Automatic Detection of Attitude in Diplomatic Speeches of the UNSC Dataset with Large Language Models

Valentina Tretti
Potsdam Universität
valentina.tretti@uni-potsdam.de

Abstract

Different studies have been carried out on the automated identification of *Appraisal Theory* in distinct textual genres. However, there is not much research on the identification of *Appraisal Theory* in diplomatic speeches. For this reason, this investigation aims to test whether *Large Language Models (LLMs)* can identify appraisal, specifically *attitude* domain, in *UNSC* speeches (Schönfeld et al., 2019). For this purpose, a sample of 87 *UNSC* speeches is annotated considering the subdomains of *attitude* (*affect, appreciation* and *judgement*) and its polarity (positive and negative), following Anisimova's (2023) guidelines and several authors (Martin and White, 2005; Oteíza and Pinuer, 2019). This annotated sample is used to perform different experiments with various *LLMs* (i.e., *BERT, BART, Flan-T5* and *Sentence Transformers*). The results obtained show that *LLMs* can identify attitude in *UNSC* speeches with an accuracy over 50% for best performing models (*Sentence Transformers*). The error analysis shows that models have difficulty identifying *appreciation* and *judgement*, while it is easier for them to identify *affect*. In addition, a linguistic analysis of patterns in fragments annotated with *attitude* is presented, which could facilitate the identification of *attitude* in these type of discourses in future research. Finally, it is confirmed that the *Lexicoder Sentiment Dictionary* (Young and Soroka, 2012) can be used as a complement to identify attitude in *UNSC* speeches.

1 Introduction

Language has been the focus of multiple investigations from diverse perspectives. Through language, speakers' express emotions and evaluations of people, events and objects. Based on the above, language can be analyzed from different methodological perspectives, one of which is *Systemic Functional Linguistics (SFL)*, which focuses on the relationship between the meaning of language and the functions of the social context (Halliday, 1996). There are several frameworks for analyzing language from this perspective, among them the *Appraisal Theory* (Martin and White, 2005). This theory was developed as an analytical approach to diverse issues of speaker/writer evaluation (i.e., certainty, commitment and knowledge) and to consider how the textual voice positions itself with respect to other voices and other positions in the discourse (Martin and White, 2005; Oteíza, 2017).

There is a wide variety of studies in which *Appraisal Theory* is applied to different textual genres such as political discourses (Siyou and Zhongwen, 2018), Facebook (Tran and Ngo, 2018) and Twitter comments (Ross and Caldwell, 2020), online reviews (Paronen, 2011) and with different methodologies, including computational (Taboada and Grieve, 2004; Bloom and Argamon, 2010; Neviarouskaya et al., 2010; Parameswaran et al., 2020).

However, there are no studies presenting an automated approach to appraisal identification in diplomatic speeches such as those delivered in *United Nations Security Council (UNSC)* [1]. Based on the above, the following research questions are posed: Can *Large Language Models (LLMs)* capture attitude in diplomatic speeches? Which models work better? In order to answer the questions, the following objectives are proposed:

1. To build an annotated sample of diplomatic speeches (from *UNSC* dataset) considering the *Appraisal Theory* (Martin and White, 2005; Oteíza, 2017), specifically *Attitude* subdomains (*affect, judgement* and *appreciation*) and their polarity (*positive* or *negative*).

2. To experiment with different *Large Language Models (LLMs)* in classification tasks.

3. To conduct an error analysis of model's results

[1] Specifically in the *UNSC* dataset (Schönfeld et al., 2019).

and to compare their performance.

To achieve these objectives a sample of 87 diplomatic speeches from the *UNSC dataset* (Schönfeld et al., 2019) will be annotated considering the *Appraisal Theory* (Martin and White, 2005; Oteíza, 2017), following the pilot study conducted by Anisimova and Zikánová (2022). Additionally, experiments (classification tasks) will be carried out with different *LLMs* (i.e., *BERT* (Devlin et al., 2019), *BART* (Lewis et al., 2019), *DistilBERT* (Sanh et al., 2019), *RoBERTa* (Liu et al., 2019) and *Sentence Transformers* such as T5 (Ni et al., 2021))[2]. Moreover, a linguistic analysis will be performed to identify patterns in the speeches containing appraisal.

Section 2 attends to a walk through of the previous work. The method including a description of the data used, the annotation process and the experiments are available in *Section* 3. *Section* 4 contains the results obtained in the different phases of the project and the error analysis. Finally, conclusions (*Section* 5) and limitations and future work (*Section* 6) are presented.

2 Previous Work

As mentioned above, the *Appraisal Theory* was developed as an approach to diverse issues of speaker/writer evaluation. This theory provides a taxonomy divided in three domains (*Attitude, Graduation* and *Engagement*) and each of them has subdomains and categories. Additionally, there are studies that focus on deeply describing (Oteíza, 2017) and complementing the theory in case studies of *historical or collective memory* [3] (Oteíza and Pinuer, 2019) and political debates (Molina and Tretti, 2021) and on the annotation process (Fuoli and Hommerberg, 2015; Fuoli, 2018). On the other hand, several studies focus on determining patterns according to this theory, specifically on the *attitude* domain (Bednarek, 2009; Su and Hunston, 2019).

Most of the studies related to automated appraisal identification are related to lexicon creation (Whitelaw et al., 2005; Argamon et al., 2007; Neviarouskaya et al., 2010). In addition, there are different researches focused on recognizing appraisal and classifying varying genre texts automatically. Taboada and Grieve (2004) work with product reviews and focus on the potential value of adjectives for attitude subdomains (affect, judgement and appreciation). On top of that, there are studies on movie reviews (specifically, the IMDb dataset) that focus on performing experiments to isolate rating features to compare which parts helped in the sentiment analysis classification task (Fletcher and Patrick, 2006), identification of rating adjectivation and training of an SVM with "bag of words" for classification (Whitelaw et al., 2005), and, supervised learning to WordNet glosses to identify attitude type and strength (Argamon et al., 2007).

In addition, there are studies developing automated approaches. These approaches can identify attitude and link it to its target (Bloom and Argamon, 2010); and, classify text with a system (*ATtitude Analysis Model (@AM)*) based on a fine-grained attitude lexicon and labels (Neviarouskaya et al., 2010). Additionally, studies focus on attitude classification using pre-trained models. Parameswaran et al. (2020) used ALBERT (Lan et al., 2019) and Aroyehun and Gelbukh (2020) used NBSVM and RoBERTa-large fine tuned (Liu et al., 2019) to predict judgement in tweets. In addition, Parameswaran et al. (2022) evaluate human annotation agreement and three judgement classifiers: *NLP-CIC*, logistic regression ensemble and RoBERTa (Aroyehun and Gelbukh, 2020), *OrangutanV2*, ensemble of two ALBERT models (Parameswaran et al., 2020) and *NITS*, *XG-Boost* and decision tree classifiers with pre-trained BERT embeddings (Khilji et al., 2020).

Further, different researchers use *Appraisal Theory* for the analysis of specific cases, such as the analysis of biographies (Su and Hunston, 2019), doctoral reports (Starfield et al., 2015) and analysis of Donald Trump's tweets (Ross and Caldwell, 2020).

