Stock Price Prediction Leveraging Reddit: The Role of Trust Filter and Sliding Window

Dennis Huynh, Department of Electrical & Computer Engineering

Garrett Audet, Faculty of Commerce

Nikolay Alabi, Faculty of Arts & Sciences

Yuan Tian, School of Computing

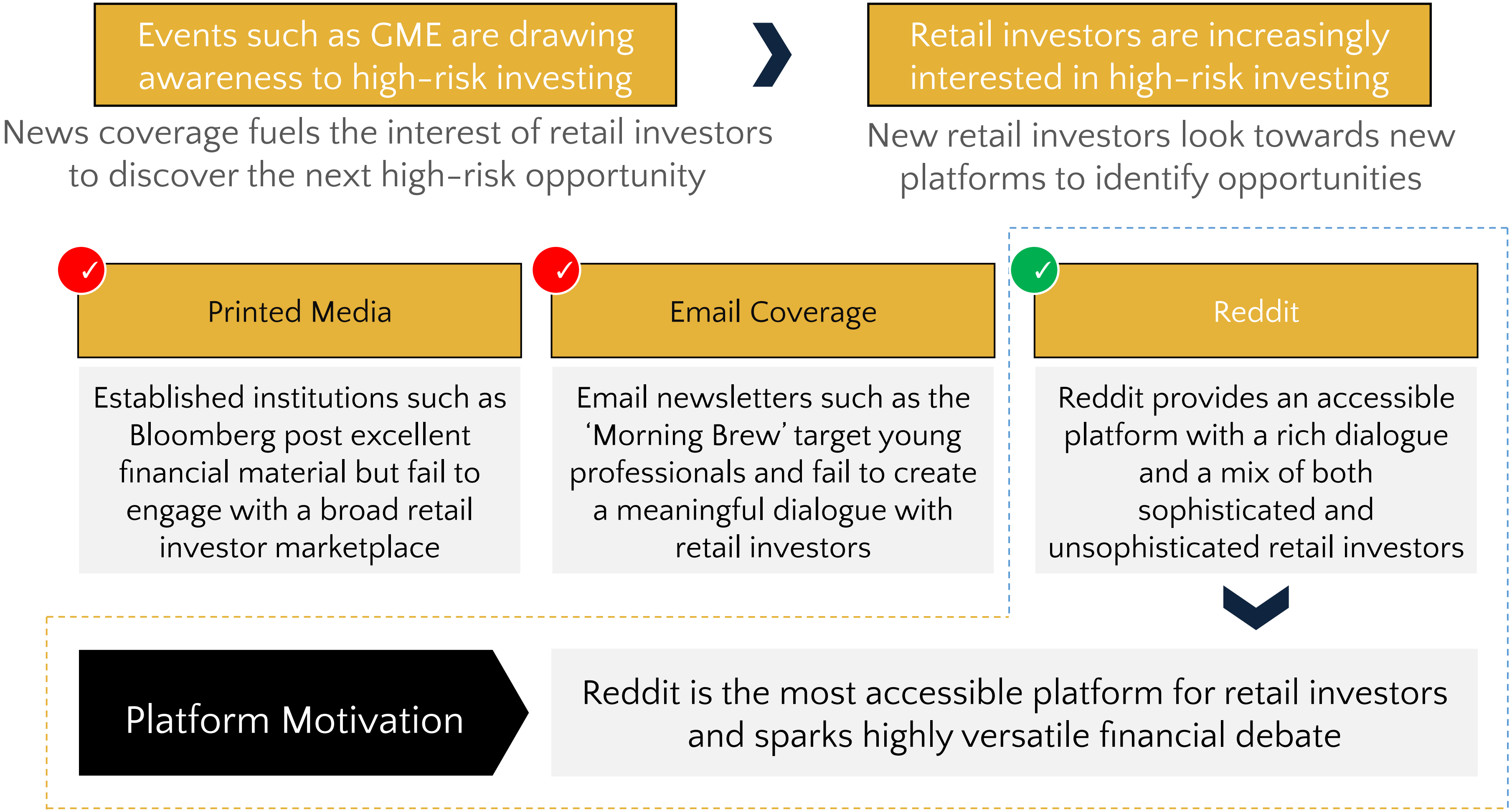
Contact: {dennis.huynh, y.tian} @queensu.ca

2021 IEEE International Conference on Big Data

Roadmap

1. Introduction
2. Methodology
3. Experiments
4. Results
5. Conclusion

Stock .. On Reddit



Events such as GME are drawing awareness to high-risk investing

News coverage fuels the interest of retail investors to discover the next high-risk opportunity

Retail investors are increasingly interested in high-risk investing

New retail investors look towards new platforms to identify opportunities

✓ Printed Media

Established institutions such as Bloomberg post excellent financial material but fail to engage with a broad retail investor marketplace

✓ Email Coverage

Email newsletters such as the ‘Morning Brew’ target young professionals and fail to create a meaningful dialogue with retail investors

