Multi-Entity Aware Sentiment Analysis for Financial Headlines

Noah Prozan, Yifan Wang

Abstract

This paper implements NLP approaches to the financial domain by processing news headlines and performing ABSA (Aspect Based Sentiment Analysis). In this paper, we integrate the two components of ABSA, namely the extraction of aspect terms and determination of sentiment polarities. We utilize the SEntFiN 1.0 dataset, a relatively new and large dataset for news headlines with human-annotated entity targets and sentiment labels. Based on recent evolutions in ABSA research, we implement machine learning approaches such as T5 generative model, CRF method, and NER-BIO tagging schemes. While we practice one-step NER (Named Entity Recognition) based approaches, we find that a two-step method that extracts an entity and then performs sentiment analysis generates the best performance, achieving an accuracy of 74.16%. Additionally, we conducted detailed analysis on the effectiveness of the aforementioned models.

1 Introduction

The Efficient Market Hypothesis (EMH) is a fundamental economics hypothesis that is critical for investment decisions. EMH stipulates that at any given time, asset prices reflects all publicly available information. According to the EMH, once new information is made available, asset prices will immediately adjust to it, and accurately reflect the information. Empirically, the time needed for such accurate adjustments depends on factors such as liquidity and transaction costs. Nevertheless, the EMH suggests that those who react to new information appropriately and quickly would monetize on the convergence towards the new price.

News headlines, in general, are one of the most common ways that information is released in the market. Third-party news providers such as Bloomberg, not only provide news headlines in a succinct, accurate, and timely manner, but also extract and break down material information from lengthy articles and financial disclosures into multiple headlines. As a result, news headlines contain a wealth of tradable information in practice, and the ability to extract and comprehend information from headlines would be of significant value to preparing for potential stock movements (Li et al., 2011, Chan, 2003).

One form of information comprehension is the classification of sentiments embedded in a piece of financial information. Such echoes the concept of Financial Sentiment Analysis (FSA). Research has supported the significance of sentiments on asset price movements (Van de Kauter et al., 2015, Bollen et al., 2011), in line with implications from the EMH hypothesis. Inspired by this, in order to generate concrete trading decisions, our paper aims to study the extraction of sentiments specific to relevant traded assets, such as stocks and bonds of corporations, from financial headlines. However, traditional FSA provides limited insights into sentiments specific to a target, as it lacks necessary granularity. Hence, our task falls in the realm of Aspect Based Sentiment Analysis (ABSA), as it relates to a) identifying key entities within a text such as company, currency, industries, etc. and b) classifying the underlying sentiments with regard to the identified entities. Compared to traditional sentiment analysis, the challenges of our task are a) deciphering the financial language b) detecting multiple entities, and c) extracting conflicting sentiments or mixed sentiments.

ABSA has two major components, namely the extraction of aspect terms and determination of sentiment polarities. Earlier researchers have studied the task of identifying each component separately (Zhang et al., 2022). More recently, with the development of deep learning models, new frameworks have been proposed to tackle the compound ABSA tasks that combine the two components. Neural networks models like MNN (Wang et al., 2018) and RNN (Li et al., 2019) marked earlier efforts to integrate the two tasks.

Recent evolution of generative language models (Raffel et al., 2020, Brown et al., 2020) inspired ABSA research to convert entity recognition and text classification tasks into text generation problems. Unified generative frameworks, such as T5-based model proposed by Zhang et al., 2021 and BART-based model proposed by Yan et al., 2021 achieved state-of-the-art results in ABSA tasks.

A few more approaches of ABSA research development have been the development of a NER-BIO tagging scheme. That is, to co-extract the sentiments and entities in one step (Luo et al., 2019). Additionally, the

implementation of conditional random fields (CRF) on top of a BERT model-architecture has shown impressive results in the ABSA space (Karimi et al., 2021).

