Analyzing BERT and RoBERTa Performance on Aspected-Based Sentiment Classification on Financial Microblog and Headlines Dataset

Yuhao Ding
New York Univeristy Shanghai
yd1158@nyu.edu

Minyi Wan
New York Univeristy Shanghai
mw3706@nyu.edu

Abstract

Aspect-based sentiment analysis deals with the scenario where a sentence may have different sentiments towards different aspects of it. In this paper, we choose a Financial Microblog and Headlines Dataset that contains aspects information, presented as FiQa 2018 Challenge, and focus on two questions: 1) How does feeding aspect features affects BERT and RoBERTa performance? 2) And how do both models perform on complicated sentences where a sentence may have different aspects and potentially different sentiments toward each of them? For the first question, our results intriguingly show that aspect-based features are helpful for the BERT model, but are not helpful to the RoBERTa model. We then introduce a "Concatenate-Sentence" testing method that naively combines two sentences with different aspects into one sentence to model complex scenarios which a sentence may have different aspects. Our results show that both models perform poorly on this task, especially when a sentence contains different aspects and different sentiments. This leads to our final conclusion that a better way to train both models with aspect features is needed.

1 Introduction

Social media has been increasingly important to forecast future outcomes in the market, and sentiment data from microblogs such as Twitter and StockTwits has been shown to be useful (Asur and Huberman, 2010). Research shows that opinions expressed in these microblogs have significant effects in financial market performance (Goonatilake et al., 2007). A more complicated sentiment classification problems considers when a sentence have different sentiments towards each different aspect of it. For sentence such as "This company's strategy is good, but its stock price is falling", it is expressing a positive sentiment towards its corporate strategy, but a negative sentiment on its stock performance. (Jangid et al., 2018).

Aspect-based sentiment analysis (ABSA) is introduced to overcome this issue (Thet et al., 2010). The sentiment is predicted by taking consideration into the aspects extracted from the sentence. Aspect data refers to a label chosen from an already defined set of aspects (e.g. Quality, Area) in the domain (Hoang et al., 2019). In the Financial Opinion Mining in FiQA (2018) Challenge, it provides a text instance in a microblog message, along with its aspect and sentiment score.[1] The original task is to detect the target aspects, and predict the sentiment of the sentence.

In this work, we simplify the task to using the correct aspects directly in predicting the sentiment. We use two models: BERT and RoBERTa for this classification task. As shown by previous researchers, rather than leaving second part of input (textB) in BERT blank as in normal sentiment classification task (Devlin et al., 2018), we can put aspect data in textB (Hoang et al., 2019). This can also be applied to RoBERTa.

Our first experiment aims to confirm our hypothesis that feeding higher level aspects will increase both models' performance. The experiments test model performance under these settings: 1) Not using any aspect data. 2) Using only Level 1 (the most general) aspect. 3) Using both levels of aspects. We then manipulate the data by assigning random labels to further explore the usefulness of aspect labels. Our results show that BERT models performs much better when both levels of aspects are used. However, RoBERTa model un-intuitively does not generate a significantly better performance after using any aspect labels. This questions whether RoBERTa actually utilizes aspect labels as desired.

Then, we move on to test both models perfor-

[1] https://sites.google.com/view/fiqa#h.p_IFVdOB6rBYc9

mance on complicated sentences where a sentence may have different aspects. However, this dataset contains only one aspect for each data instance. Therefore, we propose a new testing method "Concatenate-Sentence" testing. We naively concatenate two sentences with different aspects to form one long sentence, then this sentence will have two aspects and two sentiment scores. This can naively model the scenario when a sentence has two different aspects and potentially different sentiment towards each of the aspects. Results how that both models only have slight increase from baseline models. Detailed error analysis shows that they all have difficulties capturing different sentiments towards different aspects on one sentence. This is in fact intuitive because our training data does not contain such instances. We conclude by hypothesizing that both models, especially RoEBRTa, utilize too many linguistic factors of original sentence and tend to ignore aspect labels, and therefore a better training method or a larger dataset need to be devised for this task.

The codes are in this public repository: https://github.com/OscarWan/Yuhao-and-Oscar

2 Related Work

Sentiment analysis is widely discussed in machine learning area. Traditional machine learning approaches can get a fair result. Kumar et al. (2017) use SVC with Logistic Regression and SVR with lexicon features and then combine them together. Ghosal et al. (2017) instead use deep learning algorithm combining CNN and LSTM network with POS tagging that gets some improvement . There are other methods also using deep learning approaches with similar final result such as RNN (Cabanski et al., 2017), Bi-GRU (Kar et al., 2017), CNN (Pivovarova et al., 2017), etc. From these results, we can see that deep learning approaches generally perform better over machine learning approaches.

