The Bull and the Bear: Summarizing Stock Market Discussions

Jay Shah[1] Ayush Kumar[1] Dhyey Jani[1] Devanshu Thakar[1] Varun Jain[1] Mayank Singh[1]

[1]Indian Institute of Technology Gandhinagar, India

1. Introduction

- Stock market investors debate and heavily discuss stock ideas, investing strategies, news and market movements on social media platforms like Twitter, Reddit, Yahoo, etc. The discussions are significantly longer in length and require extensive domain expertise for understanding.
- The majority of the stock market-related datasets focus on the entire historical daily price and volume data (numerical values).
- To the best of our knowledge, we do not find any dataset that contains textual summaries of stock market discussions.
- In this paper, we curate such discussions and construct a first-of-its-kind of abstractive summarization dataset.
- Our curated dataset consists of 7888 *Reddit* posts and manually constructed summaries for 400 posts.

2. Contributions

- We propose a dataset containing 400 summaries of Reddit's stock market-related posts.
- We quantitatively and qualitatively evaluated constructed summaries.
- We test two SOTA summarization models BART and PEGASUS and discuss their limitations.

3. The Dataset

1. **The Reddit Stock Market Corpus (RSMC)**
   - We curate the relevant posts by searching Reddit's platform. Among several possible subreddits, we selected the most popular subreddit *r/wallstreetbets*. As on January 15 2022, *r/wallstreetbets* is being followed by *11,491,040* users.
   - We used Python's *PRAW* (Python Reddit API Wrapper) module for initial data curation. We used a set of keywords such as *finance* and *stocks* to search the relevant posts.
   - Overall, we curated 7888 posts comprising an average number of roughly 900 words.
   - At any time the scraper was used, the posts were selected starting from the most recent post on the subreddit and going backwards in time. The posts were curated in multiple phases in October 2021.
   - From these, we removed 400 posts and employed annotators to manually create gold summaries for them, which are compiled in a separate dataset. The remaining 7488 posts are compiled in a dataset called the *Reddit Stock Market Corpus* (**RSMC**).
2. **The Summarization Dataset (SMSC)**
   - We employ seven annotators to generate summaries for randomly selected 400 posts. We call this dataset as *Stock Market Summary Corpus* (**SMSC**).

4. Annotation Guidelines

- The generated summary is expected to contain approximately 50-70 words irrespective of the length of the post.
- The summaries are expected to be accurate and complete as possible.
- While generating the summaries, noisy contents like hyperlinks and emoticons must be discarded.
- Since the posts are related to stock markets and finance, the annotator is expected to preserve the financial information of the post in the summaries that are generated.
- The annotators are refrained from keeping abusive words in the summaries.

5. Dataset Statistics

	RSMC	**SMSC**
Total Posts	7488	400
Avg. no. of words	932.62	53.47
Max. no. of words in a post	7518	166
Avg. no. of sentences in a post	36.93	3.04

Table 1: Dataset Statistics. RSMC denotes the statistics for an individual post and SMSC for an individual summary. The RSMC statistics are for posts excluding the SMSC dataset.

6. Sample from the Dataset

Title Flawless Strategy For Printing Infinite Money
Link Reddit Link
Post text The stock market opens at 9:30am. Due to algorithms and shit and more people trading at market open, the theory is that over the long run the stock price will increase 51% or more of the time between 9:30 and an arbitrary time that is close to 9:30 like 9:45.
Steps to print money:
1. Pick any stock ideally one that has a lower chance of fluctuating 2% in the span of 15 mins
2. Buy 9:30
3. Set stop loss of 2%
4. Sell at 9:45 no matter what the price
5. Repeat every day
When you lose, you only lose 2%. When you gain, on avergae you will gain more than 2%. And since you will gain on 51% or greater of the days you will gain money guaranteed.
I have solved the stock marker your welcome.

Summary Here are the steps to print money. Pick any stock that has a low chance of fluctuating 2% in the span of 15 mins. Buy at 9:30 am, set a stop loss of 2%. Sell at 9:45 no matter what the price. Repeat every day. Over the long run stock price will increase 51% or more between time 9:30 to 9:45 am.

