Attitude Identification Through Parameter-Efficient Fine-Tuning

Mariia Anisimova[1], **Gabriella Lapesa**[2], **Šárka Zikánová**[3]
[1]Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Praha, Czech Republic
[2]GESIS – Leibniz Institute for the Social Sciences
Gereonstraße 34-36, 50670 Köln, Germany
[3]Charles University, Faculty of Mathematics and Physics, Institute of Formal and Applied Linguistics
Malostranské náměstí 25, 118 00 Praha, Czech Republic
anisimova@ufal.mff.cuni.cz, Gabriella.Lapesa@gesis.org,
zikanova@ufal.mff.cuni.cz

Abstract

We investigate automatic attitude detection in UN Security Council speeches using adapters. Following Martin and White's Appraisal Theory, we identify three types of evaluative language: affect (emotional responses such as hope or concern), judgement (ethical evaluations of behavior), and appreciation (valuations of objects or situations). Training only 0.95% of BERT-large's parameters, adapters achieve F1 scores ranging from 0.76 (affect) to 0.46 (appreciation), approaching full fine-tuning performance while enabling rapid task-specific experimentation. Differences in observed evaluation metrics mirror the pattern of the human inter-annotator agreement. This correlation suggests that computational difficulty reflects genuine linguistic ambiguity. Affect benefits from conventionalized diplomatic expressions, while appreciation faces context-dependent evaluation and severe class imbalance. Analysis demonstrates that evaluative intensity varies systematically across diplomatic contexts, with implications for corpus design in specialized discourse analysis.

Keywords: Appraisal Theory, attitude detection, diplomatic discourse, parameter-efficient fine-tuning, BERT

1. Introduction

Evaluative language plays a critical role in discussions of international military conflicts by the UN Security Council. The way states express subjectivity through evaluative language reveals their stance, alliances, and strategic intentions. However, computational approaches to fine-grained attitude detection in this domain remain limited.

Detecting attitudes in diplomatic speeches presents several challenges. Diplomatic language combines conventionalized expressions of attitude (so-called diplomatic cliches (Anisimova and Zikánová, 2024)) with subtle evaluations which may take the form of metaphors, comparisons, etc. Additionally, manually annotated corpora for specialized domains are typically small, making full model fine-tuning prone to overfitting.

We apply adapters (Pfeiffer et al., 2020) to the diplomatic attitude detection task using data from CoDipA 1.0 (Anisimova and Zikánová, 2024), a corpus of 100 UN Security Council speeches containing 1938 manually annotated attitudes, using Martin and White's Appraisal Theory (2005).

Adapters offer two key advantages for specialized discourse analysis: they train only a small fraction of model parameters (0.95% in our case), reducing overfitting risk on small specialized corpora, and enabling rapid task-specific experimentation with different training strategies without retraining the entire model. We train four task-specific adapters for affect (emotional responses), judgement (ethical evaluations of behavior), appreciation (valuations of objects or situations), and attitude in general. Our analysis demonstrates that performance correlates with both linguistic characteristics of the diplomatic discourse and distributional properties of the corpus, with implications for corpus design in specialized discourse analysis.[1]

2. Related Work

Martin and White's (Martin and White, 2005) Appraisal Theory has been applied to various domains, including educational texts (Lam and Crosthwaite, 2018; Sheyholislami and Hall, 2013), journalism (White, 2006), and translation studies (Tajvidi and Arjani, 2017). Computational work on Appraisal began with Taboada and Grieve (2004), who developed rule-based methods for appraisal detection. Later, Kolhatkar et al. (2020) have created a corpus of movie, book, and consumer product reviews annotated with Appraisal theory. Read et al. (2007) assessed the difficulty of automatic Appraisal analysis through inter-annotator agreement studies, establishing evaluation approaches for computational approaches to the framework.

Recent computational work on the UN Security Council discourse includes analysis of conflict expression (Zaczynska et al., 2024) and identifica-

[1]https://github.com/mariiaanisimova/attitude-adapters-bert-large

Proceedings of The 3rd Workshop on Natural Language Processing for Political Sciences (PoliticalNLP 2026) @ LREC 2026, pages 204–209
16 May 2026, Palma, Mallorca (Spain).

tion of argumentation structures (Poiaganova and Stede, 2025). These studies demonstrate a growing interest in computational approaches to diplomatic language, although they focus on linguistic phenomena different from evaluative language.

