POLICY STANCE DETECTION IN POLITICAL TEXTS USING NATURAL LANGUAGE PROCESSING

Hamdi Kerem Kucukengin [1,2,†]

[1] Northwestern University

Master of Science in Data Science Program

[2] Github Repository:

https://github.com/hamodikk

† Address to which correspondence should be addressed:

kkucukengin@gmail.com

Abstract

Understanding the position or "stance" that a document takes on a policy issue is crucial for tracking public discourse, combating misinformation, and supporting research in political communication. This project addresses the task of stance detection in news articles, using a transformer-based RoBERTa model to classify articles as supporting, opposing, or being neutral toward a given topic. In addition to building and evaluating a supervised model trained on a manually annotated corpus, this work also explores the use of large language models (LLMs) for zero-shot annotation and label generation. We compare model performance across two corpora: one annotated by hand, and the other using LLM-predicted stance labels. Our results show that while LLM-generated labels enable higher raw accuracy, they also introduce a strong bias toward the "support" class. We find that stance detection performance is highly sensitive to label quality and corpus balance and size, highlighting the trade-offs in relying on automated annotation methods for sensitive NLP tasks.

Keywords: Natural Language Processing, NLP, stance detection, zero-shot classification, policy analysis, LLM, language model annotation

Contents

Introduction and Problem Statement

In recent years, natural language processing (NLP) tools have played an increasingly important role in analyzing the language of politics, public health, and social policy. One particularly valuable capability is stance detection – the ability to automatically determine whether a document expresses support, opposition, or neutrality toward a given topic. While much prior work in this space has focused on social media posts and short-form opinion texts, this project investigates the more subtle and complex challenge of detecting stance in policy journalism.

This work asks: *To what extent can transformer-based models detect stance in policy-oriented texts?*

And: *How does performance differ when the model is trained on human-generated versus LLM-generated annotations?*

The motivation for this work is both technical and social. Technically, we aim to understand how well stance detection methods generalize in a low-data setting, and whether large pretrained language models – in this case BART-MNLI used for zero-shot entailment-based classification – can meaningfully contribute to annotation tasks. Socially, the ability to monitor implicit policy stances in public discourse is increasingly critical in an era of misinformation and political polarization.

This work builds two classifiers using RoBERTa: one trained on a small, carefully labeled dataset annotated by this author, and one trained on a comparable set of LLM-generated labels. We analyze the trade-offs in accuracy, generalization, and class balance between the two models. We also reflect on the practical use of LLMs in low-resource annotation settings – both as a shortcut and a potential source of bias.

Literature Review

Researchers typically frame stance detection as a supervised or semi-supervised classification task in which a model is trained to predict whether a piece of text supports, opposes or remains neutral toward a given proposition or policy. Early work in stance detection relied on feature engineering approaches using syntactic cues, sentiment indicators, and topic modeling. However, with the introduction of large pretrained transformer models like BERT (Devlin et al., 2019), more recent efforts have shifted toward fine-tuning these architectures for sentence or document level classification tasks.

Transformer-based models such as RoBERTa (Liu et al., 2019) and BART (Lewis et al., 2020) have been particularly successful in stance detection tasks, especially in low-data or domain specific settings. These models can be fine-tuned for direct classification, or used in zero-shot settings by framing the task as a natural language inference (NLI), as demonstrated in zero-shot pipelines such as BART-MNLI (Yin et al., 2019). Studies done similar to Hardalov et al. (2021) have benchmarked multiple transformer models across stance detection datasets, and show that even zero-shot models can outperform traditional supervised classifiers when labels are limited or inconsistent.

Earlier approaches to stance detection – such as Somasundaran and Wiebe (2010) and Thomas et al. (2006) – relied heavily on identifying linguistic patterns or sentiment orientation. These works laid the groundwork for understanding the relationship between argument structure and stance but are not largely superseded by neural models with deeper contextual awareness.