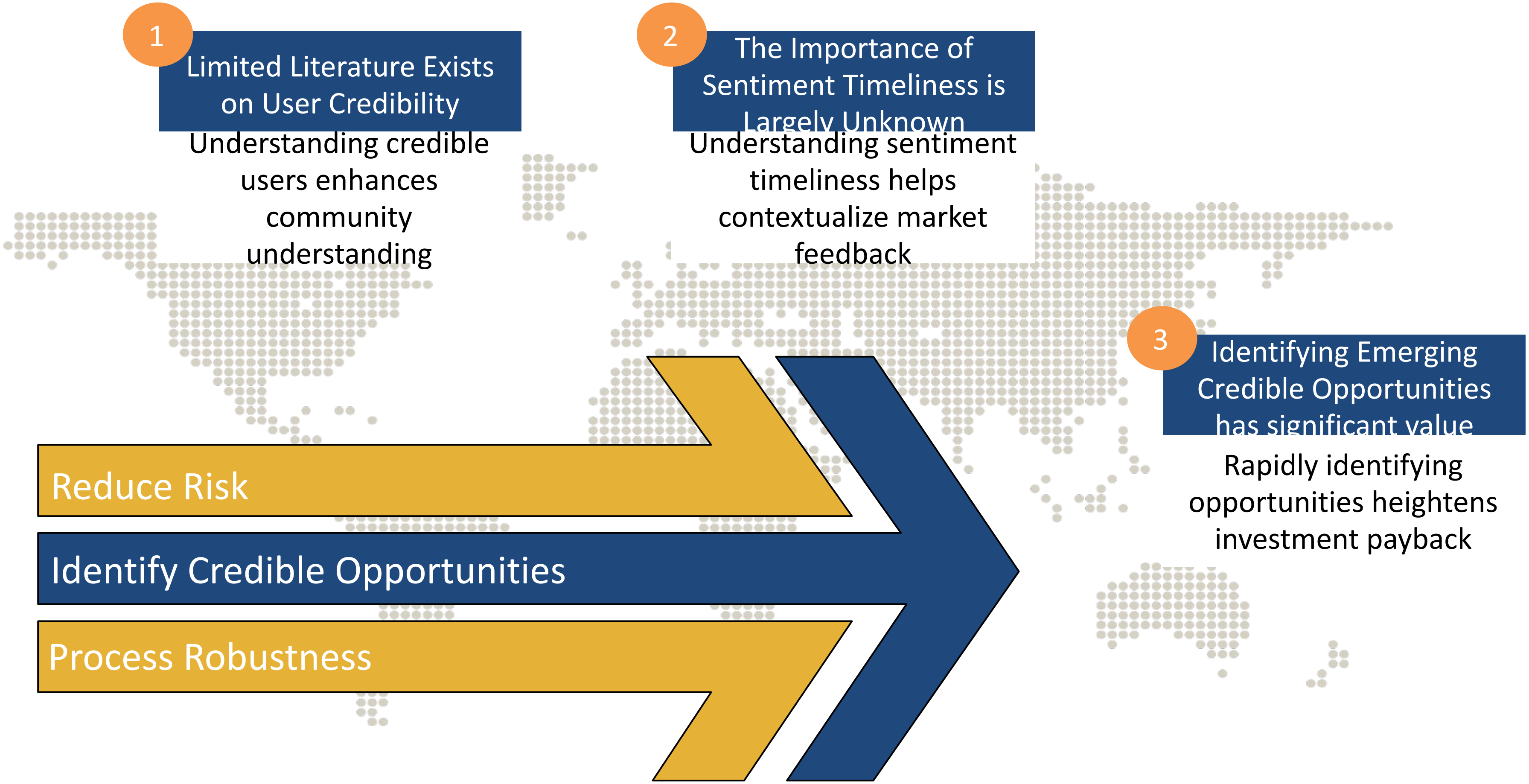
✓ Reddit

Reddit provides an accessible platform with a rich dialogue and a mix of both sophisticated and unsophisticated retail investors

Platform Motivation

Reddit is the most accessible platform for retail investors and sparks highly versatile financial debate