Finally, regarding the use of *Appraisal Theory* and the *UNSC* dataset (Schönfeld et al., 2019), there is the study by Anisimova and Zikánová (2022) which focuses on analyzing *UNSC* speeches according to the *attitude* domain, considering its different subdomains (affect, judgement and appreciation) and the different internal categorizations within these subdomains. There are other studies that work with the *UNSC* dataset, however, they focus on sentiment analysis (Scherzinger, 2022; Sakamoto, 2023).

Although much research has been done applying *Appraisal Theory* to various textual genres and with a variety of computational methodologies, cur-

[2] Code for experiments, annotated data and visualizations are available in Project's repository

[3] Also known as *the politics of memory* in Political Science.

Speech ID	Year	Topic/Agenda Item	No. of speeches
7138	2014	Ukraine	17
7154	2014	Ukraine	17
7165	2014	Ukraine	13
7219	2014	Ukraine	15
7643	2016	United Nations peacekeeping operations	16
7658	2016	Woman and Peace and Security	9

Table 1: Distribution of Annotated Data subset in terms of year, topic and number of speeches.

rently only the work of Anisimova and Zikánová (2022) on the *UNSC* dataset (Schönfeld et al., 2019) is available. However, no studies have been performed from a computational perspective considering this theory and the *UNSC* dataset.

3 Method

3.1 Original Data

For this project a sample of the *United Nations Security Council Debates (UNSC)* dataset (Schönfeld et al., 2019) was used. The original dataset, in English-language, is available through *Harvard Dataverse*. The dataset includes speeches of public meetings from 1995 until 2020 and contains 5,748 public meetings and 82,165 individual speech contributions. Additionally, the dataset includes metadata for speeches and meetings which provides information of participants, speakers, country, time, topics, date, among others.

3.2 Data Subset

Our sample contains 87 speeches in which one of the following topics or agenda items: *"Ukraine"*, *"Women and Peace and Security"* and *"United Nations peacekeeping operations"* was discussed. The speeches were pre-processed by Zaczynska et al. (2024) and two data subsets were provided:

1. speeches divided into sentences; and,
2. speeches divided considering *Elementary Discourse Units (EDUs)* [4].

The speeches included in this data subset have a total of 79,205 tokens and 3530 sentences. Table 1 presents the number of speeches (for the data subset) considering year and topic or agenda item (see *Tables* 9, 10, 11, 12, 13 and 14, in A, for more details on all speeches).

Annotations were made on the sentence split subset and then the annotated fragments were transposed with the *EDUs* subset (more details on this process in 3.3.3).

[4]Clause-like units, hereinafter referred to as *EDUs*.

3.3 Annotations

3.3.1 Annotation Framework

As mentioned in *section* 2, the *Appraisal Theory* provides a framework for the analysis of diverse issues of speaker/writer evaluation and to consider how the textual voice positions itself with respect to other voices and other positions in the discourse (Martin and White, 2005; Oteíza, 2017). This theory was chosen for the current study because it focuses on the analysis of "meanings in contexts towards rhetorical effects rather than towards grammatical forms" (Martin and White, 2005, 94). We consider this approach relevant for the case of diplomatic speeches (specifically *UNSC* speeches), since they present particular characteristics of the genre that could not be analyzed only with a grammatical approach, ignoring the context and rhetorical effects.

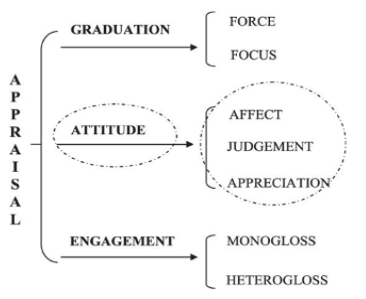

Figure 1: Domains and subdomains of *Appraisal Theory* (based on Martin and White (2005))

As shown in *Figure* 1[5], this theory has three main domains: 1) *graduation*, is concerned with the grading phenomena used by speakers/writers to graduate their interpersonal impact (Martin and White, 2005); 2) *attitude*, refers to the types of appraisals that can be attributed to the valued discursive entities (humans and things) (Martin and White, 2005; Oteíza, 2017; Oteíza and Pinuer, 2019); and, 3) *engagement*, includes the resources speakers use to take a stance on a proposition and how much personal investment they direct to it (Martin and White, 2005). For this study, only *attitude* domain is annotated because it is the one that provides us with a taxonomy for analyzing appraisals targeted to either humans or things (Martin and White, 2005; Oteíza and Pinuer, 2019). This domain is divided in three subdomains:

1. *Affect:* is the assessment as an emotional reaction (Martin and White, 2005; Oteíza and

[5]Figure taken from Zhang (2013).

Pinuer, 2019).

2. *Judgement:* related with ethics, refers to the attitudes towards human behavior or character (Martin and White, 2005; Oteíza and Pinuer, 2019).

3. *Appreciation:* related to aesthetics, involves evaluations of semiotic and natural phenomena according to the ways in which they are valued or not in a given field (Martin and White, 2005; Oteíza, 2017).

In addition, we followed the *Annotation manual for defining attitudes developed* by Anisimova (2023), which includes a description of the subdomains, its categories, specific examples from the *UNSC* dataset (see *Table* 15, in A, for more details) and recommendations for annotating this type of speeches.

3.3.2 Annotation Tool

The annotations were made on the sentence split set using the *INCePTION tool* (Klie et al., 2018). It is an open-source annotation platform for diverse tasks (i.e., interactive and semantic annotation). The tool allows you to create a project with specific features according to the annotation's needs. For the present study, a project with one layer named ***Appraisal Theory*** was created. Additionally, this layer is associated with two tag sets:

1. *attitude_labels:* which includes affect, judgement and appreciation; and,

2. *polarity_labels:* including positive and negative.

3.3.3 Annotation Process

As mentioned in the previous *subsection* 3.3.2, to annotate *attitude* a set of tags for the subdomains (*affect*, *judgement* and *appreciation*) were manually created. In contrast to Anisimova and Zikánová (2022), who annotated speeches considering all subcategories within each subdomain, in the present study the annotations were made only with the subdomains. This is because the objective of the project is to examine how the models capture the attitude (at a general level) in diplomatic discourses. For this reason, the annotations are made considering only the second level of the taxonomy and the polarity.

Following Anisimova (2023) we performed annotations in a chunk level depending on how the *attitude* was formulated (inscribed/explicit or invoked/implicit). The annotator's task consisted of:

1. Reading each sentence looking for *tokens* or *chunks* with attitude and annotating them with the *Appraisal Theory* tag. In case the annotator is certain of the category (attitude and polarity labels) it will be established in this first step, if not then the fragment will remain with the *Appraisal Theory* tag.

2. Once the annotator finishes the speech, the doubtful cases annotated with the Appraisal Theory tag will be revisited. If the annotator is certain of the category, the chunk will be annotated accordingly. If not, then these cases will be evaluated at the end of the annotation process.

The annotation process consisted of two phases:

1. ***Phase I***: annotation of 34 speeches. The unclear cases of this first annotation process were evaluated in a meeting with Mariia Anisimova and Karolina Zaczynska on 10/01/2024.

2. ***Phase II***: annotation of 53 speeches. The unclear cases of this second phase were evaluated in a session with a specialist in *Discourse Analysis*, who had previously worked with *Appraisal Theory*.

