TradeRight - Create, Share, Follow

Team Alpine (Team #100)
ajain463, amadathil7, donaldvu, egirishekar3, ssoni41, wganesh @gatech.edu

INTRODUCTION

Currently it is difficult for an investor to extract gains from stock market trading without an exhaustive market research.
The plethora of information responsible for driving the market sentiment makes things more difficult.
Personal research on stock picks often devolves into gut/emotional decisions.
Also the fact that wealth managers today charge an exorbitant fee makes it impossible for an average investor to participate in stock market trading.
This brings a need for a cost-effective platform that integrates all information driving a stock's price and applies accurate algorithms, thereby predicting a stock's performance and guiding an investor.

TradeRight mines stock related news and financial data from reputable websites and applies the latest research in language models and machine learning techniques to provide an investor with a clear signal to buy or sell a stock.
It also provides an interface where informed investors can share their portfolio development strategies with fellow investors.

DATA

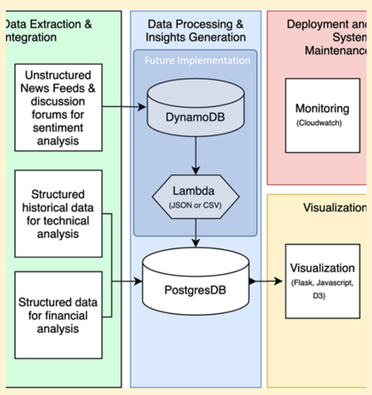
Data Collection:
Data collected is of three types:

Financial Data	News Data	Technical Data
Data Source: Yahoo	Data Source: News API	Data Source: Yahoo
1. 6000+ ticker symbols 2. 150 MB data 3. 900K + rows	1. S&P 500 news 2. 20MB data 3. 25K + news articles	1. 6000 + stock data 2. 150 MB 3. 1.2 million records

Data is downloaded using the above APIs. Post data collection, it was sanitized and processed.
Duplicate data and outliers were removed.
All records were converted into the same format (e.g formats of timestamp fields & character encoding).
Stock technical data was normalized.
Data was split into 80/20 train and test set and we used the train data for cross validation.
The train and test data were temporally split.

APPROACHES & INTUITION

Algorithm



We mine financial and news related data for stocks and feed them into our insight generation engine which predicts buy/sell signals for the stocks. These signals are then input into our visualization module.

Insight Generation Approaches
We make use of multiple approaches in determining the sentiment of text articles to determine buy/sell signals.

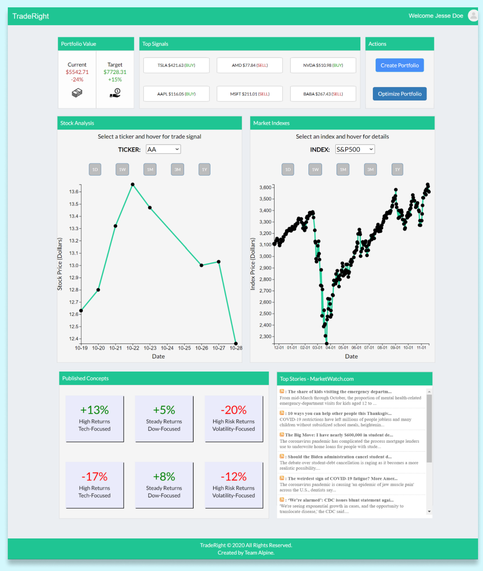
Baseline – VADER
VADER makes use of sentiment lexicon and rules to find out the sentiment of sentences and is mainly tuned for text appearing in social websites. VADER not only provides the classification of the sentence but also provides the degree of sentiment with a score. We initially just use VADER to extract the sentiment of the news articles and find best threshold for classification by doing the ROC sweep and use the threshold to report the metrics on held out test set.

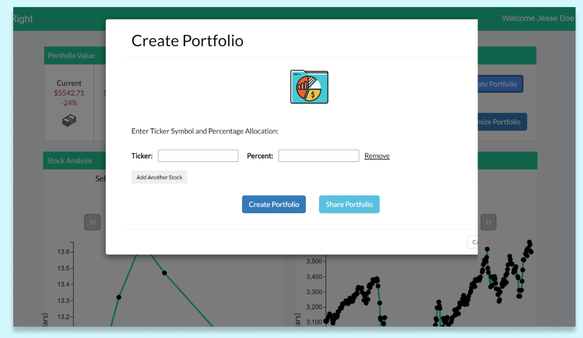
VADER + Technical Indicators(TI)
In this approach we use technical indicators along with VADER scores to predict buy/sell label. The technical indicators being used include 2/5 days SMA, volumes and dividend splits. Using these as the features we build multiple classifiers including logistic regression, multi-layer perceptron and ensemble techniques such as random forests, ada boost classifiers and gradient boosting trees.

