Financial Aspect Based Sentiment Analysis

Student Name: Tan Ching Fhen
School of Student: SCSE

URECA Supervisor Name: Prof Yan Shi Xing
Collaborating Supervisor Name: Prof Sun Aixin
School of Supervisor: School of Computer Science and Engineering (SCSE)

Abstract - Aspect-Based Sentiment Analysis (ABSA) is a form of sentiment analysis that has greater granularity compared to traditional sentiment analysis. It advantageous in applications that are focused on sentiments of short phrases. This paper explores the potential of utilizing ABSA in the financial domain. Compared to traditional sentiment analysis, research in the application of ABSA in financial domain is considerably less extensive. In this work, we annotated a financial domain ABSA dataset called Fi_ATSA. Since the number of training samples is very limited, we explored the use of data augmentation to increase dataset size. Next, we optimized a simple ABSA pipeline by experimenting with various Transformer models and aspect embedding configurations. To demonstrate the advantage of this pipeline in multi-aspect and multi-sentiment scenarios, we also performed evaluations on the MAMS dataset. Our experiments show that data augmentation techniques can be applied to ABSA to improve model generalization and our ABSA pipeline is useful when target sentences have multiple aspects and multiple sentiments.

Keywords - Aspect-based Sentiment Analysis, Financial News, Data Augmentation, Transformers, BERT

1 INTRODUCTION

Aspect-based Sentiment Analysis (ABSA) is a fine-grained sentiment analysis task, where the objective is to predict the sentiment polarity towards an aspect in a sentence. ABSA has two main subtasks namely, Aspect Term Sentiment Analysis (ATSA) and Aspect Category Sentiment Analysis (ACSA). In this work, we will focus on ATSA.

The objective of ATSA is to predict the sentiment polarity of aspect terms, which can be words or phrases, explicitly mentioned in the sentence. ATSA can be broken down into two phases, which are, aspect extraction phase and sentiment classification phase.

In recent studies, many models have been proposed to improve performances in the ATSA task. Many of the proposed models utilize Transformers in the model pipeline [1,2,4,13,15,18] or domain adaptation techniques [1,2,5]. Furthermore, data augmentation is a well-established technique shown to enhance performance and generalization of natural language models [7]. Thus, we utilize these main techniques in our work.

The domain of interest in our work is the financial domain, which has not attracted as much interest as compared to the laptop review and restaurant review domain [10]. Because there are no current ATSA datasets for this domain, we manually annotated our own financial domain ATSA dataset, Fi_ATSA.

Overall, our contributions are as follows:

- Annotated a financial domain ATSA dataset
- Proposed and optimized a simple and practical ATSA pipeline
- Illustrate the effectiveness of domain adaptation and data augmentation in the ATSA pipeline

2 ATSA PIPELINE

In this section, we describe our ATSA pipeline.

2.1 PIPELINE

For our task, we restricted the aspects to company names and symbols from the S&P 500, such as Amazon and AMZN, due to reasons which we explain in Section 3.1. Since the aspects are known and unambiguous, we use utilized string matching in the aspect extraction phase to extract the character-level aspect spans. Next, during sentiment prediction, we acquired contextualized embeddings by passing the sentences through Transformer encoders. Lastly, we utilized a neural network to perform sentiment prediction on the aspect embeddings.