Although the guidelines of Anisimova (2023) were followed, there are differences in our annotations. In this study, in addition to following Martin and White (2005), we followed Oteíza and Pinuer (2019), specifically in the case of the annotation of historical events or situations as appreciation. Given what has been said, we consider the use of the word *crisis*, when it appears referring to the Ukrainian situation, as a negative appreciation. For example:

(1) The **crisis** [appreciation-negative] in Ukraine is rooted in a complex situation [...](SPV.7219_spch008)

(2) Since the beginning of the Ukrainian **crisis** [appreciation-negative] [...] (SPV.7138_spch015).

Likewise, cases in which verbs such as *must* and *should* are used are not annotated as positive judgements because it is considered that the use of modal verbs does not necessarily imply an appraisal. Moreover, in speeches these verbs are used

to evaluate actions that should be done in the future, so the speaker is evaluating the behavior that a person could have but has not had in the present or in the past. Thus, actions are evaluated assuming that they will or should happen (see example (3)).

(3) Moscow **must** reject these latest unlawful actions, and do so publicly. (UNSC_2014_SPV.7154_spch007)

In addition, in the clarification session [6], it was discussed among the participants whether to consider the cases in which the tokens *terrorist*, *armed* + noun (e.g. *group* or *forces*) and *separatists* appeared, as negative judgement or not. In this regard, one of the participants stated that these words are common and therefore do not make a judgement. However, for this study we considered them as judgemental since they were used to state a negative appraisal of the behavior of members of the Russian Federation or Ukraine. Therefore examples such as (4) and (5) were annotated as negative judgement in the speeches:

(4) **Armed separatists, terrorists** [judgement-negative] and foreign fighters and their supporters bear the responsibility for the deaths and injuries among the civilian population, including children, women and the elderly. (SPV.7219_spch007)

(5) We urge Russia to cease its policy of supporting **armed separatist groups** [judgement-negative] [...] (SPV.7219_spch004).

Additionally, in cases where two subdomains were present in the same annotated chunk, one implicit and the other explicit, only the explicit one was annotated. For example:

(6) The Egyptian delegation chose **not to vote against** [affect-positive] today's resolution 2272 [...] (SPV.7643_spch005)

In this example a positive *affect* is explicitly presented, the speaker expresses inclination about the resolution, but implicitly there is a positive appreciation of the resolution. Given that in the tool we cannot annotate a fragment with two categories (due to the way the project was created), the decision was made to annotate the explicit appraisal because it is more likely that the computational models will be able to identify it as opposed to the implicit one.

[6]With Mariia Anisimova and Karolina Zaczynska on 10/01/2024.

Labels	Ukraine	UNPO	WPS	Total
affect-negative	123	22	3	148
affect-positive	173	58	48	279
appreciation-negative	293	17	28	338
appreciation-positive	43	22	44	109
judgement-negative	494	51	8	553
judgement-positive	120	16	41	177
Total	124	186	172	1604

Table 2: Distribution of labels in annotated speeches according to agenda item or topic.

On the other hand, differences were found in the difficulty of annotating the speeches according to their topic. For example, in the speeches with *"Women and peace and security"* and *"United Nations peacekeeping operations"* as topic or agenda item, it was more difficult to determine whether a fragment was a *judgement* or *appreciation* since the target was not explicitly presented or the appraisal was implicitly made.

Once the annotation process was completed, the annotated fragments were extracted from *INCePTION tool*. The annotated fragments were compared with the dataset divided into *EDUs*[7], in case the annotated fragment was included within the *EDU* it was marked as True and otherwise as False. This was used to determine if the *EDUs* had appraisal, yes and no labels [8]. Subsequently, a dataset is created with the *EDUs* that have appraisal to perform the classification experiments according to the attitude subdomain. After reviewing the dataset with the annotations and the *EDUs*, it was found that several of the *EDUs* were duplicated (because the annotated text was the same and it only appeared once in the *EDUs* text). In these cases, we proceeded to manually remove the duplicated *EDUs*. *Figure* 2 presents the distribution of labels in the annotated speeches after removing the duplicates [9]. Further, *Table* 2[10] presents the distribution of labels in annotated speeches in terms of topic or agenda item [11].

Finally, after conducting an analysis on dupli-

[7]This is because the annotations were made on the data separated into sentences and for comparative purposes with the work of Zaczynska et al. (2024) they needed to be in the *EDU* format.

[8]This dataset is used for the appraisal identification experiments in the *EDUs*.

[9]In contrast, *Figure* 6, in A, presents the same data before removing the duplicates and *Table* 16, the difference between before and after removing the duplicates.

[10]Using *UNPO* for United Nations peacekeeping operations and *WPS* for Women and Peace and Security.

[11]After removing duplicates.

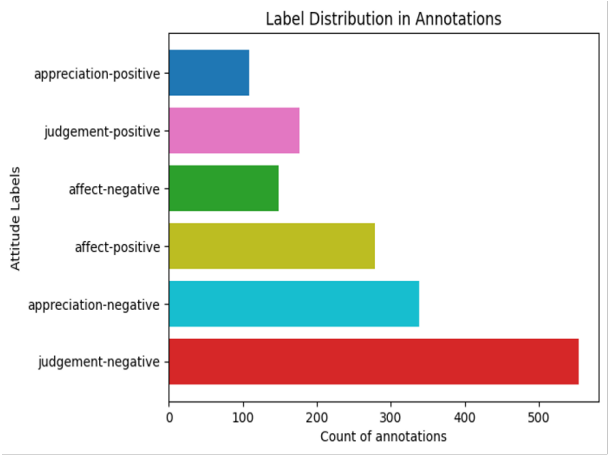

Figure 2: Distribution of attitude labels in the annotated speeches after manual removal of duplicates.

cate cases it was concluded that it was an error as a result of transposing the annotations with the *EDUs*, in those cases where a text chunk had been annotated and within that chunk tokens appeared that were the annotated text for other chunks.

3.4 Experiments

To achieve the project's goals, *Large Language Models (LLMs)*[12] were used to perform both tasks: *Binary Classification of Appraisal labels* [13] and *Multi-Class Classification of Attitude labels*[14]. A brief description of the models used is presented below:

1. **Google Flan-T5** (Chung et al., 2022):
   (a) *Question Answering with prompt:* ***google/flan-t5-large*** was used for both binary and multi-class classification. The model was provided with a prompt that included a description of each label and an example with the corresponding category.
   (b) Text to Text Zero Shot Classification: ***google/flan-t5-base*** was used and the model was not exposed to the labels during training.

2. **BART** (Lewis et al., 2019): ***facebook/bart-large-mnli*** (Lewis et al., 2019) was used with the zero shot classification pipeline.

3. **Sentence Transformers:** we used 4 sentence transformers: BERT (Devlin et al., 2019), RoBERTa (Liu et al., 2019), BART (Lewis et al., 2019) and T5 (Ni et al., 2021) with few shot classification through *SetFit* (Sentence Transformer Fine-tuning), which provides an efficient and prompt-free framework for few-shot fine-tuning of Sentence Transformers (Tunstall et al., 2022). These models were implemented in the binary classification with the annotated data divided into training set (10%) and testing set (90%); and, in the multi-class classification a dataset of 60 examples (only fragments) with *attitude* (10 per label)[15] was used for fine-tuning.

All these models are available in *Hugging Face* and were implemented in *Google Colab*. The experiments were conducted in the following phases:

1. ***Phase I:*** predictions were made on a sample of 10 annotated speeches. In this case the annotated data only included the fragment and not the sentence or *EDU* in which *attitude* was present. This was done in order to check if the selected models were appropriate for the objectives of the work.