While Appraisal Theory has been applied computationally to various domains and diplomatic discourse has received increasing computational attention, no prior work combines these strands: applying neural methods to Appraisal-based attitude detection in diplomatic language. Our work addresses this gap by applying adapters to diplomatic attitude detection and analyzing how performance patterns relate to linguistic characteristics and distributional properties of specialized corpora.

3. Corpus and Annotation

3.1. Appraisal Theory Framework

Affect refers to emotional responses to people, events, or situations. In diplomatic discourse, affect appears in conventionalized expressions such as "express profound condolences", "we are particularly *concerned*", or "we fervently *hope*". These expressions follow predictable patterns and typically cluster around specific verbs.

Judgement evaluates behavior according to normative principles, such as ethics, legality, or social norms. Diplomatic judgement appears in two main forms. Firstly, through nominalized constructions that embed evaluation in content words: "racist *colonizers*," "Thinking that war is one of the options before the Council is in itself proof of our collective *failure*". Secondly, through evaluative verbs expressing endorsement or criticism, as in "we fully *endorse* that approach", "the League *failed* to create actions from its words", or "we will continue to *pursue* every opportunity for peaceful settlement". Both patterns contrast with explicit affect markers (e.g., "we condemn," "we deplore"), making judgement more varied and context-dependent.

Appreciation evaluates the qualities of objects, situations, or phenomena rather than human behavior. In diplomatic texts, appreciation tends to be brief and context-dependent: "a historically significant change", "strategic and historic importance", or "the occupying Power". The evaluative force of adjectives depends entirely on context.

These categories differ fundamentally in their linguistic realization, which, as our results show, affects both the difficulty of human annotation and the computational detection performance.

3.2. CoDipA 1.0 and data preprocessing

The Corpus of Diplomatic Attitudes (CoDipA) (Anisimova and Zikánová, 2024) contains 100 UN Security Council speeches (105 592 tokens) from debates on five international military conflicts (Palestinian, Yugoslav, Iraqi, Ukrainian, and Georgian) spanning from 1995 to 2020. The corpus was built on the Schoenfeld et al. (2019) dataset and contains 1938 annotated attitude instances at the span level. For the presented classification task, we transformed span-level annotations into sentence-level binary labels that indicate whether each sentence contains a given attitude type.

As reported by Anisimova and Zikánová (2024), judgement is the most frequent attitude in the corpus, reflecting the normative nature of diplomatic discourse. Appreciation attitudes are the shortest (averaging 1.8 tokens), and often consist of single evaluative adjectives. Inter-annotator agreement (Cohen's κ) ranges from 0.31 to 0.44, indicating the inherent difficulty of attitude identification.

3.3. Splitting the data

The data was split on a document-level to prevent data leakage, when one text could potentially have been split between train and test sets. Therefore, 30 particular speeches are reserved as a forced test set to evaluate model performance.[2] The remaining 70 speeches are randomly split in a 90/10 fashion into training (2216 sentences) and validation (219 sentences) sets.

As we observed severe train-test distribution mismatch with this approach for the category of appreciation (train set resulted in having 7.6% positive instances while the test set - 31% positive instances), we employed stratified document-level splitting, where documents were grouped into bins by appreciation density (low/medium/high), and train/validation splits maintained proportional representation from each bin. This reduced, but did not fully eliminate the distribution gap, reflecting the forced test set's focus on conflict contexts where evaluative language is more prevalent.

Table 1 shows the resulting distributions.

Type	Train	Val	Test	Method
Attitude	37%	31%	60%	Standard
Affect	18%	18%	19%	Standard
Judgement	26%	22%	36%	Standard
Appreciation	8%	5%	31%	Stratified

Table 1: Representation of the data splits for attitude and its three types

[2] speech ids 470-499, and 1582-1583

4. Method

4.1. Model Architecture

We use BERT-large-cased (Devlin et al., 2019) (336M parameters) as the base model with adapters (Pfeiffer et al., 2020). Adapters are lightweight modules inserted after each transformer layer, while the base BERT model remains frozen during training.

