Improving Financial Sentiment Classification on ELECTRA using Adversarial Attacks.

Abstract

This paper focused on the task of sentiment analysis within the financial domain, aiming to classify text into positive, negative, or neutral sentiments. Employing an ELECTRA-small model initially pre-trained as a general sentiment classifier, a baseline model was trained on financial sentiment data, achieving an accuracy of 0.8547 on the Financial PhraseBank dataset. Misclassification of data between positive and neutral sentiment classes was the most pronounced cause of error. While attempts to augment the model's financial vocabulary using the FinRAD dataset led to decreased model accuracy, the introduction of adversarial attacks proved to be successful in improving the performance of the baseline model. Particularly, the model trained on data augmented with TextFooler-generated adversarial examples exhibited a 4.68% increase in accuracy to 0.9015. This approach also reduced misclassifications between positive-neutral classes, thus mitigating the major challenge observed in the baseline model. This result is significant in light of how well the model generalized to a challenging problem on a dataset that it never encountered before, and sets this result apart from other contemporary work in literature which uses a subset of Financial Phrasebank as training data for fine-tuning.

1. Introduction

Financial markets are profoundly influenced by investor sentiment, news, and social media discussions. Since financial language often incorporates a specialized vocabulary that relies on nuanced expressions rather than easily discernible positive or negative words, traditional sentiment analysis models often struggle to interpret this. However, a model attuned to the subtleties of financial language can provide valuable insights into market sentiment, aiding in risk management, investment strategies, and market predictions.

This paper revolves around polarity analysis, a task that involves categorizing text into positive, negative, or neutral sentiments within a particular domain. There are prominent challenges inherent in sentiment analysis within the financial domain. Firstly, the implementation of state-of-the-art classification methods, particularly those leveraging neural networks, demands extensive labeled data. However, the process of annotating financial text snippets with sentiment labels requires specialized expertise and can be a resource-intensive undertaking. Secondly, as previously mentioned, conventional sentiment analysis models trained on general corpora prove unsuitable for this task.

In this paper, an ELCTRA-small model pre-trained as a general sentiment classifier is trained on financial sentiment data. After evaluating its performance in predicting the sentiment of the sentences in the FinancialPhraseBank dataset, methods are explored to try to improve the model's performance.

2. Method

2.1 Overview

The following are the implementation details of the baseline model, methods undertaken to improve it, hyperparameters, and evaluation metrics utilized for the task. Details about the datasets used or created for training, validation, and testing of the fine-tuned model are given in section 2.4 for improved clarity.

2.2 Baseline Model

To get a baseline model, the ELECTRA-small (Clark et al.,2020) model with a sequence classification head on top (linear layer on top of pooled output) was downloaded from HuggingFace and pre-trained as a generic sentiment classifier using sentiment data from TweetEval dataset (Barbieri et al.,2020). This pre-trained model was then trained on the Twitter Financial News sentiment dataset to assess the performance for the financial sentiment classification task. The performance of this trained model was evaluated on its ability to correctly predict the sentiment of sentences in the Financial PhraseBank dataset created by Malo et al (2014). This was a dataset with financial sentiment data but one on which the model was not trained on, in an effort to analyze how well the model generalizes to understand real-world financial phrases. We used this Financial PhraseBank dataset to benchmark the performance of all models created.

2.3 Experiments for Model Improvement

Two experiments were conducted with an aim to enhance the baseline model's performance.

One of the more common types of failure of the baseline in-domain trained model was in distinguishing between positive and neutral sentiment statements. So the objective in Experiment 1 was to improve learning on this hard subset of data by improving the model's vocabulary. The FinRAD dataset (Ghosh et al.,2022), which is a collection of 13k financial terms and definitions, was used to create additional neutral sentiment data and extend the model's financial vocabulary by adding a template sentence and label as follows.