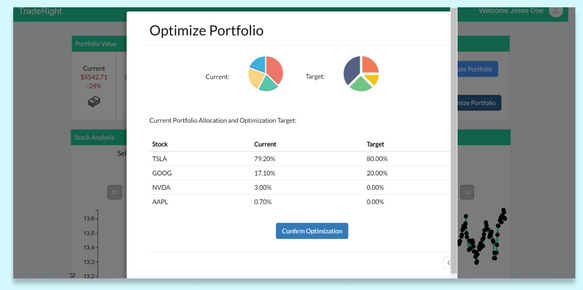
RoBERTa
Lastly, we explore latest language models (LM) to see if we can get similar performance by just using the news headlines alone. We make use of RoBERTa an optimised version of BERT. RoBERTa is trained on larger training data, longer sequences and uses dynamic masking pattern unlike BERT which uses static masking pattern. We finetune this LM to classify the news article to indicate buy/sell signal.

Intuition
The sentiment of news articles plays major role in people transacting in stock market as seen from our literature survey. While news articles is one aspect there are lot of latent factors which determine the movement of stock prices. In our approaches we make use of different NLP techniques which have been shown to perform good for sentiment detection and combine multiple technical indicators to extract the buy/sell label. LM are sparsely used for predicting the stock price movement as evidenced in our survey. Here we make use of RoBERTa LM for the purpose of predicfing the buy/sell label from news headlines.

Interactive Visualization







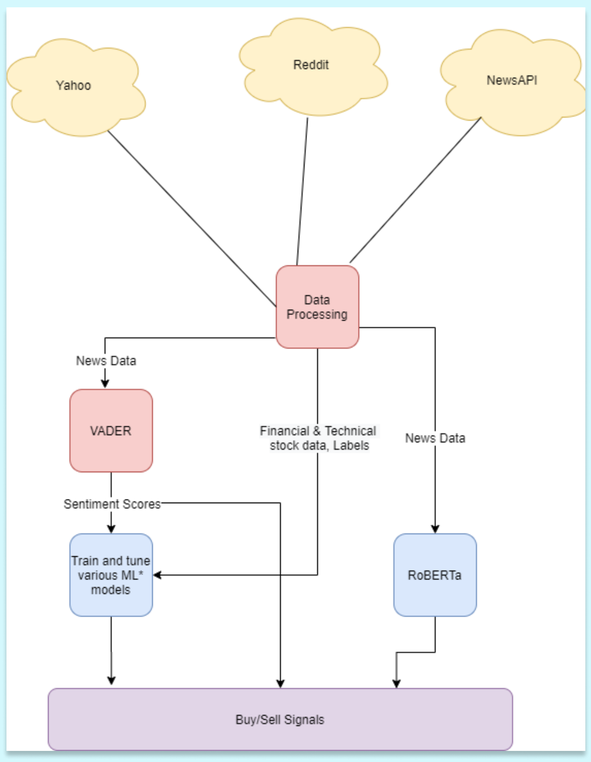
Approaches:
Key user-experience approaches are focused on enabling users to share and achieve optimal returns for their portfolio.
We use Javascript, Flask and D3.js to create the interactive user experience.
The buy/sell signals generated by the insight generation module are fed into the user interface and appear in the UI for the user to take action.

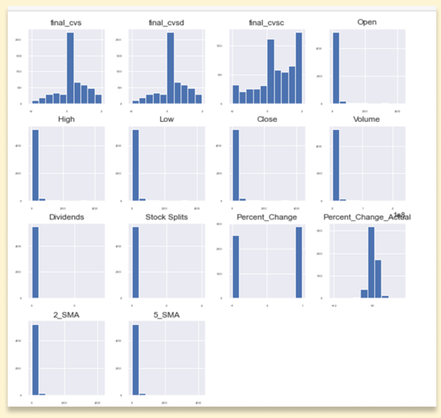
Key Visualization Components:
1. Guide investor on stock picks by providing buy/sell signals for top stocks
2. Enable the investor to quickly discern the gain improvement pattern
3. Enable investors to improve revenue stream by collaborating with others
4. Enable investors to follow published investment strategies and improve their portfolio

Intuition
Currently there are many intermediaries between investors and the stocks such as brokers, mutual fund managers. The expertise is scarce and not every successful investor is working for a financial institute. The intuition behind the concept is to break down unnecessary barriers in financial markets so that investors sitting at the desk can directly build the optimal portfolio for their needs. The key new concept in the approach is the collaboration which is backed by data mining and data science, effectively democratizing investment function of the industry.