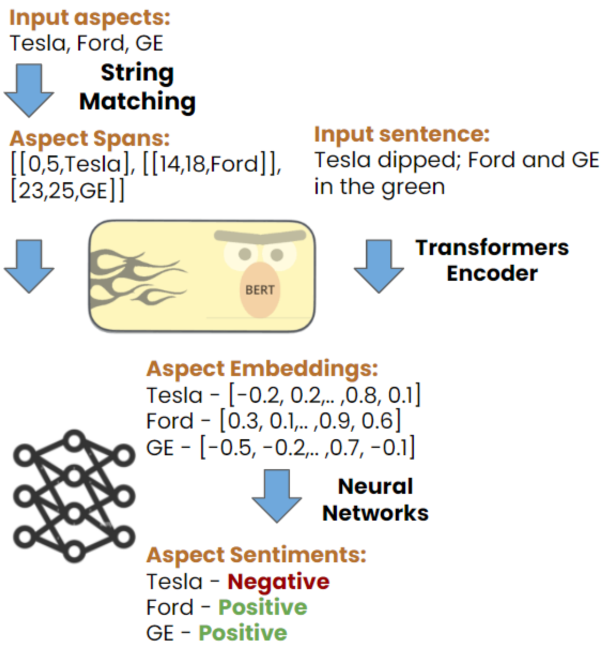

Figure 1 – Our ATSA Pipeline

3 DATASETS

We evaluated the effectiveness of our pipeline on three datasets: Fi_ATSA, MAMS and FiQA. This section gives an overview of each dataset.

3.1 FI_ATSA

To aid our task, we annotated, Fi_ATSA, a financial domain ATSA dataset of size 492. We restrict the definition of aspects to all company names or symbols from the S&P 500 such as Tesla, TSLA, Amazon and AMZN (Figure 2).

This design choice has several benefits. Firstly, it reduces the difficulty of annotation. Since the spans of known company names and symbols can be automatically labelled by string matching, only the aspect sentiments are left to annotate. Secondly, this design choice is more practical. The financial domain has many different industries and topics, each with different set of vocabulary. This makes annotation very expensive, which is impractical given our limited resources.

To acquire our raw data, we made API calls from the Financial News API. Next, we performed sentence segmentation using Spacy and extracted sentences that contain two or more aspects. Having samples with multiple aspects could alleviate the task degeneration problem observed in Restaurant dataset for ABSA [12].

Our annotation methodology was as follows: if an aspect was described with an adjective phrase in the sentence, we assign it a positive or negative sentiment depending on the sentiment of the adjective phrase. Otherwise, we will assign a neutral sentiment (Figure 2).

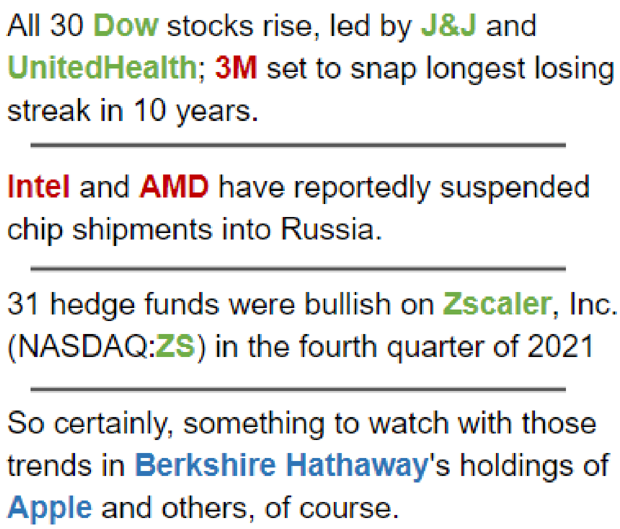

Figure 2 - The above illustrates Fi_ATSA samples. Aspect terms are bold and aspect sentiments are colored in green (positive), red (negative) and blue (neutral).

3.2 MAMS

Multi-aspect multi-sentiment analysis (MAMS) dataset is an ABSA dataset, whereby each sentence contains at least two distinct aspects with different sentiments [12]. The sentences originated from City search New York dataset [3] and annotated by three experienced researchers.

The motivation is that previous ABSA datasets such as SemEval-2014 Restaurant Review dataset, Laptop Review dataset [10] and Twitter dataset [9] contain mostly one aspect or multiple aspects with one sentiment polarity. This causes a degeneration of ABSA to sentence-level sentiment analysis [12].

3.3 FiQA

FiQA is a training set from WWW '18 [11] containing samples from financial news headlines and tweets. Each sentence is associated with the target aspect and sentiment score. The sentiment scores are continuous values ranging from -1 to 1 – most negative to most positive.