The SemEval 2017 Task 5 SubTask 2 Headlines Dataset (Cortis et al., 2017), and Financial Opinion Mining and Question Answering (FiQA) 2018 Task 1 Sentiment Scoring Dataset (Maia, 2028) offer annotated data for financial microblogs and news, with multiple entities, associated sentiment scores, including conflicting sentiments. However, the size of the two datasets are relatively small (<1000 samples), and the set of entities incorporated is limited. Our paper is inspired by a publication of a human-annotated financial headlines database SEntFiN 1.0 (Sinha et al., 2022). The value of the SEntFiN 1.0 data is the inclusion of abundant headlines (10x size of SemEval and FiQA) with multiple entities and conflicting sentiments. The SEntFiN paper (Sinha et al., 2022) experimented with SVM, RNN, LSTM, and BERT-family models to perform targeted aspect sentiment tasks. Unfortunately, the aforementioned more sophisticated and potentially better-performing approaches such as generative T5 framework and CRF method were not investigated. As the SEntFiN 1.0 dataset was recently published, no existing work has evaluated the performance of these frameworks on the dataset, to the best of our knowledge. Additionally, the SEntFiN paper (Sinha et al., 2022) relied on annotated entity dictionary for targeted sentiment classification, which represents only one component of the aforementioned compound ABSA task and renders its model less scalable.

Our paper aims to extend the investigation by Sinha et al., 2022. Firstly, we aim to examine the performance of the more sophisticated models and techniques, namely a generative T5 model and CRF method, on the SEntFiN 1.0 dataset. Secondly, in order to improve the SEntFiN paper's process, we aim to integrate the two components of ABSA, namely aspect extraction and sentiment classification, and create a more scalable framework for fine-grained headline sentiment extraction.

2 Data

2.1 SEntFiN 1.0

SEntFiN 1.0, a human annotated dataset that contains 10,753 annotated news headlines. The dataset has 2,847 headlines referring to multiple entities, and frequently the sentiments are conflicting. The authors are of Indian descent and the financial headlines are mostly for Indian companies. The sentiment tagging mechanism is 'positive', 'neutral', 'negative'. In addition, the authors provide an entity database for stocks to aid in the classification task, but the database is limited by the authors' admission.

2.2 Other Data

We utilize two smaller datasets previously referenced in the introduction: SemEval 2017 5 Subtask 2 and FiQa Task 1 Sentiment Scoring Dataset. The datasets use a numerical scoring mechanism, ranging from -1 to 1.

3 Approach

In the SEntFiN paper, there was reliance on utilizing the manually constructed entity database to aid ABSA tasks. However, the database only includes 920 companies, and the overwhelming majority of these companies are listed on the Indian stock exchange. Thus, the database option in its current state is not scalable. Given the sophistication of the large pre-trained models that currently exist, we first explored the idea of ABSA as an NER classification driven task using advanced pre-trained mechanisms. With a sophisticated neural network, we could make a model that is capable of identifying entities and sentiments independent of a database. We also explored the idea of ABSA as a text generation task, inspired by Zhang et al., 2021. Additionally, we experimented with both one-step and two-step models under our generative framework.

3.1 Named Entity Recognition

3.1.1 Generative Model

The text-to-text transformer model T5 (Raffel et al., 2020) was adapted and fine-tuned for identifying targeted entities in financial headlines. To do so, we constructed target text output as entity names separated by commas. E.g. Text input = "Suffolk raises stake in Patni Computer Services to 5.3%". Text target = "Patni Computer Services, Suffolk".

3.1.2 Part of Speech Tagging

To re-purpose the task into an NER approach, we utilize a method similar to the B-I-O tagging scheme as referenced by Luo et al., 2019. The beginning of a token would be labeled as a 2,3, or 4. A negative token would be labeled as 2, neutral as 3, and positive as 4. This way the two components of ABSA, entity identification and sentiment analysis, are done in one task. The "I" in "B-I-O" would be labeled as 1, and the "O" would be labeled as 0.

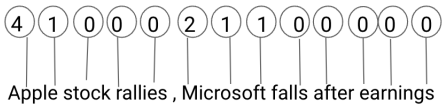

FIGURE 1: Example BIO tagging scheme

3.1.3 Conditional Random Field

In order to improve improve the NER-BIO token-classification process, we introduced a conditional random field (CRF) layer for structured predictions. Specifically, we constructed a BERT-CRF architecture for token classification, a similar approach to H-sum in Karimi et al., 2021.

