CSI 5810 Project 2: Stock Sentiment Analysis

Introduction and Domain Knowledge

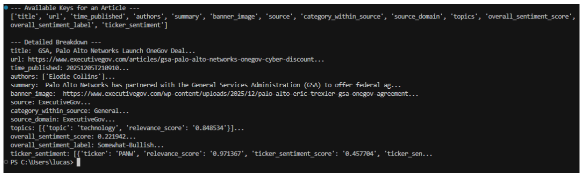
In the current market, there is much volatility surrounding certain stock options as public sentiment sways over time. This effect is recently felt particularly pronounced among tech-related stocks such as Microsoft, Nvidia, and Google, all of which are quickly becoming significant contributors to markets such as the S&P 500 and Dow Jones as well as comprising a considerable portion of the daily trade volume. Currently, tech-companies collectively hold a valuation of up to 42 Trillion Dollars [1]. As such, it is important for investors to be highly aware of changing market conditions, as well as pull on previous information to help guide future strategies.

Given the high output of data on these rapidly changing market conditions, it can be difficult to keep track of overall sentiment concerning these highly monitored stocks. Therefore, by using data scraping techniques as well as sentiment analysis tools, important analysis regarding the correlation between stock valuation and article sentiment can be determined. The articles were collected using the *'requests'* library to pull online information from AlphaVantage's online articles using its API [2]. Sentiment analysis was conducted with The Valance Aware Dictionary and sEntiment Reasoner (VADER). VADER is rule-based as well as lexicon-based, and assigns each word in the title or summary a score. This score is then compounded and used to gauge overall sentiment of the article itself. Furthermore the model takes into account other factors such as punctuation, capitalization, degree modification, conjunctions, and negation to ensure accuracy in sentiment scoring [3,4]. This was then contrasted against the built in sentiment score provided by the AlphaVantage's API to gauge effectivity and accuracy. Additionally, the *'yfinance'* library was used to pull stock-ticker information for comparison to the sentiment analysis. Other libraries used include '*tkinter'* for the GUI [5], *'pandas'* for data structuring [6], and finally *'matplotlib'* for graphing and visualization [7]. This report is structured in three parts. The first focuses on data collection and processing. This includes necessary data structures, format correction, and text aggregation. The next part focuses on the analysis of the text and sentiment score generated from each result using VADER, and comparing those results to the AlphaVantage Sentiment Analysis API. Finally, the third section contains results, conclusions, and lessons learned throughout the entire project.

I. Data Collection and Processing

A. *Collecting the data*

Data collections mainly consist of using 2 libraries and an API. The *'requests'* library pulls the articles from AlphaVantages financial feed using API calls. Afterwards, the text is obtained via a function call to the article feed. Fig 1 demonstrates the available keys for each article obtained in this way.



```
--- Available Keys for an Article ---
['title', 'url', 'time_published', 'authors', 'summary', 'banner_image', 'source', 'category_within_source', 'source_domain', 'topics', 'overall_sentiment_score',
overall_sentiment_label', 'ticker_sentiment']

--- Detailed Breakdown ---
title:  GSA, Palo Alto Networks Launch OneGov Deal...
url: https://www.executivegov.com/articles/gsa-palo-alto-networks-onegov-cyber-discount...
time_published: 20251205T210910...
authors: ['Elodie Collins']...
summary:  Palo Alto Networks has partnered with the General Services Administration (GSA) to offer federal ag...
banner_image:  https://www.executivegov.com/wp-content/uploads/2025/12/palo-alto-eric-trexler-gsa-onegov-agreement...
source: ExecutiveGov...
category_within_source: General...
source_domain: ExecutiveGov...
topics: [{'topic': 'technology', 'relevance_score': '0.848534'}]...
overall_sentiment_score: 0.221942...
overall_sentiment_label: Somewhat-Bullish...
ticker_sentiment: [{'ticker': 'PANW', 'relevance_score': '0.971367', 'ticker_sentiment_score': '0.457704', 'ticker_sen...
PS C:\Users\lucas>
```

Fig. 1. List of all keys in the article.

AlphaVantage provides sentiment scores for each article, however for the purposes of this first part, it will not be used. Instead the text will be output and analyzed by VADER. The only two relevant keys to this project are the article title and summary. The second library used to obtain data is the *'yfinance'* library, which provides current information on ticker-valuation. This is used to compare the ticker-valuation over time as sentiment shifts over time due to ever-changing market conditions.

