

Machine Learning-Based Sentiment Analysis of Financial Texts

Chuan Jia(No. 17), Bo Li(No. 9), Sheng Su(No. 10)*

Abstract

With the widespread adoption of the Internet, cloud computing, and big data, significant transformations have occurred in financial information processing. A vast amount of financial texts, such as news, reports, and social media content, have become crucial resources for formulating investment strategies, asset evaluation, and market forecasting. These texts not only contain numerical data and facts but also embody emotional signals that influence investment decisions. Integrating emotional signals into analytical frameworks helps enhance the rationality and predictability of investment strategies. However, the scarcity of Chinese financial sentiment corpora limits the application of deep learning and pre-trained models (such as BERT) in Chinese sentiment analysis. To address this, we propose a data-driven approach that first utilizes a Transformer-based neural machine translation model to convert English financial sentiment corpora into Chinese, then employs a Chinese sentiment lexicon for annotation, and finally fine-tunes the BERT model to improve Chinese sentiment classification performance. Nevertheless, the emergence of "meaningless Chinese words" during the translation process indicates deficiencies in the model's mapping of specialized terminology. We also explore the possibility of replacing or supplementing BERT with Bayesian or traditional machine learning methods, which sometimes offer advantages in terms of accuracy and interpretability.

Keywords: financial text information, sentiment analysis, Transformer model, BERT model, data augmentation, machine translation, Bayesian interpretability

1 Introduction

Financial activities lie at the heart of modern economic systems, shaping everything from micro-level investment decisions to macro-level policies. Driven by developments in the Internet and data analytics, enormous quantities of financial text—including news, reports, and social media commentary—are now readily available. While traditional analyses often rely on explicit indicators like pricing data or fundamental ratios, they may omit subtle emotional signals embedded within these texts.

*The authors are listed in surname order. All authors contributed equally to this work.

Latent sentiments, such as fear, optimism, or skepticism, can influence market trends by affecting trading decisions and shaping public opinion. For instance, a speculative rumor shared on social media regarding a potential product failure might trigger massive sell-offs, contributing to short-term volatility. More comprehensive decision-making frameworks must, therefore, incorporate these emotional cues.

Despite the abundance of English-language financial sentiment corpora, labeled Chinese financial datasets remain scarce. This gap has motivated us to translate high-quality English content into Chinese, thus enlarging our corpus and addressing data insufficiency for a specialized domain. However, translation alone is not a complete solution—certain terms do not map neatly between languages, especially when dealing with finance-specific jargon. Additionally, advanced solutions (such as graph-based methods for unified semantic representation) may be needed to further mitigate mismatches introduced by translation.

This study offers three primary contributions:

1. **Data Augmentation via Translation:** We employ a Transformer-based NMT model to convert English financial sentiment texts into Chinese, expanding the availability of labeled data for model training.
2. **Lexicon-Based Annotation:** By leveraging a domain-specific Chinese financial sentiment lexicon, we automatically assign polarity labels to large volumes of text, enriching the dataset further.
3. **BERT Fine-Tuning and Alternatives:** We fine-tune a pre-trained BERT model for Chinese financial sentiment classification. Additionally, we consider how Bayesian or traditional machine learning methods may sometimes outperform or offer clearer interpretability compared to deep neural networks in certain contexts.

The rest of this paper is organized as follows: Section 2 reviews related work in financial sentiment analysis and Transformer models. Section 3 outlines our data collection, pre-processing, and model training procedures. Section 4 presents our experimental results, and Section 5 discusses the implications of our findings. Finally, Section 6 concludes the paper and outlines future research directions.

2 Related Work

2.1 Financial Sentiment Analysis

Financial sentiment analysis aims to quantify subjective feelings or attitudes from unstructured text. Studies have shown that managerial tone, media coverage, and user-generated

content can all be predictive of stock returns and market volatility [3, 4]. Despite substantial progress in English financial text analysis, limited research has focused on Chinese corpora, due primarily to the shortage of annotated data and the linguistic complexity of Chinese.

2.2 Transformer Models in NLP

The Transformer architecture proposed by Vaswani et al. [1] uses multi-head self-attention to capture global dependencies in text. This architecture has proven adept at machine translation, sentiment analysis, and question answering. In recent years, various Transformer-based models (e.g., BERT, GPT variants) have achieved state-of-the-art results across numerous NLP benchmarks. Figure 1 illustrates the core components of Transformer models.