While much of the literature focuses on short-form opinion texts like tweets or other social media, some recent research has extended stance detection to longer and more detailed policy-oriented documents (Augenstein et al., 2016)(Allaway and McKeown, 2020), which align more closely with the scope of this project.

Data

News articles and agencies often claim neutrality, which makes stance detection and argument mining a challenge. While explicit stances in news articles might be rare, implicit stances like framing choices can still indicate the position of a news article. The corpus used for this paper consists of 49 documents that has been provided by 12 independent researchers, each providing 4 text documents on current and previous presidents of the United States, with varying topics in economy, health, education, domestic and international affairs.

The dataset includes:

- **DSI_Title:** Filename of the document
- **Text:** The full text of the document

Data Preprocessing

To prepare the data for analysis, we applied the following preprocessing steps:

- **Text cleanup**
  - NUL byte and multibyte character removal
  - Encoding normalization
  - Malformed character cleanup (â€™ → ’)

- URL and link wrapper removal

- **Tokenization and Stopword Removal:** Standard NLP techniques to clean up the text.
- **Final reassembly:** We rejoined the clean tokens into strings for downstream modeling using TF-IDF or input to transformer tokenizers.

In addition to these preprocessing steps, the documents are hand labeled by the author with stance categories based on support, opposition or neutrality. The author initially annotated a subset of 15 documents, followed by the entire corpus. Finally, BART_MNLI is utilized to annotate the same corpus for comparison.

Stance Label	*Manual Annotation*	*LLM Annotation*
Support	16	32
Oppose	16	17
Neutral	17	0

Table 1. Comparison of stance label distributions between manual and LLM-based annotations across the full corpus of 49 documents. The LLM demonstrated a notable bias toward the *Support* class, while manual annotations were more evenly distributed.

Methods

We used a transformer-based classification approach to detect the stance of policy-related articles toward specific topics. Our model is based on the roberta-base variant from Hugging Face's transformers library. The architecture included the pretrained RoBERTa encoder followed by a dense linear layer for classification into three stance categories: *support*, *oppose*, and *neutral.* The full stance detection pipeline – from preprocessing to classification – is visualized in Figure 1.

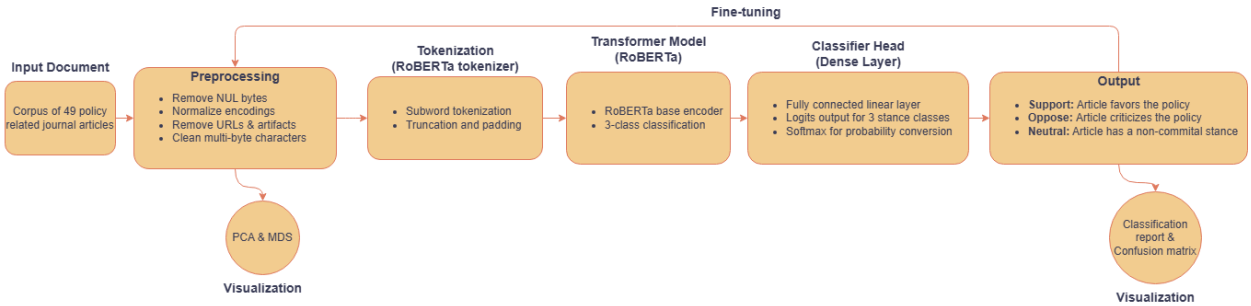

Figure 1. Overview of the stance detection pipeline. The system takes a corpus of policy-related articles and performs multi-step preprocessing, including removal of encoding errors, artifacts, and links. Tokenization is performed using the RoBERTa tokenizer, followed by fine-tuning or a RoBERTa-based classification model. Outputs are classified into one of three stance categories – Support, Oppose, or Neutral – and results are evaluated using standard classification metrics.

For preprocessing, input documents were cleaned to remove NUL bytes, multi-byte encoding artifacts, and URLs. We tokenized each cleaned document using the RoBERTa tokenizer with truncation and padding implemented to accommodate the model's input requirements.