To align with our task, we converted sentiment scores into sentiment polarities in the following manner. Sentiment scores below -0.2 were assigned negative sentiment polarity. Sentiment scores above 0.2 were assigned positive sentiment polarity. Lastly, sentiment scores between -0.2 and 0.2 were assigned neutral sentiment polarity.

3.4 DATASET STATISTICS

Dataset	Training size	Test size	Avg. #Aspects	Avg. #Sentiment
Fi_ATSA	344	148	2.93	1.16
MAMS	4296	1000	2.60	2.02
FiQA	797	341	1.05	1.01

Figure 3 - Table showing the training sizes, test sizes, average number of aspects per sample and average number of sentiments per sample on Fi_ATSA, MAMS and FiQA

4 TRANSFORMER ENCODER

In this experiment, we compare seven different Transformer encoders over the three datasets.

We train the models on the training sets, evaluate on the validation sets and take the average performance over 3 seeds. Training to validation size ratio is 7:3.

4.1 MODELS

ProsusAI-FinBERT: BERT pretrained on TRC2-financial, a financial corpus consisting of 1.8M news articles that were published by Reuters between 2008 and 2010 [1].

Bert-base-uncased: Original Bidirectional Encoder Representations from Transformers (BERT) that was pre-trained using masked language modelling (MLM) and next sentence prediction (NSP) [6].

Distilbert-base-uncased: A distilled version of BERT that is smaller, faster, and lighter [14].

Distilroberta-base: A distilled version of RoBERTa [16].

Google-electra-base-discriminator: A BERT variant that was pretrained with using the *replaced token detection* task that is more sample efficient [8].

Google-mobilebert-uncased: A BERT variant that was trained using knowledge transfer techniques [17].

Roberta-base: A BERT variant that was pre-trained using the MLM task with dynamic masking, without NSP task, on larger mini-batches and longer time [16].

4.2 PLOTS

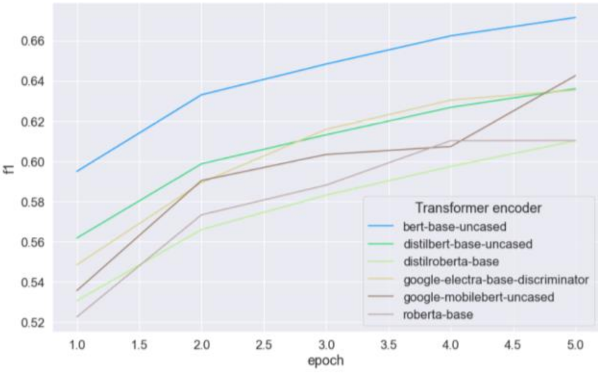

Figure 4 - Comparison of Transformer encoders. Validation F1 Score over training epochs on MAMS dataset

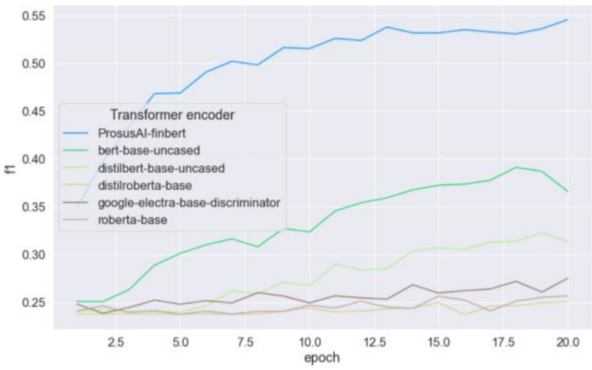

Figure 5 - Comparison of Transformer encoders, including a financial domain adapted Transformer encoder (FinBERT). Validation F1 Score over epoch on FiQA dataset