2. ***Phase II:*** experiments were conducted with the annotated data of the 87 speeches and the *EDUs*. The models had to classify the *EDU* according to two labels: **yes** (appraisal is in the *EDU*); and, **no** (*EDU* has no appraisal).

3. ***Phase III:*** the multi-class classification of the *attitude* labels was performed with the annotated data of the 87 speeches. This data had only the *EDUs* that contained appraisal.

The last phase, *Phase III*, was conducted twice because after reviewing the results obtained in the first attempt we realized that the data had errors. There were repeated entries of data, therefore, it was necessary to perform this step again with the corrected data. This was not done in *Phase II* because the duplicated data was not considered to have a significant impact on the binary classification result.

[12] For the first results we used *DistilBERT* (Sanh et al., 2019) but when performing experiments with complete data the session crashed in multiple occasions with this model. Hence, we are not considering it under the implemented models.

[13] Yes and no.

[14] Affect-positive, affect-negative, judgement-positive, judgement-negative, appreciation-positive, appreciation-negative.

[15] Created from the examples available in Anisimova's (2023) annotations guidelines. Chunks are used because this is how the examples appear in the guide.

Model	Task	Accuracy Score
Sentence-trans-RoBERTa	Few-Shot-Classification	72.58%
Sentence-trans-BART	Few-Shot-Classification	66.12%
Sentence-trans-T5	Few-Shot-Classification	63.70%
Sentence-trans-BERT	Few-Shot-Classification	57.25%
Google Flan-T5	Few-Shot-Classification	43.55%
Google Flan-T5	Zero-Shot Classification	27.42%
BART	Zero-Shot-Classification	15.32%
DistilBERT	Text Classification	13.71%

Table 3: Accuracy scores for model's predictions of *attitude* labels with data from 10 annotated speeches.

4 Results

4.1 Phase I: First Experiments.

As mentioned in *Subsection* 3.4, *Phase I* consisted of predicting attitude labels with 8 *LLMs* in a sample of 10 annotated speeches. As shown in *Table* 3, *Sentence Transformers* obtained an accuracy over 50%. In contrast, the other models had an accuracy below 45%.

These results allowed us to have an idea of which models would be the most suitable for classifying attitude labels with the 87 annotated speeches. In this case, we considered that the four *Sentence Transformers* would probably have the best results, however, we did not exclude the other models. As will be seen later, we also used them to make predictions with all the speeches.

4.2 Phase II: Binary Classification.

For the binary classification task experiments were performed with all models, except for *DistilBERT* due to technical issues. This task was performed over the data of the 87 annotated speeches considering the *EDUs*. Models had to determine if the *EDU* contained or not appraisal (yes/no).

Table 5 presents the accuracy scores for the models in this task. In general, models' performance is over 50%, except for *BART* with zero shot classification (45%), this is to be expected given that this model was not exposed to the data prior to making the predictions. Nonetheless, this is also the case with model *Flan-T5-ZSC* and it obtained an accuracy over 50%. Moreover, as hypothesized in *Phase I*, the three best performing models are *Sentence Transformers*[16]: *ST-BART, ST-BERT* and *ST-RoBERTa*. In addition, as shown in *Table* 4 , *ST-BART* is the model with best F1-scores for both labels followed by *ST-BERT*.

[16]From now on referring to *Sentence Transformers* as ST.

Model	Label-No	Label-Yes
ST-BART	77%	76%
ST-BERT	76%	74%
ST-RoBERTa	73%	75%
ST-T5	76%	71%
Flan-T5-QA	74%	37%
Flan-T5-ZSC	66%	27%
BART-ZSC	30%	55%

Table 4: F1-score for all models in binary classification.

Model	*Binary*	*Multi Class*
ST-BART	**76.59%**	**52.62%**
ST-BERT	**74.81%**	**51.18%**
ST-T5	73.82%	**51.31%**
ST-RoBERTa	**74.35%**	43.58%
BART-ZSC	45.21%	15.49%
Flan-T5-QA	63.27%	38.53%
Flan-T5-ZSC	53.52%	15.17%

Table 5: Accuracy scores for all models in binary and multi-class classification.

4.3 Phase III: Multi-class Classification.

As shown in *Table* 5, the three models with the best accuracy score for multi-class classification task are: *ST-BART* (52.62%), *ST-T5* (51.31%) and *ST-BERT* (51.18%). From these models, *ST-BART* performs the best in both tasks (binary and multi-class classification), followed by *ST-BERT*. *ST-T5* is the third best performing model on the multi-class classification task, but was the fourth best model for the binary classification task. However, the difference with the other model is not significant (see *Table* 5 for details). Additionally, in general, the accuracy scores for the best models in multi-class classification presents a 20% lower accuracy compared to the best models in binary classification.

According to the F1-scores obtained by the 3 best performing models, see *Table* 6, all three models have scores above 50% for the labels: *affect-positive* and *affect-negative* (except for *ST-T5* which obtains 49%). In addition, the *ST-BART* and *ST-T5* models present values above 60% for *judgement-negative*. Only *ST-BERT* obtains for *appreciation-positive* a score higher than 65%. On the other hand, all models show F1-scores be-

low 45% for *judgement-positive* and *appreciation-negative* labels. This is also the case for the *ST-BART* and *ST-T5* models with *appreciation-positive* label and only for *ST-BERT* with *judgement-negative*. In addition, these values are best described in the confusion matrices for each model (see *Figures* 3 for *ST-BART*, 4 for *ST-BERT* and 5) for *ST-T5* [17].

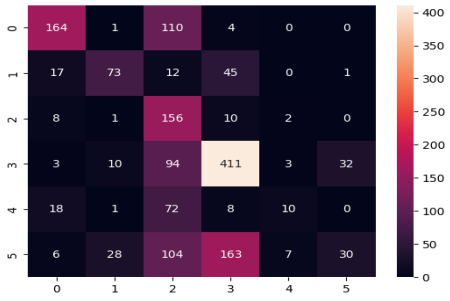

Figure 3: Confusion matrix *ST-BART* with *attitude* labels.

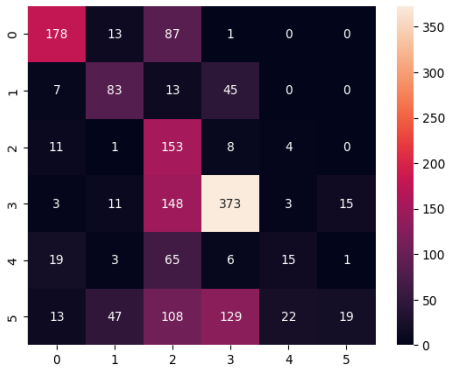

Figure 4: Confusion matrix *ST-BERT* with *attitude* labels.

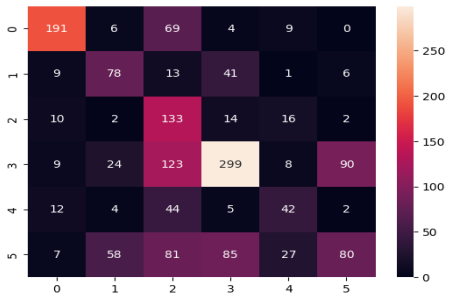

Figure 5: Confusion matrix *ST-T5* with *attitude* labels.

