PragMaBERT: Analyzing Pragmatic Markers in Political Speech

Stanford CS224N Custom Project

Matt Wise
Department of Computer Science
Stanford University
mattwise@stanford.edu

Houda Nait El Barj
Department of Economics, Computer Science
Stanford University
hnait@stanford.edu

Abstract

This project introduces a novel dataset and analytical framework aimed at identifying and evaluating pragmatic markers (PrMs) in political discourse, with a focus on markers used manipulatively. Utilizing a fine-tuned BERT model, we demonstrate the capability to detect and analyze these markers effectively within political speeches. Our approach not only enhances the understanding of PrMs' use in political contexts but also provides a foundation for addressing the manipulative use of language in politics. Preliminary results indicate that our model can accurately identify context-dependent manipulative PrMs. This work contributes to the broader field of natural language processing by providing tools to analyze the nuanced use of language in influential arenas. The dataset creation, powered by manual labeling, is ongoing and aims to further refine the model's accuracy and applicability. This work has significant implications for enhancing transparency in political communication and can be instrumental for educational purposes and media analysis.

Key Information: TA mentor: Anna Goldie. No external collaborators/mentor. Not sharing project

1 Introduction

Language is a fundamental tool for communication, persuasion, and manipulation. Beyond the explicit content of words, the subtleties of language usage significantly influence how messages are perceived and understood. Pragmatic markers (PrMs) are particularly pivotal in shaping communication. These are syntactically diverse linguistic elements that perform various attitudinal and meta-communicative functions, helping to structure discourse and convey speaker intent. Examples of pragmatic markers include phrases like "of course," "surely," "I think," and "well," which can subtly influence the interpretation of an utterance by indicating certainty, doubt, agreement, or politeness [4].

The complexity of identifying and analyzing pragmatic markers in natural language processing (NLP) reflects the dynamic nature of language evolution and usage. PrMs frequently undergo grammaticalization, a process where lexical items evolve into grammatical markers, often acquiring new meanings and functions [1, 5]. This evolution can result in semantic overlap and ambiguity, making the automated detection and interpretation of PrMs a challenging task [6].

The study of pragmatic markers is not only a linguistic concern but also bears significant social implications. In political discourse, for instance, the strategic use of PrMs can greatly affect public perception and influence. Politicians often employ these markers to hedge their statements, thereby softening claims or expressing certainty to strengthen their position without committing fully to a statement [4]. This manipulation can have profound effects, especially in an era where trust in institutions is waning.

This research addresses the need for better tools to analyze the use of pragmatic markers in political speech. We focus on two specific types of PrMs:

- **Hedging markers**, like "it seems" or "possibly," which mitigate the force of an assertion.
- **Authority markers**, like "obviously" or "in fact," which encourage an audience to trust the speaker's authority or the veracity of their statements.

We introduce a novel dataset and enhancements to a BERT-based model, which we have named PragMaBERT, to detect and categorize these markers. The dataset, derived from the MediaSum database, includes annotated instances of both hedging and authority markers, providing a rich resource for training and evaluating our model.

The contributions of this paper are organized into three primary workstreams:

1. **PragMaIMS (Pragmatic Marker Insights from MediaSum)**: Development of a comprehensive dataset annotated with instances of pragmatic markers.
2. **PragMaBERT (Pragmatic Markers in BERT)**: Adaptation and fine-tuning of a state-of-the-art language model to identify and categorize pragmatic markers in political speeches, surpassing current models like GPT-4-Turbo and Gemini-1.5-Pro in performance.
3. **Analysis**: Application of the trained model to a variety of political texts to demonstrate its practical utility and the potential for broader applications in automated discourse analysis.

Ultimately, this project aims to advance the field of NLP by developing methods that can more accurately interpret the nuanced use of language in political contexts. This has implications not only for political analysis but also for enhancing transparency and accountability in public discourse. By equipping researchers and practitioners with better tools to detect and understand pragmatic markers, we contribute to a clearer and more transparent communication landscape.

2 Approach

2.1 Dataset Creation

We use MediaSum [8] as the base for our new dataset. MediaSum is a collection of 463K dialogue transcripts from NPR and CNN representing a wide array of dialogue styles, speakers, and settings.

