trends in how stocks are discussed, shedding light on how language influences market behavior. The insights gained from this analysis are crucial for improving predictions, gauging public sentiment, and managing financial risks more effectively.

II. Overview

In this financial domain, the primary Natural Language Processing (NLP) problem is extracting and analyzing the vast amount of unstructured textual data related to global issues and macroeconomic factors to understand their influence on stock market trends. Financial news, reports, social media, and other public sources frequently contain valuable information about economic conditions, geopolitical events, and market sentiment. However, these texts are often unstructured and written in varied, sometimes complex language. The challenge lies in accurately identifying and extracting relevant economic indicators, such as inflation, GDP growth, interest rates, and geopolitical events, from this large pool of textual data, and then analyzing how these factors correlate with stock market movements.

N-gram analysis helps identify key phrases and word sequences that occur frequently in relation to economic factors like "inflation," "GDP growth," or "interest rates." By capturing these sequences, we can detect patterns and relationships that link specific economic conditions with stock market movements. POS tagging further enhances this analysis by categorizing words into parts of speech—nouns, verbs, adjectives—allowing us to pinpoint key economic terms, actions, and sentiments within the text. For example, nouns like "inflation" and verbs like "decline" help us understand the context of the discussion, while adjectives like "volatile" or "stable" provide insight into the market sentiment surrounding these factors.

By analyzing these patterns, we can develop models that link specific macroeconomic events to stock market behavior, allowing investors and analysts to make more informed predictions about market trends.

III. Scenerio

For this analysis, we collected a dataset comprising 20 articles from various online sources, focusing on financial topics. These articles primarily focus on the impact of macroeconomic factors, such as interest rates, exchange rates, inflation, and economic growth, on stock market performance. The dataset encompasses diverse perspectives from regions like South Asia, Africa, Europe, and the Middle East, providing a broad representation of global financial discourse. The goal of this analysis is to explore the structure and content of financial texts, identify key linguistic patterns, and gain insights into the most frequently discussed topics and terms.

The dataset includes articles from reputable financial websites, research papers, and news outlets, ensuring a diverse and comprehensive representation of current financial discourse. We specifically targeted articles that address market behaviors, government policies, economic indicators, and global financial trends to better understand the language used in these discussions.

Through techniques like Part-of-Speech (POS) tagging and n-gram analysis, we aim to uncover recurring phrases, patterns, and key terms that highlight the most significant topics within the financial sector. This will help in understanding how various financial concepts are represented and discussed in the texts, providing insights into trends and helping predict future movements in the market.

IV. RELATED METHODS

Natural Language Processing (NLP) provides advanced techniques to analyze and extract valuable information from unstructured text data. Among these techniques, n-gram analysis and Part-of-Speech (POS) tagging are fundamental methods that help in understanding both the structural and contextual aspects of language. These methods are widely utilized in applications like sentiment analysis, trend prediction, and linguistic pattern recognition.

A. N-gram Analysis:

An N-gram is a continuous sequence of n items (such as words, characters, or symbols) extracted from a given text. N-grams are essential for identifying recurring patterns, contextual relationships, and trends in language. They are commonly employed in natural language processing tasks like text classification, sentiment analysis, and predictive modeling.For example, consider the sentence: "The stock market is volatile":

- Unigram (n=1): ["The", "stock", "market", "is", "volatile"]
- Bigram (n=2): ["The stock", "stock market", "market is", "is volatile"]
- Trigram (n=3): ["The stock market", "stock market is", "market is volatile"]

This method enables applications like phrase detection and context-sensitive predictions by analyzing frequently occurring n-gram sequences in the text. Larger n-grams (e.g., trigrams) capture more nuanced relationships, improving the understanding of linguistic structure.

B. Part-of-Speech (POS) Tagging:

POS tagging is the process of labelling words in a text with their respective grammatical roles, such as noun, verb, adjective, or adverb, based on their context. For instance, in the sentence "The market rises," the words "The," "market," and "rises" are tagged as determiner (DT), noun (NN), and verb (VB), respectively.

POS tagging is essential for understanding the syntactic structure of text, extracting linguistic features, and analyzing patterns like frequent verb usage in financial predictions or noun occurrences in sentiment-heavy contexts.

V. SELECTED METHOD (POS TAGGING)

The POS tagging process was implemented in a Jupyter Notebook using Python, leveraging the capabilities of the SpaCy library. This environment was chosen for its interactive and efficient handling of data processing and visualization tasks, allowing seamless exploration and analysis of linguistic data.