BERT model can also be applied to sentiment analysis. Devlin et al. (2018) suggest that for sentiment classification task, we can simply leave the second sentence (textB) blank , which is also proposed by other sentiment classification task (Karimi and Shahrabadi; Hoang et al., 2019). These studies also justify that by adding aspect feature while fine-tuning BERT models, there are improvements with the final performance.

This can be applied to financial microblogs and headlines dataset. Shahid et al. (2020) use level 1&2 with 27 aspects, such as "Corporate", "Stock" for level 1 aspects, and "Corporate/Financial", "Stock/Options" and "Economy/Trade" for level 2 aspects. They merge the trained aspect embedding and word embedding to apply to CNN and then LSTM. The results show that with aspect, the model outperforms previous Bi-LSTM models (Shahid et al., 2020).

Based on that, Gaillat et al. (2018) include behaviour aspects such as buyer intention, financial results that describe more contents of the microblog. They also divide the aspects as explicit and implicit. Their results show that the number and accuracy of aspect classes will cause different performance, as explicit 32-class aspects increase about 15% accuracy compared with 7-class aspects, while implicit 32-class aspects decrease about 50% accuracy .

Similar studies includes Salunkhe and Mhaske (2019), who use BERT to extract aspects and then apply aspects to training process to make improvement. They use pre-trained BERT to get 4 parent aspect classes, and then the similar way to get 27 aspect subclasses. Finally they use Logistic Regression with aspects to get sentiment score. Our studies differ from them as they first extract the aspect labels and then predict the sentiment to improve the performance in sentiment analysis, we directly use aspect labels as a feature and our goal is to study the effect of using aspect features. We also include RoBERTa model rather than only using original BERT model.

3 Dataset

The dataset we used is a preprocessed dataset contains 1111 training examples, including 675 text instances from microblog posts and 435 text instances from news headlines. Each text instance has a related sentiment score from -1 to 1, and one aspect data like "Market/trade" for one sentence. An example data is shown in *Table 1*.

We did a study about the sentiment score distribution in this dataset, to get a good acknowledgement for our model analysis. The distribution of the dataset is focused on positive sentiment score, with 65% of data instances scoring larger than 0.

To simplify our experiment, we transform the sentiment score to a two-class classification, by only having "Positive" with score in (0,1) and "Negative" labels with score in (-1,0]. There are in total 4 Level-1 aspect labels and 27 Level-2 aspect la-

Text	Aspect	Sentiment Score
Royal Dutch Shell profit rises; dividend up 4%	Corporate/ Dividend Policy	0.65

Table 1: Example data.

bels. The full information about using aspect is included in Appendix.

All the experiment is done with 80% of the data as training data and 20% of the data as the validations set.

4 Experimental Evaluation

4.1 Experiment Setup

The baseline result of our experiment will be each model's performance when trained without feeding any aspect labels.

Our first experiment with original labels involves using different levels of aspect data, including 1) Using no aspect data (the baseline result). 2) Using only Level-1 (the most general) aspect. 3) Using both Level-1 and 2 aspects. We then experiment with randomized labels. The hypothesis is that if the model preform worse after we randomize the aspect labels, it shows that the model is utilizing the aspect labels.

Then, we use a new testing method "Concatenated-Sentence" Testing. We manually concatenate two sentences of different aspects. Therefore, each concatenated sentence will appear twice in the validation set with two different aspects and their respective sentiment. We use this testing method to naively test model performance on complicated sentences.

4.1.1 Experiment with Original Labels

Each experiment is done three times with shuffled data. We also report the standard deviation (Std) of the accuracy between those trials. A lower standard deviation indicates a more consistent result among these trials. The performance is measured using accuracy and f1 score. The result of both models is shown on *Table 2*. We also take into account that a small increase in both scores are probably insignificant and may result from random variations.

It's clear that RoBERTa performs better in both accuracy and f1-score. This is intuively easy to understand as RoBERTa is a more optimized model.

BERT			
Aspects Used	**Accuracy**	**F1**	**Std**
No Aspect	0.723	0.795	0.011
Level 1	0.726	0.798	0.049
Level 1 & 2	**0.767**	**0.820**	0.004
RoBERTa			
Aspects Used	**Accuracy**	**F1**	**Std**
No Aspect	0.855	0.889	0.007
Level 1	0.864	0.896	0.007
Level 1 & 2	0.853	0.882	0.015

Table 2: The average accuracy and F1 score for fine-tuning **BERT** and **RoBERTa** model on the aspect-based sentiment task.