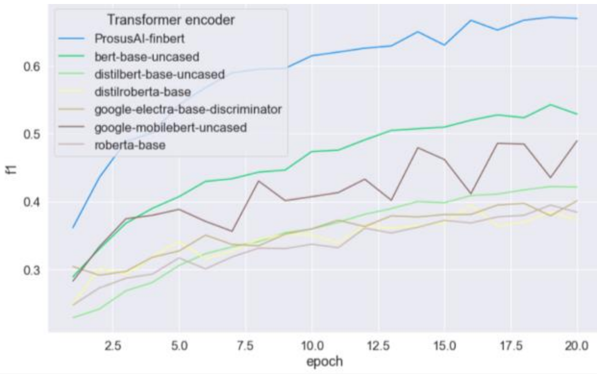

Figure 6 - Comparison of Transformer encoders, including a financial domain adapted Transformer encoder (FinBERT). Validation F1 Score over epoch on Fi_ATSA dataset

4.3 FINDINGS

Based on our experiments, FinBERT outperformed all other transformer models on Fi_ATSA and FiQA by significant margins (Figure 5 and 6). This shows that further pre-training on target domain corpus is a key contributor to performance in ABSA tasks.

Furthermore, the original BERT outperforms ELECTRA and RoBERTa, which is a surprising finding. Both ELECTRA and RoBERTa were found to perform better in other tasks like question answering and sentence level classification [8,16]. This suggests that the word embeddings of BERT contain better sentiment information. However, this is beyond the scope of our study.

5 ASPECT EMBEDDINGS

In this section, we compare the compare different aspect embedding configurations i.e., "feature type", to use as input for the neural sentiment classifier.

The purpose of this experiment was to answer these research questions:

1. *Which embeddings is best for input to the neural classifier?*
2. *Does concatenating the aspect word to the sentence improve performance?*
3. *Can CLS embeddings be used as input to the neural classifier for ABSA?*

We train the models on the training sets, evaluate on the validation sets and take the average performance over 3 seeds. Training to validation size ratio is 7:3.

5.1 FEATURE TYPES

First: The first aspect embedding was used as input for the sentiment classifier.

Last: The last aspect embedding was used as input for the sentiment classifier.

Mean: The mean of the aspect embeddings was used as input for the sentiment classifier.

Mean_two_sentence: For input to the Transformer encoders, each sentence was concatenated with the aspect word, separated by an SEP token. Then, the mean of the aspect embeddings was used as for the neural sentiment classifier.

Cls_two_sentence: As input for the Transformer encoders, each sentence was concatenated with the aspect word, separated by an SEP token. The CLS embedding was used as input to the sentiment classifier.

Cls_last: The CLS embedding, and the last aspect embedding were concatenated as input for the sentiment classifier.

5.2 PLOTS

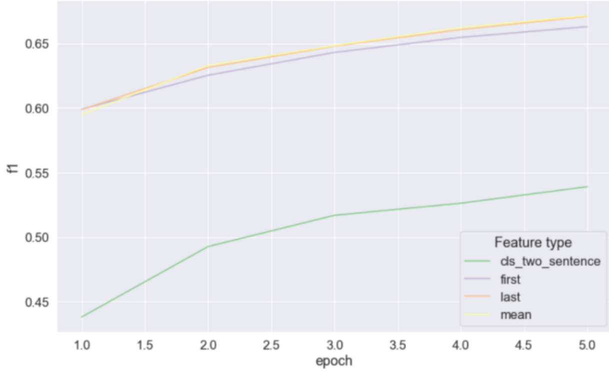

Figure 7 - Comparison of embedding inputs for aspect sentiment prediction. Validation F1 Score over epoch on MAMS dataset.

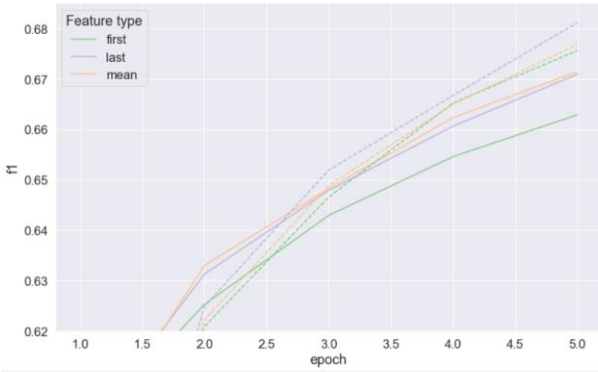

Figure 8 - Comparison of embedding inputs for aspect sentiment prediction. Training (dotted) and validation (bold) F1 Scores over epoch on MAMS dataset.