We used a list of PPrMs primarily aggregated in [2]. We supplemented with PPrMs from other authors ([5], [6]). For annotation, we selected a random sampling of utterances from the MediaSum dataset that contained PPrMs. From there we reviewed the sample of conversation and classify the PPrM as either "hedge", "authority", or "none."

We developed a rubric to standardize our responses, and our final dataset only includes examples where both authors of this paper independently agreed on the classification of the PPrM (see Appendices for examples and notes from the rubric).

2.1.1 Public Model Evaluation

We used a standardized prompt to generate 1-shot JSON responses from several leading models (see 6.1 for the prompt). We calculate F1, Precision, Recall, and Accuracy metrics for each category. Overall model performance for each of these metrics is calculated as a macro average of the metrics for 'hedge', 'authority', and 'none'. A macro average gives equal weight to each category and allows us to evaluate how well the models perform across the different categories without overweighting to higher-representative samples.

2.2 Model Development

To automatically detect and classify pragmatic markers in text, we fine-tune the BERT (Bidirectional Encoder Representations from Transformers) language model [3]. BERT is a state-of-the-art pre-trained model that has achieved high performance on a wide range of natural language processing tasks, including text classification and sequence labeling. We use the BERT-base-uncased variant as our base model, which has 12 hidden layers, 768 hidden units, and 12 attention heads, totaling 110 million parameters. The model is pre-trained on a large corpus of English text, enabling it to learn rich linguistic representations that can be fine-tuned for specific downstream tasks. For our pragmatic marker detection task, we add a token classification head on top of the pre-trained BERT model. This head consists of a linear layer that takes the hidden state of each token as input and outputs a probability distribution over the possible labels (hedge, authority, or none) for each token. During fine-tuning, the model learns to assign the correct label to each token based on the context in which it appears. Our fine-tuning process involves the following steps:

- **Data Preparation:** We tokenize the input text using the BERT tokenizer and convert the tokens to their corresponding input IDs. We also create attention masks to indicate which tokens are padding and should be ignored. In addition, [START] and [END] tokens surrounding the PPrM give the model details on which terms to train the label on. The pragmatic markers in the text are mapped to their respective labels (hedge, authority, or none) to serve as the ground truth for training and evaluation.
- **Evaluation/Test Data:** We use a 70/15/15 ratio for train/eval/test datasets
- **Loss Functions:** We experimented with 3 different weighted loss functions–see below.
- **Hyperparameter tuning:** We ran 3 sweeps using WeightsAndBiases to determine the optimal parameters. In addition to loss function, the hyperparameters we tested were number of epochs, learning rate, and batch size.
- **Evaluation:** After selecting a final version through hyperparameter fine-tuning, we use the test set for final metrics. We use standard metrics for sequence labeling tasks, including precision, recall, and F1 score, to assess the model's ability to correctly identify and classify pragmatic markers.

Our implementation leverages PyTorch, the Hugging Face Transformers library [7], and Weights and Biases.

2.2.1 Loss Functions

We experimented with three different loss functions, outlined below using the following standard notation:

- C is the total number of classes
- p_c is the predicted probability for class c
- p_y is the predicted probability of the true class label y
- y_c is a binary indicator (0 or 1) indicating if class label c is the correct classification for the observation
- N is the number of samples in the dataset
- n_c is the number of samples for class c

To select our final model, we use macro average F1 score on the eval dataset from each model's hyperparameter-tuned results.

Cross-Entropy Loss

$$\mathcal{L}_i = -\sum_{c=1}^{C} y_c \log(p_c) = -\log(p_y)$$

This encourages the model to maximize the probability of the correct label by minimizing the negative log probability of the correct class.

Balanced Weighted Cross-Entropy Loss

To address class imbalance, we calculate the weight w_c for each class c as:

$$w_c = \frac{N}{C \times n_c}$$

The weighted cross-entropy loss for a single observation with true class label y is defined as:

$$\mathcal{L}_i = -\sum_{c=1}^{C} w_c \cdot y_c \cdot \log(p_c) = -w_y \cdot \log(p_y)$$

This ensures that the loss contributed by each class is scaled by its corresponding weight, with higher weights assigned to less frequent classes.