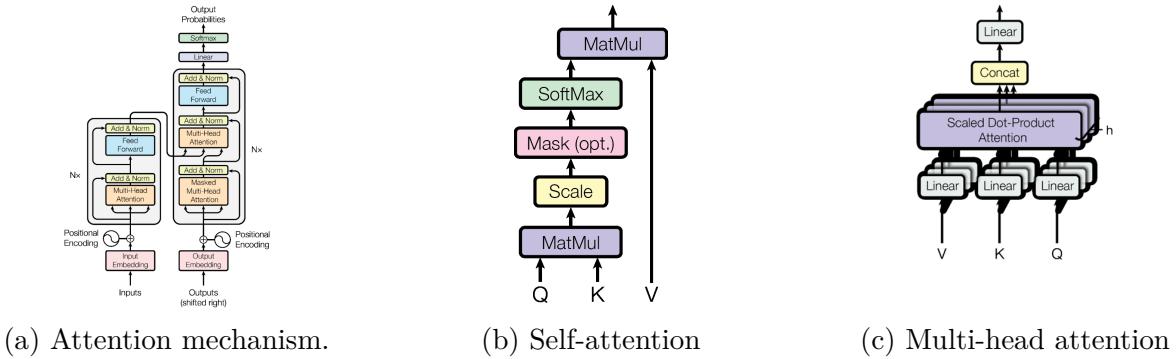


Figure 1: Illustrations of attention mechanisms in Transformer models.

2.3 BERT and Domain Adaptation

BERT [2] uses a bidirectional encoder to learn contextual representations through masked language modeling and next sentence prediction. Fine-tuning BERT on domain-specific corpora often yields significant performance improvements. In financial domains, specialized BERT variants or further pre-training on finance-related text can better capture terminology and context.

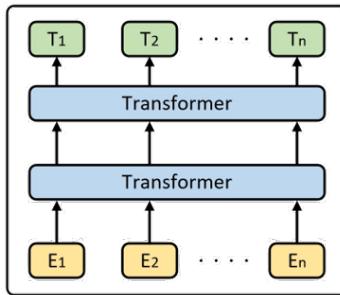


Figure 2: BERT architecture.

3 Methodology

3.1 Data Collection and Preprocessing

We gathered 54,749 Chinese financial news articles from various reputable sources, including 40,000 Chinese financial news articles and 14,749 translated English financial texts. The latter were translated using a Transformer-based NMT model. We also utilized a Chinese financial sentiment lexicon to pre-label a subset of unannotated texts, enriching the training set. Each news is labeled with a sentiment category: **positive**, **neutral**, or **negative**.

All data were tokenized using a Chinese tokenizer (*jieba*). For the unannotated subset, we employed a Chinese financial sentiment lexicon to pre-label texts. This procedure facilitated the creation of an enriched training set that better captures domain-specific language features.

Data Statistics and Visualization We conducted exploratory data analysis to understand the distribution of sentiment categories, text lengths, and word frequencies. Figures 3–5 illustrate the word clouds, top words, text length distributions, and sentiment category distributions in our dataset.

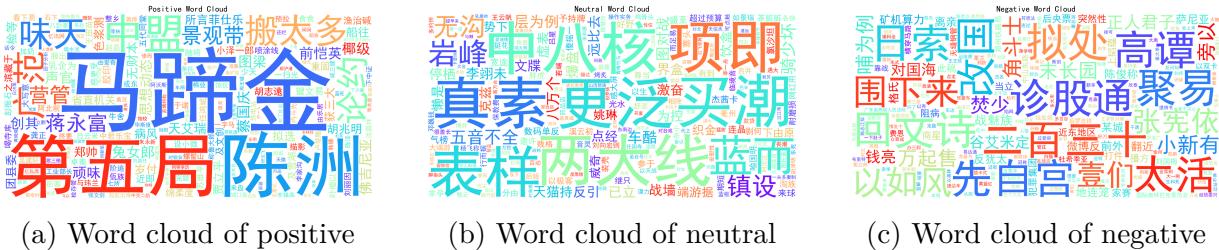
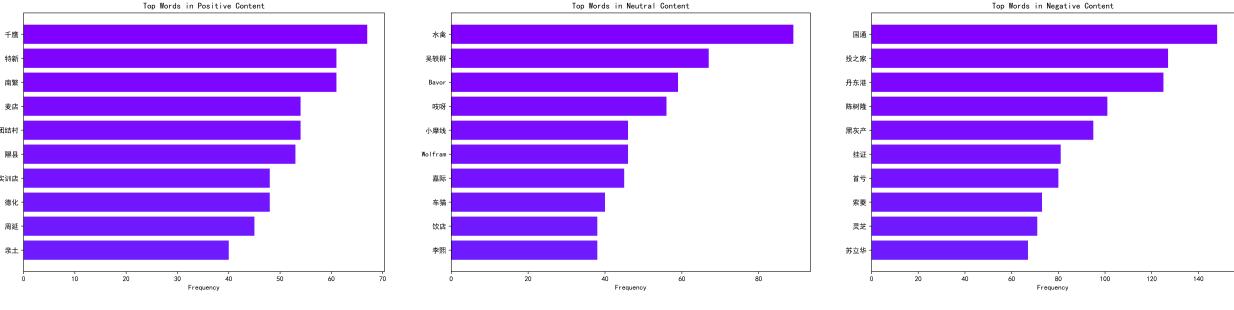


Figure 3: Word clouds for each sentiment category, illustrating the most frequent words associated with positive, neutral, and negative sentiments. These visualizations highlight the key terms that differentiate the sentiment classes in the dataset.