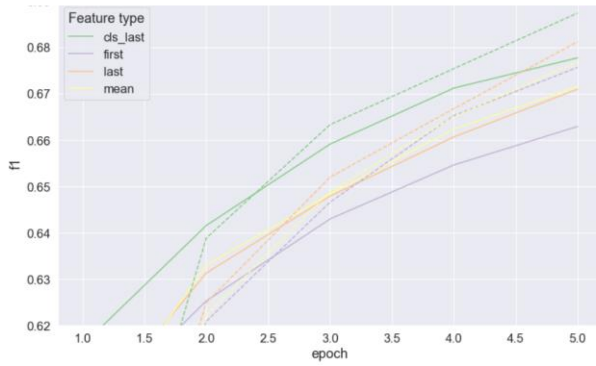

Figure 9 - Comparison of embedding inputs for aspect sentiment prediction. Training (dotted) and validation (bold) F1 Scores over epoch on MAMS dataset.

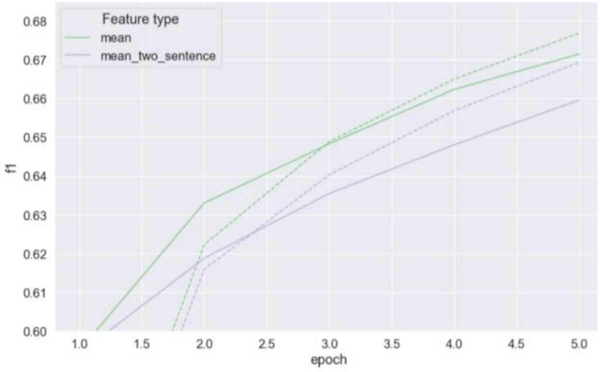

Figure 10 - Comparison of embedding inputs for aspect sentiment prediction. Training (dotted) and validation (bold) F1 Scores over epoch on MAMS dataset.

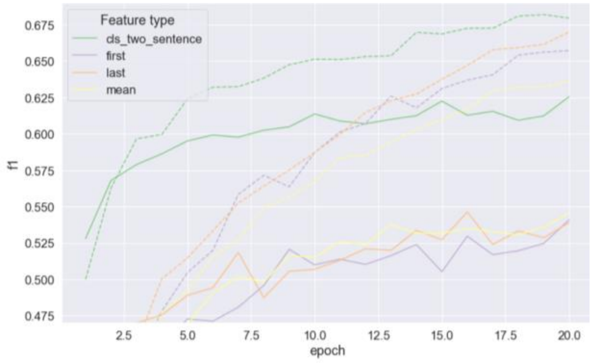

Figure 11 - Comparison of aspect embedding inputs for aspect sentiment prediction. Training (dotted) and validation (bold) F1 Scores over epochs on FiQA dataset.

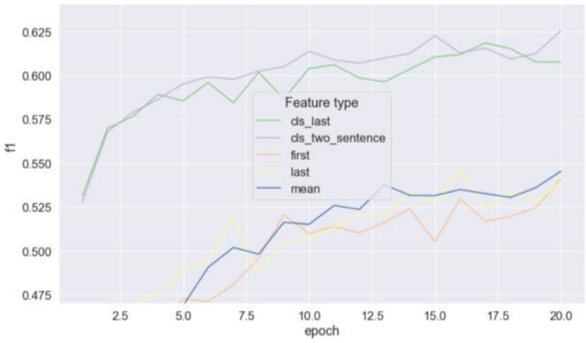

Figure 12 - Comparison of aspect embedding inputs for aspect sentiment prediction. Validation F1 Scores over epochs on FiQA dataset.