Focal Loss

Focal Loss modifies the standard cross-entropy loss to focus more on hard-to-classify examples, particularly useful for addressing class imbalance. It is calculated as:

$$\mathcal{L}_i = -\alpha(1 - p_y)^\gamma \log(p_y)$$

Where:

- α is a scalar factor to tune the weight of the positive class
- γ is a focusing parameter to adjust the rate at which easy examples are down-weighted
- p_y is the predicted probability of the true class label y

3 Results

3.1 PraMIMS: Our Human-Annotated Dataset

Our dataset contains 1161 labeled instances of PPrMs. Of these, 47% are classified as "hedge" or "authority" markers, and 53% are classified as "none." See 1 for summary statistics on the dataset. Approximately 74% of utterances reviewed had a matching PPrM in our dual-person annotation, which means that we keep about 74% of utterances we reviewed in our dataset. Note that this does not mean we aligned on every PPrM in the utterance, so this means that in our final dataset we only show markers where we matched.

Table 1: Summary of Dataset Statistics

Metric	**Value**
Utterances Reviewed	1000
PPrMs Reviewed	2146
Annotated utterances	741
Annotated PPrMs	1,337
Hedge	441
Authority	194
None	702

In a fine-tuning context for a model, we believe that these examples provide a solid basis for learning to distinguish hedge and authority markers from other uses of the same words/phrases.

3.2 Model Performance on PraMIMS

3.3 PragMaBERT Performance

See table 2 for performance of our optimal model. We note that F1 performance on the test data is 0.13 higher than any current public SoTA models we tested (see below).

We use macro averages for F1, Precision, and Recall to ensure that equal weight is given to performance on each label rather than .

3.3.1 Public Models

See 3 for F1 scores by model, category, and 4 for overall performance scores by model. `GPT-4-turbo` performs best of these models on "authority" and "none", while `gemini-1.5-pro` outperforms on "hedge". In the macro average, `GPT-4-Turbo` performs the best.

Table 2: Overall Evaluation Metrics: Weighted Loss Function

	Metric	Training	Evaluation	Test
	Loss	0.385	0.423	0.334
	Accuracy		0.871	0.891
Macro Average	F1		0.865	0.880
	Precision		0.902	0.871
	Recall		0.834	0.884

Table 3: Public Model Performance - F1 Scores by Category on Human Annotated Dataset. Top performers are highlighted in bold

Model	hedge	authority	none
gpt-4-turbo	0.7826	**0.7404**	**0.7357**
gemini-1.5-pro	**0.7870**	0.6350	0.6465
gpt-4o	0.7628	0.6703	0.6034
gemini-1.5-flash	0.7725	0.5877	0.5464
gpt-3.5-turbo	0.4880	0.4880	0.5026

3.4 Analysis

After completing model training, we used PragMaBERT to analyze a sampling of dialogue transcripts from MediaSum. These samples were not included in the training set. We present a few notable findings here but also note that we recognize many additional avenues of research that this model and dataset open up.

3.4.1 General Usage of Hedging and Authority Markers by Speaker Category:

In our analysis, we separated speakers into 5 main categories: US Politicians (Democratic vs. Republican), other government officials (international, or where political affiliation wasn't clearly available), Military, News Media (journalists, reporters), and Other. We exclude "Unknown" speakers from sample for results below.