["sentence": "The term <term> means <definition>.", "label": 1]

To expand the model's vocabulary, a custom function was written to introduce new tokens derived from financial terms that were present in FinRAD but absent in the original tokenizer's vocabulary. The new words were added to the tokenizer which increased the vocabulary size from 30522 to 40303. The model's token embeddings were resized accordingly to account for the increased vocabulary. After training the model on this augmented dataset, its performance was benchmarked again using the Financial Phrasebank dataset.

Experiment 2 consisted of performing three different adversarial attacks and training the model on the Twitter Financial News dataset augmented with each of the generated adversarial datasets. All three attack models iterated through the dataset and created an adversarial perturbation for 1000 correctly predicted sentences. The attack model created by Pruthi et al. (2019) was based on measuring visual similarity, generating perturbations that were visually similar while introducing realistic spelling mistakes or 'typos'. The other two attacks - TextFooler (Jin et al.,2020) and BAE (Garg et. al., 2020) were based on measuring semantic similarity, generating perturbations that were grammatically valid and semantically indistinguishable. Once the perturbations were generated, the baseline model was attacked and perturbations used in successful attacks were added to the Twitter Financial News dataset. TextAttack (Morris et al.,2020), a Python framework for adversarial attacks, training, and data augmentation was used to perform the attacks and generate perturbations. After training, the performance of the models was benchmarked using the Financial Phrasebank dataset.

2.4 Datasets

2.4.1 Financial Phrasebank dataset

This dataset comprises 4,845 randomly selected English sentences from financial news on the LexisNexis database. The sentences were annotated by 16 finance and business experts, categorizing each sentence into positive, negative, or neutral sentiments. The dataset is subdivided based on agreement rates (50%, 66%, 75%, 100%) among 5-8 annotators. In this study, only the subset with 100% agreement among annotators (amounting to 2,264 sentences) was utilized for testing. The testing dataset is imbalanced, with 61.4% neutral, 25.2% positive, and 13.4% negative sentiment sentences.

2.4.2 TweetEval (sentiment) dataset

The ELECTRA-small model was pre-trained using the Sentiment subset of the TweetEval dataset (Barbieri et al., 2020). This subset consists of generic tweets with associated sentiment labels. To address the minority class imbalance, where the negative class had 7,093 rows, an equal number of rows from the neutral and positive classes were randomly selected, creating a balanced dataset. This approach was favored over oversampling or other augmentation methods to avoid inflating the dataset size and increasing computational cost. Training utilized 75% of the data, while 5% and 20% were allocated to validation and test sets, respectively.

2.4.3 Twitter Financial News sentiment dataset

The Twitter Financial News sentiment dataset from HuggingFace contained an annotated corpus of 11932 finance-rated tweets and sentiments. Since the negative class which was the minority class had 1442 rows, the same number of rows from the neutral and positive classes were similarly randomly selected to create a balanced dataset for reasons as explained in the previous section.

2.4.4 FinRAD dataset

The FinRAD dataset (Ghosh et al.,2022), a collection of 13k financial terms and definitions, was used to construct sentences that were guaranteed to be neutral and increase the baseline model's financial vocabulary. An 80-20 train-validation split was used on this dataset.

2.5 Hyperparameters

The ELECTRA-small sequence classification model was trained with default hyperparameters, except for the batch size which was set to 64, as it demonstrated improved performance. The training spanned 10 epochs with a learning rate of 2e-5. The model architecture comprises 12 layers, a hidden dimension of 256, 4 attention heads, and an embedding size of 128.

2.6 Evaluation criteria

All models were assessed using the Financial PhraseBank dataset created by Malo et al (2014). No data from this dataset was utilized for training or validation to ensure the evaluation occurred on unseen data. Evaluation metrics included average accuracy, loss, and F1 score. Given the dataset's imbalance, with 60% of sentences being of neutral sentiment, and the equal importance of predictive performance across all three classes (positive, negative, and neutral), a macro-averaged F1 score was employed. Additionally, confusion matrices were computed for each model's test results to provide insights into the classes where the model exhibited the most confusion.

