Sankalp Shubham - sxs190290
Daniel Kadosh - dxk220045

Sentiment Analysis For Stock Price Prediction

Idea: We use NLP to extract information and news about certain companies for sentiment analysis to predict the stock market to maximize investment gains.

Process:

1. Data retrieval
    a. Fetch the articles using RapidAPI's NewsNow and Google news endpoints
    b. Limit the amount of articles from one publisher to get unbiased results

2. Data Pre-processing
    a. We divide each of the articles into a list of paragraphs and sentences
        i. Each paragraph corresponds to the list of sentences through the index
    b. document structure -> document = {[url]: {'paragraphs': [], 'sentences': [[], [], ...]}}
    c. We use the nltk corpus stopwords to remove unnecessary articles or words that can hinder the sentiment scoring and only store alphanumeric characters from the text
    d. We also use a lemmatizer from TextBlob's Word lemmatizer for further refinement

```python
# CHECK THE SENTIMENT SCORES AND MAKE ADJUSTMENTS. FILTER NEWS PLATFORMS (Nvidia doc at the end has ABOUT Nvidia section (bad for scoring))
def preprocess_text(text):
    stop_words = set(stopwords.words('english'))
    words = [Word(word.lower()).lemmatize() for word in text.split() if word.lower() not in stop_words and word.isalpha()]
    return " ".join(words)
```

3. Sentiment analysis
    a. Give nltk's Vader (Valence Aware Dictionary and sEntiment Reasoner) module our pre-processed data
    b. For every document, we go through the article paragraph by paragraph and append each of their sentiment scores to a list.
    c. After iterating through all the paragraphs in the list, we average the sentiment scores of all the paragraphs to obtain a reasonable and final sentiment score for the document to determine if the article has a positive or a negative connotation about the subject of interest.

4. Final Decision
    a. After computing all the sentiment scores for a specific Company, check if its score is above or below a specific threshold to buy/sell/hold the stock
    b. Threshold numbers could be anything above 0.2 buy, below -0.2 sell, and between hold

Results & Comparison:

1. Stemmer vs Lemmatizer
2. Removing stop words vs keeping them.
3. Calculating sentiment scores using two different methods:

Sankalp Shubham - sxs190290
Daniel Kadosh - dxk220045

a. Taking an average of each sentence sentiment score and then determining the paragraph sentiment through that. Then finally averaging the sentiment scores of all the paragraphs to determine the sentiment score of the entire document. (Lower performing scores)
b. Averaging the scores of the paragraphs and skipping sentiment analysis on the sentence individually. (Better performing scores)

Outputs:

Sentiment Output on an article about recent nationwide protesting:

```
[21] clean_text2(article.maintext)

     'protest roiling college campus nationwide administrator graduation ceremony next month face dema
     sted saturday campus including indiana university arizona state university washington university
     university arrested demonstrator april sign solidarity seen protest columbia university campus ne
     ampment around prompting range response arrest criminal student suspension simply continued plea
     ion legal record follow student adult faculty member university georgia texas initiated passed la
     joe biden strong would leave ma…'
```

```
[22] sid.polarity_scores(clean_text2(article.maintext))

     {'neg': 0.242, 'neu': 0.672, 'pos': 0.086, 'compound': -0.9992}
```

1. **positive sentiment:** `compound` score >= 0.05
2. **neutral sentiment:** (`compound` score > -0.05) and (`compound` score < 0.05)
3. **negative sentiment:** `compound` score <= -0.05

NOTE: The `compound` score is the one most commonly used for sentiment analysis by most researchers, including the authors.

<u>Disclaimer:</u> Usually not the case on big articles

Compound is calculated by summing all the lexicon ratings which have been normalized between -1 and +1

Sankalp Shubham - sxs190290
Daniel Kadosh - dxk220045

Before

```
Document URL: https://about.fb.com/news/20
Sentiment: 0.018978076923076924
Document URL: https://www.cnet.com/tech/co
Sentiment: 0.1926482183908046
Document URL: https://www.forbes.com/sites
Sentiment: -0.2263
Document URL: https://news.google.com/rss/
Sentiment: 0.31420188679245287
Document URL: https://apnews.com/article/m
Sentiment: 0.19149871794871792
Document URL: https://digiday.com/marketin
Sentiment: 0.1659933779761905
Document URL: https://www.theverge.com/202
Sentiment: 0.12341413043478262
Document URL: https://www.reuters.com/tech
Sentiment: 0.0
Document URL: https://about.fb.com/news/20
Sentiment: 0.2526486842105263
Document URL: https://about.fb.com/news/20
Sentiment: 0.2135890625
```

After

```
Document URL: https://about.fb.com/news/202
Sentiment: 0.06300769230769232
Document URL: https://www.cnet.com/tech/com
Sentiment: 0.3629
Document URL: https://www.forbes.com/sites/
Sentiment: -0.2263
Document URL: https://news.google.com/rss/a
Sentiment: 0.4372075471698113
Document URL: https://apnews.com/article/me
Sentiment: 0.21880769230769231
Document URL: https://digiday.com/marketing
Sentiment: 0.23042187499999997
Document URL: https://www.theverge.com/2024
Sentiment: 0.26563478260869566
Document URL: https://www.reuters.com/techn
Sentiment: 0.0
Document URL: https://about.fb.com/news/202
Sentiment: 0.44252105263157904
Document URL: https://about.fb.com/news/202
Sentiment: 0.38873125
```

Sankalp Shubham - sxs190290
Daniel Kadosh - dxk220045

Future Implementations:

1. Identifying more areas to improve scoring
    a. example: avoiding company news websites, finding more diverse news outlets
    b. remove news outlets that can't be processed correctly
2. Adding the Gemini model for final analysis for final decision making