Research Importance



1 Limited Literature Exists on User Credibility

Understanding credible users enhances community understanding

2 The Importance of Sentiment Timeliness is Largely Unknown

Understanding sentiment timeliness helps contextualize market feedback

3 Identifying Emerging Credible Opportunities has significant value

Rapidly identifying opportunities heightens investment payback

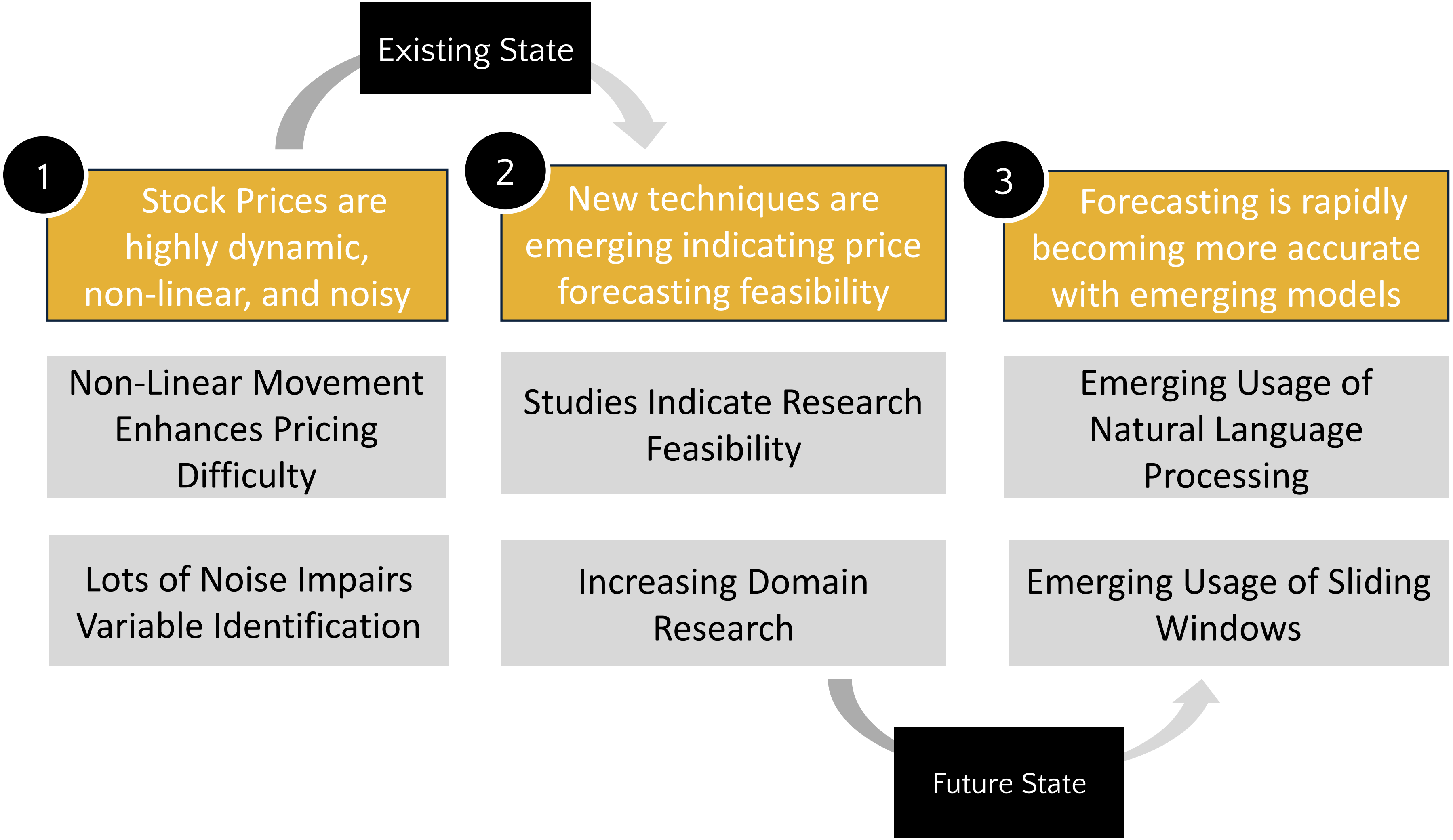
Reduce Risk

Identify Credible Opportunities

Process Robustness

Through focusing on the most credible users – we seek to facilitate a methodology that identifies most statistically credible opportunities – resulting in several added benefits to investors.

Stock Price Forecasting



Existing State

1 Stock Prices are highly dynamic, non-linear, and noisy

2 New techniques are emerging indicating price forecasting feasibility

3 Forecasting is rapidly becoming more accurate with emerging models

Non-Linear Movement Enhances Pricing Difficulty

Studies Indicate Research Feasibility

Emerging Usage of Natural Language Processing

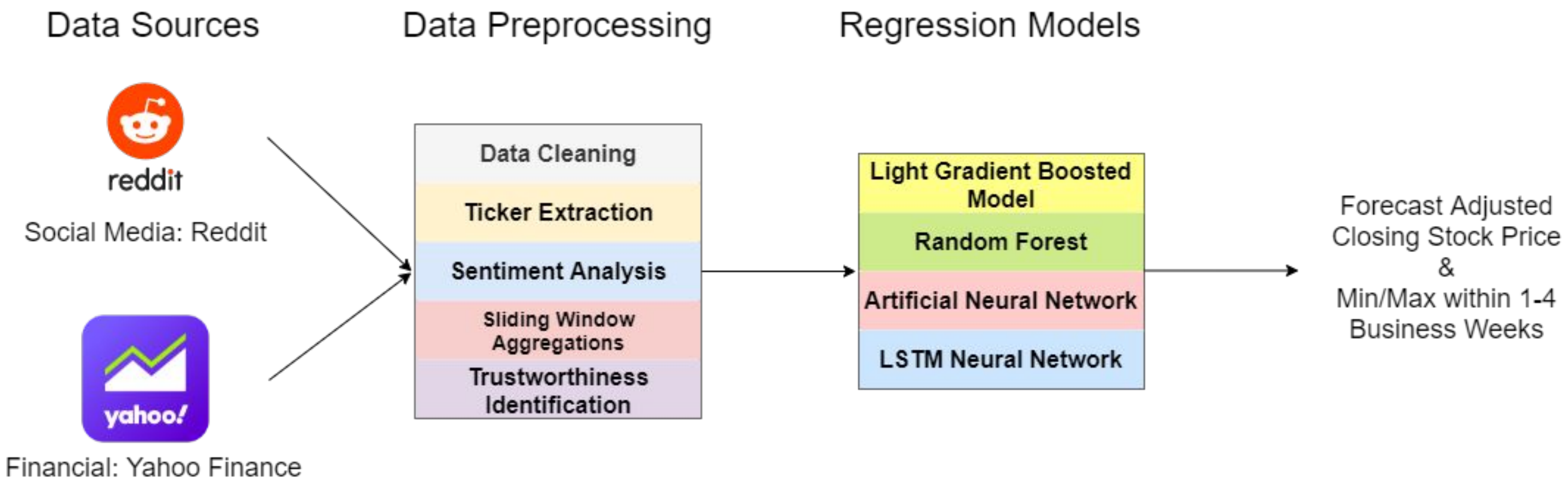
Lots of Noise Impairs Variable Identification