EXPERIMENTS & RESULTS

Data Gathering
- We explore multiple avenues to gather stock related news, financial and technical data.
- For technical and financial data we mainly rely on Yahoo API.
- For news data we explored Reddit, NEWS and Bing API.
- While reddit API provided us with lot of textual data related to stock market it is hard to associate it with specific tickers.
- For example, news articles related to Microsoft can have terms like 'MSFT', 'Microsoft', 'Nadella'. Its hard to find exhaustive list of terms related to a ticker.
- Similarly BING API provided data in raw format which could not be associated with a ticker in reliable way.
- Eventually we made use of NEWS API which provides news articles related to specific tickers by taking tickers itself as one of the argument.
- Using NEWS API we gathered news related to various tickers and merged this data with the financial data to create our dataset.





Post data collection and normlization, we visualized the distribution of these features.

Label Generation
- We experimented with using 2-SMA/5-SMA as the technique to assign buy/sell labels, but did not get desired distribution.
- The thing which worked for us was to make use of open and close signals to come up with labels.
- If the open is higher than close we would assign a label of sell so a profit is made and we take similar approach of assigning buy signal. - Although this is a straightforward approach, we still wanted to see how it affects our models' performance.
- Generally news articles tend to have an immediate effect on the stock's price and it reflects this label gathering strategy.
- As the project matures, we would like to test more sophisticated ways of assigning labels.

Hyper Parameter tuning
- Apart from the baseline VADER experiments we do extensive hyper parameter tuning for rest of the models. We explore different combination of hyper parameters including number of weak estimators, maximum number of features and depth of weak learners among others.
- For RoBERTa as well we explore multiple combination of data and hyperparameters. For the purpose of using RoBERTa for classification we add a linear classification head on top of the embeddings and train RoBERTa.
- We experiment by training with title alone and title and description of news articles.
- Adding description degrades the performance of the model indicating description might contain not very relevant information and can act as noise.
- Since RoBERTa is huge in number of parameters and we have comparatively very less data we finetune pretrained RoBERTa using very low learning rate of 2e-5.
- We stick to 3 epochs and increasing this leads to overfitting.

Results

For the purpose of evaluation of our techniques we make use of separate unseen test data which is not used for training. Test set is created by temporally splitting the data so we can use the historical news articles to train and predict for the upcoming days. We report the metrics on this data for the purpose of evaluation.

Method	Test Accuracy
VADER (Text Only)	52.89%
Logistic Regression (Text + TI)	54.89%
Multi-Layer Perceptron (Text + TI)	53.91%
Random Forests (Text + TI)	54.27%
AdaBoost (Text + TI)	55.25%
Gradient Boosting Trees (Text + TI)	60.88%
SVM (Text + TI)	57.21%
RoBERTa (Text Only)	**61.81%**

We observe unsupervised VADER to provide an accuracy of 52.89%. Different ensemble and supervised learning approaches improve upon this and provide accuracy in the range 53.91% to 60.88%. Another thing to note is these supervised learning algorithms along with sentimental scores make use of financial data as well and improve upon the baseline. Gradient Boosting trees achieves an accuracy of 60.88% and significantly improves upon the baseline VADER. We also train RoBERTa with news headlines only and it achieves a test score of 61.81%. The interesting part is the LM performs best text data alone which highlights the capability of pretrained LM.

We use the algorithms described in multiple papers including SVM, multi-layer perceptrons for the purpose of predicting buy/sell signal. Since we do not have access to the data described in the papers we surveyed, we compare the proposed approaches with our approaches on the dataset which we have created. We observe GradientBoosting trees which make use of both technical indicators and news articles to outperform the SVM and multi-layer perceptron model proposed in multiple papers. Similarly, RoBERTa outperforms all of the explored approaches as indicated in results table and we believe using latest language models for financial applications has tremendous potentia