Political Bias in Text Generation

Gianluca Cacciola

February 25, 2025

Introduction

Problem Statement and Proposed Solution:

- This study explores political bias in the DeepSeek-R1-Distill-Llama-8B model using advanced NLP techniques.
- Focuses on sentiment analysis, named entity recognition and political stance classification.
- Aims to uncover patterns and insights from textual data through different labels score correlation.

Methodology

Data Collection:

- Source and nature of the dataset used.
- Preprocessing steps applied to clean and structure the data.

Analytical Methods:

- Sentiment Analysis with Twitter-roBERTa-base.
- Named Entity Recognition with BERT-base-ner.
- Zero-Shot Stance Detection with BERT-large-mnli.

Tools and Frameworks:

- Python libraries: Transformers, Scikit-learn, Pandas, Matplotlib, Torch.
- Deep learning models for NLP.

Sentiment Analysis Approach

Implementation:

- Utilized the cardiffnlp/twitter-roberta-base-sentiment-latest transformer-based model for sentiment classification.
- Processed text in chunks to handle longer documents.
- Applied confidence scores to classify results accurately.

Goals:

- Show possible discrepancy of sentiment scores amongs different topics.
- Combine the highligthed discrepancy with other scores for a better understanding of bias in text-generation.

Named Entity Recognition (NER)

Implementation:

- Used the dslim/bert-base-NER pre-trained transformer-based model for entity recognition.
- Applied NER to detect entity related to China or his ruling party.
- Verified entity relevance and accuracy through manual checks.

Goals:

- Using the number of detected entity to understand the leaning of the model to introduce such entities when not explicitly requested.

Zero-Shot Stance Detection

Implementation:

- Used the facebook/bart-large-mnli pre-trained transformer-based model for MNLI.
- Applied the model to compute the relevance of prompts and responses towards Communism and Capitalism.

Goals:

- Discern if the model associates different sentiments to different topics accoring to economic and political ideologies.

Visualization of Results

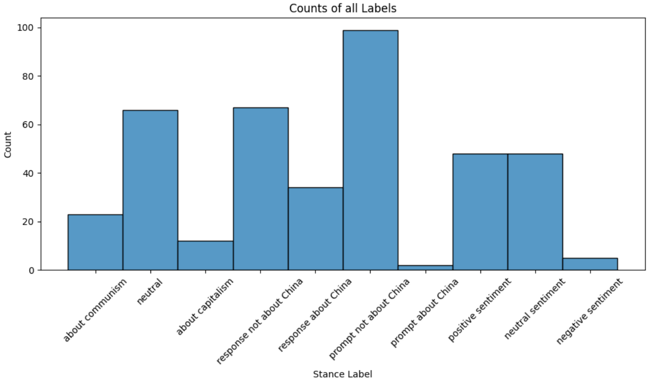


Figure: labels occurencies distribution

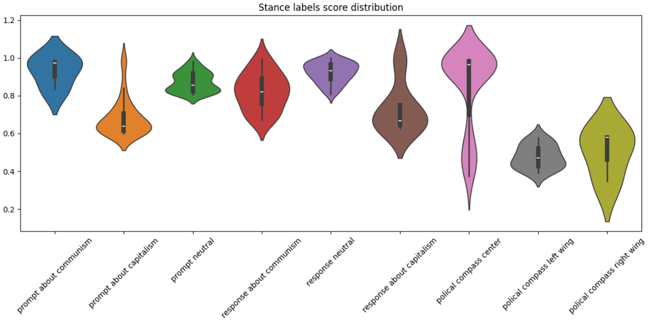


Figure: labels score distribution

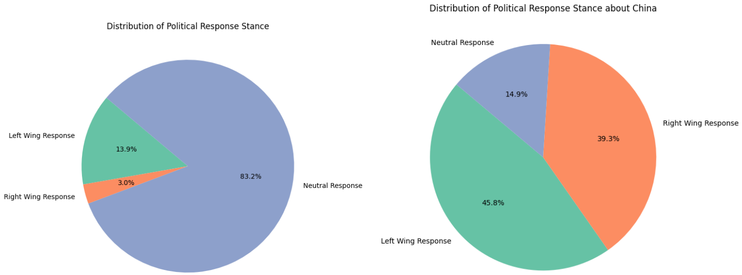
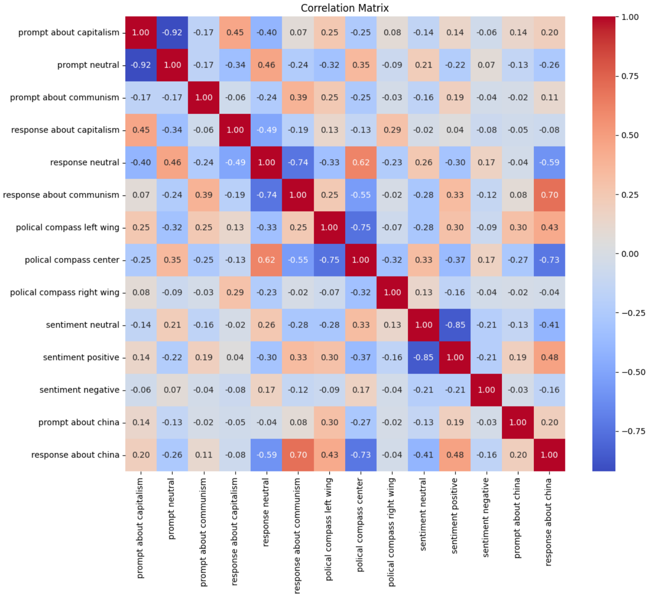


