# 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.

## Sentiment Distribution:

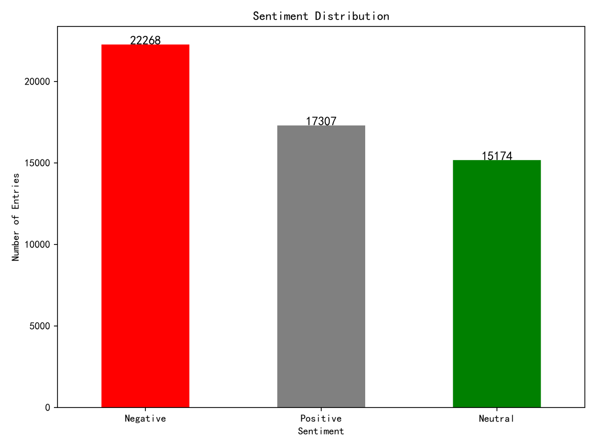


Figure: Distribution of Sentiments

# Dataset Visualization (Continue)

**Word Clouds:**

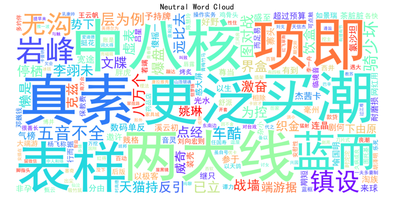


Figure: Neutral Sentiment


Figure: Positive Sentiment

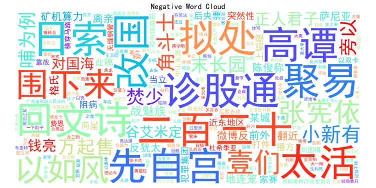


Figure: Negative Sentiment

# 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.

## Transformer Model Results:

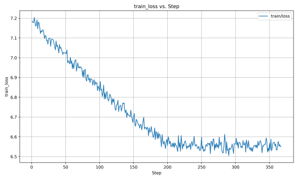


Figure: Training Loss: 200 steps stable

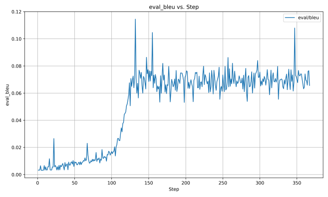


Figure: Evaluation BLEU Score: 130 steps stable

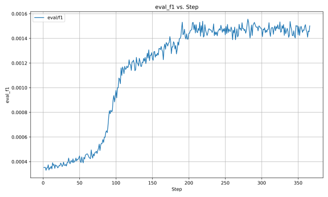


Figure: Evaluation F1 Score: 200 steps stable

**BERT Model Training Results:**

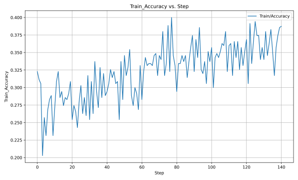


Figure: Training Accuracy: Increasing

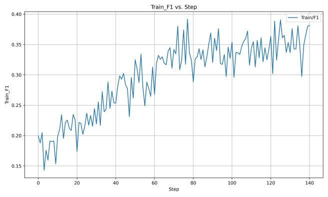


Figure: Training F1 Score: 60 steps stable

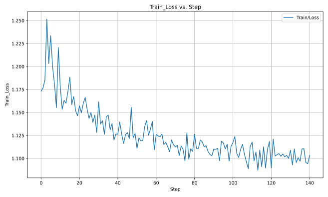


Figure: Training Loss: 120 steps stable

# Results - Domain Adaptation

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

# 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.