

An abstract geometric design on the left side of the slide. It features a dark blue background with various geometric shapes and patterns. A white circle is positioned near the top left. Below it, a light blue semi-circle is visible. To the right of the semi-circle, there are concentric circles. Further down, there are several overlapping squares and triangles in shades of blue, purple, and pink. Some of these shapes contain patterns of concentric lines or dots. A white diagonal line runs from the top left towards the bottom right, separating the abstract design from the text area.

RELATING SOCIAL MEDIA TO STOCK MOVEMENT



INTRODUCTION

- Risk is related with uncertainty and so is stock market.
- Here we will Analyse by considering all of these things and try to predict the exact price fluctuation.
- The objective is to find strong correlation between the movement of the stock prices and sentiment on Social media .

MOTIVATION & RELATED WORK

Efficient Market Hypothesis- outperforming the market in the long run is near impossible.

- We observe whether efficient market hypothesis holds through the discovery of Analysing / Wallstreetbet market sentiment for reddit.
- Perhaps we even contribute to the hypothesis itself with empirical evidence.
- According to (Granger & Timmermann, 2002), degrees of accuracy of 56% hit rate are often reported as satisfying results for stock prediction.
- But if we relate it with topic sentiment and analyse it through different ML models accuracy can be increased.

DATASET AND FEATURES

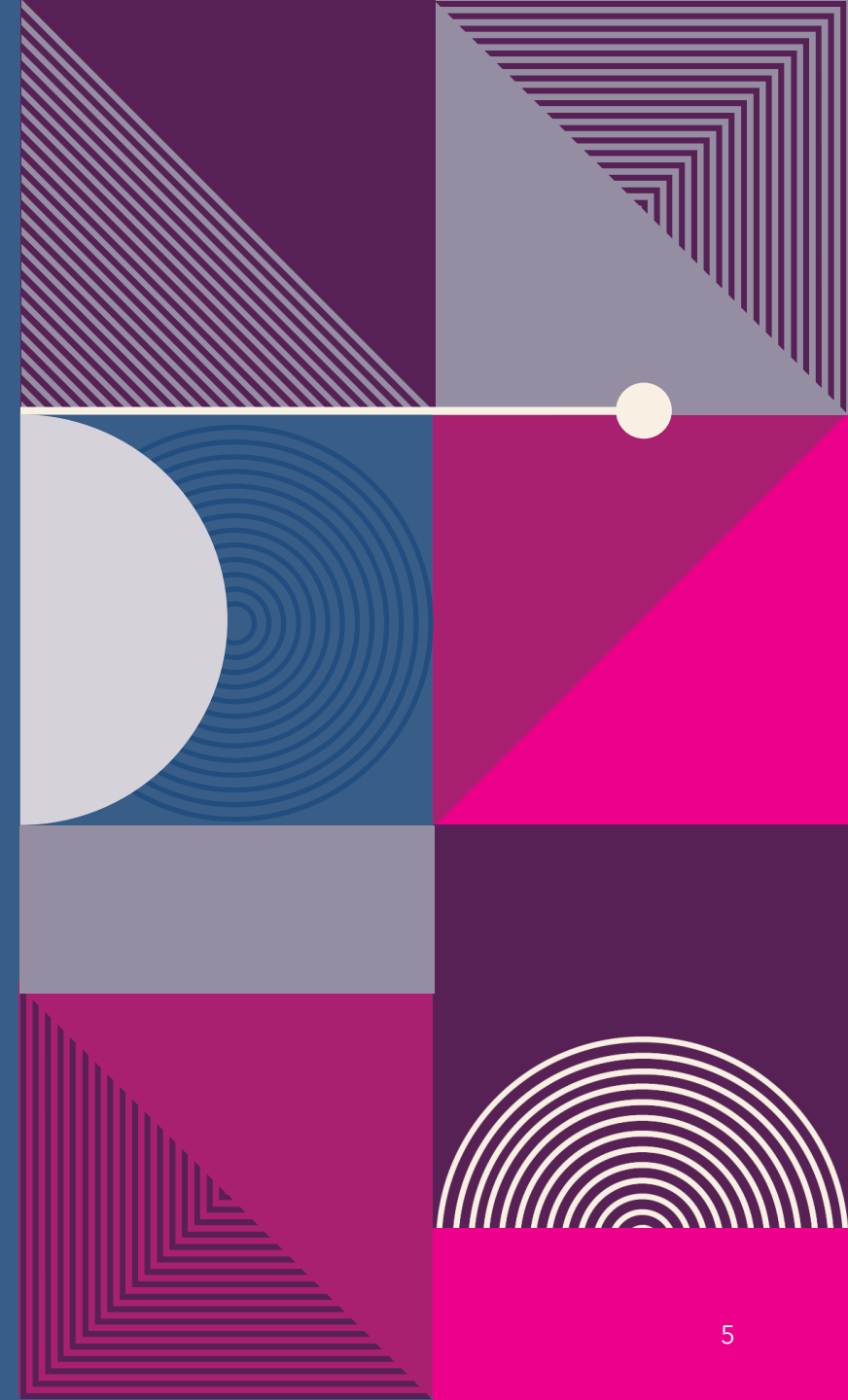
For data processing, we pulled data primarily from wallstreetbets.

