Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 6909–6913
Marseille, 20-25 June 2022

The Bull and the Bear: Summarizing Stock Market Discussions

Ayush Kumar*, Dhyey Jani*, Jay Shah*, Devanshu Thakar*, Varun Jain, Mayank Singh
Indian Institute of Technology Gandhinagar
Palaj, Gandhinagar, Gujarat, India 382355
{ayush.kumar, jani.dhyey, shah.jay, nilesh.thakar, varun.jain, singh.mayank}@iitgn.ac.in

Abstract

Stock market investors debate and heavily discuss stock ideas, investing strategies, news and market movements on social media platforms. The discussions are significantly longer in length and require extensive domain expertise for understanding. 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. We robustly evaluate the summaries and conduct experiments on SOTA summarization tools to showcase their limitations. We plan to make the dataset publicly available. The sample dataset is available here: `https://dhyeyjani.github.io/RSMC`.

Keywords: Stock Market, Summarization, Reddit

1. Introduction

Commodity and equity trading on the open market is often viewed as a demanding and time-consuming activity. Stock markets have grown in popularity in recent years, with a dramatic increase in daily active users. With an ever-increasing volume of daily transactions, the stock market is becoming highly volatile to various fundamental features of the entity that underpins the trading commodity. To keep up with this pace, the investor community discusses trading strategies, investment and global news, pitfalls and movements, market sentiments, etc., on various social media platforms like Twitter, Reddit, Yahoo.com, etc. However, understanding and participating in these discussions require significant domain expertise and knowledge.

Figure 1 presents a Reddit post that technically analyses the stock market crash of Feb/March 2020. As expected, it contains several domain-specific entities like *VIX, SPX, RBLX, GME, NASDAQ, Microsoft, Apple, Amazon, Alphabet* and *Facebook*. Entities such as VIX, SPX, and RBLX are stock market indexes, whereas GME is the stock acronym of video game retailer GameStop Corp. NASDAQ, Microsoft, Apple, Amazon, Alphabet and Facebook are the names of organizations. The post also contains several date and time instances such as *March 2021, Nov 2020, May 2020, Jan 2022 Nov 2018, Feb 2020*, etc. In addition, it contains relative trend markers like *40%, 50%*, etc. The post illustrates the textual complexities due to domain-specific terms with usual complexities associated with social media data such as shorthand notations, usage of symbols, emojis, etc.

The majority of the stock market-related datasets[1][2] focus on the entire historical daily price and volume data

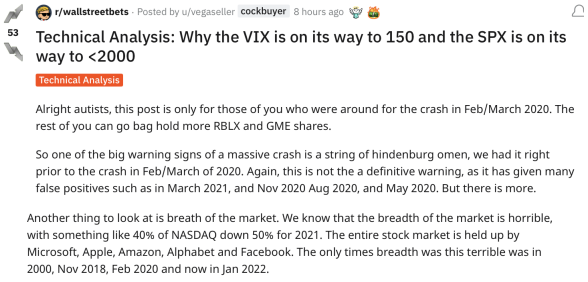

Figure 1: A stock market discussion post in Reddit.

(numerical values). It has shown that the tone of discussions on the subreddit *r/wallstreetbets* displays significant predictive associations with GameStop returns, volatility, bid-ask spreads as well as volume (Anand and Pathak, 2021). However, to the best of our knowledge, we do not find any dataset that contains textual summaries of stock market discussions. The closest work to the current proposal is Reddit TIFU dataset (Kim et al., 2019) that contains general Reddit's post summaries.

In this paper, we collected 7888 stock market posts from Reddit and created abstractive summaries for 400 posts. We conducted an automatic and human evaluation to evaluate the quality of the summaries robustly. Towards the end, we leverage two SOTA summarization models, BART (Lewis et al., 2020) and PEGASUS (Zhang et al., 2020), to generate summaries and showcase their limitations even though the proposed dataset contains a small set of manually generated summaries. We envisage several lines of work on stock market summarization in future. The proposed dataset can be useful in two scenarios: (i) evaluation of existing generic summarizers and (ii) constructing domain-specific summarizers requiring a small amount of fine-tuning dataset.

Our main contributions are:

*These authors contributed equally.

[1]https://www.kaggle.com/borismarjanovic/price-volume-data-for-all-us-stocks-etfs

[2]https://www.kaggle.com/rohanrao/nifty50-stock-market-data

	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
Max. no. of sentences in a post	427	8

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.

