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 of 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 2020.
• 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 refrain from keeping abusive words in the summaries.

5. Dataset Statistics

<table>
<thead>
<tr>
<th>Total Posts</th>
<th>RSMC</th>
<th>SMSC</th>
</tr>
</thead>
<tbody>
<tr>
<td>Avg. no. of words</td>
<td>154788</td>
<td>1680</td>
</tr>
<tr>
<td>Max. no. of words in a post</td>
<td>154788</td>
<td>1680</td>
</tr>
<tr>
<td>Avg. no. of sentences in a post</td>
<td>35.93</td>
<td>3.01</td>
</tr>
</tbody>
</table>

Table 1: Dataset Statistics. RSMC denotes the statistics for the 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 URL
Post text: The stock market opens at 9:30 am. 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 9:45.
Steps to print money:
1. Pick any stock ideally that has a lower chance of fluctuating 2% in the span of 15 mins.
2. Buy at 9:30 am.
3. Set stop loss of 2%.
4. Sell at 9:45 am no matter what the price.
5. Repeat every day.

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 am 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.

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.

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

• We evaluate two state-of-the-art summarization models: (i) BART and (ii) PEGASUS, on SMSC dataset. We leverage three variants of ROUGE metric, ROUGE-1, ROUGE-2 and ROUGE-L for the evaluation.
• The BART and PEGASUS models are fine-tuned on the CNN/DailyMail dataset 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.

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

- BART
- PEGASUS
- ROUGE-1: 0.68 (0.19) 0.42 (0.20)
- ROUGE-2: 0.30 (0.22) 0.30 (0.22)
- ROUGE-L: 0.45 (0.19) 0.40 (0.20)

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.

References