3. Results

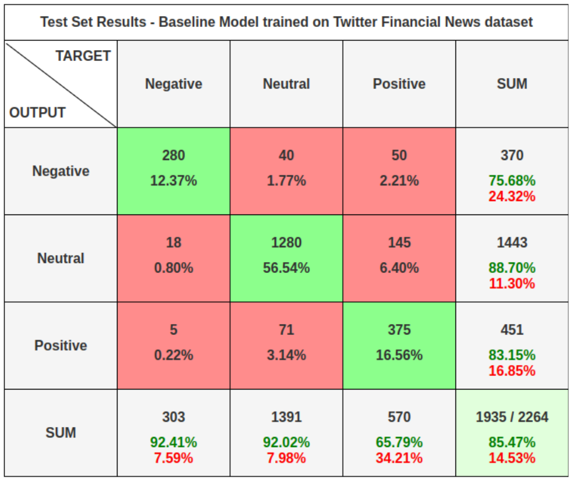
Table 1 summarizes the performance of various models on the Financial PhraseBank dataset. The evaluated models include the ELECTRA-small model pre-trained as a general sentiment classifier, the baseline model trained on the in-domain dataset, the FinRAD model trained on the dataset augmented with FinRAD data, the Pruthi model trained on the dataset augmented with adversarial data generated by Pruthi et al. (2019), the BAE model trained on the dataset augmented with adversarial data using BAE, and the TextFooler model trained on the dataset augmented with adversarial data generated by TextFooler.

Table 1: Performance of models on test set

Model Name	Accuracy	Loss	Macro F1 score
Pretrained Electra	0.7005	0.7465	0.5678
Baseline model	0.8547	0.4495	0.8233
FinRAD model	0.7703	0.6777	0.6659
Pruthi model	0.8975	0.3318	0.8763
BAE model	0.8874	0.3536	0.8472
TextFooler model	**0.9015**	**0.3142**	**0.8776**

In Table 1, boldface highlights the best results in each corresponding metric. The baseline model, trained on the Twitter Financial News sentiment dataset augmented with adversarial data generated using the TextFooler attack, outperformed others. The model achieved an accuracy of 0.9015, a 4.68% improvement over the baseline model's accuracy of 0.8547. Notably, the accuracy of the majority class classifier on the Financial PhraseBank dataset was substantially lower at 0.6140 and that of the ELECTRA-small model pre-trained as a general sentiment classifier was at 0.7005. While the Pruthi model and BAE model also enhanced accuracy by 4.28% and 3.27%, respectively, augmenting the dataset with FinRAD data resulted in an 8.44% decrease in model accuracy.

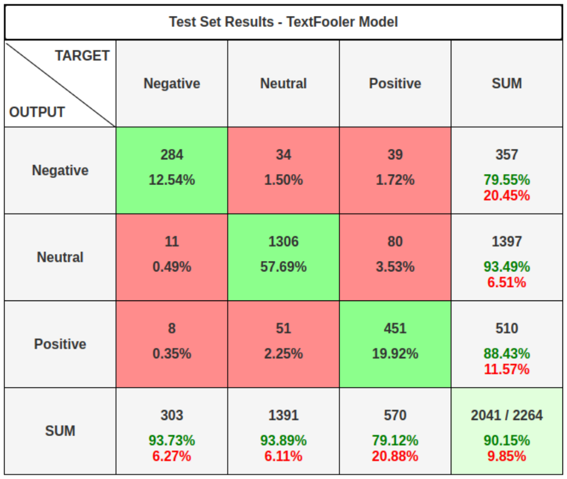
Figure 1 presents the confusion matrices for the test results of the baseline model, highlighting the classes where the model tends to make the most errors, particularly between positive and neutral sentiments. In Figure 2, the test results from the best model (TextFooler model), both overall accuracy and predictive performance on the more challenging subset of data are depicted. The 'Target' denotes the actual sentiment labels, while the 'Output' represents the predicted labels.