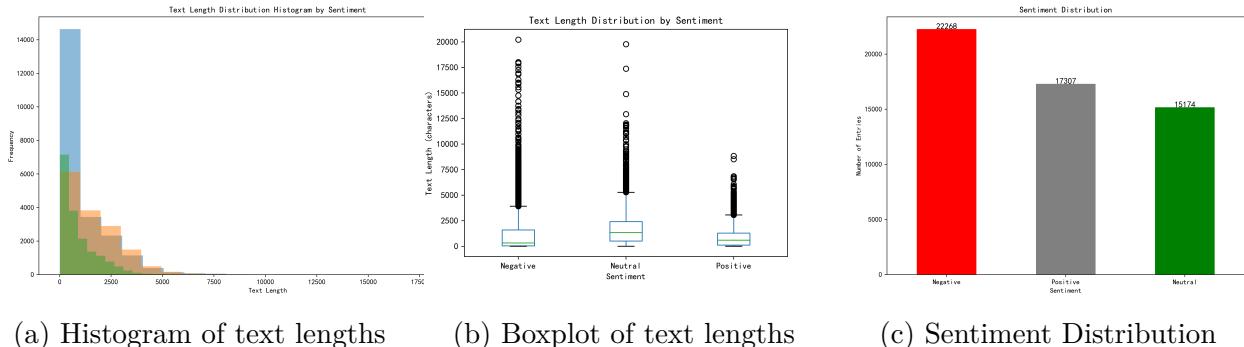


(a) Top positive words

(b) Top neutral words

(c) Top negative words

Figure 4: Most frequent words in each sentiment category, showcasing the top terms that contribute to positive, neutral, and negative sentiments. These charts provide insight into the language patterns prevalent in different sentiment classes.



(a) Histogram of text lengths

(b) Boxplot of text lengths

(c) Sentiment Distribution

Figure 5: Statistical analysis of the dataset: (a) Histogram displaying the distribution of text lengths, (b) Boxplot illustrating the variability and outliers in text lengths, and (c) Distribution of sentiment categories showing the proportion of positive, neutral, and negative texts. These analyses provide a foundational understanding of the dataset’s characteristics.

Meaningless Chinese Words in Word Clouds During word-cloud visualization, several tokens appeared meaningless or out of context. Closer investigation revealed that these originated from awkward translations of specialized English finance terms. This underscores the limitations of our translation approach, wherein certain English words map improperly into Chinese hyperspace, yielding unnatural tokens that nonetheless appear in frequency-based visualizations.

3.2 Lexicon-Based Annotation

Lexicon-based annotation matches words or phrases from a predefined dictionary to text segments and assigns associated sentiment scores. Jiang et al. [3, 4] proposed a Chinese financial sentiment lexicon that In the Chinese financial domain, terminologies and colloqui-

alisms evolve rapidly, complicating this step. Yet, even partial coverage by a domain-specific lexicon can significantly accelerate the annotation process.

3.3 Model Training

3.3.1 TNMT for Corpus Expansion

We trained a TNMT model to translate English financial texts into Chinese. Evaluated using BLEU scores and F1 metrics, the model effectively enlarged our Chinese dataset. However, some translated words remain semantically mismatched. Future enhancements could involve domain-specific dictionaries or graph-based mapping approaches to reduce translation noise.

3.3.2 BERT Fine-Tuning

Our main classification model builds upon a pre-trained Chinese BERT. We fine-tuned it by feeding the annotated data as input, monitoring metrics like accuracy, F1 score, and training loss. While BERT generally outperforms older generation RNN-based methods, traditional or Bayesian models may, in certain scenarios with less data or higher interpretability demands, provide more transparent and sometimes even superior performance.

The whole algorithm pipeline is illustrated in Figure 6.