We train four binary classifiers, one per attitude or its type, with independent adapters and classification heads. This setup allows for task-specific training strategies without retraining the entire model.

4.2. Training Configuration

Training employs early stopping (patience 3-4 epochs) with F1-score as the selection metric. We employ task-specific training strategies based on the distributional characteristics observed in preliminary experiments. Affect, showing balanced distributions (18% positive across all splits), uses standard configurations without oversampling. Attitude and judgement, with moderate imbalance, employ light oversampling (exponents 1.1-1.2) to improve minority class representation. Appreciation, exhibiting severe imbalance (7.6% training vs 31% test), requires comprehensive intervention: stratified splitting, aggressive oversampling (exponent 1.5), reduced batch size (8), and extended training (15 epochs) to provide sufficient learning signal for the minority class. This progressive approach demonstrates that intervention intensity should match task difficulty rather than applying uniform strategies across all categories.

Table 2 shows the final configurations for each task.

Task	LR	Batch	Epochs	Os
Affect	1e-4	16	10	None
Attitude	5e-5	16	10	1.1
Judgement	5e-5	16	10	1.2
Appreciation	3e-5	8	15	1.5

Table 2: Training configurations per task. Oversampling (Os) column represents exponent for weighted random sampling, where None corresponds to standard AdapterTrainer without oversampling

For attitude and judgement, we also performed threshold optimization on the validation set, by testing thresholds from 0.30 to 0.70 in 0.05 increments. For each threshold, we calculated the F1-score and selected the threshold with the highest metric. The selected threshold is then applied to the test set. This procedure allows to assess whether validation-based optimization transfers to test data with different distributions.

5. Results

The evaluation metrics selected to evaluate adapter performance include precision, recall, F1-score, and accuracy calculated on the test set. The F1-score is the primary metric as it balances precision and recall, which is important for imbalanced classification tasks.

Table 3 shows the evaluation results on the test set for all labels. Performance ranges from F1=0.76 for affect to F1 = 0.46 for appreciation.

Label	F1	Prec	Rec	Acc
Attitude	0.65	0.76	0.58	63.5%
Affect	0.76	0.78	0.74	91.4%
Judgement	0.57	0.53	0.61	66.6%
Appreciation	0.46	0.49	0.43	68.3%

Table 3: Test set results. Prec corresponds to precision, Rec - to recall, Acc - to accuracy

The most general task of *attitude* detection achieved F1=0.65 despite substantial distribution mismatch. The model shows high precision (0.76) but lower recall (0.58), indicating conservative predictions.

Affect achieved the highest F1-score (0.76), with balanced precision (0.78) and recall (0.74). In diplomatic speeches of the UNSC, affect expressions typically use explicit verbal markers like "hope," "welcome", "join", "grieve" and "regret".

Judgement shows F1=0.57 with balanced precision (0.53) and recall (0.61). Judgement attitudes are often embedded in nominalized constructions, vague formulations, and modality-based urges and often lack explicit evaluative markers, making them harder to identify.

Appreciation identification presented the greatest challenge (F1=0.46). This task combines severe class imbalance with linguistic characteristics that further complicate label detection, as appreciation attitudes are brief and often consist of implicit evaluations requiring contextual interpretation.

6. Discussion

The results demonstrate a clear performance hierarchy across attitude types that persists despite different training strategies: affect achieved the highest performance despite using only standard configurations, while appreciation achieved the lowest performance despite comprehensive interventions (stratified splitting, aggressive oversampling, extended training). This suggests that performance differences stem from fundamental task characteristics, such as linguistic characteristics of the diplomatic speeches and distributional properties of the chosen language resource, rather than inadequate training procedures.

Performance differences also correlate with the extent of interventions required. Although affect did not require special improvements, attitude and judgement benefited from light oversampling, while appreciation required data-focused strategies. This pattern suggests that some categories present challenges that cannot be fully addressed through standard fine-tuning approaches alone, pointing to the need for either larger annotated corpora or alternative modeling strategies for severely imbalanced specialized discourse analysis.