Increasing Domain Research

Emerging Usage of Sliding Windows

Future State

Our Approach



Data Sources

Data Preprocessing

Regression Models

Social Media: Reddit

Financial: Yahoo Finance

Data Cleaning

Ticker Extraction

Sentiment Analysis

Sliding Window Aggregations

Trustworthiness Identification

Light Gradient Boosted Model

Random Forest

Artificial Neural Network

LSTM Neural Network

Forecast Adjusted Closing Stock Price & Min/Max within 1-4 Business Weeks

Data Source: Collecting Raw Data

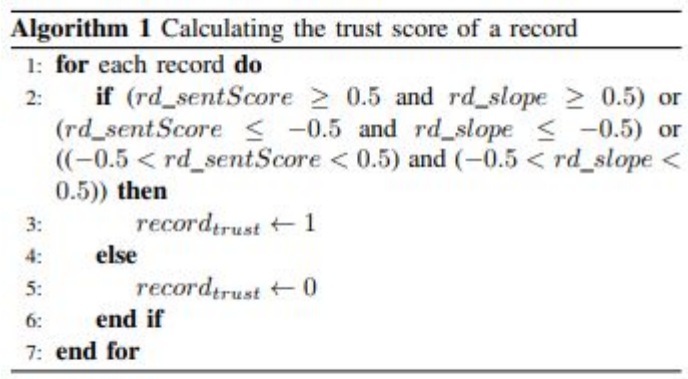
Data Preprocessing: Cleaning Data

Regression Models: Testing Models

Forecast Results: Analyzing Results

Design 1: Trust Filter on Records

Main assumption: a discussion (with opinion) comment is trustable if the sentiment expressed by the comment author aligns with the general trend.



Algorithm 1 Calculating the trust score of a record

```
for each record do
    if ($rd\_sentScore \geq 0.5$ and $rd\_slope \geq 0.5$) or ($rd\_sentScore \leq -0.5$ and $rd\_slope \leq -0.5$) or (($-0.5 < rd\_sentScore < 0.5$) and ($-0.5 < rd\_slope < 0.5$)) then
        $record_{trust} \leftarrow 1$
    else
        $record_{trust} \leftarrow 0$
    end if
end for
```

1. **Two Trust Scores:** One trust score is assigned for each record and another for the author of a record

2. **Agreement Assignment:** Scores are assigned based on the author's sentiment relative to the post/comment and general trend

3. **Author Trustworthiness Calculation:** Authors are assigned a trust score based on their average trust score / total posts

A threshold of 80% accuracy was selected to indicate if a Reddit member was considered trustworthy in their stock price predictions/sentiment

Design 2: Sliding Window

Testing Window Sizes

Determining the Effect Each Window Size has on Performance

- The sliding window technique aggregates values over a continuous sequence

Determining Optimum Window Size

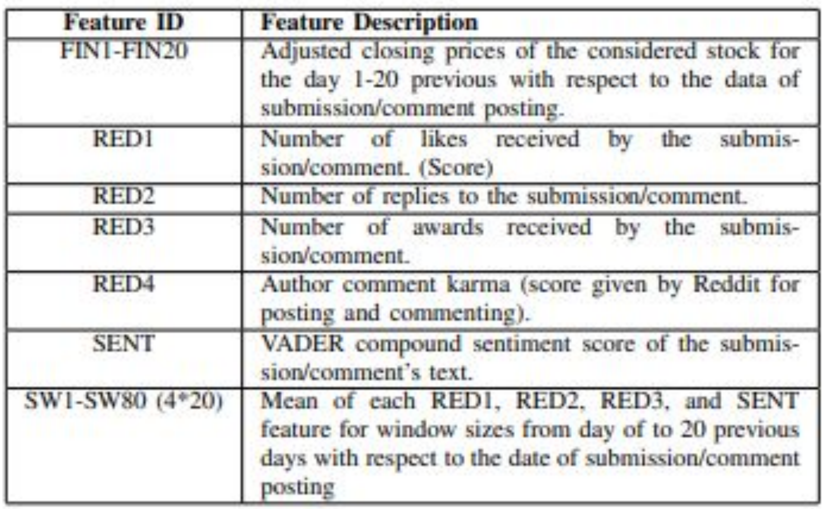
Applying the Sliding Window Technique to Aggregate Features

- We try varying window sizes from 1 to 20 days of the date of Reddit post and aggregate using the mean function

For instance, setting the sliding window parameter to five means that for each record, we select all records within the past 5 days that correspond to the same stock and aggregate these records to determine the final Reddit features for that record.