3.2 Pre-training

While 10,753 news headlines are extensive, it is by no means exhaustive. To combat this, we utilized secondary datasets for pre-training. We cleaned 1,296 usable headlines for transfer-learning tasks from the aforementioned SemEval and FiQa data. We used the same NER-BIO tagging mechanism for relevant models. To be consistent with SEntFiN 1.0, we transformed the numerical sentiment scores in transfer-learning datasets to sentiment categories via a) "negative": $score \leq -0.2$ b) "neutral": $-0.2 < score < 0.2$ c) "positive": $score \geq 0.2$.

3.3 Target Sentiment Analysis

3.3.1 Generative Model

In addition to utilizing T5 model to the NER process, we explored its application to sentiment classification in a unified approach that combines the task of entity recognition with sentiment classification in a one-step T5 model. To do so, we constructed text output in the format of "⟨ Entity Name⟩: ⟨Sentiment Category⟩". E.g. Text input = "Suffolk raises stake in Patni Computer Services to 5.3%". Text target = "Patni Computer Services: positive, Suffolk: neutral".

3.3.2 Sequence Classification

For our non-unified/two-step models, after the NER process, we transform the input headlines as follows. First, we identify all the entities in an input headline. For each target entity, we label the target entity as "TGT" and other entities identified unchanged, and create a single-entity input headline with one target entity. We subsequently classify the sentiments of the transformed single-entity input headline sequences using standard BERT-based architectures.

4 Models

4.1 Setup and Baseline

For our main dataset SEntFiN 1.0, we used a 70%, 20%, 10% Train-Val-Test spilt. For all of our models, we set the max length to 30. The SEntFiN paper models run on top of manually annotated entities for input headlines. As a result, the model results are inflated by perfect entity identification. We view the manual annotation process as inefficient and intend to replace it with a modeled NER process as described earlier. Hence, we keep the best-performing model re-constructed according to the SEntFiN paper as only a reference. Our baseline model uses a T5 architecture and combines entity identification with sentiment classification in one step, i.e. a unified T5 model. We find a T5 structure appropriate for the task as it is versatile and could be conveniently adapted for our task.

4.2 Token Classification Models

We explored other unified models that are based on BERT architectures, which are reported to be best-performing by the SEntFiN paper. The aforementioned NER-BIO tagging scheme allows us to construct unified BERT-family-based models that tackles entity identification and sentiment analysis at once, similar to our baseline T5. We found the best results for the models utilized a dropout rate of $p = 0.15$, had a batch size of 16, and had two dense layers attached to the last hidden state (layer 1 = 120, layer 2 = 40).

FinBERT Naturally, FinBERT fits our goal well as it is specifically designed for financial context. As a result, we built FinBERT-based token classification model, using the aforementioned NER-BIO scheme.

RoBERTa While FinBERT is specifically designed for financial context, it frequently underperformed RoBERTa (Arslan et al., 2021). RoBERTa is much larger in size and also has specific pre-training in financial context. Since the simple version of RoBERTa dramatically outperformed FinBERT, consistent with SEntFiN paper, we decided to build the rest of our models using RoBERTa.

RoBERTa+CRF Architecture For CRF, we used a higher dropout rate of $p = 0.20$, a batch size of 32, and a single dense layer of size 120. We chose a higher dropout rate, a larger batch size, and a single dense layer because our initial models overfitted.

4.3 A Two-Step Approach

In addition to unified models, we constructed two-step models that tackle NER first and classification subsequently, based on step-1 NER tags.