A. Environment Setup

The implementation of Part-of-Speech (POS) tagging was carried out using the **SpaCy** library, which is a robust and widely-used NLP toolkit. First, the necessary dependencies were installed, including **SpaCy** and the pre-trained **en_core_web_sm** model. This model is designed for general-purpose English text analysis, making it suitable for tokenizing text and identifying grammatical structures. Additionally, to ensure compatibility and avoid potential conflicts between different libraries, specific versions of **numpy**, **blis**, and **thinc** were installed. The environment was further verified by confirming that **SpaCy** and **numpy** were correctly set up.

B. Text Processing

After setting up the environment, the main task was processing the input text for POS tagging. Using **SpaCy**, the text was tokenized, splitting it into individual words. Each word was then tagged with its respective POS label, such as noun, verb, or adjective. The tokens were processed in a loop, where for each token, its POS tag was retrieved and mapped to a human-readable description (e.g., “NN” for noun, “VB” for verb) using a predefined mapping dictionary. Words, punctuation, and symbols were all handled differently to ensure proper categorization, with specific conditions applied for punctuation and numbers. This step facilitated the extraction of essential linguistic features from the text.

C. Visualization

To enhance the analysis, the results were visualized in two ways. First, the data was tabulated using the tabulate library, which displayed each word, its POS tag, and the corresponding description in a clear and organized table format. This made it easier to understand the distribution of grammatical elements in the text. Next, a bar chart was generated to show the frequency distribution of POS tags, helping to identify patterns in the use of different parts of speech. The bar chart visually highlighted the most common POS tags, which is useful for understanding language trends in specific contexts, such as financial text analysis.

VI . SELECTED METHOD (N-GRAM ANALYSIS)

The identification of n-grams and trigrams was conducted using AntConc, a powerful corpus analysis toolkit designed for text analysis. AntConc enables the extraction of contiguous sequences of words, providing valuable insights into the structure and context of text.

A. Environment Setup

AntConc, a standalone tool, requires no installation of additional libraries or dependencies, making it accessible and straightforward for linguistic analysis. The text data is loaded into AntConc as a plain text file, ready for analysis.

B. Configuration of Trigram Parameters

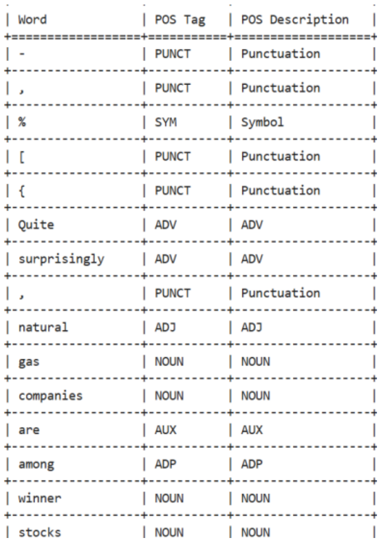
Within AntConc's N-Grams feature, the parameters were configured specifically for trigram analysis. Both the minimum and maximum N-gram size were set to 3, ensuring that only three-word sequences (trigrams) were extracted. Additional settings, such as enabling case-insensitivity and excluding stop words, were adjusted based on the analysis needs. Once the parameters were configured, AntConc processed the text data and generated a frequency list of all trigrams present in the text. For example, from the input sentence, **"The stock market is volatile,"** the tool identified trigrams like "The stock market," "stock market is," and "market is volatile." The resulting trigram list, along with their occurrence counts, provided insights into recurring patterns.

C. Visualization

Using AntConc, we generated a frequency list of trigrams, identifying the most common three-word sequences in the text. The tool’s word cloud visualization highlighted these trigrams, with their size corresponding to their frequency, offering an intuitive overview of dominant phrases. This feature was particularly useful for identifying key patterns and trends in financial text.