The classificer head consists of a single fully connected linear layer. This layer output raw logits for each of the three stance classes, which were then passed through a softmax function to produce probability distributions. We trained the model for 5 epochs with the default learning rate of the Trainer class and evaluated it using a stratified 60/40 train-test split.

Two training sets were used: one labeled manually by the author, and one labeled by using LLM predictions from the facebook/bart-large-mnli zero-shot classifier. To assess model performance, we computed precision, recall, and F1-score for each stance class, and visualized results using confusion matrices.

This dual approach allowed us to evaluate how well the model learns from small human-annotated corpora versus automatically generated labels, and to examine the effects of label quality and downstream classifier performance.

Results

Exploratory Document Clustering

Prior to training any stance classification models, unsupervised clustering was used to explore patterns in the text corpus. We applied two methods:

- TF-IDF vectorization + PCA projection
- Doc2Vec embeddings + MDS projection

Each document was clustered into one of eight groups using KMeans clustering, and the top terms from each cluster were extracted to examine theme separation.

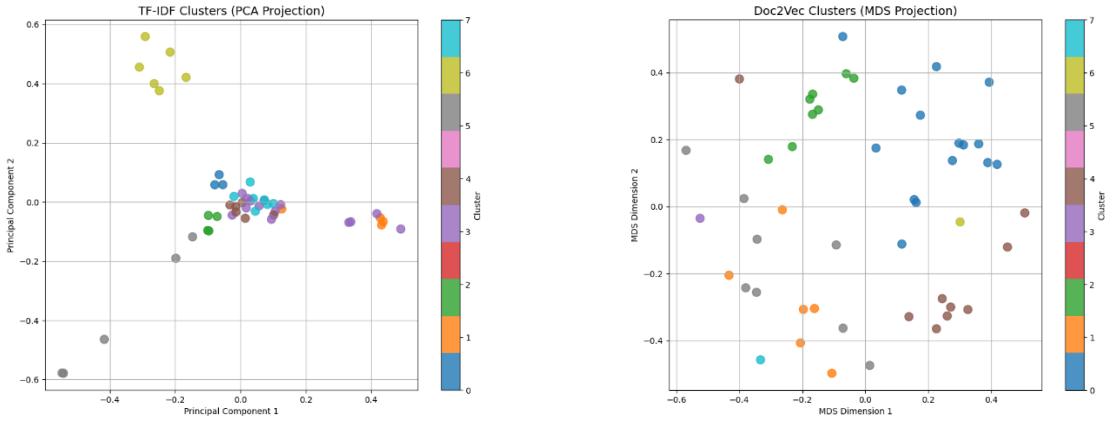

Figure 2. PCA projection of TF-IDF features and MDS projection of Doc2Vec embeddings colored by KMeans cluster label.

The visualizations show some thematic separation among articles, although clusters do not clearly align with stance labels. This reinforces the need for supervised models to capture nuanced stance information.

Performance on Manually Annotated Subset

A RoBERTa-based classifier was trained and evaluated using the 15 manually annotated documents. The classification results are shown in **Error! Reference source not found.**(A).