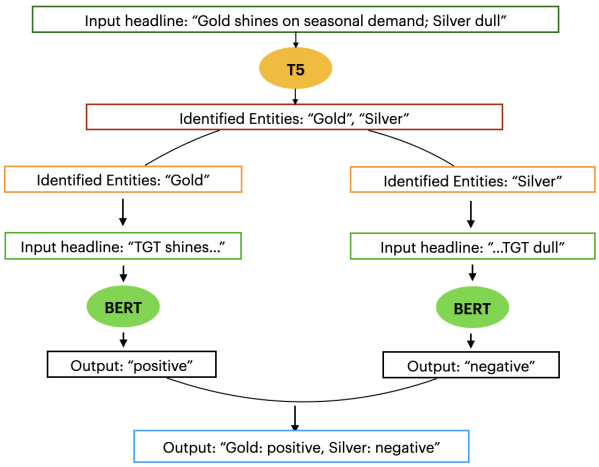

FIGURE 2: Two-step Approach

4.3.1 Step-1 Entity Recognition

We utilized a standard T5 model for NER. The "t5-base" model from HuggingFace was adopted, and trained with a batch size of 32, learning rate of 0.0001 and weight decay of 0.02. We achieved an full result accuracy of 80.78%, i.e. correctly identifying all entities in a headline for 80.78% of test headlines.

4.3.2 Step-2 Sequence Classification

BERT To identify the best classification model, we examined the performance of several BERT family models on headlines, with entities tagged using the SEntFiN manually annotated dictionary, i.e. perfect NER. We trained the HuggingFace "bert-base-cased" model with a hidden layer size of 30, learning rate of 0.00002, dropout rate of 0.2, and batch size of 32. The classification-only task performed upon perfectly labeled entities achieved an accuracy of 84.68% and F1-score of 0.86.

FinBERT We examined the sequence classification performance of FinBERT, trained with same set of hyperparameters as BERT model. We achieved a small improvement from BERT results, at an accuracy of 85.42% and a F1-score of 0.87.

RoBERTa We examined the RoBERTa model performance on the classification-only task. Training with the same set of hyperparamerters achieved the highest accuracy and F1-score among the three, at 86.44% and 0.89. As a result, we chose our RoBERTa model as the step-2 classification model. We thus combined RoBERTa with the step-1 NER T5 model, i.e., the T5+RoBERTa architecture.

Pre-training+RoBERTa We explored the impact of the aforementioned transfer learning using the SemEval and FiQa datasets. We first trained a RoBERTa model on classification-only task on the combined SemEval and FiQa data, using the same pre-processing method. The obtained model weights were then used to initialize our SEntFiN RoBERTa model to conduct final training. For the training of SemEval and FiQa data, we chose hidden size of 30, dropout rate of 0.15, and learning rate of 0.000005. The final RoBERTa model was then combined with T5 as illustrated in Figure 2.

5 Results and Analysis

5.1 Evaluation Metrics

We defined full-text accuracy as our main evaluation metric. For text-generative models, we examine whether the output text is the same as target text in its entirety. For the token-based classification tasks, we examine whether a model correctly classifies all tokens in an input. We chose such metric because a) it is applicable to all model architectures b) of greater business significance especially in the case of conflicting sentiments with respect to different entities.

As an example, imagine a headline was “Apple Inc. did well relative to big tech stocks”. If the model predicted that both Apple and big tech did well, we would wrongly invest in big tech companies like Microsoft and Facebook, when in reality we should sell the rest of big tech and buy Apple. Thus, this scoring mechanism emphasizes fully understanding the news headline.

While it is desirable to fully predict the entire classification task target output, the most important words/tokens to predict are those at the beginning, as entity names are usually short and highly likely unique in a headline. IE, did the model accurately predict the start of the correct entity, and consequently, the accurate sentiment? Under this adjusted classification accuracy framework, “Apple: positive” would be a correct output, but incorrect under the full accuracy framework. Meanwhile, predicting “Apple inc., positive” would be considered correct in both scoring mechanisms. This adjusted metric is particularly useful in a business setting with an entity-linking database aid, while the full accuracy metric might be too punitive.