VII . Text Exploration

For this analysis, we gathered 20 datasets from the internet, primarily articles related to finance and economics. Each article was summarized to highlight key points and trends relevant to our study. We then applied Part-of-Speech (POS) tagging using the SpaCy library in Jupyter Notebook. The text was tokenized, and each word was assigned a POS label, which was then mapped to its human-readable description. The results were displayed in a table format, showing each word alongside its POS tag and description. This approach provided a clear view of the grammatical structure of the text, helping to identify patterns in word usage and the overall distribution of parts of speech within the articles.



```
| Word              | POS Tag   | POS Description    |
+===================+===========+====================+
| -                 | PUNCT     | Punctuation        |
+-------------------+-----------+--------------------+
| ,                 | PUNCT     | Punctuation        |
+-------------------+-----------+--------------------+
| %                 | SYM       | Symbol             |
+-------------------+-----------+--------------------+
| [                 | PUNCT     | Punctuation        |
+-------------------+-----------+--------------------+
| {                 | PUNCT     | Punctuation        |
+-------------------+-----------+--------------------+
| Quite             | ADV       | ADV                |
+-------------------+-----------+--------------------+
| surprisingly      | ADV       | ADV                |
+-------------------+-----------+--------------------+
| ,                 | PUNCT     | Punctuation        |
+-------------------+-----------+--------------------+
| natural           | ADJ       | ADJ                |
+-------------------+-----------+--------------------+
| gas               | NOUN      | NOUN               |
+-------------------+-----------+--------------------+
| companies         | NOUN      | NOUN               |
+-------------------+-----------+--------------------+
| are               | AUX       | AUX                |
+-------------------+-----------+--------------------+
| among             | ADP       | ADP                |
+-------------------+-----------+--------------------+
| winner            | NOUN      | NOUN               |
+-------------------+-----------+--------------------+
| stocks            | NOUN      | NOUN               |
```

Figure 1: Output of POS Tagging Table

Additionally, we visualized the frequency distribution of Part-of-Speech (POS) tags using a bar chart. The chart revealed that nouns had the highest frequency, followed by punctuation and adjectives.

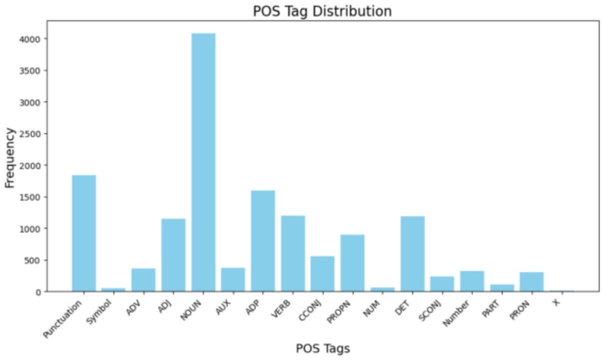
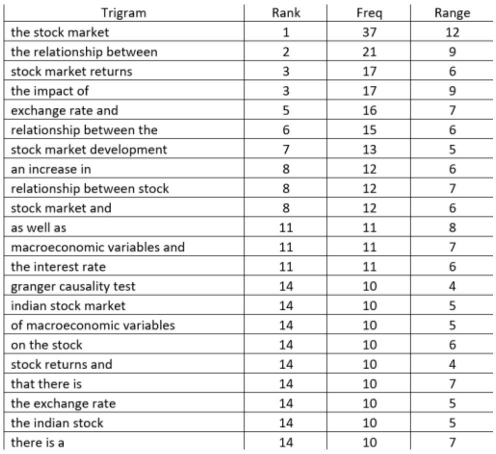


Figure 2: Bar chart showing the frequency distribution of Part-of-Speech (POS) tags.

The chart reveals that nouns have the highest frequency, with terms like "stock" and "market" appearing most frequently. This is expected as these nouns are central to financial discussions, such as "stock market," which is a key topic in the dataset. Understanding the prominence of nouns helps identify the core subjects being discussed, allowing for a more focused analysis of the main themes, such as market trends and financial indicators, in the text. This insight is valuable for refining predictive models and improving content relevance.

For the N-gram analysis, we used the same summarized dataset and set it to analyze trigrams. By focusing on trigrams, we captured three-word sequences, which provided a deeper understanding of common phrases or terminology used within the text.



Trigram	Rank	Freq	Range
the stock market	1	37	12
the relationship between	2	21	9
stock market returns	3	17	6
the impact of	3	17	9
exchange rate and	5	16	7
relationship between the	6	15	6
stock market development	7	13	5
an increase in	8	12	6
relationship between stock	8	12	7
stock market and	8	12	6
as well as	11	11	8
macroeconomic variables and	11	11	7
the interest rate	11	11	6
granger causality test	14	10	4
indian stock market	14	10	5
of macroeconomic variables	14	10	5
on the stock	14	10	6
stock returns and	14	10	4
that there is	14	10	7
the exchange rate	14	10	5
the indian stock	14	10	5
there is a	14	10	7

Figure 3: Output of Trigram Analysis Table

The output of the trigram analysis is displayed in a table that includes the trigram word, rank, frequency, and range. The rank refers to the position of a trigram in the list based on its frequency, with rank 1 indicating the most frequent trigram. In our analysis, the trigram **"the stock market"** has the highest frequency, appearing 37 times, followed by "**the relationship between"** with 21 occurrences, and **"stock market returns"** with 17 occurrences.