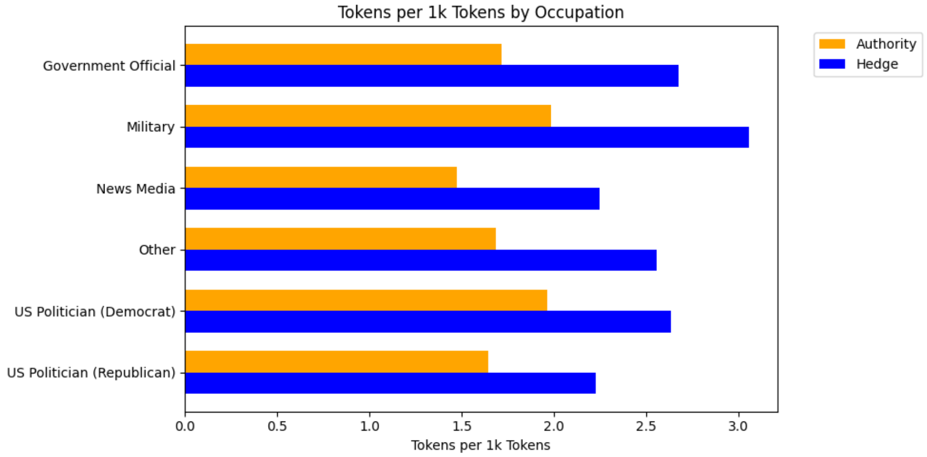

Figure 1: Hedge/Authority Tokens per 1K Tokens by Speaker Occupation

For our general insights, we treat the "other" bucket as a somewhat representative sampling of the general population–these are generally people being interviewed by the news media, so it includes some more formal discussion but also includes many samples of informal, casual conversation (see Appendices for examples).

Table 4: Public Model Performance - Overall Performance on Human Annotated Dataset

Model	F1-Score	Precision	Recall	Accuracy
gpt-4-turbo	**0.7529**	**0.8024**	**0.7546**	**0.7554**
gemini-1.5-pro	0.6955	0.7843	0.4125	0.7046
gpt-4o	0.6788	0.7813	0.7124	0.6836
gemini-1.5-flash	0.6355	0.7527	0.6812	0.6470
gpt-3.5-turbo	0.5560	0.6681	0.6199	0.5639

3.4.2 Politics: PrM Adoption by Speaker

Note that we endeavor to strip any political biases of the authors from this section, but acknowledge that it's impossible to eliminate all bias when selecting highlights from a large corpus of results. As a result, we try to highlight the most clear trends in the data.

Notable Politicians: Table 2 shows hedging adoption by several prominent politicians over the last few years. These are the speakers who were most represented in our random sampling of the MediaSum dataset, so their sampling should somewhat reflect the news coverage they received.

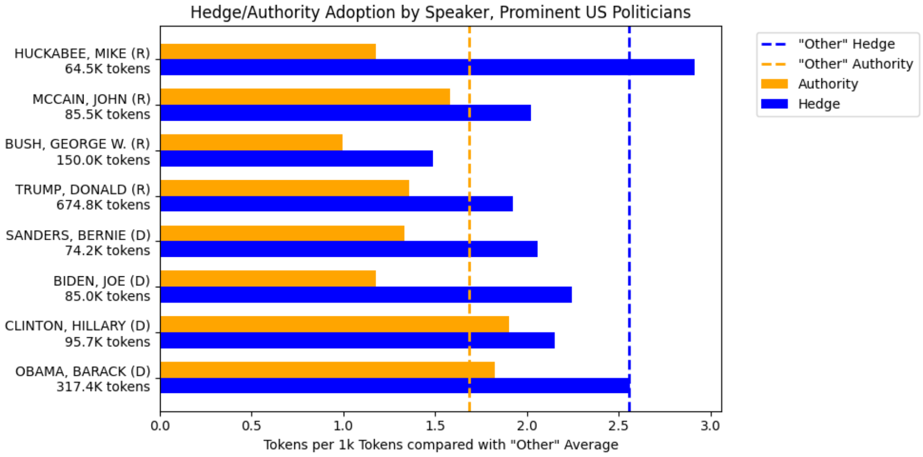

Figure 2: Hedge/Authority Tokens per 1K Tokens for Prominent US Politicians

In our sample set, we observe the following speakers with the most extreme levels of adoption of pragmatic markers.

3.4.3 Pragmatic Adoption in Notable American Speeches

In Fig 4, we also highlight a few speeches from the last 100 years as examples of hedge and authority use in speeches. These are instances where the speaker is known to be using dishonest speech. Notably in all instances we see high adoption of authority markers, often accompanies by above-average hedge markers. We hypothesize that in instances of known deceipt you may see high instances of both, but this is an area for future analysis (see Future Steps).