Table 2: A representative post and its corresponding summary from SMSC. Blue colour text represents the original post's text, and Orange colour text represents the summary text.

7. Evaluating the Summaries

We evaluate SMSC summaries under two settings: (i) large-scale automatic evaluation and (ii) small-scale manual evaluation.

- **Automatic Evaluation:** The human-generated summaries were evaluated on the metric of overall Grammarly score provided by *Grammarly*. It's value can range from 0 to 100. The mean Grammarly score of SMSC is 89.79.

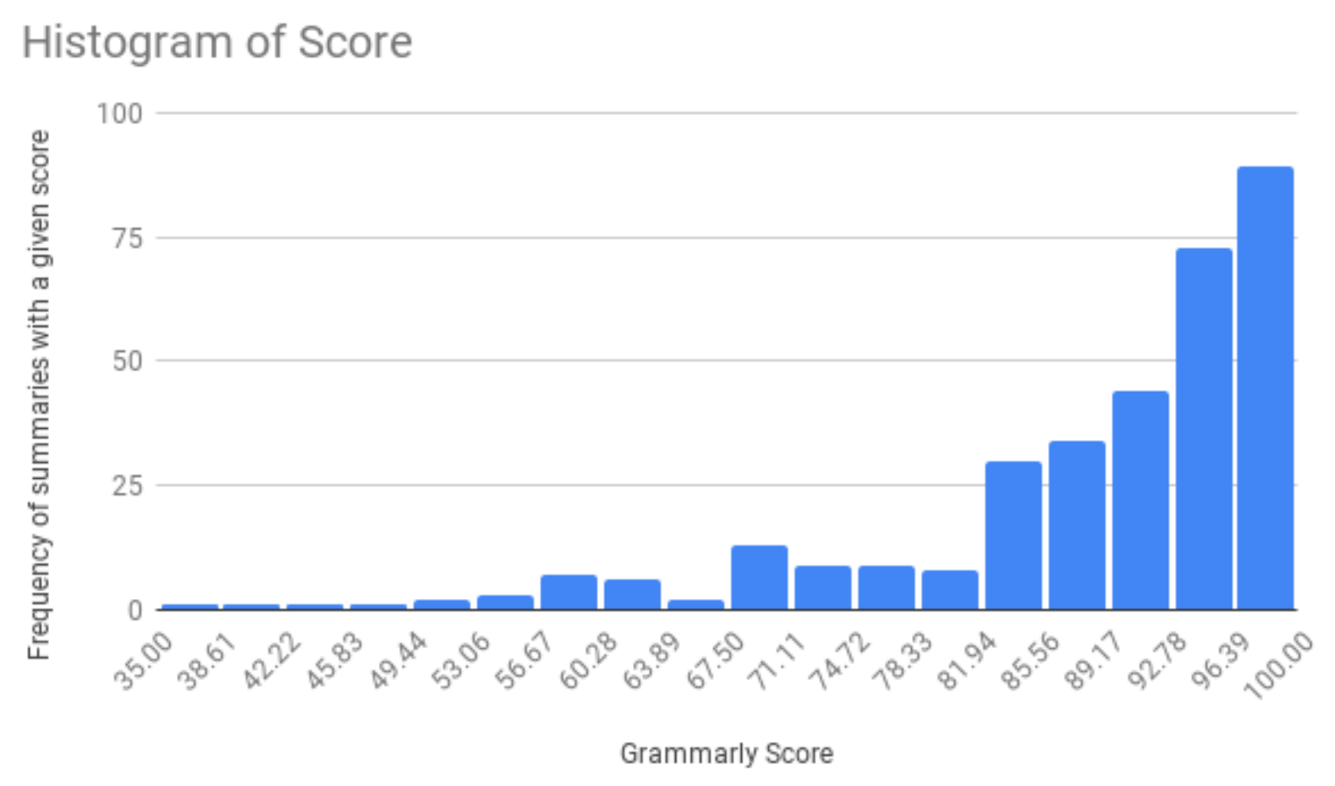

Figure 1: Number of summaries with an overall Grammarly score. The majority of the summaries (91.05%) have a score greater than 70.