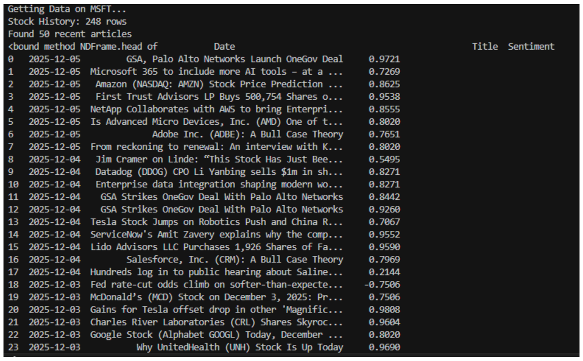
B. *Data Cleaning and Aggregation*

Due to the specific information needed for sentiment analysis, not much data cleaning had to be done in regard to articles themselves. In order to compare the two sources of information chronologically, certain reformatting had to be taken to ensure that the time returned by *'yfinance'* and the time returned by *'requests'* were compatible. This was done using the built-in *'DateTime'* library in python to ensure they could be compared properly. Furthermore data had to be carefully aggregated due to multiple articles being published on the same day. In order to preserve variation, when multiple articles were published in the same day, their mean scores used due to the comparison using the singular closing price of each ticker. Finally, to avoid headline ambiguity, the summary was also included in the data capture to ensure reliable sentiment evaluation. This was done via standard python concatenation. Finally, due to limitations in the API, when sentiment score is unavailable over a period of time, it will be assigned a score of 0 indicating neutrality. This is done to preserve the variation in the articles that are able to be accessed by the API.

II. Models Used

A. *VADER analysis*

The Valance Aware Dictionary and sEntiment Reasoner was trained on social media posts and gauges analysis by assigning each word a predetermined value downloaded from a lexicon. Beyond this, VADER takes into account grammar and other elements of conversation to modify the sentiment assigned to each word. VADER then compounds the score using sigmoid function to normalize it to between -1 and 1.



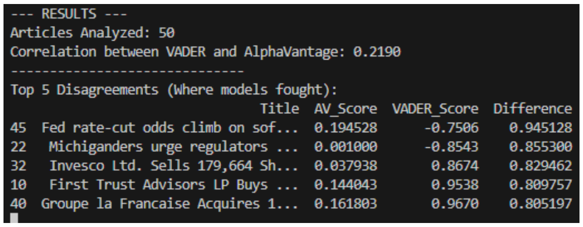
```
Getting Data on MSFT...
Stock History: 248 rows
Found 50 recent articles
<bound method NDFrame.head of          Date                                              Title  Sentiment
0   2025-12-05        GSA, Palo Alto Networks Launch OneGov Deal     0.9721
1   2025-12-05  Microsoft 365 to include more AI tools – at a ...     0.7269
2   2025-12-05   Amazon (NASDAQ: AMZN) Stock Price Prediction ...     0.8625
3   2025-12-05   First Trust Advisors LP Buys 500,754 Shares o...     0.9538
4   2025-12-05  NetApp Collaborates with AWS to bring Enterpri...     0.8555
5   2025-12-05  Is Advanced Micro Devices, Inc. (AMD) One of t...     0.8020
6   2025-12-05              Adobe Inc. (ADBE): A Bull Case Theory     0.7651
7   2025-12-05  From reckoning to renewal: An interview with K...     0.8020
8   2025-12-04   Jim Cramer on Linde: “This Stock Has Just Bee...     0.5495
9   2025-12-04   Datadog (DDOG) CPO Li Yanbing sells $1m in sh...     0.8271
10  2025-12-04   Enterprise data integration shaping modern wo...     0.8271
11  2025-12-04    GSA Strikes OneGov Deal With Palo Alto Networks     0.8442
12  2025-12-04    GSA Strikes OneGov Deal With Palo Alto Networks     0.9260
13  2025-12-04  Tesla Stock Jumps on Robotics Push and China R...     0.7067
14  2025-12-04  ServiceNow's Amit Zavery explains why the comp...     0.9552
15  2025-12-04  Lido Advisors LLC Purchases 1,926 Shares of Fa...     0.9590
16  2025-12-04         Salesforce, Inc. (CRM): A Bull Case Theory     0.7969
17  2025-12-04  Hundreds log in to public hearing about Saline...     0.2144
18  2025-12-03  Fed rate-cut odds climb on softer-than-expecte...    -0.7506
19  2025-12-03  McDonald’s (MCD) Stock on December 3, 2025: Pr...     0.7506
20  2025-12-03  Gains for Tesla offset drop in other 'Magnific...     0.9808
21  2025-12-03  Charles River Laboratories (CRL) Shares Skyroc...     0.9604
22  2025-12-03  Google Stock (Alphabet GOOGL) Today, December ...     0.8020
23  2025-12-03           Why UnitedHealth (UNH) Stock Is Up Today     0.9690
```