[17]The order of the labels in the matrices is: affect-positive, affect-negative, judgement-positive, judgement-negative, appreciation-positive, appreciation-negative from top to bottom.

4.3.1 Error Analysis

As shown in *Table* 6[18] and considering the analysis of the confusion matrices of each model (see figures 3, 4 and 5) it is determined that for models it is difficult to correctly classify the following labels: **appreciation-negative, appreciation-positive, judgement positive** (common in 3 models) and **judgement-negative** (only in *ST-BERT*).

In all 3 models the label **appreciation-negative** is incorrectly classified as one of the following: *judgement-negative:* 163[19] (ST-BART), 129 (ST-BERT) and 85 (ST-T5); *judgement-positive:* 104 (ST-BART), 108 (ST-BERT) and 81 (ST-T5); and, *affect-negative:* 28 (ST-BART), 47 (ST-BERT) and 58 (ST-T5)[20].

On the other hand, the label **judgement-positive** is confounded with: *judgement-negative:* 10 (ST-BART), 8 (ST-BERT) and 14 (ST-T5); *affect-positive:* 8 (ST-BART), 11 (ST-BERT) and 10 (ST-T5); and, *appreciation-positive:* 2 (ST-BART), 4 (ST-BERT) and 16 (ST-T5)[21].

Both *ST-BART* and *ST-T5* have a low F1-score for **appreciation-positive** (13% and 30%, respectively). This is not the case for *ST-BERT* which scored 67%, making it the second best ranked label for this model. This label was misclassified, in the *ST-BART* and *ST-T5* models, as: *judgement-positive*: 72 (ST-BART) and 44 (ST-T5), *affect-positive:* 18 (ST-BART) and 12 (ST-T5); and, *judgement-negative:* 8 (ST-BART) and 5 (ST-T5)[22].

Finally, both *ST-BART* and *ST-T5* obtained high F1-scores for **judgement-negative** (69% and 60%, respectively). However, in the case of *ST-BERT*, this was the label with the second lowest F1-score (20%) and was misclassified as positive judgement (148), negative appreciation (15), and negative affect (11) [23].

4.3.2 Analysis of Misclassified Labels

Considering the results presented in *Subsection* 4.3.1, we decided to analyze a sample of examples for labels that were misclassified. The sample consisted of 60 examples, 15 per label. The results obtained from this analysis will be presented in this section.

[18]In bold are the labels ranked lowest by the models.
[19]Number of examples incorrectly classified.
[20]For more details see *Figures* 3, 4, 5.
[21]For more details see *Figures* 3, 4, 5.
[22]See *Figures* 3 and 5 for more details.
[23]See *Figure* 4.

Model	*ST-BART*	*ST-BERT*	*ST-T5*
affect-positive	66%	54%	74%
affect-negative	56%	70%	49%
judgement-positive	**43%**	**10%**	**42%**
judgement-negative	69%	**20%**	60%
appreciation-positive	**15%**	67%	**40%**
appreciation-negative	**15%**	**41%**	**31%**

Table 6: F1-score for all models in multi-class classification.

In some cases, the models misclassified the *EDUs* as **judgement-negative** because apart from the positive judgement it included tokens that frequently are considered to be negative words (e.g., *terrorists, violently, attacked, gangs* and *threat*). Example (1) was probably classified as negative because of the presence of the token *terrorist* which is mostly used as a negative word; however, in this context it is preceded by *anti* which changes the polarity of this token. Similarly, in example (2) the models do not consider the appraisal "*peaceful demonstrators*" and probably incorrectly classified the *EDU* considering the negative words (*violently attacked* and *gangs*).

(1) The **anti-terrorist forces** [judgement-positive] in Ukraine have been repeatedly shelled from the territory of Russia. (UNSC_2014_SPV.7219_spch019)

(2) [...] just yesterday **peaceful demonstrators** [judgement-positive] in favour of the unity of the Ukraine were violently attacked by armed pro-Russian gangs armed with clubs and baseball bats in the city of Donetsk. (UNSC_2014_SPV.7165_spch011)

In other cases, speakers made positive judgements through the negation -with an adverb- of negative actions. This is the case of examples (3) and (4) in which the speaker uses the adverbs *not* and *never* to evaluate as positive, actions carried out by themselves (3) and by other entities (4).

(3) **We are not to blame** [judgement-positive] for what has produced the results [...]. (UNSC_2014_SPV.7154_spch021)

(4) During the entire course of the crisis in Ukraine, Russia has **never advocated aggravating it or destabilizing the country** [judgement-positive]. (UNSC_2014_SPV.7154_spch021)

In addition, model *ST-T5* classified incorrectly as judgement negative cases in which the tokens *de-escalate* or *de-escalation* were present. In this case, we consider that the model is probably considering *de-escalate/-ion* as a negative token. Nonetheless, this is not the case with models *ST-BART* and *ST-BERT* which classified examples (5) and (6) correctly as positive judgement. Moreover, this also happens with example (7), which was incorrectly classified as *judgement-negativ*e by models *ST-BERT* and *ST-T5*.

(5) First, the United States **has constantly called for de-escalation and urged restraint** [judgement-positive]. (UNSC_2014_SPV.7154_spch023)

(6) **Ukraine has done its utmost to de-escalate the situation** [judgement-positive] [...]. (UNSC_2014_SPV.7165_spch019)

(7) [...] to **de-escalate the situation** [appreciation-positive]. (UNSC_2014_SPV.7154_spch002)

In some cases, *EDUs* contained more than one appraisal. In these cases, models classified the fragment considering the appraisal that appeared first. This is what happened with examples (8) and (9) were the models classified them as *affect-positive* instead of *judgement-positive* and *appreciation-positive*. In addition, in example (10) the models considered the appraisal "*strongly condemns*" and classified this fragment as *affect-negative* instead of *appreciation-negative*.

(8) *I would like to express my sincere gratitude* [affect-positive] [24] to the members of the Security Council for their **overwhelming support**[25] [judgement-positive] of the draft resolution contained in document S/2104/l89 [...]. (UNSC_2014_SPV.7138_spch018)

(9) Spain *thanks* [affect-positive] the United States for this **important initiative** [appreciation-positive]. (UNSC_2016_SPV.7643_spch014)

(10) Brazil *strongly condemns* [affect-negative]

[24] Italics is used in the examples that contain more than one appraisal to identify the second appraisal.

[25] Bold is used to identify the appraisal of interest (main one) on the example.

these **abhorrent violations** [appreciation-negative]. (UNSC_2016_SPV.7658_spch029)

Moreover, it was difficult for models to accurately classify the *EDUs* in which the target of the appraisal was not explicitly mentioned. In example (11) speaker refers to women and in (12) and (13) to a resolution.

(11) [...] and a **role mode**l [judgement-positive] not just in Africa, [...]. (UNSC_2016_SPV.7658_spch029)

(12) We have made **significant progress** [appreciation-positive] in the 15 years [...] (UNSC_2016_SPV.7658_spch017)

(13) **That has never really been operational** [appreciation-negative]. (UNSC_2016_SPV.7643_spch011)

These targets of the appraisals were mentioned in previous lines of the speeches; therefore, it is not possible to classify the *EDU* without having knowledge of the context and of what has been mentioned in the speech.