Features and Models

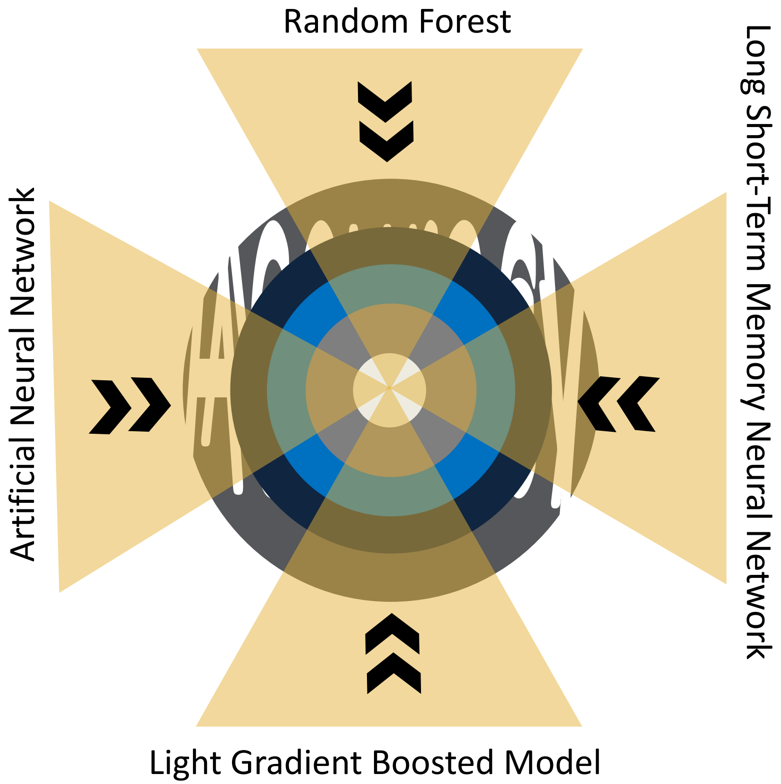
Proposed Features



Feature ID	Feature Description
FIN1-FIN20	Adjusted closing prices of the considered stock for the day 1-20 previous with respect to the data of submission/comment posting.
RED1	Number of likes received by the submission/comment. (Score)
RED2	Number of replies to the submission/comment.
RED3	Number of awards received by the submission/comment.
RED4	Author comment karma (score given by Reddit for posting and commenting).
SENT	VADER compound sentiment score of the submission/comment's text.
SW1-SW80 (4*20)	Mean of each RED1, RED2, RED3, and SENT feature for window sizes from day of to 20 previous days with respect to the date of submission/comment posting

Financial	FIN1-FIN20
Reddit	RED1-4, SENT
Aggregate	SW1-SW80

Regression Models



Random Forest

Artificial Neural Network

Long Short-Term Memory Neural Network

Light Gradient Boosted Model

The Random Forest was utilized as the baseline due to its prominence in recent model studies that utilized social media, news, and financial data in conjunction where it had the best performance of 13 models.

Experiments

1 Which regression model performs best?

2 How do features contribute to model performance?

3 How would the sliding window design affect regression models?

4 How would the trust filter affect regression models?

Model + Different sets of features

- Financial attributes only (fin)
- Reddit attributes only (red)

- Reddit and financial attributes (redFin)
- Reddit and financial attributes after the trust filter is applied (redFinT)

- Reddit attributes aggregated for sliding window sizes of 1-20 with financial attributes (sw#)
- Reddit attributes aggregated for sliding window sizes of 1-20 with financial attributes after the trust filter is applied (sw#T)

A total of 44 (4+20+20) trials per model！

Results

1

Best Regression Model

The RF and LSTM models consistently outperformed other regression models

2

Feature Impact on Models

The most accurate minimum and maximum stock predictions utilized Reddit data

3

Sliding Window Impact

The sliding window improves the performance for all considered models in comparison to the financial only

4

Trust Filter Impact

The most noticeable performance increase resulted from the usage of trust filters