- 0 represented "strong buy", where the stock had a percentage increase of over 3 percent, 1 represented "buy", where the stock increased between 1 and 3 percent, 2 represented "hold", where the stock increased between 0 and 1 percent, 3 represented "sell" where the stock decreased between 1 and 3 percent, and 4 represented "strong sell", where the stock decreased by over 3 percent.
- The data was crawled through Reddit API.
- Then, in order to get our trailing and forward 5-day stock price change, we cross-referenced this ticker We cross referenced the data through web scrapping of Yahoo finance API's historical data.
- Overall, we had about 181,000 datapoints which we randomly split into train, dev, and test data with an 80/10/10 split.

METHODS

we break our model down into three main components:

- embedding/encoding layer,
- sentiment analysis,
- multimodal output layer.






EMBEDDING

To preprocess our data, we convert raw text into vectors using pre-trained GloVe embeddings, and then we use these embeddings to build a vocabulary. We use GloVe embeddings with 6B tokens and 100-dimensional vectors, and by using these embeddings.



SENTIMENT ANALYSIS

Our sentiment analysis model, SentimentLSTM, predicts whether a given text has negative (0) or positive (1) sentiment.



• THIS MAKES BILSTMS PARTICULARLY POWERFUL FOR LANGUAGE PROCESSING, AS SHOWN BY OUR HIGH ACCURACY ON OUR SENTIMENT ANALYSIS

• THE PREDICTIONS SO THAT THEY LIE BETWEEN 0 AND L, AND THE LOSS OF A SINGLE SAMPLE (x, y) IS DEFINED AS

• BCEWITHLOGITS = $-\left[y \cdot \log A(x) + (1 - y) \cdot \log(L - A(x))\right]$;

• IN OUR PRETRAINING CASE, WE WERE ABLE TO ACHIEVE A SENTIMENT PREDICTION ACCURACY OF 99.5%.

$a(\exp)$

MULTI MODAL OUTPUT LAYER

our multimodal output layer takes our output of our sentiment analysis, and concatenates it with the score from the last five trading days along with net upvotes/downvotes

To recall, we define our ELU non-linearity as

$$\text{ELU}(x) = a(\exp(x) - 1)$$

where a is a hyperparameter in our neural network. Furthermore, our softmax layer is, given an input vector z .

$$\sigma(\vec{z})_i = \frac{e^{z_i}}{\sum_j e^{z_j}}.$$

EXPERIMENTS, RESULTS, AND DISCUSSION

and the proximity from the prediction bucket to the correct label.

More formally, accuracy and closeness are defined as

correct predictions

acc =

total predictions

1. $\frac{\text{# correct predictions}}{\text{# total predictions}} = \text{acc}$
2. $\frac{\text{# correct predictions}}{\text{# total predictions}} = \text{acc} + 1$

$$\begin{aligned} &= \frac{TP + 1}{TP + FP + 1} \\ &= \frac{TP + 1}{TP + FN + 1} \\ &= \frac{2(\text{precision}_i)(\text{recall}_i) + 1}{2} \end{aligned}$$

In addition to our self-defined accuracy and closeness scores, we also used the standard classification metrics: precision, recall, F 1, and Matthews Correlation Coefficient .

EXPERIMENTS AND RESULTS

After training all of the models, we ran both baselines (linear regression and neural network without sentiment) and our proposed model (neural network with sentiment) using the optimal hyperparameters $LR = 0.001$, dropout probability 0.5, and $a = 1.0$ on the test set.



TABLE 1. EVALUATION METRICS WITH DIFFERENT LEARNING RATES AND A

	Precision	Recall		MCC
Strong Buy	0.313	0.245	0.481	0.001
Buy	0.582	0.009	0.501	0.394
Hold	0.296	0.013	0.304	-0.128
Sell	0.411	0.003	0.342	0.681
Strong Sell	0.139	0.001	0.501	0.427

NEURAL NETWORK W/OUT SENTIMENT BASELINE METRICS

	Precision	Recall		MCC
Strong Buy	0.474	0.419	0.733	0.590
Buy	0.494	0.406	0.505	1.0
Hold	0.612	0.351	0.602	1.0
Sell	0.5	0.034	0.668	-0.019
Strong Sell	0.363	0.671	0.731	0.441

LINEAR REGRESSION BASELINE METRICS

	Precision	Recall		MCC
Strong Buy	0.043	0.524	0.153	0.190
Buy	0.384	0.004	0.273	-0.451
Hold	0.591	0.001	0.501	-0.102
Sell	0.326	0.801	0.641	0.568
Strong Sell	0.219	0.001	0.501	0.391



CONCLUSION AND FUTURE WORK

Finally, we noticed that our sentiment analysis model had an accuracy of around 99 percent, it is unlikely that our sentiment analysis component is erroneous. Considering the unpredictable nature of the stock market and the lack of additional information we provide our model, we can improve our model by providing more information. For example, instead of providing a single Reddit post and mapping it to a label, we can feed in the concatenation of multiple Reddit posts from the same day. (Note that this would demand significantly more time and computational power, as this would imply that each Daily Discussion thread could contribute at most one data point for each ticker.) We could also provide additional information by including the trailing change over the past 30 days or frequency of mentions of the ticker in the Daily Discussion thread to the Multimodal Output Layer number of upvotes

An abstract geometric design on the left side of the slide. It features a dark blue background with various geometric shapes and patterns. A white circle is positioned near the top left. Below it, a light blue semi-circle is visible. To the right of the semi-circle, there is a pink triangle with diagonal lines. Below the semi-circle, there is a pink square with a pattern of concentric lines. To the right of the square, there is a light blue triangle. Below the square, there is a pink triangle. To the right of the triangle, there is a dark blue triangle. The overall design is modern and minimalist.

THANK YOU