- **Manual Evaluation:** We employ two independent annotators to evaluate randomly selected 20 summaries from the SMSC dataset. We define two subjective metrics, *Readability* and *Completeness*, on a scale of 1–5. *Readability* implies that the summary is readable and does not contain grammatical inaccuracies. A Completeness score measures the extent of summary capturing the information in the original post.

	Anno-1	Anno-2	Avg.	IAA
Readability	4.15	4.55	4.35	0.368
Completeness	4.40	4.35	4.375	0.368

Table 3: Statistics of manual evaluation of SMSC dataset (on a scale of 1-5). Anno-1 and Anno-2 represent average scores for annotators 1 and 2 respectively. Avg. corresponds to the average score computed for annotators 1 and 2. IAA is an inter-annotator agreement score.

8. How Good are State-of-the-Art Summarization Models?

- We evaluate two state-of-the-art summarization models: (i) **BART** [2] and (ii) **PEGASUS** [4], on SMSC dataset. We leverage three variants of **ROUGE** metric [3], ROUGE-1, ROUGE-2 and ROUGE-L, for the evaluation.
- The BART and the PEGASUS models are fine-tuned on the CNN/DailyMail dataset [1] for the summarization task.
- We generated abstractive summaries using the BART and PEGASUS models for every individual post in the RSMC and SMSC datasets and compared these machine summaries of SMSC dataset with their corresponding manual summaries.

	BART	PEGASUS
ROUGE-1	0.46 (0.19)	0.42 (0.20)
ROUGE-2	0.30 (0.22)	0.26 (0.22)
ROUGE-L	0.45 (0.19)	0.40 (0.20)

Table 4: Comparing BART and PEGASUS on SMSC dataset against ROUGE. Values in the bracket represent standard deviation in ROUGE scores.

- The high standard deviation values show a significant fluctuation in ROUGE scores for both BART and PEGASUS.

Post text Why isnt Volkswagen stock going up like the rest of the automotive industry? Ive done well with some EV stocks and have been looking at traditional automotive manufacturers such as GM, Ford, and VW. Unfortunately, I feel like I am too late on GM and Ford now. I still see some upside with them but not as much as theyve experienced this past year. On the other hand, VW stock seems to have plenty of room for growth. In the past two year they are only up 8% compared to 85% and 170% from GM and Ford. There P/E is only 5.3 compared to 8.9 (GM) and 34 (Ford). Most of the increase in stock price for GM and Ford has had to do with news surrounding the EVs they are making (Im using GM and Ford as examples but VW has lagged pretty much all the large car manufacturers). VW is arguably one of the best positioned for EVs. They are investing $100 billion on EVs, their CEO is all bought in, they are looking at manufacturing their own batteries, and their ID4 has done pretty well with over 70,000 purchases. What am I missing? Why hasnt their stock been as popular as others in the industry?
SMSC In past two years Volkswagen (VW) stocks are only up 8% compared to 85% and 170% from GM and Ford. VW's P/E is only 5.3 compared to 8.9 (GM) and 34 (Ford). The increase in GM and Ford is due to the EV they are making. It is expected that VW will rise as they are investing heavily in EV manufacturer sector.
BART VW is investing $100 billion on EVs, their CEO is all bought in, they are looking at manufacturing their own batteries, and their ID4 has done pretty well with over 70,000 purchases. Most of the increase in stock price for GM and Ford has had to do with news surrounding the EVs they are making. VW is arguably one of the best positioned for EVs.
PEGASUS VW is arguably one of the best positioned for EVs. They are investing $100 billion on EVs, their CEO is all bought in, they are looking at manufacturing their own batteries, and their ID4 has done pretty well with over 70,000 purchases.

Table 5: A representative Reddit post with its manual, BART and PEGASUS summaries.

- In the example shown above, critical information about the *Volkswagen* stock is missing from the BART and PEGASUS summaries.

References

[1] Karl Moritz Hermann, Tomas Kocisky, Edward Grefenstette, Lasse Espeholt, Will Kay, Mustafa Suleyman, and Phil Blunsom. Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701, 2015.

[2] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online, July 2020. Association for Computational Linguistics.

[3] Chin-Yew Lin. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81, 2004.

[4] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.