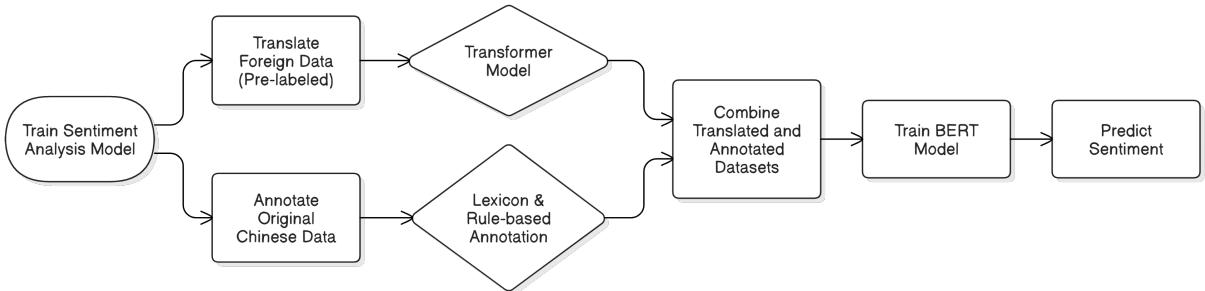


Figure 6: Algorithm pipeline for Chinese financial sentiment analysis. The workflow encompasses data collection, translation, lexicon-based annotation, and model training.

4 Experiments and Results

4.1 Transformer-Based Translation

Figures 7 shows the training loss, evaluation BLEU score, F1 score, accuracy, precision, and recall of the TNMT model over 350 steps. The model exhibits a steady decline in training loss and a corresponding increase in evaluation BLEU, F1, and accuracy. Precision and recall

curves also show consistent improvement, indicating the model’s ability to capture nuanced sentiment cues in the translated texts. However, the presence of ”meaningless” tokens in word clouds suggests that the model may struggle with specialized financial terminology, necessitating further refinement. Furthermore, although the performance metrics show improvement, the model’s overall accuracy and F1 score remain suboptimal, indicating the need for additional enhancements.

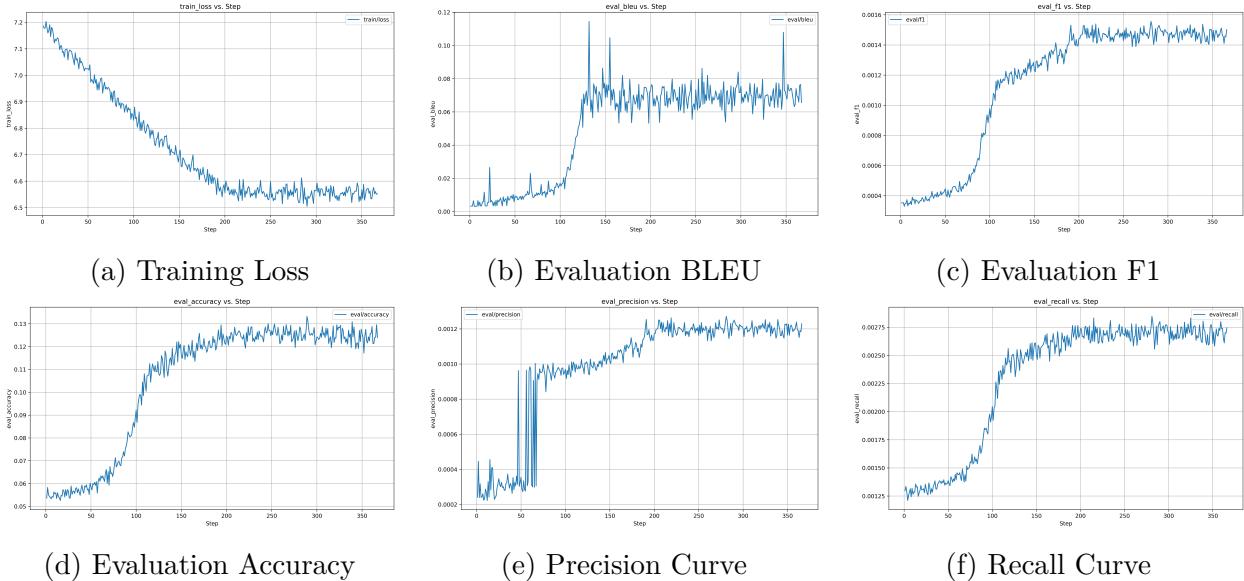


Figure 7: Performance metrics of the TNMT model.

Next, we evaluated a fine-tuned Chinese BERT model. Figures 8 show the training accuracy, F1 score, loss, learning rate, precision, and recall over epochs. The model exhibits a steady increase in accuracy and F1 score, with a corresponding decrease in loss. Precision and recall curves also show consistent improvement, indicating the model’s ability to capture nuanced sentiment cues in the translated texts. However, overall this model perform bad in the financial sentiment analysis task. This may due to the translation quality and the lexicon-based annotation.

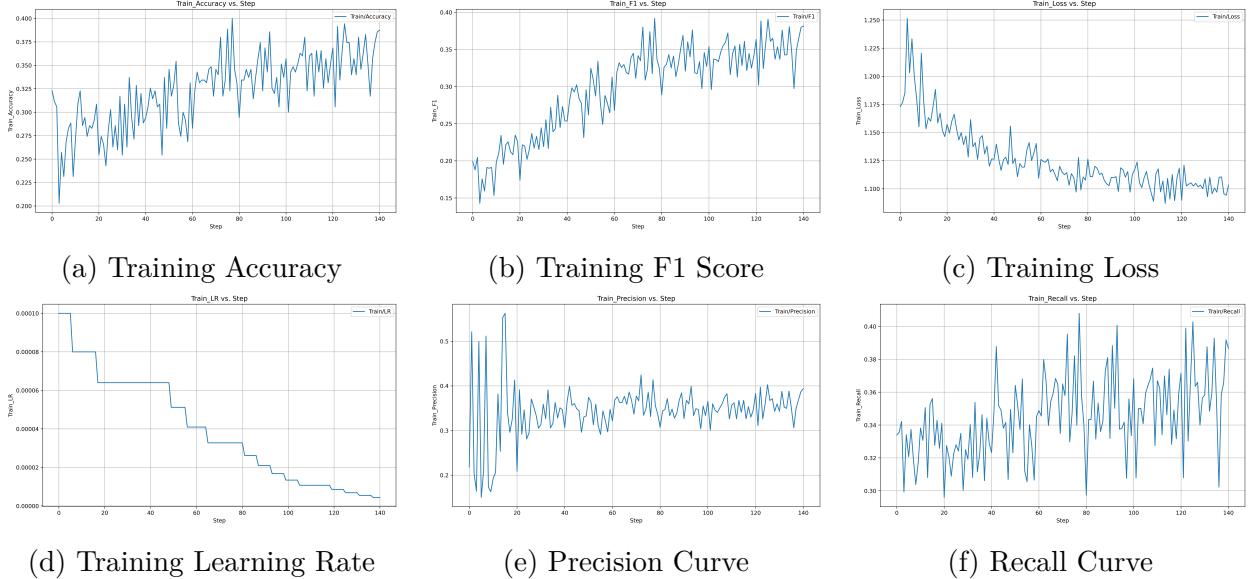


Figure 8: BERT model training performance.

Although BERT ultimately achieves our goal of improving Chinese financial sentiment classification, the presence of "meaningless" tokens in word clouds suggests that the model may struggle with specialized financial terminology. Furthermore, the low accuracy and F1 score indicate that the model may not be well-suited or well-tuned for the financial sentiment analysis task. Future work will focus on refining translation techniques, exploring Bayesian or traditional machine learning methods, and enhancing data filtering mechanisms to improve classification accuracy and interpretability.

5 Discussion

Our multi-stage pipeline leverages translation to alleviate data scarcity, yet the performance and interpretability challenges remain. "Meaningless" tokens in word clouds reflect inadequate mapping of specialized English financial phrases into Chinese. Potential remedies include domain-specific dictionaries and graph-based methods that unify semantic representations across languages.

BERT vs. Bayesian Methods While BERT excels at capturing subtle sentiment cues, Bayesian and other classic machine learning models (e.g., random forests, SVM) can outperform deep neural networks, particularly when data are limited or transparency is critical. Bayesian frameworks, for example, provide posterior probabilities that quantify uncertainty, enhancing the trustworthiness of modeling outcomes in high-stakes environments like finance.

Alternative Approaches and Future Outlook We also foresee the growing importance of knowledge graphs or ontology-based approaches to better unify semantic representations, particularly when bridging linguistic gaps between English and Chinese. Future work will address these limitations by refining translation modules, exploring multi-lingual pre-training strategies, and examining simpler but interpretable classification algorithms.

6 Conclusion and Future Work

This paper proposed a comprehensive solution to Chinese financial sentiment analysis, underscoring the value of translation-based data augmentation, lexicon-assisted labeling, and BERT fine-tuning. Our empirical investigations confirm that the workflow substantially improves classification metrics compared to purely lexicon-based or simpler machine learning baselines. Nonetheless, translation quality shortfalls, noisy lexicon coverage, and interpretability gaps suggest the need for continued innovation:

- **Refined Translation Techniques:** Incorporating domain-specific glossaries or graph-based representations may minimize semantic distortions.
- **Bayesian Interpretability:** Further exploration of Bayesian or conventional ML methods could yield comparable or superior performance while offering clearer insights into model uncertainty.
- **Advanced Data Filtering:** Enhanced mechanisms for filtering, verifying, and curating translated texts will boost final classification accuracy.
- **Expanded Modalities:** Future research might integrate other data sources (e.g., images, audio from earnings calls, or social media metrics) into a multi-modal sentiment model.

By addressing these fronts, we aim to develop robust automated sentiment systems that not only detect subtle emotional signals but also support transparent, accountable decision-making for financial market participants. For curiosities, the code and data can be found at <https://github.com/Lemon-gpu/DataScienceFinalProject>. A recorded presentation is available at <https://www.bilibili.com/video/BV1UAkRYmEij>.

Acknowledgements: Contribution of Authors

The authors extend their profound gratitude to the instructors for their invaluable guidance and unwavering support throughout the duration of the course. Furthermore, they express

sincere appreciation for the constructive feedback provided by their peers and the intellectually stimulating discussions that have significantly deepened their comprehension of the subject matter. In the course of this project, all members contributed equally to its completion. Specifically, Chuan Jia was primarily responsible for drafting and structuring the report/paper, ensuring the coherence and clarity of the written work. Bo Li played a central role in synthesizing relevant materials, offering domain-specific insights, organizing interim results, and delivering the final presentation. Sheng Su focused on the implementation of the code, the preparation of the presentation slides, and the iterative revision and refinement of the report/paper. Notably, no group leader was designated for this project, as all members collaborated on an equal footing.

Acknowledgements: Observation

The course itself is a truly vibrant and intellectually stimulating journey, offering us a wealth of knowledge and fresh perspectives that we deeply appreciate. It is evident that the instructor has invested tremendous effort and passion into crafting such a rich and comprehensive curriculum, and we feel privileged to have been part of this exceptional learning experience. That said, we must admit that the sheer breadth and density of the material, while undeniably impressive, have occasionally left us feeling a bit overwhelmed. Coming from diverse academic backgrounds and possessing varying levels of prior knowledge, some of us have found it challenging to fully keep up with the pace and depth of the course. This is, of course, a testament to the ambitious scope of the curriculum, but it has also led to moments of fatigue as we strive to meet its high standards. Nonetheless, we remain profoundly grateful for the opportunity to engage with such a thoughtfully designed course. We humbly suggest that, perhaps in the future, dividing the material into two courses might allow students to delve even deeper into the subject matter and absorb the content more thoroughly. This small adjustment could further enhance the already remarkable learning experience that this course provides. We are confident that the knowledge and insights we have gained here will serve as invaluable assets in our academic and professional journeys, and we sincerely thank the instructor for their dedication and guidance.

Chuan Jia brings forth unique ideas of his own that he wishes to highlight. Below, we present his original words in full:

Dear Professor Lu,

First and foremost, I would like to extend my heartfelt gratitude for designing such an intellectually enriching and forward-looking curriculum for us. Your meticulous dedication and profound commitment to teaching have left a deep

impression on all of us. In particular, the inclusion of emerging fields and cutting-edge tools in the course has not only broadened our horizons but also provided us with valuable insights into the trends shaping the industry. We have truly benefited immensely from your efforts.

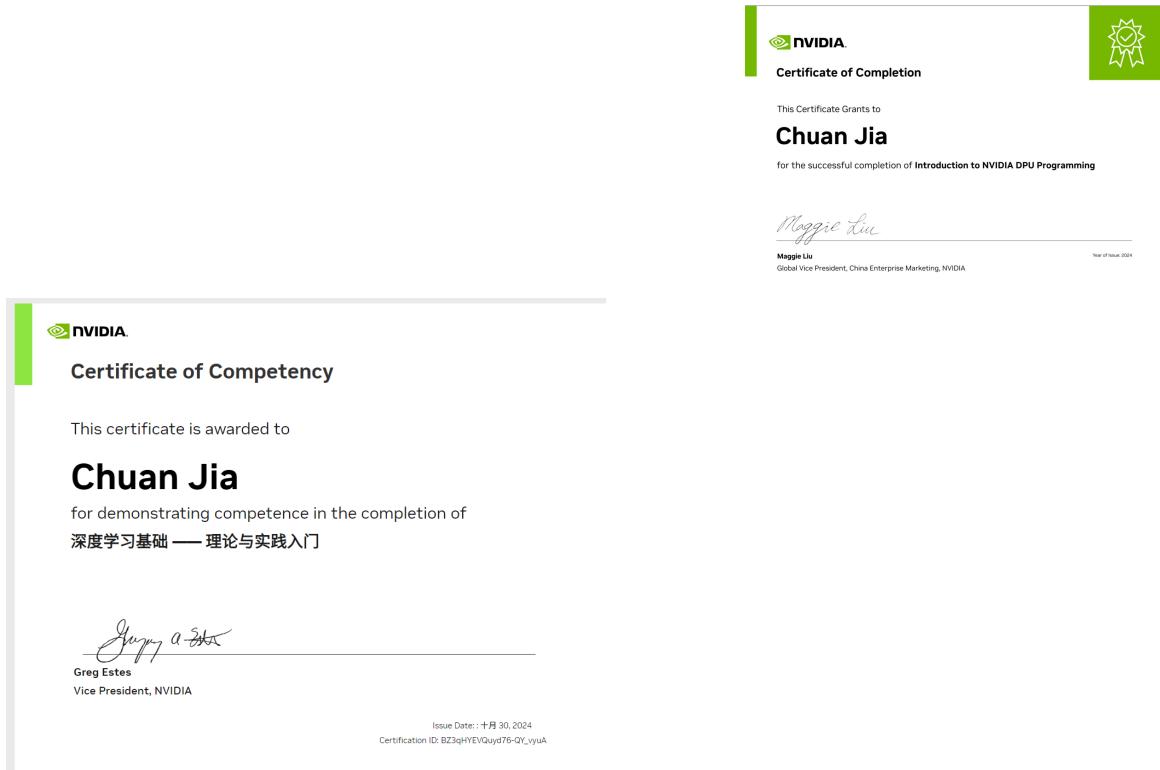
That said, as a student, I humbly wish to share a few immature reflections, acknowledging the limitations of my own abilities and understanding. For instance, some of the more macro-level content covered in the course, such as the sections related to NVIDIA, while undoubtedly visionary and enlightening, occasionally feels somewhat challenging to grasp. At times, I find myself struggling to discern how to effectively connect these concepts to practical applications. Additionally, regarding the assignments on Datacamp, while this self-directed learning approach has indeed proven to be highly beneficial, our underdeveloped time management skills occasionally result in a sense of being pressed for time. Consequently, we may not always achieve the level of learning outcomes that you had envisioned for us.