Figure 13 - Comparison of aspect embedding inputs for aspect sentiment prediction. Validation F1 Scores over epochs on Fi_ATSA dataset.

5.3 FINDINGS

Using the mean or the last embedding of the aspect words perform equally well (Figures 7, 8, 9, 11, 12, 13).

Concatenating the aspect word to the sentence as input to the Transformer encoders does not improve performance (Figure 10).

For the sentences that have multiple sentiments, using only the CLS embedding performs much worse than if aspect embeddings were used (Figure 7). However, when sentences generally have single sentiments, using CLS embedding is performs better than using aspect embeddings (Figures 12 and 13).

On another note, we found that concatenating both the CLS embedding and the aspect embedding performed the best (Figure 9, 12 and 13).

6 DATA AUGMENTATION

In this section, we explain how we used data augmentation to improve the performance of our ATSA model on the Fi_ATSA dataset.

6.1 EDA

Due to a relatively small dataset, we perform data augmentation to increase dataset size, which can improve model generalization. Specifically, we utilized Easy Data Augmentation (EDA) [7]. EDA has four operations: synonym replacement, random insertion, random swap, and random deletion (Figure 14).

However, the annotated labels will be lost when performing vanilla EDA. Thus, we adapt the EDA algorithm for ATSA in the following manner. During data augmentation, the aspects words were made invulnerable to EDA operations and the order of aspects in the sentence remain unchanged. In this way, augmented samples can retain their annotated labels, even after data augmentation.

A possible downside of data augmentation is that the aspect sentiments might change due to the EDA operations. Nonetheless, the authors who developed EDA show that “for the most part, sentences augmented with EDA conserved the labels of their original sentences” [7]. Thus, adapting to EDA for ATSA will be immensely beneficial for our case due to the lack of resources.

2.2 EDA SAMPLES

Original:
Canada Pension Plan Investiture Board quadrupled its stake in electric-vehicle giant Tesla, slashed stakes in GM and Nvidia, and trimmed its GE stake.

Synonym Replacement:
Canada Pension Plan **Investment control panel** Board quadrupled its stake in electric-vehicle giant Tesla, slashed stakes in GM and Nvidia, and trimmed its GE stake.

Random Swap:
Canada **electric-vehicle** Plan Investment Board quadrupled its stake in **Pension** giant Tesla, slashed stakes in GM and Nvidia, and trimmed its GE stake.

Random Deletion:
Canada Pension Plan **<deleted “Investiture”>** Board quadrupled its stake in electric-vehicle giant Tesla, slashed stakes in GM and Nvidia, and trimmed GE stake

Random Insertion:
Canada Pension Plan Investment **assistance** Board quadrupled its stake in electric-vehicle giant Tesla, slashed stakes in GM and Nvidia, and trimmed its GE stake.

Figure 14 - An illustration of the four operations in Easy Data Augmentation. Changes are highlighted in orange.

6.2 EDA EXPERIMENTS

We conducted experiments to determine if EDA could improve the performance and generalization of our model.

To do so, we trained one ATSA model on the original Fi_ATSA training set over multiple epochs for a total of roughly 16000 samples. We trained another ATSA model on an augmented Fi_ATSA training set for the same number of samples. For the augmentation parameters, we used the configurations suggested by the authors - 16 augmented samples per original sample at an alpha value of 0.05 [7].

Lastly, we evaluated the performance of each model on the original unaltered Fi_ATSA test set.

Based on our experiments, applying EDA does improve the performance after training on enough samples (Figure 15).

Figure 15 - Comparing the use of EDA against without EDA. Test F1 Score against number of training samples (roughly 16000 training samples).

6 HYPERPARAMETER OPTIMIZATION

Finally, we optimized the hyperparameters of our neural sentiment classifier on the Fi_ATSA dataset.

We used the “cls_last” as input to the neural network because it was found to perform best for this task. We trained the neural network on the Fi_ATSA training set with EDA and evaluated on the Fi_ATSA test set.