5.2 Results

Model Type	Model	Full-Sentence Accuracy	Adj. Full-Sentence Accuracy
Baseline	*T5 Unified*	*61.01%*	*64.81%*
NER-BIO-Token-Based Classification	FinBERT	65.18%	69.17%
	RoBERTa	67.60%	72.42%
	Pretrain+RoBERTa	68.62%	72.61%
	RoBERTa+CRF	70.84%	74.65%
Two-Step Classification	**T5+RoBERTa**	**72.79%**	**76.14%**
	T5+Pretrain+RoBERTa	71.96%	75.12%

FIGURE 3: Model Results

Figure 3 shows the results evaluated on both full-sentence accuracy and adjusted full-sentence accuracy basis for our SEntFiN task. Our baseline T5 unified model has an accuracy of 61.01%. For our token-based unified models, RoBERTa+CRF achieved the best result, with a 70.84% accuracy. Our best two-step classification model, T5+RoBERTa, performed better than token-based unified models by 1.12%, as measured by full-sentence accuracy, and 1.49% by adjusted full-sentence accuracy. Overall, relaxing the standard for entity labeling boost accuracy by around 4% across our models. For both types of model structures, transfer learning or pre-training did not improve the model performance.

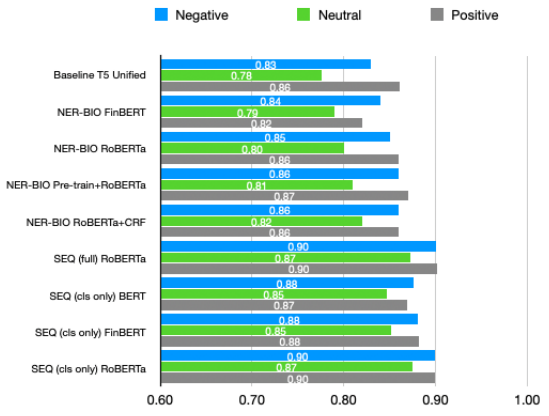

FIGURE 4: F1 Scores

We observed from our results that the "neutral" sentiment received meaningfully lower F1-scores compared

to the more polarized sentiments. Such results are not surprising given the ambiguity in neutral texts.

The baseline T5 model performed decently well for positive sentiments, with a F1 score of 0.86. The more significant improvement in model performance to the baseline was for "neutral" sentiments, which had almost 10 points improvements. Interestingly, noisier entity tagging did not meaningfully affect downstream classification performance, as evidenced by similar F1 scores across all sentiment categories between the classification-only RoBERTa model and T5+RoBERTa model, but with a slightly larger difference in "neutral" category.

Of particular note is the comparison of the NER-BIO pre-train model relative to other NER-BIO models. The pre-train model was the highest scoring F1 score for the positive token. This likely has to do with the way we converted our numerical label pre-training to a categorical positive/neutral/negative. We likely overestimated the score that the human annotators used for neutral sentiments, and therefore only the highest scoring sentiments would be labeled positive. We minimized the space available to positive scores and overestimated the space for neutral scores.

5.3 Multiple Entities and Conflicting Sentiments

We further evaluated model performance in two cases, when only one entity is present in the headline, and when multiple entities are present in the headline.

	Single-Entity Accuracy	Multi—Entity Accuracy
T5 Unified (Baseline)	*75.92%*	*34.37%*
RoBERTa+CRF	**81.62%**	55.56%
T5+RoBERTa	81.37%	**61.81%**

FIGURE 5: Single vs. Multiple Entity Model Performance

The authors of SEntFiN make a note about the difficulty of predicting multiple entities in one headline. Our results confirm that, with our best models performing $20 - 25\%$ worse on multiple entity sentiments. Although T5+RoBERTa is the best performing model, the RoBERTa+CRF performs slightly better in single-entity accuracy; RoBERTa+CRF underperforms in multi-entity accuracy. Our best guess for this is due to class imbalance of the dataset. SEntFiN has 26.47% of their labels as multi-entity, so the maximum likelihood is going to predict one entity. The CRF model may underestimate the likelihood of n >1 entities. Given the dependence of one word to another word in a sentence, the joint probability of n-entities is likely lower than the product of n-entity marginal probabilities. Meanwhile, the T5+RoBERTa model isolates the likelihood of each entity and then predicts the sentiment. We were unable to implement a CRF model in the 2-step approach due to time constraints of the project but we encourage future researchers to do so.