Dataset Statistics & Evaluation Metrics

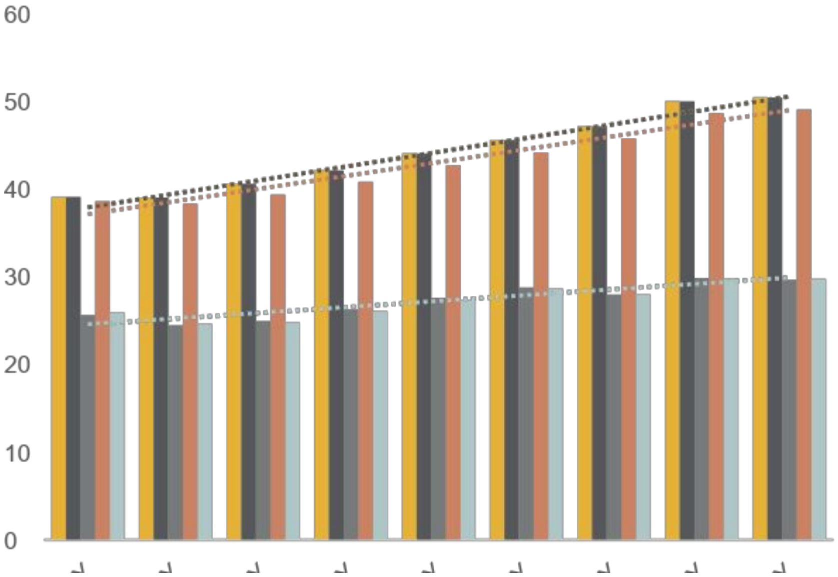
From 1 million posts and 3.5 million comments – 2,371 unique stocks were identified as having been discussed on r/WSB – as mentioned in the collected posts/comments

	All	With Trust Filter
# of Records	940,785	36,578
# of Unique Posts	484,171	15,577
# of Submissions	40,882	819
# of Unique Submissions	18,310	545
# of Comments	899,903	35,759
# of Unique Comments	465,861	15,032
# of Unique Authors	67,092	8,930
# of Unique Stocks	2,371	1,421

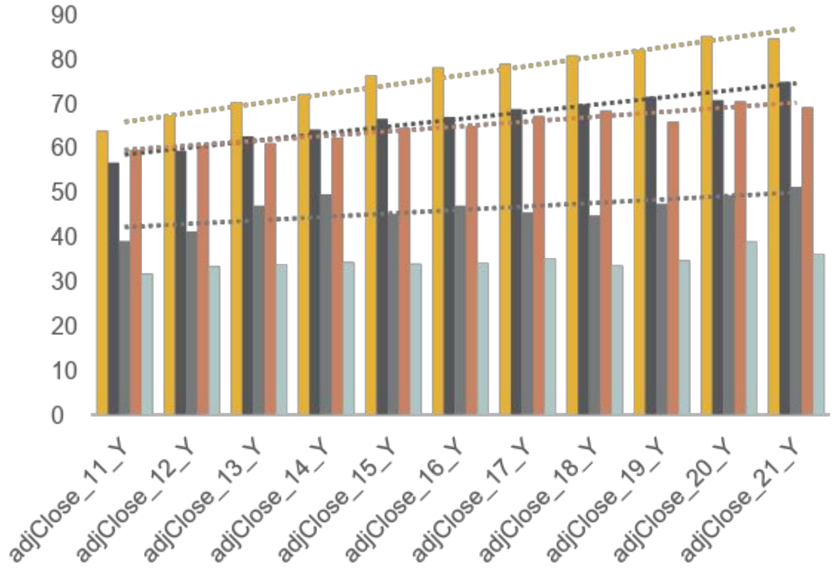
To evaluate model performance, Root Mean Squared Error and Mean Absolute Error was utilized, to calculate the error % on a per model basis