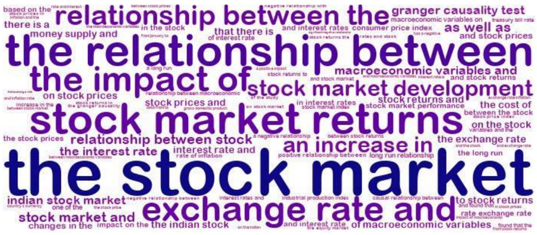


Figure 4: WordCloud Visualization for Trigram

Using the WordCloud feature in AntConc, we visualized the trigram analysis by generating a word cloud. In this visualization, the size of each trigram is proportional to its frequency in the dataset. The largest letters represent the most frequent trigrams, while smaller letters correspond to those with lower frequencies. For example, the trigram **"the stock market"** would appear in the largest font, indicating it has the highest occurrence, followed by other trigrams like **"the relationship between"** and **"stock market returns"**, with progressively smaller sizes. This visual representation makes it easy to identify key phrases that are most prevalent in the text.

In our analysis, the trigrams **"the stock market"**, **"the relationship between"**, and **"stock market returns"** appear frequently because they represent key concepts and commonly discussed topics within the dataset. The phrase **"the stock market"** is often used together as it refers to a significant financial entity, and discussions around it naturally follow this sequence. Similarly, **"stock market returns"** is a common term in financial contexts when discussing performance metrics or investment outcomes, which explains its frequency. The trigram **"the relationship between"** typically occurs when explaining correlations or relationships in financial discussions, such as between market factors, indicators, or trends.

These frequent trigrams are helpful in the analysis as they highlight the most significant topics and relationships in the dataset. By identifying these key phrases, we can gain insights into the focus areas of the text, such as market performance, relationships between economic factors, or investment strategies. This allows for a deeper understanding of the central themes in the dataset and aids in identifying trends, making the analysis more targeted and informative.

VII. CONCLUSION

In conclusion, the findings from the POS tagging and n-gram analysis provide valuable insights into the financial text dataset. The POS tagging revealed key grammatical structures, with nouns and punctuation being the most frequent, indicating the central role of tangible concepts like "stock market" and terms that define economic conditions. The frequent co-occurrence of trigrams like "stock market," "interest rates," and "exchange rate" highlights recurring themes in the dataset. These trigrams are crucial because they reflect the primary factors influencing financial discussions, such as market conditions, policy changes, and economic indicators.

By identifying these key phrases, we can better understand the language patterns driving financial conversations. The insights gained can be used to inform predictive models, sentiment analysis, or economic forecasting. For instance, recognizing that terms like "interest rates" and "exchange rate" are often mentioned together helps to anticipate trends and policy impacts, making the analysis beneficial for both financial decision-making and content creation. This deeper understanding can improve financial models, refine strategies, and guide more informed predictions based on the linguistic patterns present in the data.

REFERENCES

[1] Ali, M. (2018). Inflation, Interest and Exchange Rate Effect of the Stock Market Prices. Journal of Business and Economic Options,1(2),38-43. https://resdojournals.com/index.php/jbeo/article/view/6Traveltom.

[2] Kyereboah-Coleman, A. and Agyire-Tettey, K.F. (2008), "Impact of macroeconomic indicators on stock market performance: The case of the Ghana Stock Exchange", Journal of Risk Finance, Vol. 9 No. 4, pp. 365-378. https://doi.org/10.1108/15265940810895025

[3] Radonjić.M,& Đurišić.V,&Rogić.S,&Đurović.A.(2019)."The Impact Of Macroeconomic Factors On Real Estate Prices: Evidence From Montenegro," Ekonomski pregled, Hrvatsko društvo ekonomista (Croatian Society of Economists), vol. 70(4), pages 603-626.

https://ideas.repec.org/a/hde/epregl/v70y2019i4p603-626.html

[4] Sivasubramaniam Balagobei, Saseela & Bandara, D.. (2022). Impact of Macroeconomic Variables on Stock Market Performance: Evidence from Sri Lanka. Wayamba Journal of Management. 13. 28. 10.4038/wjm.v13i1.7551.