6.1 PLOTS

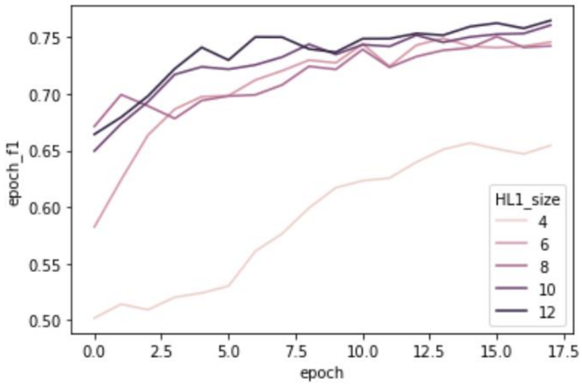

Figure 4 - Comparison of different hidden layer sizes. Test F1 Score over epoch on Fi_ATSA dataset with EDA.

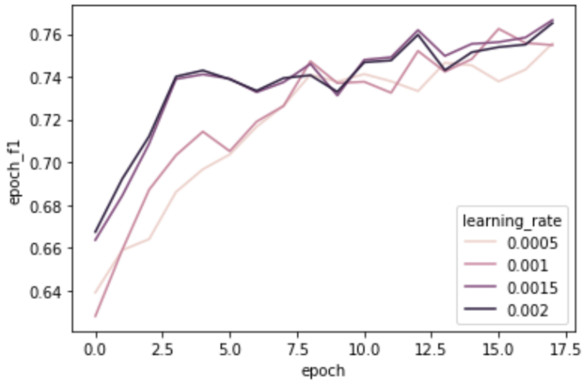

Figure 5 - Comparison of different learning rate for sentiment classifier. Test F1 Score over epoch on Fi_ATSA dataset with EDA.

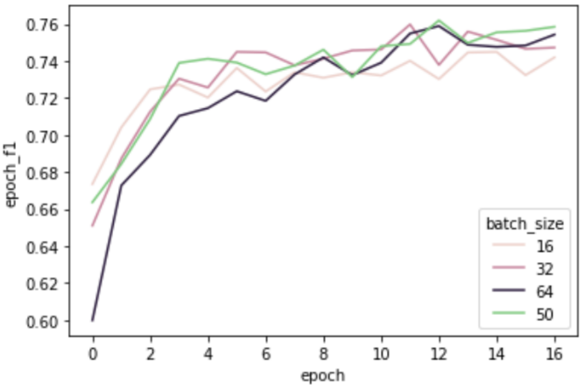

Figure 6 - Comparison of different batch sizes. Test F1 Score over epoch on Fi_ATSA dataset with EDA.

7 DISCUSSION

Our ATSA pipeline (Figure 1) has various benefits. First, sentiment prediction is only performed on the aspect embeddings, this leads to a lower computational requirement during training and inference time. In contrast, other ATSA pipelines jointly train aspect extraction and sentiment prediction [4,13,15,18], which increases task complexities and computational requirements. Second, our pipeline allows users to specify the desired aspects to perform sentiment prediction during inference time, this allows greater flexibility and control for the user.

However, there are various pitfalls. First, our pipeline is only suited for very specific applications where the aspect terms under interest are known and unambiguous, thus can be reliably extracted by string matching. Also, we find that in practice, particularly in the financial domain, sentences that contain multiple aspects rare and sentences that contain multiple sentiments are even more so. In such scenarios, ATSA models are unsuitable and sentence-level sentiment classification models would suffice.

7 CONCLUSION

In this paper, we proposed an aspect-term sentiment analysis (ATSA) pipeline and optimized its module components and configurations. We annotated a financial domain ATSA dataset, Fi_ATSA. Lastly, we show that domain adaptation techniques and data augmentation techniques are both beneficial to ATSA. The codes are available at chingfhen/Financial-Aspect-Based-Sentiment-Analysis (github.com).

ACKNOWLEDGMENT

I would like to acknowledge the funding support from Nanyang Technological University – URECA Undergraduate Research Programme for this research project. Next, I would also like to express my appreciation Prof Sun Aixin for his valuable advice and guidance.