Adjusted Closing Price Prediction Regression Results

RF RMSE of Adjusted Closing Prices Predictions



LSTM RMSE of Adjusted Closing Prices Predictions

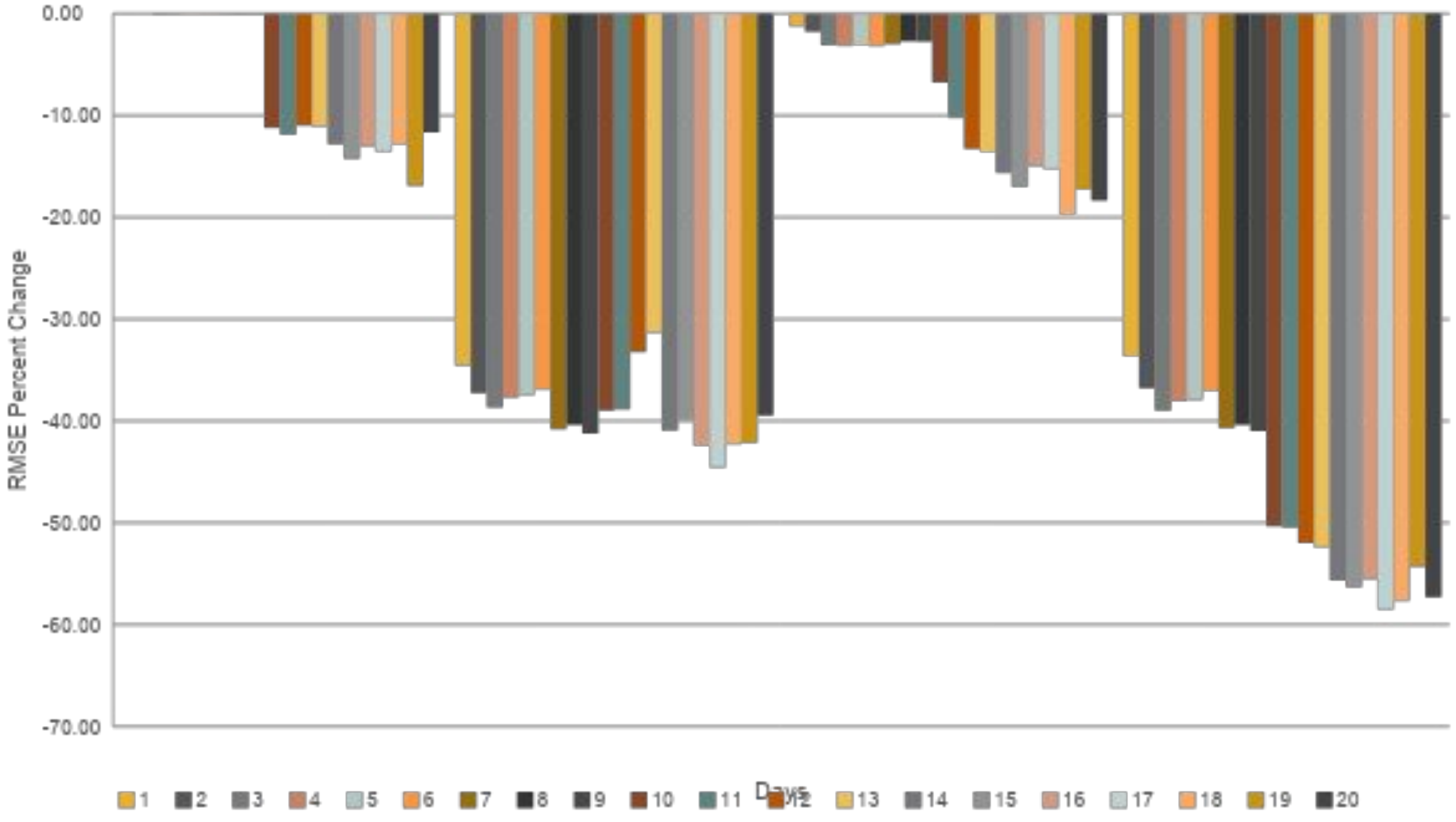


- 🟨 Financial only
- ■ Reddit and Financial
- ■ Best Sliding Window
- 🟫 Reddit and Financial, trust filtered
- ▨ Best Sliding Window, trust filtered