1. We propose a dataset containing 400 summaries of Reddit's stock market-related posts (see Section 2).

2. We quantitatively and qualitatively evaluated constructed summaries (see Section 3).

3. We tested two SOTA summarization models BART and PEGASUS and discuss their limitations (see Section 4).

2. The Dataset

In this section, we discuss the curation of the Reddit posts and, thereafter, the construction of the summary dataset.

2.1. The Reddit Stock Market Corpus

We, first of all, 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. Note that Reddit posts contain multimodal datatypes such as images and videos along with the text. Therefore, we filtered out posts containing less than 50 words. Overall, we curated 7888 posts comprising an average number of roughly 900 words (see Table 1 for more statistics). 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**). Here, along with the posts' textual content, each instance also contains the unique post id, the title of the post and a hyperlink to it.

2.2. The Summarization Dataset

We employ seven annotators to generate summaries for randomly selected 400 posts. Each annotator is an undergraduate student. All the annotators have reading and writing proficiency in English language. Even though, none of the annotators are stock market domain experts, all annotators know stock market-related jargons. They regularly participate in stock market discussions. The 400 posts were equally distributed among annotators[3]. The following policies and information were discussed with the annotators before the annotation exercise:

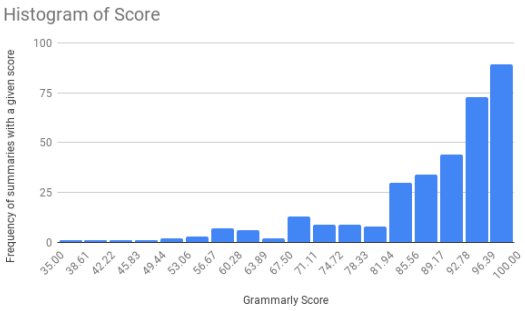

Figure 2: Number of summaries with an overall Grammarly score. The majority of the summaries (91.05%) have a score greater than 70.

- 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.

We call this dataset as *Stock Market Summary Corpus* (**SMSC**). Table 2 shows two representative posts and their manual summaries. Table 1 presents statistics of SMSC.

3. Evaluating the Summaries

Next, 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

[3]The annotators are different from the authors of the paper.

Example 1	
Title	Recent developments on $HIMX
Link	Reddit Link
Post text	As of today, the stock price of Himax has had an enormous run from 10 to 16$ in less than a month which was followed by a massive drop of 25% to the 12$ line in just three days. Regarding this unexpected drop I wanted to update my view on the stock and the company itself. As of now, I don't (and probably can't) know why exactly this immense price decrease took place. The most common theories are that it was either a whale exiting positions and thereby triggering stop losses and panic selling or it was short sellers doubling down and having to bring an end to the rally we saw end of year. Nevertheless, both of these explanations don't change anything about the stocks potential in the long term. What a lot of trading/investing pros keep repeating is that when a stock isn't doing what you expect it to do, you have to sit back and look at it's fundamentals again. That's exactly what I've been doing for the last days and I've found that the company is in the exact same position (If not a better one) than three weeks ago. Fundamentals still check out and the opportunites behind this haven't vanished at all. (You can check out my earlier posts to learn more about my investment thesis) You have to keep in mind that this is a long term play. I'm personally not expecting us to hit 25$ before mid 2022. The volatility that we've seen in the last weeks is normal and it will probably continue like this for the next months. Because of the volatility we've been seeing and the big opportunities being long term I highly adivse you to either buy LEAPS or stocks as you don't want to see your whole portfolio wiped away by badly placed options (or maybe you do, idk, this is wsb). This is pretty much all I've got to say about the recent price development. The stock stays massively undervalued and the market just hasn't priced in the recent earnings growth and the company's huge business opportunites which basically makes Himax a no-brainer. I wish us all good luck on this one and we'll see us again in a few months when LEAPS will be printing. Cheers.
Summary	I wanted to update my view on Himax and its stock. Himax stock has seen a massive drop of 25% in less than one month. I am personally not expecting Himax to hit 25$ before mid 2022. Because of the volatility, I would highly advise you to buy LEAPS or stocks to diversify your portfolio.

Example 2	
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: Two representative posts and their corresponding summaries from SMSC. Blue colour text represents the original post's text, and Orange colour text represents the summary text.