I deeply appreciate the foresight and intentionality behind your efforts to expose us to cutting-edge knowledge and skills. Such dedication and vision inspire nothing but the utmost respect and admiration. Once again, thank you for your tireless guidance and unwavering commitment to our growth.

Sincerely,
Chuan Jia

Certification

Chuan Jia's certifications for "Fundamentals of Deep Learning" and "Introduction to Nvidia DPU Programming", as well as the classification score on Testset in the final assessment project of the course "Fundamentals of Deep Learning".



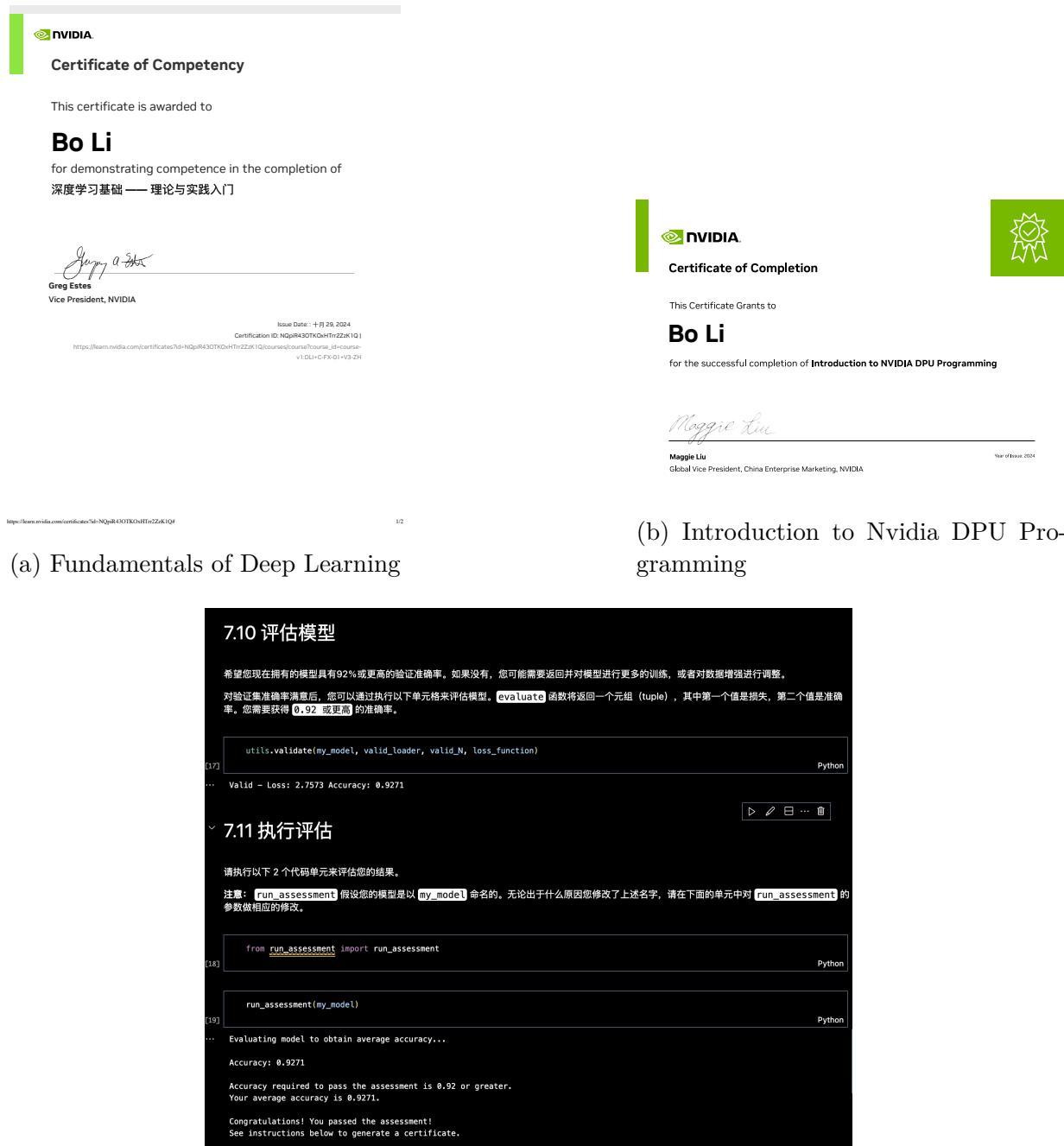
(a) Fundamentals of Deep Learning (b) Introduction to Nvidia DPU Programming



