Machine Learning Based Sentiment Analysis of Financial Texts

Chuan Jia, Bo Li and Sheng Su*

Macau University of Science and Technology

December 14, 2024

*Note: *The order of authors' names follows the alphabetical order of their last names.*

Problem Statement

Background:

- Massive volumes of financial text (news, reports, social media) influence investment decisions.
- Traditional analysis often overlooks subtle emotional cues embedded in textual data.
- Objective: Integrate sentiment analysis into financial decision-making to enhance rationality and predictive accuracy.

Objectives:

- Challenge: Limited high-quality, annotated Chinese financial sentiment corpora.
- Need: A robust methodology to generate and leverage domain-specific labeled datasets.
- Goal: Improve Chinese financial sentiment classification via data augmentation and Transformer-based models.

Dataset Describtion

Dataset Overview:

- Extended Chinese financial news articles from *Various Sources*.
- Sentiment Labels: Positive, Neutral, Negative.
- Size: 54749 labeled articles.

Data Preprocessing:

- Extension: Translated high-quality labeled English financial sentiment corpora into Chinese.
- Prelabel: Used a Chinese financial sentiment lexicon for non-labeled data.
- Tokenization: Jieba Chinese tokenizer.

Dataset Visualization

Sentiment Distribution:

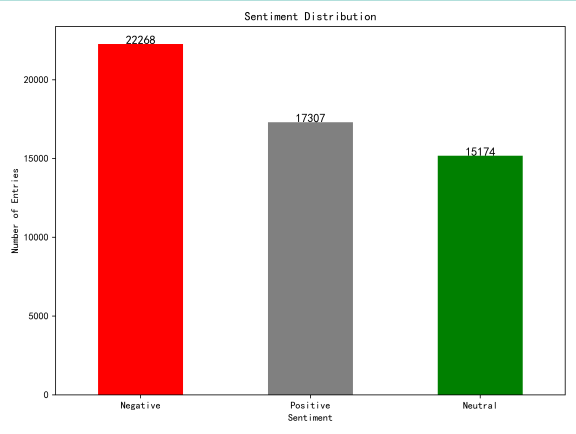

图: Distribution of Sentiments

Dataset Visualization: Neutral Sentiment

Word Cloud: Neutral Sentiment