We examine these patterns in detail below, first analyzing linguistic factors, then distributional effects, and finally discussing implications for corpus design in specialized domains.

6.1. Linguistic Factors

A notable correlation was observed between the computational performance of our adapter and human annotation difficulty. Affect achieves both the highest F1 score (0.76) and the highest inter-annotator agreement (κ=0.44), while appreciation shows the lowest F1 (0.46) and lower agreement (κ=0.32) (Anisimova and Zikánová, 2024). This correlation suggests that computational difficulty reflects genuine linguistic ambiguity rather than modeling limitations alone.

We assume that detection of affect benefits from conventionalized expressions that provide clear lexical signals. Diplomatic speeches regularly use phrases like "we welcome," "we regret," and "we are concerned," which function as an inscribed affect. Such expressions follow predictable patterns, lexically clustering around specific verbs, narratively occurring in specific parts of a speech, such as greetings, initial expression of gratitude, etc.

Judgement attitudes show more variation. These evaluations appear in nominalized ("violation of international law") and verbalized constructions, and through implicit verbal framing (describing an action without explicit judgement markers, as in *Some continue to confuse the proposed review list with a denial list*). To identify judgement, the model often must recognize the evaluation embedded in the content words rather than relying on explicit markers, contributing to moderate adapter performance and annotation difficulty.

Appreciation poses the greatest challenge for several reasons: the evaluative nature of "complex" or "significant" depends entirely on what is being evaluated and in what context. This characteristic manifests itself in both lower human agreement and lower computational performance. Additionally, appreciation attitudes are brief (averaging 1.8 tokens), often consisting of single adjectives, which limits contextual information.

6.2. Distribution Effects

The distribution differences between the training and test sets significantly affect the training results, particularly in terms of appreciation and attitude. The forced test set, which consists of specific speeches, contains substantially more attitudes than the other speeches in the corpus used for training. This reflects a genuine characteristic of diplomatic discourse, namely that evaluative language varies systematically across contexts: diplomats may change their preferred means of attitude expression based on the topic, as well as within the time frame of the debates.

For appreciation, we attempted to mitigate this through stratified document-level splitting by attitude density. This reduced but did not eliminate the mismatch. The fundamental issue could not have been fully addressed through splitting strategies alone, which reflects in general on highly specialized and unbalanced corpora of smaller sizes.

This finding reveals a methodological tension in the design of specialized corpora. Selecting test data to represent specific contexts of interest (e.g., multiple international military conflicts) naturally creates distribution differences from general training samples. Stratification can reduce, but not eliminate these differences. This suggests that for specialized discourse analysis, researchers must choose between sets that match training distributions, or targeted test sets designed to avoid context leakage between the documents at the cost of distribution mismatch.

7. Conclusion

This paper presented results of applying adapters to diplomatic attitude detection. Using CoDipA 1.0, a corpus of 100 annotated UN Security Council speeches, we trained task-specific adapters for affect, judgement, appreciation, and attitude detection in general. Training only 0.95% of model parameters, adapters achieved F1 scores ranging from 0.76 (affect) to 0.46 (appreciation).

The results of our study show that the performance of the said adapters is correlated both with previously observed linguistic characteristics of the separate attitude types and with the data distribution. Affect detection benefits from conventionalized expressions and balanced data. Appreciation faces challenges from severe class imbalance, linguistic brevity, and implicit evaluation. Threshold optimization experiments demonstrate that train-test distribution mismatch affects model calibration, with implications for corpus design in specialized domains.

In our future work, the presented adapters will be applied to a larger set of diplomatic speeches

from Schoenfeld et al. (2019) to analyze how attitude usage changes within the timelines of the said conflicts.

8. Acknowledgements

The research presented in this paper was supported by a GESIS Junior Visiting Researcher Grant for a fully funded research visit to GESIS in Cologne, Germany from 12.5.2025 to 20.6.2025, and partially supported by HVar project (Disagreement in corpus annotation and variation of human understanding of text, GAČR project 24-11132S). The CoDipA corpus used in the experiments is hosted by LINDAT/CLARIAH-CZ.

8.1. Ethical considerations

This study analyzes publicly available UN Security Council speeches from the CoDipA UNSC 1.0 corpus. The data consists of official diplomatic statements delivered by state representatives in public sessions; therefore, no personal or private data are processed.

Nevertheless, modeling evaluative language in diplomatic discourse raises broader ethical considerations. Automatic attitude detection only captures linguistic patterns within a specific theoretical framework, and the results of such analysis should not be misinterpreted as reflecting clear political intent or sincerity. Results should therefore not be used to draw normative judgments about states or political actors.

Finally, models trained on limited historical conflicts (1995-2020) may not generalize to other geopolitical contexts. We encourage cautious, context-aware application of such models in political analysis and public discourse research.

8.2. Limitations

Our approach has several limitations.

Firstly, the corpus with which we work is limited to five conflicts from 1995-2020, which may not generalize to other diplomatic contexts.

Secondly, the annotation of attitudes often requires subjective interpretation, as evidenced by moderate inter-annotator agreement (κ=0.31-0.44).

Thirdly, the forced test set design creates distribution differences between training and test data, complicating model calibration and evaluation.

And finally, to prove the efficiency of using adapters for attitude analysis in diplomatic discourse, the experiment setup would benefit from comparing against a fully fine-tuned language model.

9. Bibliographical References

Mariia Anisimova and Šárka Zikánová. 2024. Attitudes in diplomatic speeches: Introducing the CoDipA UNSC 1.0. In *Proceedings of the 20th Joint ACL - ISO Workshop on Interoperable Semantic Annotation @ LREC-COLING 2024*, pages 17–26, Torino, Italia. ELRA and ICCL.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Minneapolis, Minnesota. Association for Computational Linguistics.

Varada Kolhatkar, Han Wu, Luca Cavasso, Emilie Francis, Kavan Shukla, and Maite Taboada. 2020. The SFU opinion and comments corpus: A corpus for the analysis of online news comments. *Corpus Pragmatics*, 4:155–190.

Suet Ling Lam and Peter Crosthwaite. 2018. Appraisal resources in l1 and l2 argumentative essays: A contrastive learner corpus-informed study of evaluative stance. *Journal of Corpora and Discourse Studies*.

James R. Martin and Peter R. R. White. 2005. *The Language of Evaluation: Appraisal in English*. Springer.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. MAD-X: An Adapter-Based Framework for Multi-Task Cross-Lingual Transfer. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7654–7673, Online. Association for Computational Linguistics.

Maria Poiaganova and Manfred Stede. 2025. From debates to diplomacy: Argument mining across political registers. In *Proceedings of the 12th Argument mining Workshop*, pages 205–216, Vienna, Austria. Association for Computational Linguistics.

Jonathon Read, David Hope, and John Carroll. 2007. Annotating expressions of appraisal in english. In *Proceedings of the Linguistic Annotation Workshop*, LAW '07, page 93–100, USA. Association for Computational Linguistics.

Jaffer Sheyholislami and Carla Hall. 2013. Using appraisal theory to understand rater values: An

examination of rater comments on esl test essays. *The Journal of Writing Assessment*, 6.

Maite Taboada and Jack Grieve. 2004. Analysing appraisal automatically. *In Proceeding of AAAI spring symposium on exploring attitude and affect in text*, pages 158–161.

Gholam-Reza Tajvidi and S. Hossein Arjani. 2017. Appraisal theory in translation studies: An introduction and review of studies of evaluation in translation. *Journal of Research in Applied Linguistics*, 8:3–30.

Peter White. 2006. *Evaluative semantics and ideological positioning in journalistic discourse*, pages 37–67.

Karolina Zaczynska, Peter Bourgonje, and Manfred Stede. 2024. How diplomats dispute: The UN security council conflict corpus. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8173–8183, Torino, Italia. ELRA and ICCL.

10. Language Resource References

Anisimova, Mariia and Zikánová, Šárka. 2024. *CoDipA UNSC 1.0*. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL).

Schoenfeld, Mirco, and Eckhard, Steffen, and Patz, Ronny, and van Meegdenburg, Hilde and Pires, Antonio. 2019. *The UN Security Council Debates*. Harvard Dataverse.