Grammarly score (Figure 2) provided by *Grammarly*. It's value can range from 0 to 100. The mean Grammarly score of SMSC is 89.79. The Grammarly scores are available for 380 out of 400 samples as the remaining samples were smaller in length to provide a score by Grammarly. As per Grammarly's documentation[4], a score of nearly 90 is a good score as it implies that when compared with 90% of documents by all the users worldwide with similar Grammarly goals, the grammatical correctness of the summaries are bet-

[4]https://support.grammarly.com/hc/en-us/articles/360007144751-What-is-Performance-and-how-is-it-calculated-

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 3: A representative Reddit post (in Blue color). Orange, Green and Violet colors representing its manual, BART and PEGASUS summaries respectively.

	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 4: 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.

ter than them.

- **Manual Evaluation:** We employ two independent annotators[5] to evaluate randomly selected 20 summaries from the SMSC dataset. Each annotator is an undergraduate student. All the annotators have reading and writing proficiency in English language. Even though, none of the annotators are stock market domain experts, all annotators know stock market-related jargons. They regularly participate in stock market discussions. 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 Readability score of one indicates an unreadable summary, whereas a score of five indicates no grammatical inaccuracies in the summary. A Completeness score measures the extent of summary capturing the information in the original post. A high score ($\sim$ 5) represents that the summary misses no information present in the original post. Table 4 shows high Readability and Completeness scores with a fair amount of inter-annotator agreement (Krippendorff's Alpha (ordinal)).

[5]Different from annotators that created SMSC dataset and authors of this paper

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

In this section, we evaluate two state-of-the-art summarization models: (i) **BART** (Lewis et al., 2020) and (ii) **PEGASUS** (Zhang et al., 2020), on SMSC dataset. We leverage three variants of **ROUGE** metric (Lin, 2004), ROUGE-1, ROUGE-2 and ROUGE-L, for the evaluation.

BART is a denoising autoencoder for pretraining sequence-to-sequence models. It is implemented as a bidirectional encoder and an auto regressive decoder. Evaluations show that BART performs significantly good on both text generation and comprehension tasks. In comparison, PEGASUS uses a pre-training self-supervised objective (gap-sentence generation). This objective is particular to creating text summarizers that aim to fine-tune the performance and even work with smaller datasets.

The BART and the PEGASUS models are fine-tuned on the CNN/DailyMail dataset (Hermann et al., 2015) for the summarization task.

Evaluating BART and PEGASUS: We generated abstractive summaries using the BART and PEGASUS models for every individual post in the RSMC and SMSC datasets and compared these machine sum-

maries of SMSC dataset with their corresponding manual summaries. Table 5 presents the average ROUGE scores for BART and PEGASUS. The high standard deviation values show a significant fluctuation in ROUGE scores for both BART and PEGASUS.

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

In the example shown in Table 3, critical information about the *Volkswagen* stock is missing from the BART and PEGASUS summaries. The rise in *Volkswagen* stock is only 8% which is very less than its contemporaries GM (85%) and Ford (170%), and its P/E is only 5.3 compared to GM's 8.9 and *Ford's* 34. This is potentially the most important span of the post because a casual user will immediately want to know about the trends in the stock price change. The above critical information is available in the manually created summary.

5. Conclusion

In this paper, we introduce a summarization corpus that encompasses the field of finance by scraping the social media platform *Reddit*. We robustly evaluate the quality of the summaries by conducting manual and automatic evaluations. We also evaluated popular summarization models like PEGASUS and BART and showcased their inefficiencies in generating stock market-related discussion summaries. Even though small in volume, our proposed dataset will be beneficial in fine-tuning existing neural summarization models.

6. Bibliographical References

Anand, A. and Pathak, J. (2021). The role of reddit in the gamestop short squeeze. *Economics Letters*, page 110249.

Hermann, K. M., Kocisky, T., Grefenstette, E., Espeholt, L., Kay, W., Suleyman, M., and Blunsom, P. (2015). Teaching machines to read and comprehend. *Advances in neural information processing systems*, 28:1693–1701.

Kim, B., Kim, H., and Kim, G. (2019). Abstractive summarization of reddit posts with multi-level memory networks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2519–2531.

Lewis, M., Liu, Y., Goyal, N., Ghazvininejad, M., Mohamed, A., Levy, O., Stoyanov, V., and Zettlemoyer, L. (2020). 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. Association for Computational Linguistics.

Lin, C.-Y. (2004). Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Zhang, J., Zhao, Y., Saleh, M., and Liu, P. (2020). Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR.