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, model, 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. 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

Team Contributions: Both: Manual Annotations, Writeup. Matt: Analysis, Customized Loss Functions and Hyperparameter Tuning. Houda: PragMaBERT setup, ethics and policy implications

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 [5].

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 [2, 6]. This evolution can result in semantic overlap and ambiguity, making the automated detection and interpretation of PrMs a challenging task [7].

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 [5]. 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. **Pragmatic Marker Dataset:** 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 Related Work

Pragmatic markers (PrMs) have been studied extensively in the context of linguistics and communication. These markers play a crucial role in structuring discourse and signaling speaker attitudes, intentions, or emotions. Our work is particularly inspired by the studies of Furko [5], who explored the impact of PrMs on manipulation in political discourse. Building upon these qualitative analyses, our approach leverages advanced NLP techniques to quantitatively assess PrMs' usage across a large dataset.

Previous work in the field has predominantly focused on qualitative analysis, often within limited linguistic and cultural contexts. For example, Brinton [2] and Traugott [6] discuss the grammaticalization of PrMs, which involves the evolution of lexical items into grammatical markers. These studies highlight the semantic and functional shifts that PrMs undergo, making their automated detection and interpretation a challenging task.

Furthermore, the work by McWhorter [7] examines the ambiguity and overlap in the meanings and uses of PrMs, underscoring the complexity of determining their presence and function in spoken discourse. These insights are crucial for our development of PragMaBERT, where understanding the subtleties of PrMs can significantly enhance the model's accuracy and applicability.

In the political realm, the strategic use of PrMs such as hedging and authority markers can significantly influence public perception, as noted by Furko [5]. Politicians use these markers to navigate the delicate balance of expressing certainty and maintaining flexibility, often to manipulate public opinion. Our research aims to extend these findings by providing a computational method to analyze and interpret the use of PrMs in a broader and more systematic manner than previously possible.

Our dataset, derived from the MediaSum database [11], includes a wide range of political speeches annotated for pragmatic markers. This allows us to train PragMaBERT not only to detect these markers but also to understand their contextual usage, paving the way for more nuanced analyses of political rhetoric.

The existing literature thus provides a solid foundation for our study, highlighting both the potential and the challenges of working with PrMs. By integrating these insights with state-of-the-art machine

learning models, we aim to advance the field of NLP and contribute to the development of tools that can enhance transparency and accountability in political discourse.

3 Approach

3.1 Dataset Creation

We use MediaSum [11] 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 [3]. We supplemented with PPrMs from other authors ([6], [7]). 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 more).

3.1.1 Public Model Evaluation

We used a standardized prompt to generate 1-shot JSON responses from several leading models. 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.

3.2 Model Development

To automatically detect and classify pragmatic markers in text, we fine-tune the BERT (Bidirectional Encoder Representations from Transformers) language model [4]. We also make use of the Hugging Face Transformers library [10], and Weights and Biases (www.weightandbiases.com) for hyperparameter tuning.

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:** [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
- **Hyperparameter tuning:** We ran 3 sweeps using WeightsAndBiases to determine the optimal parameters, tuning on 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.

3.2.1 Loss Functions

In addition to standard Cross Entropy Loss, we experimented with two additional 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.

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

4 Experiments and Results

4.1 Human-Annotated Dataset

Our dataset contains 1,337 labeled instances of PPrMs. Of these, 47% are classified as "hedge" or "authority" markers, and 53% are classified as "none." See Table 3 in the appendices for more detailed summary statistics. 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.

Our model performance suggests that these examples provide a solid basis for learning to distinguish hedge and authority markers from other uses of the same words/phrases.

4.2 Model Performance on PraMIMS

4.3 PragMaBERT Performance

After hyperparameter tuning, we selected a final model based on training/eval metrics. See table 1 for performance, including test data. We note that F1 performance on the test data is 0.13 higher than any current public SoTA models we tested (see below).

Table 1: Overall Evaluation Metrics: Weighted Loss Function

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

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

4.3.1 Public Models

See Table 2 for overall performance scores by model, and note additional metrics in the appendix. `GPT-4-Turbo` outperforms other models on all summary metrics, but we note that there is still material room for improvement from current models.

Table 2: 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

4.4 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. We present these charts and insights as initial suggestions for the types of insights that can result from adoption of our dataset and model, and we believe that there are far more insights to come in this field (see Future Work). For our initial analysis, we applied the model to transcripts from the MediaSum database.

Caveats

- Pragmatic markers are not direct, conclusive indicators of deceit; they are common linguistic tools used for various communicative purposes. Analysis of these markers should be part of a balanced effort–their presence alone does not confirm manipulative or deceiptful intent.
- In addition we note the existence of personal bias in political discussion, even within this paper. We have endeavored to remove political bias from our choice of topics, speakers, and examples, but we recognize that some bias is inherent in any such decision.

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

We first explore differences in speech patterns across occupations, so Figure 1, we split out the primary occupations in the dataset.

For a benchmark for general behavior, it wasn't appropriate to take take an average from the population sample–News Media is so heavily represented in the database that the overall averages are very close to the News Media. Instead, we treat the "Other" occupation bucket as a somewhat representative sampling of the general population–these are people being interviewed by the news media, so it includes some more formal discussion but also includes many samples of informal, casual conversation.

We highlight a couple of insights:

- As might be expected, News Media has perhaps the most measured dialogue, with the lowest adoption of both hedges and authority markers
- Military has the highest use of both "authority" and "hedge" markers
- Average adoption for "Other" category falls between average adoption for Republicans and Democrats on both markers

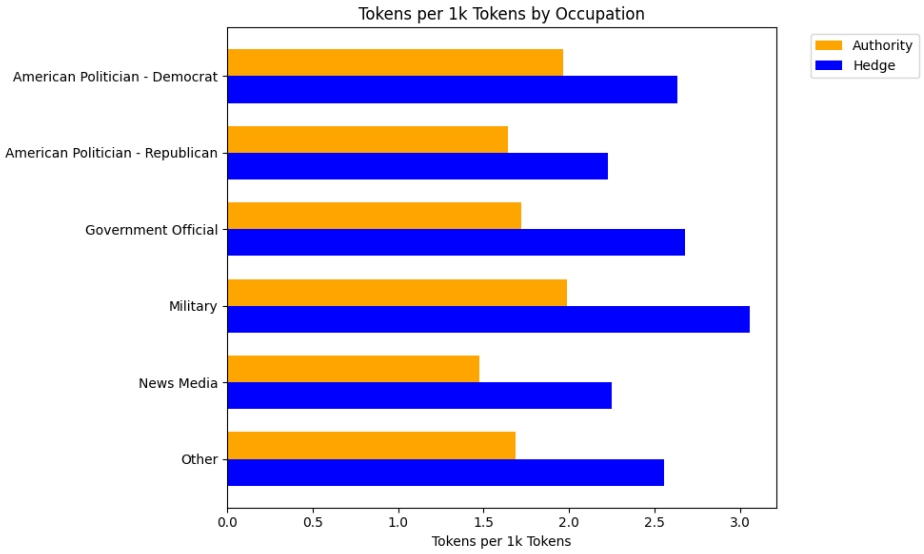

Figure 1: Hedge/Authority Tokens per 1K Tokens by Speaker Occupation

4.4.2 Politics: PrM Adoption by Speaker

We see much more variation when looking at Inter-Speaker Patterns–Figure 2 shows hedging/authority adoption by the politicians who were most represented in our sampling.

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

4.4.3 PrM Deviations in Speaking Style

For a last form of analysis, we highlight samples from the previous two US Presidents that show deviation from their typical speaking style under certain circumstances. For President Trump, we analyze his speech on Election Night 2020 ([9]), in which he first made the spurious claim that the election had been stolen. In the case of President Obama, we analyze his speech from June 2009 in which he said “"If you like the plan you have, you can keep it. If you like the doctor you have, you can keep your doctor, too. The only change you’ll see are falling costs as our reforms take hold” ([8], which was voted Politifact’s “Lie of the Year” in 2009 ([1]). In both cases, we observe a similar deviation from their typical speech patterns. Again we note that this is illustrative and not conclusive, but we believe there is material opportunity for further analysis in this space.

5 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

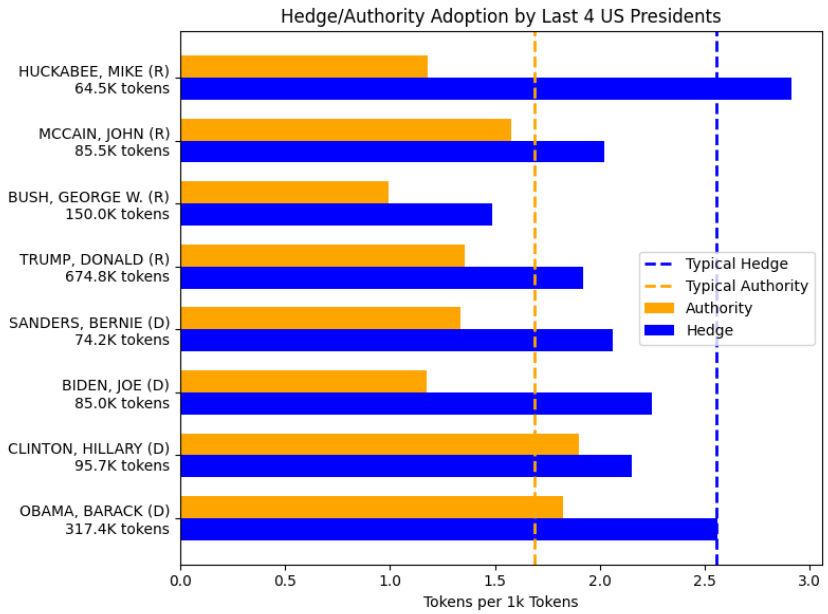

Figure 2: Hedge/Authority Tokens per 1K Tokens for Prominent US Politicians–Typical refers to the "Other" category