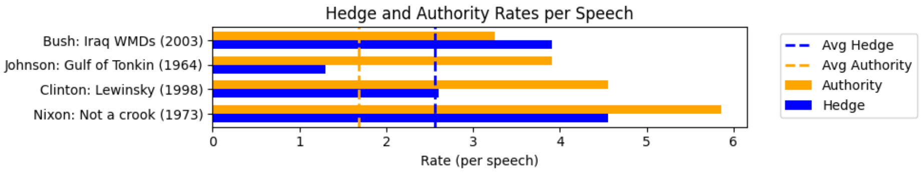

Figure 3: Hedge/Authority Tokens per 1K Tokens from Famous Speeches

3.4.4 Pragmatic Adoption Deviation by Speaker Topic

For a last form of analysis, we highlight an example when a speaker deviates from their standard, which may suggest something different about their beliefs. As with all of the above analyses, this is simply a representative example and many more samples (and code to duplicate) can be found at our github repo: https://github.com/Houdanait/PoliticalTextandAttitudes

Note that this is an illustrative example that we use to simply demonstrate that a speaker's language patterns change under different circumstances. It remains for future work to draw conclusive relationships between language pattern deviations and the underlying beliefs/psychology of the speaker.

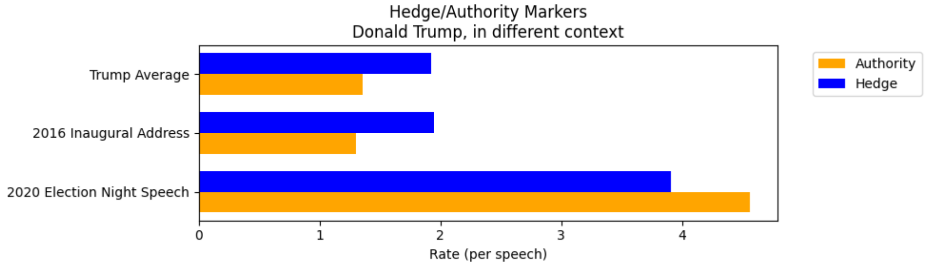

Figure 4: Hedge/Authority Tokens per 1K Tokens from Famous Speeches

4 Ethical Considerations

This project raises a couple key ethical challenges that needs consideration. First, the technology we are developing to automatically detect pragmatic markers could potentially be misused for harmful political purposes. For example, bad actors could use such a system to hone manipulative language and make their communication even more persuasive and misleading. There's an inherent dual-use risk in building powerful language understanding technology.

Additionally, while our intent is to ultimately promote linguistic transparency and combat subtle manipulation, the line between persuasion and manipulation is arguably subjective. We must be cautious about making value judgements on how language "should" be used, as all communication contains some bias and subjectivity. Labeling a politician's speech as manipulative risks its own form of manipulation.

To mitigate these concerns, responsible development and deployment of any pragmatic marker detection system is essential. Access to the technology should be carefully controlled and monitored, especially as it pertains to use in the political domain. Clear communication about the limitations and potential flaws of the AI system is also important to prevent over-reliance or misplaced trust in its outputs. Users should understand that it is ultimately just one analytical tool to augment but not replace human judgement.

Finally, if this technology progresses to a point of analyzing political discourse at a large scale, we believe it's critical that it be used in an nonpartisan way to examine language usage across the political spectrum. Cherry-picking analysis to disparage political opponents would be inappropriate and unethical. Responsible stewardship focused on increasing linguistic understanding rather than point-scoring should be the guiding principle.

While pragmatic marker detection holds promise for promoting clearer communication, we must remain vigilant to the ethical pitfalls as the technology advances and deploy it with great care. Establishing clear ethical guidelines and oversight will be essential for mitigating risks and ensuring it provides a societal benefit.