[5] Shoukat, iqbal, Malik, F., Kanwal, N., & Khan, N. A. (2024). The Possessions of Financial & Macroeconomic Variables ON Pakistan Stock Market A Case from Pakistan : Pakistan Stock Market. Asian Finance Research Journal (AFRJ), 6(6). https://journals.uol.edu.pk/afrj/article/view/2991

[6] Kaur, H. & Singh, Jagdeep & Gupta, N.. (2016). Impact of macroeconomic variables on stock market: A review of literature. 14. 167-196.

https://www.researchgate.net/profile/Jagdeep-Singh-26/publication/316588319_Impact_of_macroeconomic_variables_on_stock_market_A_review_of_literature/links/5f154ef1a6fdcc3ed718b20a/Impact-of-macroeconomic-variables-on-stock-market-A-review-of-literature.pdf

[7] Cherif, M., & Gazdar, K. (2010). Institutional and Macroeconomic Determinants of Stock Market Development in Mena Region: New Results From a Panel Data Analysis. International Journal of Banking and Finance, 7(1), 139–159. https://e-journal.uum.edu.my/index.php/ijbf/article/view/8403

[8] Mokhova,N,& Zinecker,M.(2014). Macroeconomic Factors and Corporate Capital Structure,Procedia - Social and Behavioral Sciences,Volume 110,Pages 530-540,ISSN 1877-0428,

https://doi.org/10.1016/j.sbspro.2013.12.897

[9] Šimáková, J., Stavárek, D., Pražák, T. and Ligocká, M. (2019), "Macroeconomic factors and stock prices in the food and drink industry", British Food Journal, Vol. 121 No. 7, pp. 1627-1641. https://doi.org/10.1108/BFJ-12-2018-0839

[10] Makan,& Chandni&Ahuja,& Kaur,A,&Chauhan,& Saakshi (2012), A Study of the Effect of Macroeconomic Variables on Stock Market: Indian Perspective.

https://mpra.ub.uni-muenchen.de/43313/?trk=public_profile_project-title

[11] Olokoyo, F. O., Ibhagui, O. W., & Babajide, A. (2020). Macroeconomic indicators and capital market performance: Are the links sustainable? Cogent Business & Management, 7(1). https://doi.org/10.1080/23311975.2020.1792258

[12] Aggarwal, P., & Saqib, N. (2017). Impact of Macro Economic Variables of India and USA on Indian Stock Market. International Journal of Economics and Financial Issues, 7(4), 10-14.

https://dergipark.org.tr/en/pub/ijefi/issue/32006/353502

[13] Omoruyi, Aigbovo & Osaretin, Andrew. (2015). THE IMPACT OF MACROECONOMIC VARIABLES ON STOCK MARKET INDEX IN NIGERIA.

https://www.researchgate.net/profile/Aigbovo-Omoruyi/publication/326403154_THE_IMPACT_OF_MACROECONOMIC_VARIABLES_ON_STOCK_MARKET_INDEX_IN_NIGERIA/links/5b4ac3bf0f7e9b4637d9f513/THE-IMPACT-OF-MACROECONOMIC-VARIABLES-ON-STOCK-MARKET-INDEX-IN-NIGERIA.pdf

[14] Marsel,&Toni,N,&Simorangkir,E,D. (2022). ANALYSIS OF THE EFFECT OF EXCHANGE RATE, INTEREST RATE, INFLATION, ANDGDP GROWTH ON PROPERTY AND REAL ESTATE STOCK PRICE INDEX LISTED ON IDX IN 2011-2019.

https://ijbel.com/wp-content/uploads/2022/07/IJBEL26.ISU-2_319.pdf

[15] Famubode,O,& Hafidh,H. (2024). The impact of the exchange rate, interest rate, and inflation on stock price. 6. 70-75.

https://www.researchgate.net/profile/Hafidh_Hafidh/publication/383266141_The_impact_of_the_exchange_rate_interest_rate_and_inflation_on_stock_price/links/66c54e66ccd355055fe172c7/The-impact-of-the-exchange-rate-interest-rate-and-inflation-on-stock-price.pdf

[16] Mazur,M,& Dang,M,& Vega,M. (2021).COVID-19 and the march 2020 stock market crash. Evidence from S&P1500,Finance Research Letters,Volume 38,101690,ISSN 1544-6123.

https://doi.org/10.1016/j.frl.2020.101690

[17] Ahmad M. Al-Kandari,&Sadeq J. Abul (2019). The Impact of Macroeconomic Variables on Stock Prices in Kuwait

https://www.ccsenet.org/journal/index.php/ijbm/article/view/0/39478