	Conflicting Sentiments Accuracy	Consistent Sentiments Accuracy	Multi—Entity Accuracy
T5 Unified (Baseline)	*25.38%*	*41.77%*	*34.37%*
RoBERTa+CRF	58.46%	53.16%	55.56%
T5+RoBERTa	**66.92%**	**57.59%**	**61.81%**

FIGURE 6: Conflicting vs. Consistent Sentiments Model Performance

In a multi-entity setting, the models are able to identify conflicting sentiments at a higher accuracy than consistent sentiments. This builds off the notion of polarity, with the neutral sentiment under-performing its polar counterparts (see the discussion of f1-scores). It is more difficult to identify similar sentiments at high conviction, particularly in condensed text. The models can predict polarity as opinionated words are further away from a sentiment-decision boundary.

5.4 Model Error Analysis

Our models incorrectly predicted these sentences, and we have a general explanation why. These incorrect predictions underscore the difficulty and idiosyncrasies of financial text. Sometimes in finance, words with negative connotations can be interpreted in a positive fashion. They also underscore the difficulty of how the dataset was labeled, and how our models could be decieved by such a labeling scheme.

Example Headline	Model Outputs	Target Outputs	Description
Sterling hits 2-week low vs dollar, as housing fervour cools	{'Sterling': 'negative', 'dollar': 'neutral'}	{'Sterling': 'negative'}	*Wrong NER Tagging*
Gold's longest run in a year ends as US inflation picks up	{'Gold': 'neutral'}	{'Gold': 'negative'}	*Ambiguous sentiment*
Standard Chartered axes 15,000 jobs, raises $5.1 billion in capital	{'Standard Chartered': 'negative'}	{'Standard Chartered': 'positive'}	*Failure to digest finance specific information*
Kwality among cheapest plays in dairy sector: Ashish Maheshwari	{'Kwality': 'negative'}	{'Kwality': 'positive'}	*Failure to understand word in a financial context*
I do not see too much of a downside for Infy: Sandeep Wagle	{'Infy': 'positive'}	{'Infy': 'negative'}	*Target output incorrectly annotated*
Negative on Tata Motors: Ambareesh Baliga, Way2Wealth Brokers Pvt. Ltd	{'Tata Motors': 'neutral', 'Way2Wealth': 'neutral'}	{'Tata Motors': 'neutral', 'Way2Wealth Brokers Pvt. Ltd': 'neutral'}	*NER tagging essentially correct*
A $25 fall in crude is like a $10-billion stimulus for Indian economy, say experts; top stock bets	{'crude': 'positive'}	{'crude': 'neutral'}	*Failure to isolate target-specific sentiment from overall sequence sentiment*

FIGURE 7: Examples of Incorrect Results

6 Discussion and Next Steps

6.1 Entity Linking

We were able to generate strong results by matching tokens and classifying them. However, from a business setting, we would need to link these entities to their stock tickers so that we can complete a fully automated task. Additionally, we would never mix up "Way2Wealth brokers" from "Way2Wealth Brokers Pvt. Ltd" (Figure 7) as they would map to the same thing in the database. Entity-linking would also help consolidate labels as they show up in unique formats in our text like "CitiBank", "Citi", or "Citigroup".

We were not able to find a robust naming convention database, particularly for the Indian stock market.

6.2 Back-testing Business Significance

One crucial next step is to examine the practical value of our proposed framework. To do so, we would need to perform back-testing on our model's profit-generation. Based on the approach proposed by Lopez-Lira and Tang, 2023, we intend to evaluate the aggregate sentiment of news on specific stocks in one day, and establish a prediction model using the change in price during the same day. Multiple factors would be adjusted. Firstly, the idiosyncrasy of stock-specific news needed to be isolated from overall market sentiments and movements. We intend to use the CAPM financial model to separate out such factor. Secondly, we need to take into consideration after-market news, which would be likely reflected in the next day's price movements. Thirdly, transaction volume and liquidity could affect the speed of price adjustments to news headlines.