Looking closely to each result separately, we can see that using more detailed aspect data generally gain a better performance compared to not using any on a BERT model. Using both Level 1 & 2 aspects outperforms other methods, which confirms that using aspect features is helping while using BERT models.

However, the RoBERTa model shows an unintuitive result. With Level 1 aspect subtly affecting the model, using both Level 1 and 2 aspects harms the performance. This goes against our hypothesis, showing that aspect features are not seemingly utilized by RoBERTa.

4.1.2 Experiment with Randomized Labels

We then move on to confirm our conclusion from the previous section that BERT is utilizing aspects, but RoBERTa fails to utilize aspects. We randomly assign aspect labels to each instance. If the model performs worse with this randomized data, it may indicate that the model is capturing the aspect features in the previous section. The results for BERT and RoBERTa are presented in *Table 3*.

For the BERT model, results show that after randomly assigning the labels, the average accuracy drops to nearly without using any aspect data. There is also a rise in standard deviation which shows that the BERT model is performing less robustly as the labels are randomized. This confirms that BERT is utilizing aspect data during sentiment classification task.

From RoBERTa, we can see that the result still change rather subtlely after we randomly assign the labels, which can still be reasonably attribute to random variations. This shows that the RoBERTa model tends to ignore the aspect labels. We think that such result may occur because that RoBERTa

BERT			
Aspects Used	**Accuracy**	**F1**	**Std**
No Aspect	0.723	0.795	0.011
Level 1 & 2, original	0.767	0.820	0.004
Level 1 & 2, random	0.734	0.813	0.027
RoBERTa			
Aspects Used	**Accuracy**	**F1**	**Std**
No Aspect	0.855	0.889	0.007
Level 1 & 2, original	0.853	0.882	0.015
Level 1 & 2, random	0.857	0.890	0.012

Table 3: The validation results for **BERT** and **RoBERTa** model on randomized Level 1 & 2 Aspect Labels. It includes "No Aspect" results and the "Level 1 & 2, original" results trained with original untouched data for comparison.

BERT			
Aspects Used	**Accuracy**	**F1**	**Std**
No Aspect	0.658	0.722	0.020
Level 1 & 2	0.692	0.757	0.029
RoBERTa			
Aspects Used	**Accuracy**	**F1**	**Std**
No Aspect	0.704	0.768	0.027
Level 1 & 2	0.725	0. 788	0.007

Table 4: The "Concatenate-Sentence" testing result on **BERT** and **RoBERTa** models. It shows that both models only slightly captures the aspect models and gains some performance in this "Concatenate-Sentence" task.

is a more optimized model, thus it can capture more linguistic properties in the original sentence, and thus may ignore aspect labels. And also as our dataset does not contain sentences with two different aspects and two different labels, RoBERTa is not forced to learn aspect features.

4.1.3 "Concatenated-Sentence" Testing

Our final testing aims to provide a general idea on how the two models perform on complicated sentences with different aspects and potentially different sentiments towards each sentiment. As our dataset does not contain such instances, we arbitrarily concatenate sentence with different aspects. We understand that these concatenated sentences may not be making sense in real world, so this result only serves as an intuition. The results of both models are shown on *Table 4*.

For both models, the increase in model performance after feeding in aspect levels is not significant. For RoBERTa, the performance increases by feeding aspect is less than BERT, confirming that RoBERTa doesn't capture aspect information as desired effectively.

Our further error analysis shows that the majority of errors comes from sentences with different sentiment towards different aspects. BERT only successfully distinguishes 6% of these sentences, and RoBERTa only distinguishes 4%. This clearly shows that both models fail to utilize aspect labels in the way we desire.

This pushes us to think about the way we trained both models. The most probable reason is that these complicated scenarios are not included in the training set. Therefore, our models do not fully capture the property of those aspect labels, i.e. the models do not realize that each aspect is referring to part of the sentences. We could also think about whether there is a better way to train aspect features when we used BERT or RoBERTa rather than only putting aspects in seperately, in order to let the model understand that aspects are referring to different parts of the sentences.

5 Conclusion

From our experimental results, we find that though generally RoBERTa outperforms BERT, BERT can use aspect labels to increase its performance on this task, but RoBERTa seemingly ignores these aspect labels during training process. We think that this can be explained by that more optimized RoBERTa can learn the sentiment from the original sentence. Although RoBERTa gets a better validation accuracy, it does not capture the linguistic properties of aspect labels in the way we desire.

Compared to models in previous researches, this may explain that CNN and Bi-LSTM (Shahid et al., 2020) can capture aspect labels because these models are not so optimized, thus they tend to put weights on aspect features. Therefore, we need new techniques to train instances with aspect features. The most straightforward way is to have a dataset that includes sentences that contain different aspects and different sentiment towards each aspect. When the model find out the sentence itself is not enough for sentiment classification, it might be forced to utilize aspect labels.