In the case of fragments classified as positive appreciation it is interesting that the models do not classify them correctly although in the *EDU* the target of the appraisal appears explicitly. In example (14) the models may have considered that it was a positive appraisal because of the construction "*allows us to*" in which it can be interpreted that the resolution acts as a means for the speaker to give a clear message. On the other hand, in example (15) the models may have considered "*Angola*" as the target and; therefore, as the one who performs the action of "*convened this extremely important debate*".

(14) The resolution allows us to **send a clear message** [appreciation-positive] to all actors: [...] (UNSC_2016_SPV.7643_spch009)

(15) [...] and to thank Angola for having convened **this extremely important debate** [appreciation-positive] on women and peace and security in Africa. (UNSC_2016_SPV.7658_spch026)

Additionally, models incorrectly classified *EDUs* as *judgement-positive* in cases where tokens with positive polarity appear in the *EDU*. It could be the case that the models considered "*to do something*" and "*prevent*" as positive actions and therefore classified examples (16) and (17) as a positive judgement.

(16) My ulterior motive is actually, finally, to do something about a **cancer** [judgement-negative]: [...]. (UNSC_2016_SPV.7643_spch010)

(17) [...] to prevent conflict and resolve **crisis** [appreciation-negative] and represent the entire world, [...] (UNSC_2016_SPV.7658_spch012)

Moreover, models present a difficulty in identifying implicit appraisals. In example (18) the speaker makes an evaluation about Russia, in this case "*it cannot deny the truth*" implicitly expresses that Russia lies. It is possible that the models have considered the combination "*cannot deny*" as positive and that it referred to an evaluation of a human entity because they are the ones who can tell the truth.

(18) [...] and **it cannot deny the truth** [judgement-negative], [...]. (UNSC_2014_SPV.7138_spch004)

Further, there are cases where it is possible that the models have considered one of the *EDU* tokens as the target of another token with negative polarity. This is because in examples (19) and (20), in the *EDU*, the target of the appraisal does not appear and nouns *force/-s* and *actions* can be considered as non-human situations or entities.

(19) [...] **separatist forces** [judgement-negative] [...] (UNSC_2014_SPV.7219_spch006)

(20) [...] that the **reckless actions** [judgement-negative] [...] (UNSC_2014_SPV.7154_spch004)

Finally, there are cases in which the models incorrectly classified *EDUs* as *affect-negative* when a word appears in the *EDU* that may represent or may be associated with an emotion (e.g., *awful*, *unacceptable* and *rejected*). This is the case in examples (21), (22) and (23).

(21) It is **awful** [judgement-negative] that the other one, [...] (UNSC_2014_SPV.7138_spch018)

(22) That is **completely unacceptable** [judgement-negative].(UNSC_2014_SPV.7154_spch007)

(23) [. . .] must be **categorically rejected** [judgement-negative]. (UNSC_2016_SPV.7643_spch011)

The results obtained show that the models can identify positive or negative affect without much difficulty. This may be a result of the data with which the models were previously trained, since the appraisals that present *affect* refer to emotions and in most cases the speaker makes use of words (i.e., *adjectives*, *nouns* and *verbs*) with an emotional charge and with a marked polarity (positive or negative). The presence of these linguistic forms is common in sentiment analysis datasets that are mostly used to train these text classification *LLMs*. However, this is not the case for the other two *attitude* subdomains: *appreciation* and *judgement*. In most cases, these appraisals are constructed through negation and more complex linguistic constructions. Moreover, they may appear in a fragment with more than one appraisal, the targets may be implicit or mentioned outside the analyzed *EDU*, and the appraisal may be implicitly made. These features make it difficult for the models to correctly classify *EDUs* for these labels.

4.3.3 Linguistic Pattern Analysis in Attitude Annotations

A sample of 300 examples (50 per label) of *EDUs* containing appraisal was analyzed. This analysis was conducted with the goal of identifying linguistics patterns. Su and Hunston (2019) present an analysis of patterns for *attitude*; however, this analysis is limited to the identification of patterns with respect to adjectives. For the present study, we considered Su and Hunston's (2019) adjective patterns and we created new ones according to those identified in the sample (not limited to adjectives).

As shown in *Table* 17 in A there are 5 patterns that are shared between two or more *attitude* labels. From these patterns, two correspond to those defined by Su and Hunston (2019): *it v-link ADJ that* (examples (7) and (8)) and *it v-link ADJ to-info* (examples (9) and (10)). The rest of the patterns identified in the sample were defined considering different *Part-of-Speech (POS) Tags* combinations. First, pattern *ADV + ADJ* (see examples (1) and (2), in *Table* 17) appears in both negative affect and appreciation.

Moreover, labels *appreciation* and *judgement* with both polarities (positive and negative) share the pattern *ADJ + (ADJ) + NOUN*. This pattern occurred with some variations:

1. *ADJ + NOUN:* as shown in examples (4) and (5) (see *Table* 17).

2. *ADJ + ADJ + NOUN:* in judgement positive and appreciation-negative, examples (3)) and (6), respectively.

The use of a negative noun (NEG NOUN) to express appraisal was identified as a pattern in all labels with negative polarity (*affect*, *appreciation* and *judgement*) (see examples (11), (12) and (13) in *Table* 17).

Finally, the pattern *PRONOUN or NOUN + (ADV + ADJ) + VERB* was identified in affect-positive with verbs *thanks* and *welcome* (see example (14) in *Table* 17) and with some variation *PRONOUN + NOUN + VERB* in one case in judgement-positive (see example (15) in *Table* 17)

Three patterns were identified only in *affect* label with both negative and positive polarities:

1. *PRONOUN + V-link + (ADV/NEG ADV)+ ADJ:* see examples (1) and (2) in *Table* 18.

2. *PRONOUN + (ADV) + VERB:* see examples (3) and (4) in *Table* 18.

3. *PRONOUN + V-modal + (ADV) + like + to-info:* see examples (5) and (6) in *Table* 18.

In addition, some patterns were identified only in specific labels:

1. As shown in *Table* 19 three patterns were recognized for *affect-negative*: only an adverb (ADV); *NOUN + VERB + ADJ + CONJ + ADJ*; and, *DET + NOUN + (ADV) + VERB* (see examples (1), (2) and (3) in *Table* 19).

2. For *affect-positive* the following three patterns appeared: *PRONOUN + VERB + (to-info)*; *PRONOUN + VERB + ADJ + that-info*; and, *PRONOUN + VERB (wish) + to + thank* (see examples (1), (2) and (3) in *Table* 20).

3. In the case of positive appreciation only pattern *VERB + ADV + ADJ* was identified (see example (1) in *Table* 21).

4. For *judgement-positive* pattern *to+ VERB+ that-info* in example (1) (see *Table* 22).

Lexicoder Results	Annotated Speeches
2432	1751

Table 7: Lexicoder results and annotated speeches compared samples.

As mentioned above, within the analyzed sample, patterns similar to those established by Su and Hunston (2019) were found. However, most of them did not coincide and it was necessary to define new patterns, given that in the fragment the appraisal was not necessarily constructed with an adjective. In addition, the patterns defined by Su and Hunston (2019) were defined on a span of the analyzed text that sometimes does not match ours, because our annotations were performed on tokens or small chunks. Further, not all the examples analyzed could be grouped within any of the defined patterns because they were complex cases which could not be categorized considering only *POS* tags. Finally, it is important to note that this was a pilot analysis and that we intend to further explore it in future research.

4.3.4 Lexicoder Results vs. Attitude Annotations

As mentioned in the 3.4 *Subsection*, a comparison was made between the results (Zaczynska et al., 2024) obtained with the *Lexicoder Sentiment Dictionary*[26] (Young and Soroka, 2012) and the sample of 87 annotated speeches. *Table* 7 shows the size of both samples.

The comparison was carried out by considering the words identified by the *Lexicoder* and the annotated fragments (i.e., tokens and chunks). For this purpose, we compared whether at least one of the words returned by the *Lexicoder* was contained in the annotated fragment. In this way an initial sample of 3830 matches was obtained. Since the results of the *Lexicoder* are based on sentences and those of the annotated speeches are on *EDUs*, we proceeded to perform a manual review of the sample. This manual review confirmed repeated cases that were eliminated and cases in which the *Lexicoder* result did not correspond to the annotated fragment according to the sentence. This step allowed us to obtain a sample of 1366 examples with matches.

In general, there is a 53.99% accuracy rate in the analyzed sample. As presented in *Table* 8, most of the annotated fragments contain one word from the *Lexicoder* results (66%), followed by the cases in which two words were identified (20%). The three and four word matches are below 10%. Likewise, the cases in which more than five words were equivalent do not exceed 1%. In most cases, the number of tokens in the sentence outdo the number of tokens in the annotated fragment, which explains why the higher the number of words identified by the *Lexicoder*, the lower the number of words recognized in the annotated fragment.

No. of matches	Total No. in Sample	Percentage
0 words	4	0%
1 word	901	66%
2 words	279	20%
3 words	96	7%
4 words	43	3%
5 words	20	1%
6 words	13	1%
7 words	4	0%
8 words	2	0%
9 words	3	0%
10 words	1	0%
Total	1366	100%

Table 8: Number of matched words between *Lexicoder's* results and annotated fragments.

These results allow us to conclude that the *Lexicoder Sentiment Dictionary* (Young and Soroka, 2012) can be used as a complement to the appraisal analysis, specifically in identifying *attitude* in *UNSC* speeches because from 1751 *EDUs* containing *attitude* (from our annotated sample) 1366 had at least one word that matched those in the *Lexicoder's* results. Further, the percentage of accuracy of 50% of the analyzed sample contained at least one word. However, the use of the *Lexicoder* in this sense should be considered only as a complement since not all the words identified by the *Lexicoder* are necessarily used to express an appraisal.

5 Conclusions

In this study, we investigated if *Large Language Models (LLMs)* were able to detect appraisals, particularly *attitude*, in diplomatic speeches, specifically in the *UNSC* dataset (Schönfeld et al., 2019). To carry out this investigation, we performed several tasks including a manual annotation of a sample of *UNSC* speeches with *attitude* labels (*affect, appreciation* and *judgement*) and different experiments with various *LLMs*. The annotated sample was used with *LLMs* in three phases of experiments: 1) *phase I: multi-class classification task* (*attitude*

[26]From now on referring to it only as *Lexicoder*.

labels) over *attitude* annotated fragments (a sample of 10 speeches); 2) *phase II: binary classification task* (yes-no labels) over *EDUs*; and, 3) *phase III: multi-class classification task* (*attitude* labels) over *EDUs*.

Additionally, we analyzed a sample of misclassified *attitude* labels from *Phase III* and conducted a linguistic analysis of a sample of *attitude* annotations in order to identify patterns. Moreover, we compared the results obtained with the *Lexicoder Sentiment Dictionary* (Zaczynska et al., 2024) with our *attitude* annotations.

Considering the accuracy and f1-scores obtained by the models in *Phases I* and *III* we confirmed that *Sentence Transformers* and other *LLMs* that were exposed to the data are the most adequate models to fulfill our objective. In addition, results from *Phases II* and *III* show that models, particularly *Sentence Transformers*, can identify *appraisals* (in general) and *affect* without much difficulty. However, they present difficulties in correctly categorizing the subdomains of *appreciation* and *judgement*. Nevertheless, we believe that by fine-tuning the models with a more robust dataset containing specific cases and complete sentences or *EDUs* (e.g., those identified in the misclassified label analysis in 4.3.2) of these subdomains will improve results.

On the other hand, our linguistic analysis of the patterns in *attitude* gives us an idea of the way in which speakers perform appraisals in *UNSC* speeches, which can be used in future research to identify them more easily. However, we believe that to confirm whether these patterns would really make the identification process easier, it is necessary to perform the analysis on a more robust sample. Additionally, our results allow us to confirm that the *Lexicoder Sentiment Dictionary* (Young and Soroka, 2012) can be used as a complement for the identification of *attitudes* in this particular dataset.

Finally, several researches have been conducted aiming at developing automated approaches to identify appraisal, mostly on datasets of social media comments. While this investigation sheds light on the ability of *LLMs* to identify appraisals in diplomatic speeches, specifically *UNSC* speeches, more research is needed to better fine-tune the models and include *attitude* subcategories.

6 Limitations and Recommendations

Although we were able to carry out the different experiments and tasks in this project, we consider that one of the main limitations was to annotate, perform the experiments and the analysis on data with different formats (sentences and *EDUs*). This resulted in a large number of hours spent on manually cleaning the data for two processes: a) the transposition of the annotated fragments and *EDUs*; and, b) the comparison of the *Lexicoder* results (in sentence format) with the annotated fragments (in *EDU* format).

On the one hand, since the annotation of the sample of 87 speeches was done manually it was not possible, due to time constraints, to annotate a larger sample. This meant that we did not have a large annotated sample from which we could create a sub-sample for fine-tuning the models. For this reason, a sample was created with examples of fragments with *attitude* obtained from Anisimova's 2023 guide. However, these examples did not contain the context of occurrence of the annotated fragments (e.g., sentence) and were sometimes only tokens. We believe that this may have negative effects on the fine-tuning of the models and that it is ideal to perform this process with a sample of annotated *EDUs* or sentences and not just the fragment. For this reason, we consider that future research should use for fine-tuning a sample of examples containing the context or at least the same format as the samples that model will make predictions on.

On the other hand, although the accuracy scores of several models are above 50% for the *multi-class classification*, we consider it necessary to continue working on the fine-tuning of these models to improve these scores as well as the f1-scores. Moreover, it is important to contemplate in future research the specific cases (identified in the misclassified labels analysis) to provide the models with more examples of these types so that they are able to identify them with less difficulty. In addition, it is pertinent to deepen the analysis of linguistic patterns in a larger sample in order to have more clarity and significant values of the use of the patterns in *UNSC* speeches to express *attitude*, in order to use them in automated approaches for appraisal identification.

Finally, although this first approach shows that *LLMs* can identify *attitude* in *UNSC* speeches, the annotations were made only at the second level

of the *Appraisal Theory*. Therefore, we consider necessary increasing the sample from 87 annotated speeches to 100 speeches and to annotate them considering the subcategories within each subdomain. Further, use this new sample to perform experiments with the *LLMs* used in this research and others.

References

Mariia Anisimova. 2023. Argumentation in speeches of the security council of the united nations organizations: Annotation manual for defining attitudes. Unpublished manuscript.

Mariia Anisimova and Sárka Zikánová. 2022. Attitude in diplomatic speeches: a pilot study. *Application and Theory*, pages 23–27.

Shlomo Argamon, Kenneth Bloom, Andrea Esuli, and Fabrizio Sebastiani. 2007. Automatically determining attitude type and force for sentiment analysis. In *Language and Technology Conference*, pages 218–231. Springer.