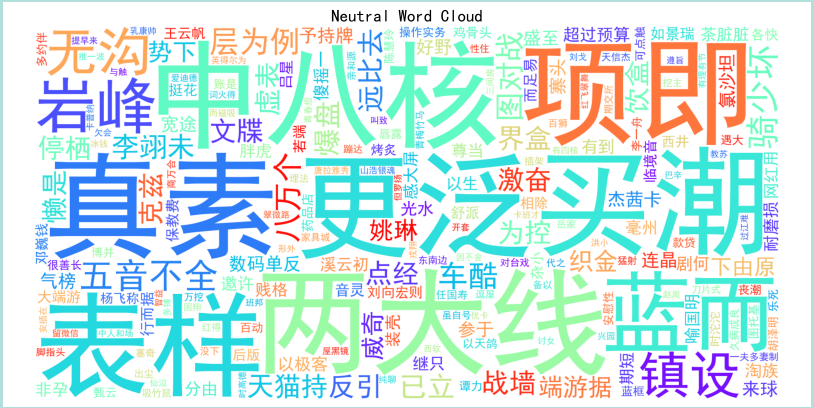

图: Neutral Sentiment

Word Cloud: Positive Sentiment

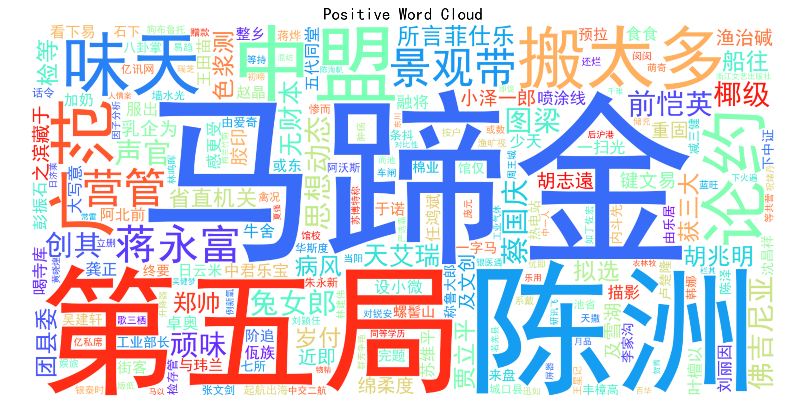

图: Positive Sentiment

Dataset Visualization: Negative Sentiment

Word Cloud: Negative Sentiment

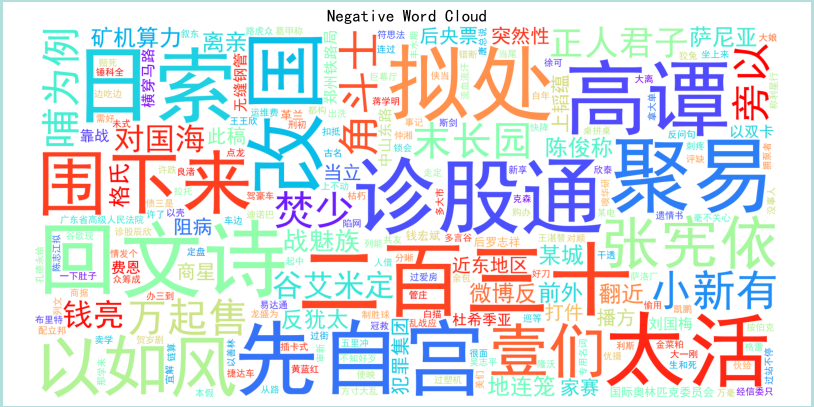

图: Negative Sentiment

Methods Overview

1. **Data Preparation and Augmentation:**
   - Translate high-quality English financial sentiment corpora into Chinese using a Transformer-based NMT model (TNMT).
   - Result: Enlarged Chinese corpus for sentiment analysis.
2. **Lexicon-Based Annotation:**
   - Use a Chinese financial sentiment lexicon for labeling, inspired by Jiang et al. (2019)[1].
   - Assign sentiments (positive, neutral, negative) via domain-specific terms.
   - Outputs: Automatically annotated Chinese dataset.
3. **Model Training: BERT Fine-Tuning:**
   - Fine-tune a pre-trained Chinese BERT model with the annotated dataset.
   - BERT captures contextual nuances, surpassing traditional methods.

[1] Jiang et al., *J. Fin. Econ.*, 132(1), 126-149.

Results: Training Loss (TNMT)

Training Loss

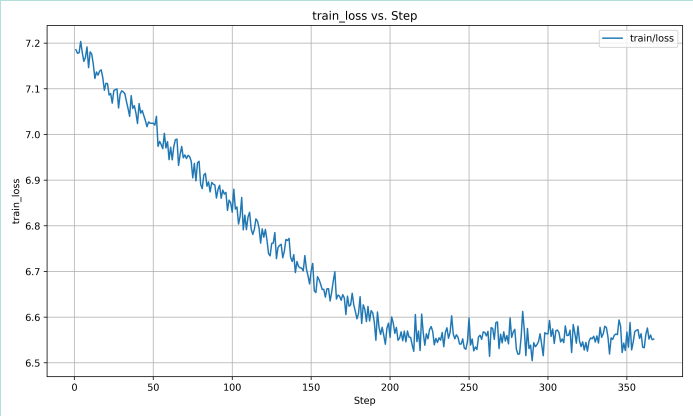

图: Training Loss: 200 steps stable, 6.55

Results: BLEU Score (TNMT)

Evaluation BLEU Score

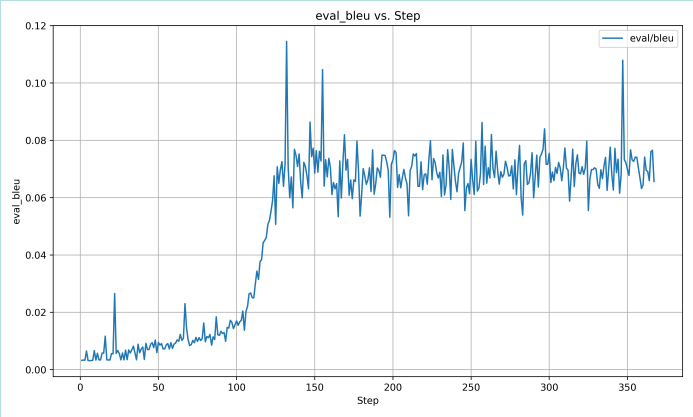

图: Evaluation BLEU Score: 130 steps stable, 0.08

Results: F1 Score (TNMT)

Evaluation F1 Score

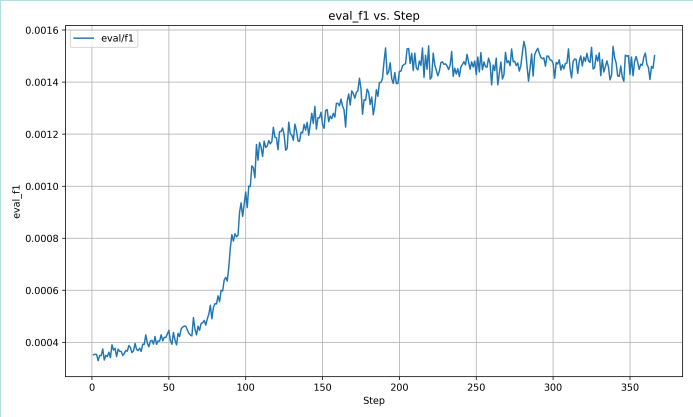

图: Evaluation F1 Score: 200 steps stable, 0.0015

Results: Training Accuracy (BERT)