Test Set Results - Baseline Model trained on Twitter Financial News dataset

TARGET OUTPUT	Negative	Neutral	Positive	SUM
Negative	280 12.37%	40 1.77%	50 2.21%	370 75.68% 24.32%
Neutral	18 0.80%	1280 56.54%	145 6.40%	1443 88.70% 11.30%
Positive	5 0.22%	71 3.14%	375 16.56%	451 83.15% 16.85%
SUM	303 92.41% 7.59%	1391 92.02% 7.98%	570 65.79% 34.21%	1935 / 2264 85.47% 14.53%

Figure 1: Confusion Matrix of Baseline model



Test Set Results - TextFooler Model

TARGET OUTPUT	Negative	Neutral	Positive	SUM
Negative	284 12.54%	34 1.50%	39 1.72%	357 79.55% 20.45%
Neutral	11 0.49%	1306 57.69%	80 3.53%	1397 93.49% 6.51%
Positive	8 0.35%	51 2.25%	451 19.92%	510 88.43% 11.57%
SUM	303 93.73% 6.27%	1391 93.89% 6.11%	570 79.12% 20.88%	2041 / 2264 90.15% 9.85%

Figure 2: Confusion matrix of TextFooler model

4. Analysis

The baseline model, trained on the Twitter Financial News sentiment dataset, achieved an accuracy of 0.8547 when tested on the Financial PhraseBank dataset. This aligns with the 0.86 accuracy reported by Araci (2019) in their evaluation of a BERT model fine-tuned for financial sentiment analysis on the same dataset. Araci (2019) further improved the model's accuracy to 0.97 by training it on a subset of data taken from the test dataset (Financial PhraseBank). However, in this paper, the Financial PhraseBank dataset is solely utilized as a test set, and the emphasis is on enhancing the model performance from the baseline of 0.8547 through alternative methods.

Most disagreements between the human annotators annotating the Financial Phrasebank dataset were also between positive-neutral labels (Malo et al.,2014). Looking at the confusion matrix of the test results of the baseline model, it is clear that the model also makes the most mistakes in predicting positive-neutral labels (6.4% of positives were predicted incorrectly as neutral and 3.14% of the neutrals were incorrectly predicted as positive).

This finding is consistent with results found by Araci (2019) when evaluating a BERT model fine-tuned for financial sentiment analysis who hypothesized that this might be due to the difficulty in distinguishing between the statements with actual positive sentiment and the statements employed with corporate embellishments that frequently occur in financial documents. Another 2% of errors each were between positive-negative and negative-neutral classes of errors.

To understand the fine-grained flaws in the model, below are some examples of incorrect predictions representing different generic types of mistakes that the model makes.

Example 1: Finnish textiles and clothing group Marimekko Oyj posted a net profit of 7.99 mln euro $ 10.4 mln for 2006, compared to 8.4 mln euro $ 10.9 mln for 2005
True Label: Negative; Predicted label: Positive

Example 2: Operating loss was EUR 179mn, compared to a loss of EUR 188mn in the second quarter of 2009.
True Label: Positive; Predicted label: Negative

Example 3: Scanfil issued a profit warning on 10 April 2006.
True Label: Negative; Predicted label: Neutral

Example 4: The group reiterated its forecast that handset manufacturers will sell around 915 mln units this year globally.
True Label: Neutral; Predicted label: Negative

Example 5: EQ Bank forecasts Olvi 's net sales at EUR 67mn in the second quarter of 2009, and operating profit at EUR 6.4 mn
True Label: Neutral; Predicted label: Positive

Example 6: Finnish steel maker Rautaruukki Oyj (Ruukki) said on July 7, 2008 that it won a 9.0 mln euro ($ 14.1 mln) contract to supply and install steel superstructures for Partihallsforbindelsen bridge project in Gothenburg, western Sweden.
True Label: Positive; Predicted label: Neutral