Fig. 2. VADER sentiment scoring for a series of articles.

Fig. 2 depicts the sentiment score of articles listed under the 'MSFT' ticker. The scores are normalized to ensure that future calculations are simple as a correlation will be drawn between sentiment and ticker valuation.

B. *AlphaVantage Sentiment Analysis*

One potential drawback of using VADER over the AlphaVantage Sentiment Analysis is that VADER is a general purpose English sentiment analyzer, whereas more specialized models are likely trained exclusively on financial articles. As such, VADER is more liable to incorrectly assign sentiment when the opposite effect may be intended by the authors. When using AlphaVantages built-in sentiment analysis, a clear conclusion could be drawn between general vs hyper-specific models. An experiment was run where the sentiment of 50 articles were analyzed by each model, and then compared to each other.



```
--- RESULTS ---
Articles Analyzed: 50
Correlation between VADER and AlphaVantage: 0.2190
------------------------------
Top 5 Disagreements (Where models fought):
                                Title  AV_Score  VADER_Score  Difference
45  Fed rate-cut odds climb on sof...  0.194528      -0.7506    0.945128
22   Michiganders urge regulators ...  0.001000      -0.8543    0.855300
32   Invesco Ltd. Sells 179,664 Sh...  0.037938       0.8674    0.829462
10   First Trust Advisors LP Buys ...  0.144043       0.9538    0.809757
40  Groupe la Francaise Acquires 1...  0.161803       0.9670    0.805197
```

Fig. 3. Comparison between AlphaVantage and VADER scoring.

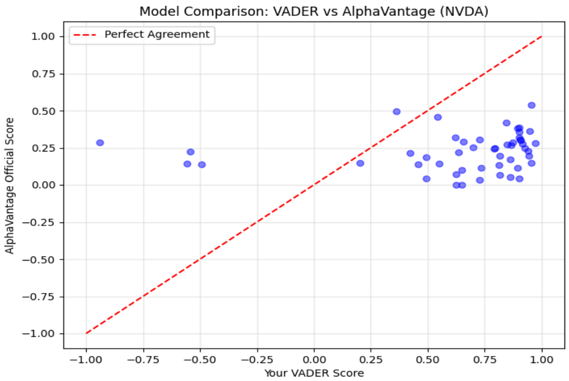
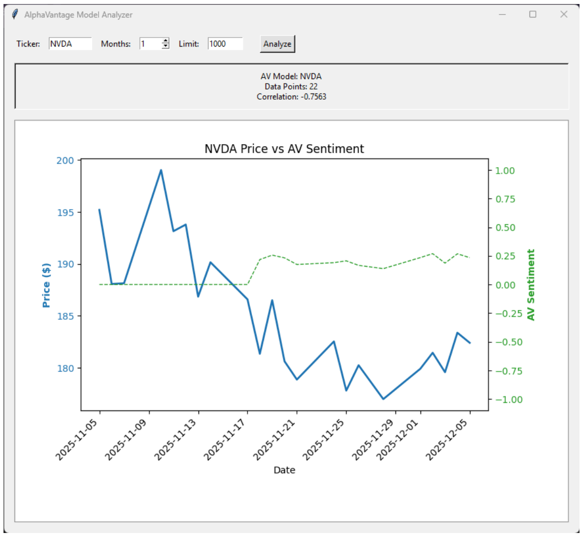


Fig. 4. AlphaVantage vs VADER Sentiment Scoring

The discrepancy between the two models is highlighted profoundly in figures 3 and 4. As visible in figure 3, those five articles had the largest differences between them, with the largest discrepancy being 0.945. The graph provided in figure 4 goes on to reveal that of the 50 articles compared, only 3 articles agreed almost perfectly. VADER tended to skew more positively than AlphaVantage. In this case, NVDA or Nvidia is being analyzed; which as of recently is a high performing stock. AlphaVantage as a model is more specific to financial change and so runs more negatively, or bullish. In this case, it rated NVIDIA sentiment as lower than VADER, which reflects a recent small drop in stock valuation. This is the main difference between the two models, as AlphaVantage measures market larger implication whereas VADER measures word and grammar choice. This is also another limitation in the API, as only 50 articles may be examined in one instance which hurts the accuracy of the correlation measure.