Finally, we are also aware that we are doing a small-scale study. For simplicity, our sentiment classification is simplified as two-class classification, and a more broader study can be conducted on a larger dataset. These questions can be explored by future researches.

6 Collaboration Statement

Oscar Wan is responsible for data pre-processing and datasets generating. He also works on literature reviews and Github updating. Yuhao Ding focuses on model fine-tuning, testing and reporting of the final results. Both of them write the final analysis.

References

Sitaram Asur and Bernardo A Huberman. 2010. Predicting the future with social media. In *2010 IEEE/WIC/ACM international conference on web intelligence and intelligent agent technology*, volume 1, pages 492–499. IEEE.

Tobias Cabanski, Julia Romberg, and Stefan Conrad. 2017. Hhu at semeval-2017 task 5: Fine-grained sentiment analysis on financial data using machine learning methods. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 832–836.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Thomas Gaillat, Bernardo Stearns, Ross McDermott, Gopal Sridhar, Manel Zarrouk, and Brian Davis. 2018. Implicit and explicit aspect extraction in financial microblogs. In *Annual Meeting of the Association for Computational Linguistics)*. Association for Computational Linguistics.

Deepanway Ghosal, Shobhit Bhatnagar, Md Shad Akhtar, Asif Ekbal, and Pushpak Bhattacharyya. 2017. Iitp at semeval-2017 task 5: an ensemble of deep learning and feature based models for financial sentiment analysis. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 899–903.

Rohitha Goonatilake, Ajantha Herath, Suvineetha Herath, Susantha Herath, and Jayantha Herath. 2007. Intrusion detection using the chi-square goodness-of-fit test for information assurance, network, forensics and software security. *Journal of Computing Sciences in Colleges*, 23(1):255–263.

Mickel Hoang, Oskar Alija Bihorac, and Jacobo Rouces. 2019. Aspect-based sentiment analysis using bert. In *NEAL Proceedings of the 22nd Nordic Conference on Computional Linguistics (NoDaLiDa), September 30-October 2, Turku, Finland*, 167, pages 187–196. Linköping University Electronic Press.

Hitkul Jangid, Shivangi Singhal, Rajiv Ratn Shah, and Roger Zimmermann. 2018. Aspect-based financial sentiment analysis using deep learning. In *Companion Proceedings of the The Web Conference 2018*, pages 1961–1966.

Sudipta Kar, Suraj Maharjan, and Thamar Solorio. 2017. Ritual-uh at semeval-2017 task 5: Sentiment analysis on financial data using neural networks. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 877–882.

Soroush Karimi and Fatemeh Sadat Shahrabadi. Sentiment analysis using bert (pre-training language representations) and deep learning on persian texts.

Abhishek Kumar, Abhishek Sethi, Md Shad Akhtar, Asif Ekbal, Chris Biemann, and Pushpak Bhattacharyya. 2017. Iitpb at semeval-2017 task 5: Sentiment prediction in financial text. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 894–898.

Lidia Pivovarova, Llorenç Escoter, Arto Klami, and Roman Yangarber. 2017. Hcs at semeval-2017 task 5: Polarity detection in business news using convolutional neural networks. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 842–846.

Ashish Salunkhe and Shubham Mhaske. 2019. Aspect based sentiment analysis on financial data using transferred learning approach using pre-trained bert and regressor model.

Simra Shahid, Shivangi Singhal, Debanjan Mahata, Ponnurangam Kumaraguru, Rajiv Ratn Shah, et al. 2020. Aspect-based sentiment analysis of financial headlines and microblogs. In *Deep Learning-Based Approaches for Sentiment Analysis*, pages 111–137. Springer.

Tun Thura Thet, Jin-Cheon Na, and Christopher SG Khoo. 2010. Aspect-based sentiment analysis of movie reviews on discussion boards. *Journal of information science*, 36(6):823–848.

A Appendix: Full Aspects List

We provide the full list of aspects here, the left most "·" means aspects for level 1; "-" means aspects for level 2:

- Corporate
  - Appointment
  - Company Communication
  - Dividend Policy
  - Financial
  - Legal
  - M&A
  - Regulatory
  - Reputation
  - Risks
  - Rumors
  - Sales

    - Strategy

- Economy
    - Central Banks
    - Trade

- Market
    - Conditions
    - Currency
    - Market
    - Volatility

- Stock
    - Buyside
    - Coverage
    - Fundamentals
    - Insider Activity
    - IPO
    - Options
    - Price Action
    - Signal
    - Technical Analysis