Example 7: In December alone, the members of the Lithuanian Brewers ' Association sold a total of 20.3 million liters of beer, an increase of 1.9 percent from the sales of 19.92 million liters in December 2004.
True Label: Positive; Predicted label: Neutral

Example 8: The disposal of Autotank will also strengthen Aspo's capital structure," commented Gustav Nyberg, CEO of Aspo.
True Label: Positive; Predicted label: Neutral

One class of errors in the model stems from its inability to grasp mathematical relationships, particularly in discerning which numerical values are higher. This happens, for instance, when a sentence laden with numerical figures contains the term 'net profit,' yet lacks any overtly negative words. Example 1 falls within this specific group. Likewise, another cluster of misclassifications involves sentences featuring numbers and the term 'loss,' mistakenly labeled as negative due to the model's challenge in determining the comparative magnitude of numerical values (as illustrated in Example 2).

Yet another set of errors arises from the model's limited comprehension of certain financial terms. For instance, sentences containing the term 'profit warning' (Example 3) are wrongly classified as neutral rather than negative. This misclassification arises because 'profit' holds a positive attribution score, while 'warning' carries a negative attribution score. Similarly, the model designates Example 8 as neutral due to its lack of recognition of the term 'capital structure,' having only a positive attribution score for the word 'capital' and a negative score for 'disposal.'

Another class of errors comes from the assumptions the model makes, due to a failure to understand context in long sentences. A substantial number of positive sentences featuring expressions like 'awarded', ‘won’' 'signed agreement,' or 'made deals' (Example 6) were erroneously classified as neutral because having many neutral tokens such as “superstructure” and “bridge” influenced the overall sentiment to lean towards neutrality.

Finally, neutral sentences incorporating the term 'sell' (Example 4) were misclassified as negative, driven by the strong negative attribution score associated with this word. Conversely, sentences containing words such as 'net sales,' 'profit,' and 'valued at' (Example 5)were misclassified as positive rather than neutral due to the strong positive attribution scores associated with these words.

The rationale for expanding the model's financial vocabulary through the integration of the FinRAD dataset was to enhance the model's recognition of financial terminology, with the expectation that this augmentation would lead to improved classification and a reduction in misclassifications as neutral. However, this strategy backfired, resulting in an increase in sentences misclassified as neutral and an 8% drop in accuracy.

This outcome might be attributed to the fact that, although the model learned additional financial terms, it encountered these terms primarily in the context of neutral sentiment sentences within the FinRAD training data(since all sentences generated using FinRAD were strictly labeled as neutral). Therefore, the model may have developed an association between these terms and neutral sentiment, contributing to the observed increase in misclassifications and the decline in overall accuracy. In the absence of these terms, the tokenizer simply converted these tokens into UNK (unknown) token, which has neither positive, negative or neutral connotations as was ignored by the model in its entirety.

In contrast, adversarial attacks proved significantly more effective in enhancing the model's performance. The model trained on the dataset augmented with adversarial data generated from TextFooler demonstrated the most substantial increase in accuracy, achieving a 4.68% improvement and emerging as the top-performing model. Notably, this model excelled in mitigating the misclassification of positive-neutral classes, which exhibited the most significant misclassification in the baseline model.

Examining the confusion matrices of test results highlights the model's advancements. In the baseline model (Figure 1), 6.4% of sentences were erroneously predicted as neutral instead of positive, and 3.14% of data were inaccurately predicted as positive instead of neutral. In contrast, the best model (Figure 2) reduced these misclassifications, with only 3.53% of sentences incorrectly predicted as neutral instead of positive and 2.25% of data inaccurately predicted as positive instead of neutral. Consequently, the overall misclassification between positive and neutral classes decreased from 9.54% to 5.78%, reflecting a notable reduction of 3.76%.

Furthermore, this improved model exhibited a reduction in misclassification across other classes as well. Misclassification in neutral-negative and positive-negative classes decreased by 0.58% and 0.36%, respectively, underscoring the broader efficacy of the model in refining predictions across various sentiment classes.