Figure: distribution skews of labels occurency in China related responses

Putting the pieces togher

Correlation Matrix:

- The correlation matrix about the different lables can be used to infer bias in the responses of the models.
- Asymmetric correlations can imply bias towards particular topics.
- The scores are analyzed to try to answer at the key question of this research.


Correlation Matrix
prompt about capitalism 1.00 -0.92 -0.17 0.45 -0.40 0.07 0.25 -0.25 0.08 -0.14 0.14 -0.06 0.14 0.20
prompt neutral -0.92 1.00 -0.17 -0.34 0.46 -0.24 -0.32 0.35 -0.09 0.21 -0.22 0.07 -0.13 -0.26
prompt about communism -0.17 -0.17 1.00 -0.06 -0.24 0.39 0.25 -0.25 -0.03 -0.16 0.19 -0.04 -0.02 0.11
response about capitalism 0.45 -0.34 -0.06 1.00 -0.49 -0.19 0.13 -0.13 0.29 -0.02 0.04 -0.08 -0.05 -0.08
response neutral -0.40 0.46 -0.24 -0.49 1.00 -0.74 -0.33 0.62 -0.23 0.26 -0.30 0.17 -0.04 -0.59
response about communism 0.07 -0.24 0.39 -0.19 -0.74 1.00 0.25 -0.55 -0.02 -0.28 0.33 -0.12 0.08 0.70
polical compass left wing 0.25 -0.32 0.25 0.13 -0.33 0.25 1.00 -0.75 -0.07 -0.28 0.30 -0.09 0.30 0.43
polical compass center -0.25 0.35 -0.25 -0.13 0.62 -0.55 -0.75 1.00 -0.32 0.33 -0.37 0.17 -0.27 -0.73
polical compass right wing 0.08 -0.09 -0.03 0.29 -0.23 -0.02 -0.07 -0.32 1.00 0.13 -0.16 -0.04 -0.02 -0.04
sentiment neutral -0.14 0.21 -0.16 -0.02 0.26 -0.28 -0.28 0.33 0.13 1.00 -0.85 -0.21 -0.13 -0.41
sentiment positive 0.14 -0.22 0.19 0.04 -0.30 0.33 0.30 -0.37 -0.16 -0.85 1.00 -0.21 0.19 0.48
sentiment negative -0.06 0.07 -0.04 -0.08 0.17 -0.12 -0.09 0.17 -0.04 -0.21 -0.21 1.00 -0.03 -0.16
prompt about china 0.14 -0.13 -0.02 -0.05 -0.04 0.08 0.30 -0.27 -0.02 -0.13 0.19 -0.03 1.00 0.20
response about china 0.20 -0.26 0.11 -0.08 -0.59 0.70 0.43 -0.73 -0.04 -0.41 0.48 -0.16 0.20 1.00
prompt about capitalism
prompt neutral
prompt about communism
response about capitalism
response neutral
response about communism
polical compass left wing
polical compass center
polical compass right wing
sentiment neutral
sentiment positive
sentiment negative
prompt about china
response about china
1.00
0.75
0.50
0.25
0.00
−0.25
−0.50
−0.75

Key Findings

The correlation analysis highlights some key areas where political bias may manifest in the generated text:

- Capitalism and communism discussions elicit opinionated responses, making neutrality difficult to maintain.
- Left-leaning responses tend to have slightly more positive sentiment (0.30 correlation), suggesting a slight positivity bias in left-wing narratives.
- Responses about China are moderately correlated with positive sentiment (0.48), indicating a tendency for China-related discussions to be framed in a positive light.

Conclusion

Final Thoughts:

- This study highlights the power of NLP techniques in extracting valuable insights from textual data.
- Sentiment analysism NER and Stance Recognition together offer a comprehensive understanding of textual trends and relationships.

Future Directions:

- Improving prompt list used to generate the texts.
- Expanding analysis to different domains and larger datasets.
- Enhancing stance recognition with Few-Shot Classification.

Thank You for Your Attention!