5 Future Work

With the introduction of a new dataset and model, there are many opportunities for future work in this area. We highlight a few areas of most interest:

- **Further Annotation:** If we are able to train a larger corpus of human-annotated samples, we believe model performance could improve enough to use a model to annotation more samples–we set a threshold F1 score of 95% for a model to be sufficiently strong to use for dataset creation.
- **Real-time analysis of political debates, etc.:** With a US presidential election this year, real-time analysis could be useful for media outlets and as educational tools that aim to provide live linguistic analysis.
- **Expansion of Pragmatic Marker Categories:** Broaden the types of pragmatic markers studied beyond hedging and authority markers to include other markers. This can provide a more comprehensive view of linguistic strategies in political rhetoric.
- **Temporal Analysis of Pragmatic Marker Usage:** Investigate changes in the use of pragmatic markers over time, particularly through different political eras or leadership changes. There have been notable shifts in political discourse over the last two decades, and this tool could provide insights into the effect of those shifts.
- **Public Policy Impact Studies:** Collaborate with political scientists and policymakers to study the impact of pragmatic marker usage on public opinion and policy making. This could include analyzing how different markers influence voting behavior or public trust.

References

[1] Laurel Brinton. *Pathways in the Development of Pragmatic Markers in English*, chapter 13, pages 306–334. John Wiley Sons, Ltd, 2006.

[2] Laurel J. Brinton. *The Evolution of Pragmatic Markers in English: Pathways of Change*. Cambridge University Press, 2017.

[3] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, 2019.

[4] Peter Furko. Manipulative uses of pragmatic markers in political discourse. *Palgrave Commun*, 2017.

[5] Paul J. Hopper and Elizabeth Closs Traugott. *Grammaticalization*. Cambridge University Press, Cambridge, 2nd edition, 2003.

[6] John H. McWhorter. *Words on the Move: Why English Won't - and Can't - Sit Still (Like, Literally)*. Henry Holt and Co., New York, 2016.

[7] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*, 2019.

[8] Chenguang Zhu, Yang Liu, Jie Mei, and Michael Zeng. Mediasum: A large-scale media interview dataset for dialogue summarization. *arXiv preprint arXiv:2103.06410*, 2021.

6 Appendices

Matt to add examples of data source, label, and output from model classification. Move one of the larger, less important tables to appendix

6.1 Public Model Prompt

We use the following prompt for the public model response generation:

```
system_prompt = """You are a linguist analyzing dialogue
transcripts for examples of "hedging" or "authority" pragmatic
markers. The terms are provided in a list, and in the
transcript the terms are capitalized within <> (e.g.
<ABSOLUTELY>). Provide an answer for every single term in the
list of "terms" you are given. Return your answer in the
following JSON format:

---BEGIN SAMPLE INPUT---
terms: ["believe", "maybe", "certainly", "best"]
transcript:
Speaker 1: "In the run for the presidency."
Speaker 2: "I knew about one incident. Understand the whole
time that he ran for office, I knew that he had had one liaison.
It still -- it still tore me up, I mean, personally tore me up.
Did I think that one liaison would disqualify him to be the
president? You know, we've had great presidents who I would
hope one liaison would not have -- have stopped from serving us.
That's what I believed. And I believed that until, golly,
<MAYBE> long after it made any sense to but, <CERTAINLY> long
after -- I mean, long after he was out of the race. And so
sometimes I had to, you know, bite my tongue. I talked a lot
about his policies, which I still <BELIEVE> were the <BEST>
policies and set the standard for the other candidates on a
lot of issues -- health care being one of them, but environment
and poverty and corporate interference with government. And I
really believed that that I could talk about those things and
mean every word that I was saying, and have him as an advocate
for those issues and meaning that as well."
---END SAMPLE INPUT---

---BEGIN SAMPLE RESPONSE---
    {{
        "believe": "hedge",
        "maybe": "hedge",
        "certainly": "authority",
        "best": "none"
    }}
---END SAMPLE RESPONSE---"""
```

6.2 Labeling of Additional statements

We use the model to classify a broad list of potential PrMs. An utterance is classified as an authority or hedge depending on the number of authority/hedge tokens in the text.

6.3 Inference

Aside from simple label inference, the model can also suggest new, unidentified PrMs. In a small experiment, the model suggested new hedging PPrMs: “maybe” and “might have”.