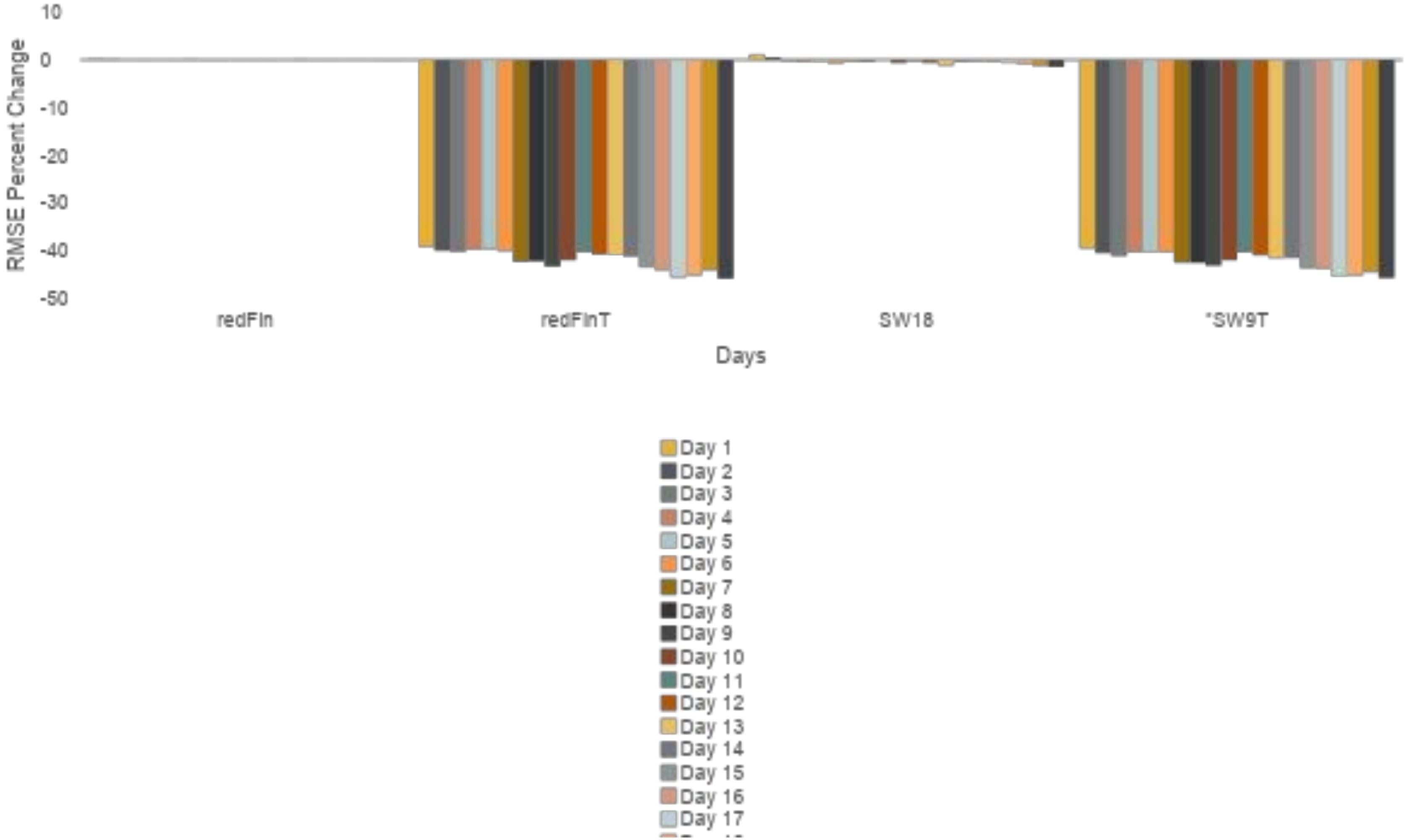
Important Takeaways

▶ The most accurate models across both the RF and LSTM utilized trust filters and a sliding window

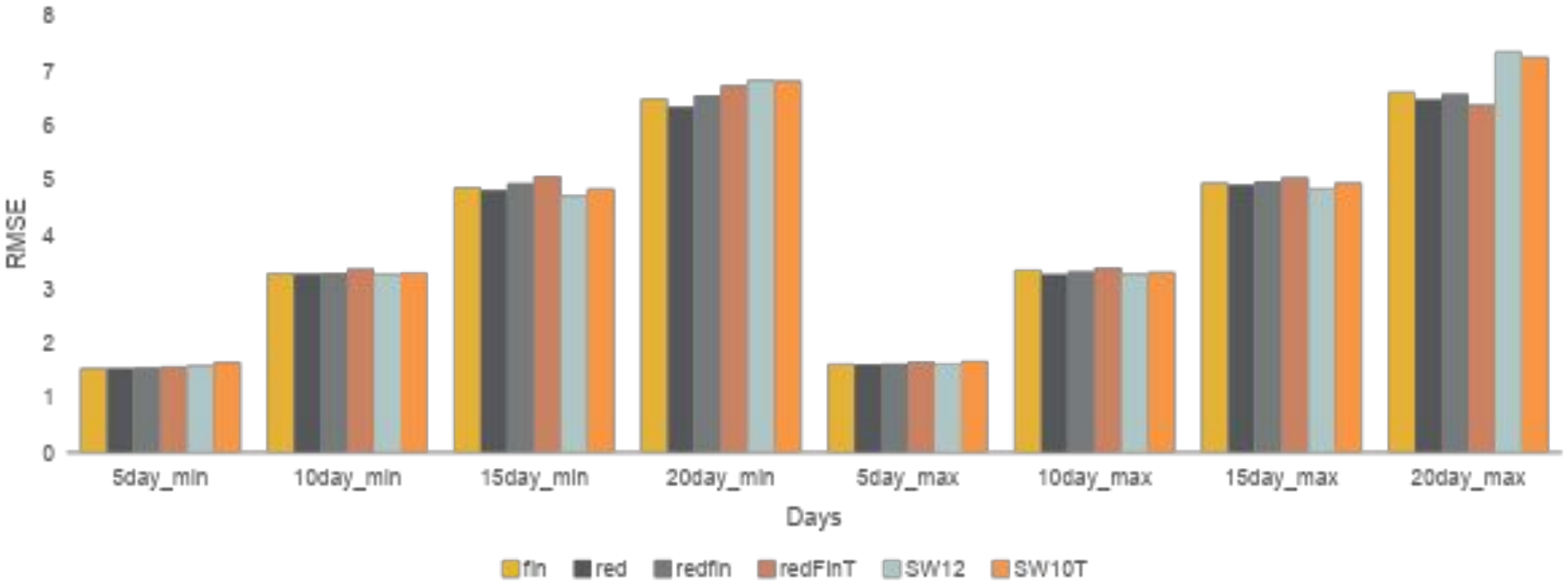
Error reductions by RF and LSTM over projected time



RF error reduction over projected time



NN - Optimal Selling Period Prediction Results



Perceived Trend Takeaways

- Social media is a distillation of retail investors interests and are primarily oriented towards purchase as opposed to sell behaviors.
- Retail investors primarily learn stocks from buying as opposed to selling and have more access to long-oriented knowledge.
- More social media data helps build consensus among community members in the absence of filtering whereas it serves as ’noise’ when outdated data is considered by reputable users.

Conclusion

Contact: {dennis.huyn, y.tian} @queensu.ca

Demonstrably Improved Performance

1 Each model was improved with the sliding window and credibility filter demonstrating that the credible individual users of r/WSB are more accurate than the community on aggregate

Both Proposed Techniques Significantly Reduce RMSE Errors

2 The LSTM, when predicting 10-20 days into the future, experienced a 50-58% performance boost in the reduction of RMSE errors whereas a sliding window of size 9 and credibility filter improves the RMSE most notably, ranging from 39-46%, with the RF

Stock Minimums are Easier to Predict than Maximums

3 The NN performed best at predicting minimums indicating that r/WSB most commonly discusses entry as opposed to exit opportunities