6.3 Conclusion

In this paper, we try a number of approaches to augment the SEntFiN dataset. Utilizing silver training data, multiple NER-BIO approaches and T5 generative models, we generate a score of 74.16% that can identify a financial entity and then perform sentiment analysis. Our highest scoring model improves on the baseline by 11.13%. Our findings conclude that a two step solution using T5+RoBERTA approach is likely best suited for this task. Experimenting with T5+RoBERTA approaches and a CRF layer would be compelling for future research.

After analyzing the SEntFiN dataset rigorously, we explore a number of reasons where the training set can be improved. Building a larger dataset, and developing an entity-linking database, will greatly improve the ability of the models in the future. Moreover, training on other forms of financial data can greatly advance these models.

References

Yusuf Arslan, Kevin Allix, Lisa Veiber, Cedric Lothritz, Tegawendé F. Bissyandé, Jacques Klein, and Anne Goujon. 2021. A comparison of pre-trained language models for multi-class text classification in the financial domain. In *Companion Proceedings of the Web Conference 2021*, WWW '21, page 260–268, New York, NY, USA. Association for Computing Machinery.

Johan Bollen, Huina Mao, and Xiaojun Zeng. 2011. Twitter mood predicts the stock market. *Journal of Computational Science*, 2(1):1–8.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners.

Wesley S Chan. 2003. Stock price reaction to news and no-news: drift and reversal after headlines. *Journal of Financial Economics*, 70(2):223–260.

Keith Cortis, André Freitas, Tobias Daudert, Manuela Huerlimann, Manel Zarrouk, Siegfried Handschuh, and Brian Davis. 2017. SemEval-2017 task 5: Fine-grained sentiment analysis on financial microblogs and news. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 519–535, Vancouver, Canada. Association for Computational Linguistics.

Akbar Karimi, Leonardo Rossi, and Andrea Prati. 2021. Improving bert performance for aspect-based sentiment analysis.

Xiaodong Li, Chao Wang, Jiawei Dong, Feng Wang, Xiaotie Deng, and Shanfeng Zhu. 2011. Improving stock market prediction by integrating both market news and stock prices. In *Database and Expert Systems Applications*, pages 279–293, Berlin, Heidelberg. Springer Berlin Heidelberg.

Xin Li, Lidong Bing, Piji Li, and Wai Lam. 2019. A unified model for opinion target extraction and target sentiment prediction. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33(01):6714–6721.

Alejandro Lopez-Lira and Yuehua Tang. 2023. Can chatgpt forecast stock price movements? return predictability and large language models. *SSRN*.

Huaishao Luo, Tianrui Li, Bing Liu, and Junbo Zhang. 2019. DOER: Dual cross-shared RNN for aspect term-polarity co-extraction. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 591–601, Florence, Italy. Association for Computational Linguistics.

Handschuh S. Freitas A. Davis B. McDermott R. Zarrouk M. Balahur A Maia, M. 2028. Www'18 open challenge: Financial opinion mining and question answering. *Companion Proceedings of the The Web Conference 2018*, pages 1941–1942.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer.

Ankur Sinha, Satishwar Kedas, Rishu Kumar, and Pekka Malo. 2022. Sentfin 1.0: Entity-aware sentiment analysis for financial news. *Journal of the Association for Information Science and Technology*, 73(9):1314–1335.

Marjan Van de Kauter, Diane Breesch, and Véronique Hoste. 2015. Fine-grained analysis of explicit and implicit sentiment in financial news articles. *Expert Systems with Applications*, 42.

Feixiang Wang, Man Lan, and Wenting Wang. 2018. Towards a one-stop solution to both aspect extraction and sentiment analysis tasks with neural multi-task learning. pages 1–8.

Hang Yan, Junqi Dai, Tuo Ji, Xipeng Qiu, and Zheng Zhang. 2021. A unified generative framework for aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2416–2429, Online. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2021. Towards generative aspect-based sentiment analysis. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 504–510, Online. Association for Computational Linguistics.

Wenxuan Zhang, Xin Li, Yang Deng, Lidong Bing, and Wai Lam. 2022. A survey on aspect-based sentiment analysis: Tasks, methods, and challenges. *IEEE Transactions on Knowledge and Data Engineering*, pages 1–20.