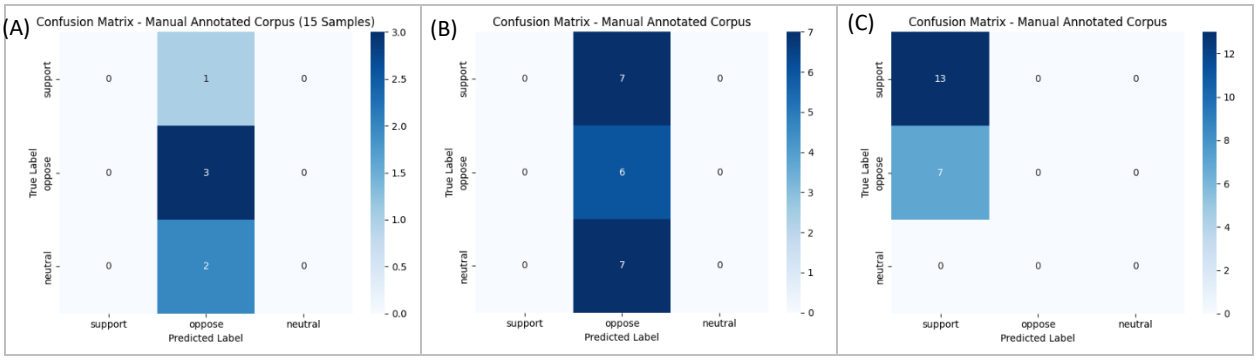

Figure 3. (A) Confusion matrix for model predictions on 15-document manually annotated subset. (B) Confusion matrix for model trained and tested on manually annotated full corpus of 49 documents. (C) Confusion matrix for model trained on LLM-annotated corpus of 49 documents.

The model struggled with this small dataset, defaulting toward a dominant class. The classification report (see Appendix A) confirms that while precision on certain labels was high, recall was poor for most.

Performance on Fully Annotated Corpus

We re-trained the classifier using the full set of 49 manually annotated documents. The updated performance is summarized in Figure 3(B). The overall accuracy improved slightly, but the model continued to overpredict the "oppose" label. The detailed classification metrics are provided in Appendix B.

Comparison to LLM-Annotated Corpus

To evaluate whether large language models (LLMs) could help scale annotation, a second version of the corpus was annotated using the facebook/bart-large-mnli model via Hugging Face's zero-shot classification pipeline. We used this data to train a new model with the same architecture as the previous ones. The results of this model can be found in Figure 3(C).

Despite higher accuracy, the model overfit to the dominant "support" label predicted by the LLM. This raises concerns about the reliability of fully automated labeling for nuanced stance detection tasks. Additionally, to compare the annotated corpus between LLM and manual, we looked at the Cohen's Kappa value. A Cohen's Kappa value of 0.121 points to a low agreement score, confirming that while LLMs can speed up the annotation process, the annotations may require manual verification.

Discussion

The results highlight the difficulty of stance detection in longer-form, policy-oriented text. Even with fine-tuning on manually labeled documents, the transformer-based model exhibited label prediction skew. On the full manually labeled corpus, the model showed a strong bias toward "oppose" class, while the model trained on LLM-annotated data heavily favored "support". Additionally, all three classification reports show the potential effects of corpus size on exacerbating the model bias.

This reversal in dominant class prediction between the two models reflects not only label distribution bias, but also the different ways stance is expressed and interpreted for human annotation, LLM annotation and for classification model. The author adopted a more cautious reading of policy positions, while the LLM more frequently inferred implicit support. Additionally, the author or any human annotator could have personal political views, which can create bias in annotation compared to LLM annotation.

The unsupervised clustering results further confirm that semantic similarity alone does not account for stance class separation. Clusters based on TF-IDF or Doc2Vec did not align with annotated stance labels, indicating that lexical and topic similarities are insufficient for accurate stance classification.

Although the model trained on LLM annotations achieved a higher accuracy overall, it failed to generalize across all stance categories. The confusion matrix shows that this model overwhelmingly predicted "support", and failed to identify "oppose" or "neutral" correctly. This imbalance, along with a low Cohen's Kappa score of 0.121, shows the risk of using zero-shot LLM outputs as training labels without validation.

Ultimately, the manually labeled corpus – though more labor intensive – enabled a more balanced and interpretable evaluation. These results suggest that while LLMs are helpful in accelerating annotation, their output may reflect oversimplified or biased interpretations of stance.

Conclusions

This project explored the challenge of stance detection in policy-oriented documents using transformer-based models. We constructed a small, manually annotated corpus and evaluated a RoBERTa classifier trained on these labels. In parallel, we generated an alternate version of the corpus labeled using a zero-shot LLM and trained an identical model for comparison.

While the LLM-based model achieved higher raw accuracy, it exhibited strong class imbalance, defaulting to the "support" label. The manually annotated mode, by contrast, offered a more balanced classification despite overall lower performance. These differences emphasize the importance of label quality and distribution, particularly in nuanced tasks like stance detection, where implicit language and rhetorical framing play key roles.