C. *AV vs VADER against NVDA*

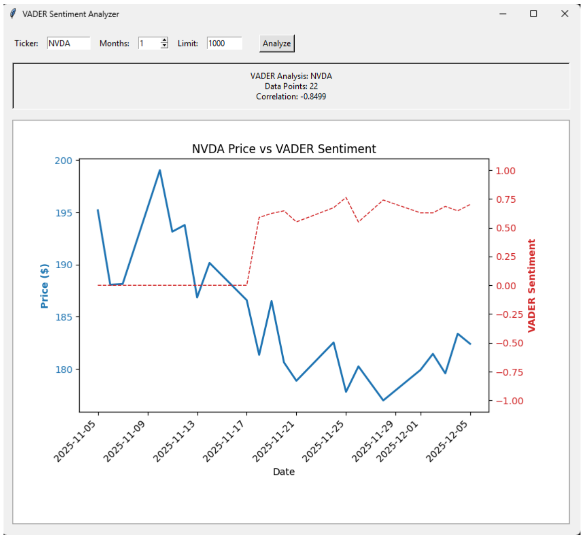


AlphaVantage Model Analyzer

Ticker: NVDA Months: 1 Limit: 1000

AV Model: NVDA
Data Points: 22
Correlation: -0.7563

Fig. 5. NVDA stock price vs AV sentiment



VADER Sentiment Analyzer

Ticker: NVDA Months: 1 Limit: 1000

VADER Analysis: NVDA
Data Points: 22
Correlation: -0.8499

Fig. 6. NVDA stock price vs VADER sentiment

Here we can see that with the same parameters, AV was more accurate to actual market conditions vs VADER sentiment analysis on its own. The conditions used in the figures were 1000 API tokens over the period of 1 month for the NVDA ticker. VADER has a correlation of -0.8499, while AV has a correlation of -0.7563. This demonstrates that the AV model has slightly more positive correlation with the relevant stock price.

III. Conclusions

A. *Overview*

In summary, the two models used both analyzed sentiment, but often differed. When VADER was run off pure lexiconical assessment, it often prescribed a positive bias, overestimating the sentiment present whereas the AV remained more bearish. That means, at least syntactically, sentiment about Nvidia remains high while financial sentiment remains high, but slightly less optimistic. This more closely matched reality and was an improvement over pure lexiconical sentiment analysis.

B. *Lessons Learned and Knowledge to Carry Over*

There was a steep learning curve during the project. Learning how to use APIs was an interesting challenge, and sentiment analysis is an extremely powerful tool that can help model and predict future events. Work I would like to continue in the future involves estimating future market conditions and seeing how accurately sentiment analysis can model and predict different market conditions. An issue that had to be overcome was the limitations of certain libraries, as well as a more thorough read of library documentation. Progress had to be reset after it was realized that a certain library would not be able to pull articles, and so the switch to the AV API was made. This required a significant restructuring of code.

Appendix A.

https://github.com/Cris-Coding/CSI-5810-Final-Project
Source Code Available Here

Appendix B.

https://drive.google.com/drive/folders/1lUA1HydjCbG-RpTOZU_27iyHswUYzxY4?usp=drive_link

Applications Available Here

References

[1]https://companiesmarketcap.com/tech/largest-tech-companies-by-market-cap/

[2]https://www.alphavantage.co/

[3]https://www.geeksforgeeks.org/python/python-sentiment-analysis-using-vader/

[4]https://github.com/cjhutto/vaderSentiment

[5]https://www.geeksforgeeks.org/python/python-gui-tkinter/

[6]https://www.geeksforgeeks.org/pandas/pandas-tutorial/

[7]https://www.geeksforgeeks.org/python/matplotlib-tutorial/