REFERENCES

[1] Dogu Tan Araci. FinBERT: Financial Sentiment Analysis with Pre-trained Language Models. 2019. arXiv:1908.10063v1.

[2] Entony Lekhtman, Yftah Ziser, Roi Reichart. DILBERT: Customized Pre-Training for Domain Adaptation withCategory Shift, with an Application to Aspect Extraction. 2021. arXiv:2109.00571.

[3] Gayatree Ganu, Noemie Elhadad, and Amelie Marian. 2009. Beyond the stars: Improving rating predictions using review text content. In Proceedings of WebDB, Providence, Rhode Island, USA.

[4] Guimin Chen, Yuanhe Tian, and Yan Song. 2020. Joint Aspect Extraction and Sentiment Analysis with Directional Graph Convolutional Networks. In Proceedings of the 28th International Conference on Computational Linguistics, pages 272–279, Barcelona, Spain (Online). International Committee on Computational Linguistics.

[5] Hao Tian, Can Gao, Xinyan Xiao, Hao Liu, Bolei He, Hua Wu, Haifeng Wang, Feng Wu. SKEP: Sentiment Knowledge Enhanced Pre-training for Sentiment Analysis. 2020. arXiv:2005.05635.

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

[7] Jason Wei, Kai Zou. EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks. 2019. arXiv:1901.11196v2.

[8] Kevin Clark, Minh-Thang Luong, Quoc V. Le, Christopher D. Manning. ELECTRA: Pre-training Text Encoders as Discriminators Rather Than Generators. 2020. arXiv:2003.10555.

[9] Li Dong, Furu Wei, Chuanqi Tan, Duyu Tang, Ming Zhou, and Ke Xu. 2014. Adaptive recursive neural network for target-dependent twitter sentiment classification. In Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 2: Short papers), volume 2, pages 49–54.

[10] Maria Pontiki, Dimitris Galanis, John Pavlopoulos, Harris Papageorgiou, Ion Androutsopoulos, and Suresh Manandhar. 2014. Semeval-2014 task 4: Aspect based sentiment analysis. In Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014), pages 27–35.

[11] Macedo Maia, Siegfried Handschuh, Andre Freitas, ́ Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In Companion of the The Web Conference 2018 on The Web Conference 2018. International World Wide Web Conferences Steering Committee, pages 1941–1942.

[12] Qingnan Jiang, Lei Chen, Ruifeng Xu, Xiang Ao, and Min Yang. 2019. A Challenge Dataset and Effective Models for Aspect-Based Sentiment Analysis. In Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6280–6285, Hong Kong, China. Association for Computational Linguistics.

[13] Xin Li, Lidong Bing, Wenxuan Zhang, Wai Lam. Exploiting BERT for End-to-End Aspect-based Sentiment Analysis. 2019. arXiv:1910.00883.

[14] Victor Sanh, Lysandre Debut, Julien Chaumond, Thomas Wolf. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. 2020. arXiv:1910.01108.

[15] Yue Mao, Yi Shen, Chao Yu, Longjun Cai. A Joint Training Dual-MRC Framework for Aspect Based Sentiment Analysis. 2021. arXiv:2101.00816.

[16] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. 2019. arXiv:1907.11692.

[17] Zhiqing Sun, Hongkun Yu, Xiaodan Song, Renjie Liu, Yiming Yang, Denny Zhou. MobileBERT: a Compact Task-Agnostic BERT for Resource-Limited Devices. 2020. arXiv:2004.02984.

[18] Zhen Wu, Chengcan Ying, Xinyu Dai, Shujian Huang, Jiajun Chen. Transformer-based Multi-Aspect Modeling for Multi-Aspect Multi-Sentiment Analysis. 2020. arXiv:2011.00476.

Websites

[19] J. Wei, "EDA: Easy Data Augmentation Techniques for Boosting Performance on Text Classification Tasks", June 20; http://www.github.com/jasonwei20/eda_nlp.