Our findings show that LLMs can accelerate the annotation process but should be used cautiously – especially when their outputs are used to train downstream models. Ultimately, stance detection remains a complex NLP task that benefits from thoughtful, human-involved annotation and continued exploration of models that can better capture rhetorical and contextual nuance.

Directions for Future Work

Future work could explore hybrid annotation workflows in which LLM-generated labels are validated or refined by human annotators or integrated with semi-supervised learning strategies. Incorporating argument mining or rhetorical role detection may help differentiate support from opposition more reliably in policy-oriented texts. Additional data collection and annotation could also help improve model performance by increasing the number of annotated documents available for training and evaluation. This could indirectly contribute to reducing the potential discrepancy between stance classes for LLM annotations.

Acknowledgements

I would like to express my gratitude to Dr. Alianna Maren for her continued support, insightful feedback, and for fostering an engaging learning environment. I would also like to thank the independent researchers for their contributions to the corpus. Their input played a key role in enabling the work accomplished here.

Data & Code Availability

Data and the code for this work can be found in the github repository here.

References

- Allaway, Emily, and Kathleen McKeown. 2020. "Zero-Shot Stance Detection: A Dataset and Model Using Generalized Topic Representations." In Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), 8913–8931. Association for Computational Linguistics. https://doi.org/10.18653/v1/2020.emnlp-main.717.
- Augenstein, Isabelle, Tim Rocktäschel, Andreas Vlachos, and Kalina Bontcheva. 2016. "Stance Detection with Bidirectional Conditional Encoding." In Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP), 876–885. Association for Computational Linguistics. https://doi.org/10.18653/v1/D16-1084.
- Devlin, Jacob, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. "BERT: Pre-Training of Deep Bidirectional Transformers for Language Understanding." arXiv preprint. https://arxiv.org/abs/1810.04805.
- Hardalov, Momchil, Ivan Koychev, and Preslav Nakov. 2021. "A Survey on Stance Detection for Mis- and Disinformation Identification." arXiv preprint. https://arxiv.org/abs/2103.00242.
- Lewis, Mike, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. "BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension." arXiv preprint. https://arxiv.org/abs/1910.13461.
- Liu, Yinhan, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, et al. 2019. "RoBERTa: A Robustly Optimized BERT Pretraining Approach." arXiv preprint. https://arxiv.org/abs/1907.11692.
- Somasundaran, Swapna, and Janyce Wiebe. 2009. "Recognizing Stances in Online Debates." In Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP, 226–234. Association for Computational Linguistics. https://aclanthology.org/P09-1026/.
- Thomas, Matt, Bo Pang, and Lillian Lee. 2006. "Get out the Vote: Determining Support or Opposition from Congressional Floor-Debate Transcripts." In Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing, 327–335. Association for Computational Linguistics. https://aclanthology.org/W06-1639/.

Appendix A

Classification Report – Manual Subset Model

	Precision	Recall	F1-score	Support
Support	1.00	0.00	0.00	1
Oppose	0.50	1.00	0.67	3
Neutral	1.00	0.00	0.00	2
Accuracy			0.50	6
Macro avg	0.83	0.33	0.22	6
Weighted avg	0.75	0.50	0.33	6

Appendix B

Classification Report – Full Manual Corpus

	Precision	Recall	F1-score	Support
Support	1.00	0.00	0.00	7
Oppose	0.30	1.00	0.46	6
Neutral	1.00	0.00	0.00	7
Accuracy			0.30	20
Macro avg	0.77	0.33	0.15	20
Weighted avg	0.79	0.30	0.14	20

Appendix C

Classification Report – LLM-Labeled Corpus

	Precision	Recall	F1-score	Support
Support	0.65	1.00	0.79	13
Oppose	0.00	0.00	0.00	7
Accuracy			0.65	20
Macro avg	0.33	0.50	0.39	20
Weighted avg	0.42	0.65	0.51	20