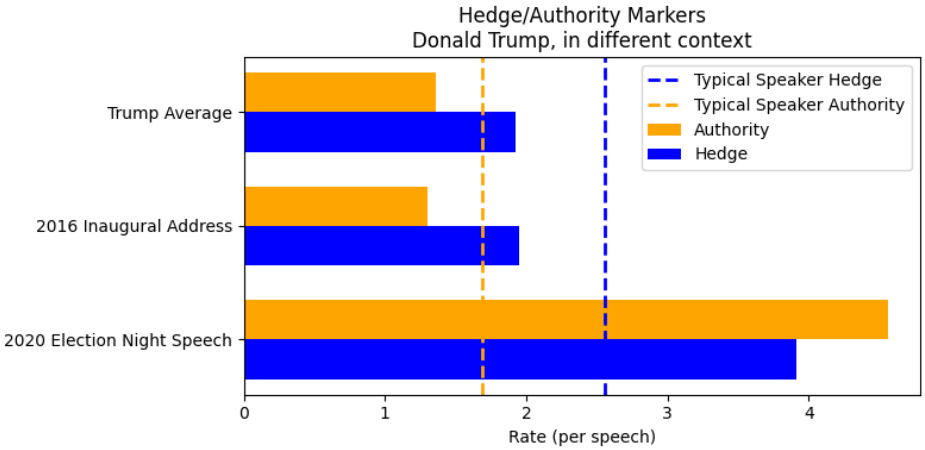

Figure 3: Hedge/Authority Tokens per 1K Tokens from Famous Speeches

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

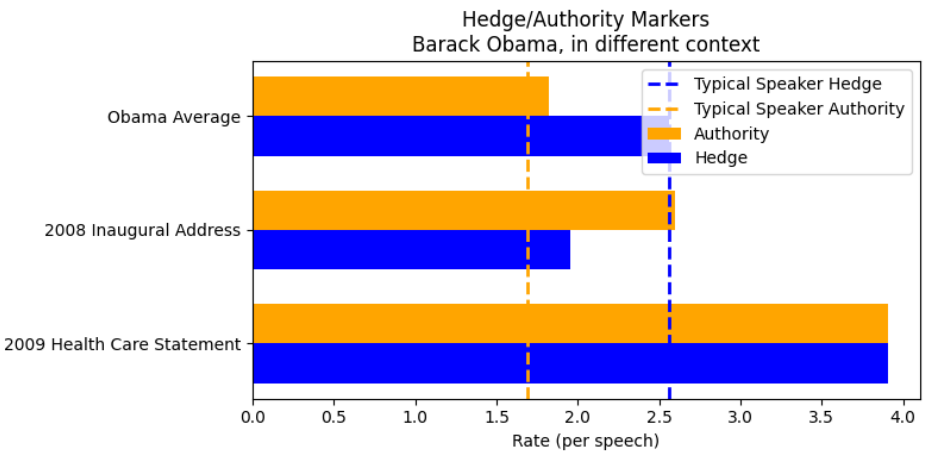

Figure 4: Hedge/Authority Tokens per 1K Tokens from Famous Speeches

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.

6 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] Obama's 'you can keep it' promise is 'lie of the year'. NPR, 2013. Accessed: 2024-06-08.

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

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

[4] 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.

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

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

[7] 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.

[8] Barack Obama. Weekly address: President obama outlines goals for health care reform. The White House, 2009. Accessed: 2024-06-08.

[9] Donald Trump. Donald trump 2020 election night speech transcript. Rev, 2020. Accessed: 2024-06-08.

[10] 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.

[11] 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.

7 Appendices

7.1 Additional Metrics

Table 3: 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

Table 4: 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

See https://github.com/Houdanait/PoliticalTextandAttitudes for additional details on prompts and to download the model

7.2 Sample Human Annotation

Annotation was completed by looking at a list of PPrMs in the text and labeling the first instance of each of those PPrMs. The previous statement was included for additional context.


```
{
    "transcript_id": "CNN-145497",
    "matched_terms": {
        "definitely": [
            "authority"
        ],
        "maybe": [
            "hedge"
        ]
    },
    "previous_statement": "All the other currencies that the dollar
        is trading with. Where is the bet going against them, do you
        think?",
    "statement": "Well, again, the U.S. dollar is <DEFINITELY> lower.
        I mean, people are talking about parity against the
        Australian dollar. We are expecting, <MAYBE>, the Australians
        see an increase in interest rates again on Monday."
}
```


7.3 Grading Rubric

This was written somewhat informally between the two of us, but we leave it here for reference:

Constitution for how to evaluate pragmatic markers

Run through all the examples. For each statement, words which belong to a pre-established list of pragmatic markers are identified. However, their use within the context of the sentence, may not be related to an authority nor hedging instance. Thus, evaluate whether the use of these words correspond to authority, hedge, or none.

First, note that some words are identified because they belong to fragments of other words. For example in the sentence “There’s maybe a few people in every crowd”, the word <may> is identified, but it actually is only a part of the word “maybe”. In this case, do not consider this instance, and label the word <MAY> as “ungraded”.

Second, if a word appears multiple time in a sequence, for example in the sentence “As you know, Catherine and Prince William have been going out together for quite a number of years, which is great for us, because we have got to know William very well”, the word <KNOW> is identified, but appears twice in the sentence. In those cases, please evaluate only the first instance. For example, in this case the first instance is “As you know, Catherine and Prince William” and in that case correspond to an instance of authority, and thus <KNOW> should be labeled as “authority”.

As a reminder, here is how we define “hedging”: A hedge softens the strength of a statement, showing uncertainty, willingness to compromise, etc. We also include instances of overly vague or obfuscating statements (For example, “We are exploring all options”, <ALL> here is vague and refers to an instance of hedging, so we label <ALL> as “hedging”.) Similarly, “authority” is defined as the below: Authority PPrMs are intended to assert the authority of the speaker, show a common understanding with an audience, and generally get an audience to agree with the speaker.

Examples of Mis-Matched Grades and how we reconcile them now: Statement: “So fascinating to see the work going on all <AROUND> you” This is not an example of authority or hedging Beginning a statement with “Well” Example Statement: <WELL>, I think it is illustrated by this memory Beginning a statement with "well" is unclear on its own and may be “none” or “hedge”, but coupling it with “I think” shows that “well” is part of a hedging statement Example Statement: "<WELL>, I can’t divulge information that is classified” This instance is categorized as “none” because there isn’t clear context that this is a hedge and is likely a discourse marker