Technology Review Presentation

$Dinero$ (A stock analysis Platform)

Cindy Lyu
Navya Edula
Shivam Agarwal
Yasovar Tammareddy

Background & Use Case

- Our objective is to create a stock analysis platform that seamlessly integrates historical stock data from five chosen companies while incorporating essential technical indicators. Additionally, we aim to perform sentiment analysis on news headlines/events pertaining to any events during sudden fluctuations in stock prices.
- **Use Case:** Perform sentiment analysis on financial news headlines/articles.
- Python has an extensive set of accessible, well maintained, and open source libraries for the purpose of Natural Language Processing (NLP). Since NLP models are complex and need to be extensively training on data, we can leverage the existing libraries in Python and pre-trained models in Python and this would save us a lot of time

Python Package Choices

- **NLTK**
  - Natural Language Toolkit
  - Suite of open source modules and datasets widely used for Natural Language Processing
  - Authors: Steven Bird, Edward Loper, Ewan Klein
- **Spacy**
  - SpaCy is a cutting-edge Python library tailored for efficient and practical NLP tasks.
  - Maintained by Explosion AI, a company focused on AI and NLP. It is open source and available on GitHub, allowing for community contributions and improvements.
  - Matthew Honnibal and Ines Montani, pioneers in computational linguistics, created SpaCy.
- **VADER**
  - VADER (Valence Aware Dictionary and sEntiment Reasoner)
  - Authors: C.J. Hutto and Eric Gilbert
  - It is fully open-sourced under the [MIT License]
  - A lexicon and parsimonious rule-based sentiment analysis tool that is specifically attuned to sentiments expressed in social media.

Background on NLTK

NLTK NLP models typically demand pre-processed text, often tokenized strings. Common steps include...

1. Removal of stopwords and Punctuation
    a. High frequency of stopwords increase computation time
2. Lowercasing the string
3. Tokenization
    a. Converting each string to a list of words
4. Stemming and Lemmatization *(optional)*

Since our project doesn't require complex deep learning models, NLTK's efficient library design makes it a great fit for our use case.

However, going through the traditional route by using NLTK pre-processing and building a model on top of it requires a lot of pre-processing!

Package Comparison - SpaCy

SpaCy offers advanced features with minimal setup:

1. Automatic Text Processing: Tokenization, stopwords/punctuation removal, and text normalization.
2. Efficiency: SpaCy is faster, ideal for large datasets, while NLTK is comprehensive, suited for detailed preprocessing.

SpaCy is favored for its rapid processing in demanding industrial applications but it has its drawbacks:

It’s emphasis on analyzing text structure rather than emotional content can hinder its effectiveness in the sentiment analysis of titles, affecting the overall insight gained from news titles.

Requires considerable expertise in NLP to tweak the custom sentiment analysis model of Spacy to take into account the emotional meaning of text.

Package Comparison - VADER

VADER offers features that fulfill our specific demands with minimal setup.

1. VADER is specialized in analyzing the polarity of a given piece of text
    - It computes a compound sentiment score for a given piece of text by aggregating the polarity scores of individual words
    - The compound score represents the overall sentiment of the text, ranging from -1 (extremely negative) to 1 (extremely positive), with 0 indicating a neutral sentiment.
    - It tells about the semantic orientation as well as how positive or negative a sentiment is. (more in demo)
2. Its rule-based heuristics ensures its accuracy and effectiveness in sentiment analysis of social media.
    - The model is based on a pre-built lexicon (a list of words and their polarity scores that indicate whether they are positive, negative, or neutral)
    - It can recognize context-specific nuances and adjust sentiment scores accordingly to their intensity and the grammatical rules surrounding them.
    - It is designed to handle informal language (emojis, emoticons, and slang) commonly found in social media.

Our Choice - VADER!

We were looking for a pre-trained model that would require no/minimal data preprocessing of the string, easy to use and would save us time.

Most of the NLP models in the NLTK library require pre-processing of strings which can be a tedious process. For example, to use Spacy, we must remove stopwords, punctuation, convert the letters to lowercase, and finally tokenize them too.

While VADER can take strings of the text directly and adjust sentiment scores accordingly with pre-trained syntactic and grammatical rules surrounding them.

Drawbacks/Remaining Concerns

1. **Black box nature**
    a. it difficult to understand why it assigns certain sentiment scores
    b. limits opportunities for fine-tuning or customization.

2. **Limited vocabulary**
    a. relies on a predefined lexicon of words and phrases
    b. may not capture the nuances, especially slang, or sarcasm

Package Comparison - Summary

VADER

Pros:

1. Minimal Setup: the analyzer can process the original string with built-in text processing functions.
- Compatibility: VADER returns semantic polarity of the text with enough details AND is compatible with our data source: news/articles from social media.
- Availability of Examples & Readable Documentation

Cons:

- Tradeoff between accuracy and efficiency (more in drawbacks)

SpaCy / NLTK

Pros:

- Well-known pre-trained model
- Computational Efficiency

Cons:

- Require pre-processing of input text
- Lack in Compatibility with the project
- Complex Documentation

Package Demonstration - VADER

[1]:
```
from vaderSentiment.vaderSentiment import SentimentIntensityAnalyzer
```

[2]:
```
titles = ["Dow Jones Futures: Fed Meeting On Deck; AI Stock Supermicro Surges 10% On Earnings Beat.",
"Tech earnings, JOLTS data: What to Watch.",
"Stocks Pick Up Afternoon Steam To End At Day's Highs; Dow And S&P Close At New Records",
"Microsoft names 'Call of Duty' executive Johanna Faries as Blizzard's president.",
'These Stocks Are Moving the Most Today: SoFi, iRobot, Tesla, Archer Daniels, McGrath RentCorp, ZoomInfo, and More',
'Cathie Wood Likes UiPath, IBD Stock Of The Day, As AI Play. But Some Analysts See Microsoft Risk.',
'Forget The Magnificent Seven. Focus On These Fab Five.',
'Magnificent Seven Stocks To Buy And Watch: Nvidia, Tesla Rally']
```

[3]:
```
analyzer = SentimentIntensityAnalyzer()
for s in titles:
    vs = analyzer.polarity_scores(s)
    print("{:-<65} {}".format(s, str(vs)))
```

Package Demonstration - VADER

```
Dow Jones Futures: Fed Meeting On Deck; AI Stock Supermicro Surges 10% On Earnings Beat. {'neg': 0.0, 'neu': 1.0, 'pos': 0.
0, 'compound': 0.0}
Tech earnings, JOLTS data: What to Watch.----------------------- {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'compound': 0.0}
Stocks Pick Up Afternoon Steam To End At Day's Highs; Dow And S&P Close At New Records {'neg': 0.0, 'neu': 1.0, 'pos': 0.0,
'compound': 0.0}
Microsoft names 'Call of Duty' executive Johanna Faries as Blizzard's president. {'neg': 0.0, 'neu': 1.0, 'pos': 0.0, 'comp
ound': 0.0}
```

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