Training Accuracy

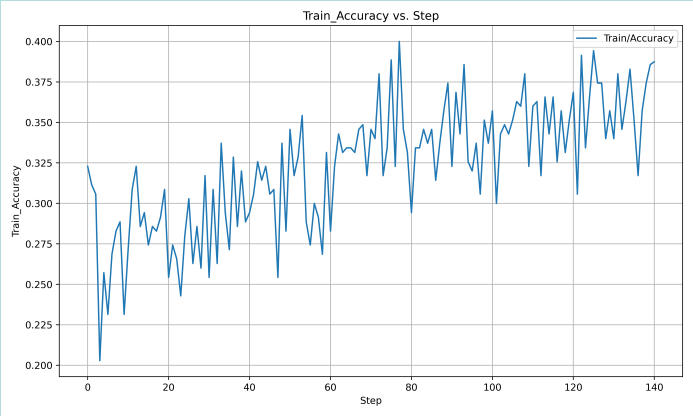

图: Training Accuracy: Increasing, ending 0.46

Results: Training F1 Score (BERT)

Training F1 Score

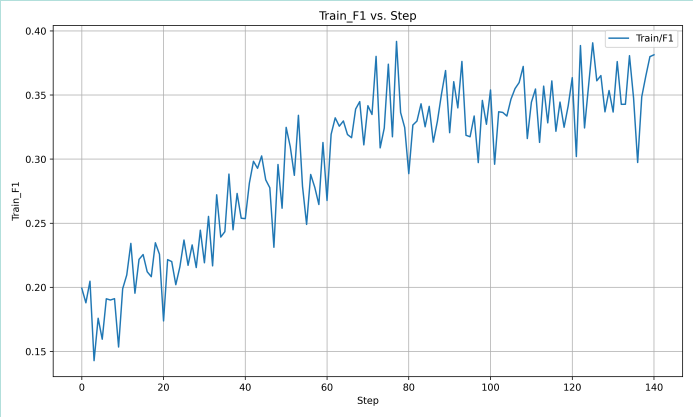

图: Training F1 Score: 60 steps stable, 0.39

Results: Training Loss (BERT)

Training Loss

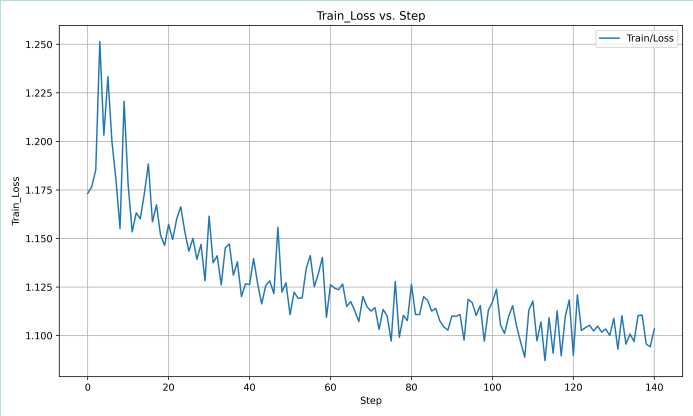

图: Training Loss: 120 steps stable, 1.06

Results - Domain Adaptation

What we learn:

- Effectively adapted English-origin sentiment cues to Chinese texts.
- Enhanced model robustness in dynamic financial contexts.
- Demonstrated scalable approach to other languages/domains.

Example:

- Chinese Text: "在英国央行本次议息会议上，市场焦点可能会落在表决结果和政策制定者的沟通上。如果投票结果显示是一个势均力敌的决定，并且会议声明再次指出不急于进一步降息，可能打压未来的宽松预期。"
- Sentiment Label: Negative

Conclusion: Contributions & Background

Contributions:

- Used TNMT to create Chinese sentiment data.
- Applied lexicon-based annotation to expand datasets.
- Fine-tuned BERT for better classification.

Background:

- Financial Texts: Market analyses, social media, reports.
- Emotional Info: Key for sentiment, trends, opportunities.
- Challenge: Manual analysis too slow.
- Solution: Potential Automated systems.

Conclusion: Improvements and Future Directions

- Current System Limitations:
  - Post-translation quality issues.
  - Ineffective use of dictionary information.
  - Need for better data filtering and quality control.
- Future Directions:
  - Fine-tuning emotional analysis modules.
  - Leveraging knowledge graphs for semantic alignment.
  - Exploring diffusion models for sentiment prediction.

Thank You!

Code Resources: `https://github.com/Lemon-gpu/DataScienceFinalProject`
Video Resources: Will be uploaded to YouTube.