The improved model demonstrated enhanced accuracy in classifying sentences. It correctly identified sentences with words such as 'profit warning' as negative sentiment and sentences with words such as 'awarded or won contracts,' and 'signed agreement' as positive sentiment, avoiding the previous misclassification as neutral. As the model learns to classify both genuine and adversarial examples, it develops a more nuanced understanding of the underlying sentiment patterns in the data rather than focusing attention on specific words. This encourages the model to learn robust features that generalize well to various linguistic nuances present in financial language. Despite these improvements, challenges persisted in cases where the model encountered sentences laden with numerical figures containing only pivotal terms like 'profit' or 'loss' (as observed in Example 1 and 2). In such instances, the model still exhibited misclassifications due to its ongoing struggle with understanding mathematical relationships.

5. Conclusion

This paper focused on the task of sentiment analysis within the financial domain, aiming to classify text into positive, negative, or neutral sentiments. An ELECTRA-small model was initially pre-trained as a general sentiment classifier. Then this model was trained on a financial sentiment dataset to create a baseline model that had an accuracy of 0.8547 when evaluated on the Financial phrasebank dataset. It was observed that the maximum amount of misclassification happened between the positive- and neutral classes. The attempt to enhance the model's performance by extending the financial vocabulary of the model using the FinRAD dataset resulted in a decrease in accuracy, whereas adversarial attacks proved successful in improving the model's performance. The model trained on the dataset augmented with adversarial data generated from TextFooler had the best performance, improving the model's accuracy by 4.68% to 0.9015. It also effectively mitigated the misclassifications, particularly in the positive-neutral classes, addressing a major challenge observed in the baseline model. This result is significant in light of how well the model generalized to a challenging problem on a dataset that it never encountered before, and sets this result apart from other contemporary work in literature which uses a subset of Financial Phrasebank as training data for fine-tuning.

References

Araci, D. (2019). Finbert: Financial sentiment analysis with pre-trained language models. arXiv preprint arXiv:1908.10063

Barbieri, F., Camacho-Collados, J., Neves, L., & Espinosa-Anke, L. (2020). Tweeteval: Unified benchmark and comparative evaluation for tweet classification. *arXiv preprint arXiv:2010.12421.*

Clark, K., Luong, M. T., Le, Q. V., & Manning, C. D. (2020). Electra: Pre-training text encoders as discriminators rather than generators. *arXiv preprint arXiv:2003.10555.*

Garg, S., & Ramakrishnan, G. (2020). Bae: Bert-based adversarial examples for text classification. *arXiv preprint arXiv:2004.01970.*

Ghosh, S., Sengupta, S., Naskar, S. K., & Singh, S. K. (2022, June). FinRAD: Financial Readability Assessment Dataset-13,000+ Definitions of Financial Terms for Measuring Readability. In *Proceedings of the 4th Financial Narrative Processing Workshop@ LREC2022* (pp. 1-9).

Jin, D., Jin, Z., Zhou, J. T., & Szolovits, P. (2020, April). Is bert really robust? a strong baseline for natural language attack on text classification and entailment. In *Proceedings of the AAAI conference on artificial intelligence* (Vol. 34, No. 05, pp. 8018-8025).

Malo, P., Sinha, A., Korhonen, P., Wallenius, J., & Takala, P. (2014). Good debt or bad debt: Detecting semantic orientations in economic texts. *Journal of the Association for Information Science and Technology*, *65*(4), 782-796.

Morris, J. X., Lifland, E., Yoo, J. Y., Grigsby, J., Jin, D., & Qi, Y. (2020). Textattack: A framework for adversarial attacks, data augmentation, and adversarial training in nlp. *arXiv preprint arXiv:2005.05909.*

Pruthi, D., Dhingra, B., & Lipton, Z. C. (2019). Combating adversarial misspellings with robust word recognition. *arXiv preprint arXiv:1905.11268.*