(c) Classification Score

Figure 9: Chuan Jia's certifications and classification score.

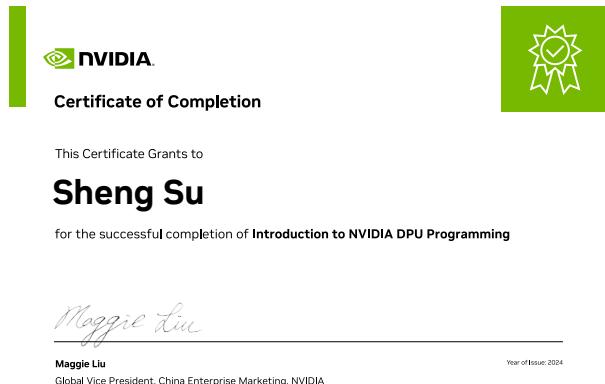
Bo Li's certifications for "Fundamentals of Deep Learning" and "Introduction to Nvidia DPU Programming", as well as the classification score on Testset in the final assessment project of the course "Fundamentals of Deep Learning".



Sheng Su's certifications for "Fundamentals of Deep Learning" and "Introduction to Nvidia DPU Programming", as well as the classification score on Testset in the final assessment project of the course "Fundamentals of Deep Learning".



(a) Fundamentals of Deep Learning



(b) Introduction to Nvidia DPU Programming

```

epochs = 12

for epoch in range(epochs):
    print('Epoch: {}'.format(epoch))
    utils.train(my_model, train_loader, train_N, random_trans, optimizer, loss_function)
    utils.validate(my_model, valid_loader, valid_N, loss_function)

...
Epoch: 0
Train - Loss: 8.7970 Accuracy: 0.8257
Valid - Loss: 2.1507 Accuracy: 0.9301
Epoch: 1
Train - Loss: 2.1987 Accuracy: 0.9569
Valid - Loss: 2.6220 Accuracy: 0.9392
Epoch: 2
Train - Loss: 1.4395 Accuracy: 0.9729
Valid - Loss: 0.9255 Accuracy: 0.9605
Epoch: 3
Train - Loss: 0.9600 Accuracy: 0.9788
Valid - Loss: 0.7705 Accuracy: 0.9726
Epoch: 4
Train - Loss: 0.7924 Accuracy: 0.9839
Valid - Loss: 1.0592 Accuracy: 0.9574
...
Valid - Loss: 0.8776 Accuracy: 0.9787
Epoch: 11
Train - Loss: 0.4606 Accuracy: 0.9907
Valid - Loss: 0.7814 Accuracy: 0.9726
Output is truncated. View as a scrollable element or open in a text editor. Adjust cell output settings.

```

(c) Classification Score

Figure 11: Sheng Su's certifications and classification score.

References

- [1] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, I. Polosukhin, Attention is all you need, in: Advances in Neural Information Processing Systems, 2017.
- [2] J. Devlin, M. W. Chang, K. Lee, K. Toutanova, BERT: Pre-training of deep bidirectional transformers for language understanding, in: Proceedings of the 2019 Conference

of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL HLT), Minneapolis, MN, USA, 2–7 June 2019.

- [3] F. Jiang, J. A. Lee, X. Martin, G. Zhou, Manager sentiment and stock returns, *J. Financ. Econ.* 132(1) (2019) 126–149.
- [4] F. Jiang, L. Meng, G. Tang, Prediction of stock returns using media text sentiment, *Econ. Q.* 20(4) (2021) 1323–1344.