Segun Taofeek Aroyehun and Alexander Gelbukh. 2020. Automatically predicting judgement dimensions of human behaviour. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, pages 131–134.

Monika Bednarek. 2009. Language patterns and attitude. *Functions of Language*, 16(2):165–192.

Kenneth Bloom and Shlomo Argamon. 2010. Unsupervised extraction of appraisal expressions. In *Canadian Conference on Artificial Intelligence*, pages 290–294. Springer.

Hyung Won Chung, Le Hou, S. Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Wei Yu, Vincent Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed Huai Hsin Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *ArXiv*, 1.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *North American Chapter of the Association for Computational Linguistics*.

Jeremy Fletcher and Jon Patrick. 2006. Evaluating the utility of appraisal hierarchies as a method for sentiment classification. In *Proceedings of the Australasian Language Technology Workshop 2005*, pages 134–142, Sydney, Australia.

Matteo Fuoli. 2018. A stepwise method for annotating appraisal. *Functions of Language*, 25(2):229–258.

Matteo Fuoli and Charlotte Hommerberg. 2015. Optimising transparency, reliability and replicability: annotation principles and inter-coder agreement in the quantification of evaluative expressions. *Corpora*, 10(3):315–349.

M. A. K. Halliday. 1996. Linguistics and literacy: A functional perspective. In R. Hasan and G. Williams, editors, *Literacy in society*, pages 339–376. Longman, Harlow & New York.

Abdullah Faiz Ur Rahman Khilji, Rituparna Khaund, and Utkarsh Sinha. 2020. Human behavior assessment using ensemble models. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, pages 140–144.

Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. 2018. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of the 27th International Conference on Computational Linguistics: System Demonstrations*, pages 5–9, Santa Fe, New Mexico. Association for Computational Linguistics.

Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. 2019. ALBERT: A lite BERT for self-supervised learning of language representations. *ArXiv*, abs/1909.11942.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. In *Annual Meeting of the Association for Computational Linguistics*.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*.

J.R. Martin and P. White. 2005. *The Language of Evaluation: Appraisal in English*.

N. Molina and V. Tretti. 2021. Evaluación en tiempos electorales: un acercamiento al proceso electoral desde el sistema de valoración. In L. Álvarez, editor, *Imaginarios, subjetividades y democracia: estudios sobre el proceso electoral del 2018 en Costa Rica*. CIEP, Costa Rica.

Alena Neviarouskaya, Helmut Prendinger, and Mitsuru Ishizuka. 2010. Recognition of affect, judgment, and appreciation in text. In *Proceedings of the 23rd international conference on computational linguistics (COLING2010)*, pages 806–814.

Jianing Ni, Gonzalo Hernández Abrego, Noah Constant, Jing Ma, Keith B. Hall, Daniel Marcu Cer, and Yinfei Yang. 2021. Sentence-T5: Scalable Sentence Encoders from Pre-trained Text-to-Text Models. *ArXiv*.

Teresa Oteíza. 2017. The appraisal framework and discourse analysis. In T. Bartlett and G. O'Grady, editors, *The Routledge Handbook of Systemic Functional Linguistics*, pages 457–472. Routledge, London.

Teresa Oteíza and Claudio Pinuer. 2019. El sistema de valoración como herramienta teórico-metodológica para el estudio social e ideológico del discurso. *Logos*, 29(2):207–229.

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, and David Eyers. 2020. Classifying judgements using transfer learning. In *Proceedings of the 18th Annual Workshop of the Australasian Language Technology Association*, pages 135–139.

Pradeesh Parameswaran, Andrew Trotman, Veronica Liesaputra, and David Eyers. 2022. Reproducibility and automation of the appraisal taxonomy. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3731–3740, Gyeongju, Republic of Korea. International Committee on Computational Linguistics.

Tiina Paronen. 2011. *Appraisal in Online Reviews of South Park: A study of engagement resources used in online reviews*. Ph.D. thesis, University of Jyväskylä.

Andrew S. Ross and David Caldwell. 2020. 'going negative': An appraisal analysis of the rhetoric of donald trump on twitter. *Language and Communication*, 70:13–27.

Takuto Sakamoto. 2023. Threat conceptions in global security discourse: Analyzing the speech records of the united nations security council, 1990–2019. *International Studies Quarterly*, 67(3).

Victor Sanh, Lysandre Debut, Julien Chaumond, and Thomas Wolf. 2019. DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter. *ArXiv*.

Johannes Scherzinger. 2022. Unbowed, unbent, unbroken? examining the validity of the responsibility to protect. *Sage Journal Cooperation and Conflict*, 58(1):81–101.

Mirco Schönfeld, Steffen Eckhard, Ronny Patz, and Hilde Meegdenburg. 2019. The un security council debates. Harvard Dataverse, V5.

Z. Siyou and P. Zhongwen. 2018. Analysis of political language based on appraisal theory: The mutual construction of language and power—taking xi jinping and donald trump's speeches at world economic forum as examples. *Advances in Social Science, Education and Humanities Research (ASSEHR)*, 248. International Conference on Social Science and Education Reform (ICSSER 2018).

Sue Starfield, Brian Paltridge, Robert McMurtrie, Allyson Holbrook, Sid Bourke, Hedy Fairbairn, Margaret Kiley, and Terry Lovat. 2015. Understanding the language of evaluation in examiners' reports on doctoral theses. *Linguistics and Education*, 31:130–144.

Hang Su and Susan Hunston. 2019. Language patterns and attitude revisited: Adjective patterns, attitude and appraisal. *Functions of Language*, 26(3):343–371.

Maite Taboada and Jack Grieve. 2004. Analyzing appraisal automatically. In *American Association for Artificial Intelligence Spring Symposium on Exploring Attitude and Affect in Text*, pages 158–161, Stanford. AAAI Technical Report SS-04-07.

G. Tran and X. Ngo. 2018. News comments on facebook – a systemic functional linguistic analysis of moves and appraisal language in reader-reader interaction. *Journal of World Languages*, 5(1):46–80.

L. Tunstall, N. Reimers, U.E. Jo, L. Bates, D. Korat, M. Wasserblat, and O. Pereg. 2022. Efficient few-shot learning without prompts. *ArXiv*.

Casey Whitelaw, Shlomo , and Navendu Garg. 2005. Using appraisal taxonomies for sentiment analysis. In *Proceedings of the First Computational Systemic Functional Grammar Conference*, University of Sydney, Sydney, Australia.

Lori Young and Stuart Soroka. 2012. Affective news: The automated coding of sentiment in political texts. *Political Communication*, 29(2):205–231.

Karolina Zaczynska, Peter Bourgonje, and Manfred Stede. 2024. How Diplomats Dispute: The UN Security Council Conflict Corpus. In *Proceedings of the Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, Turin, Italy. To appear.

Meifang Zhang. 2013. Stance and mediation in transediting news headlines as paratexts. *Perspectives: Studies in Translatology*, 21(3):396–411.

A Appendix

Speech ID	Year	Topic/Agenda Item	Country
7138_002	2014	Ukraine	Russian Federation
7138_004	2014	Ukraine	USA
7138_005	2014	Ukraine	France
7138_006	2014	Ukraine	UK
7138_007	2014	Ukraine	Lithuania
7138_008	2014	Ukraine	Rwanda
7138_009	2014	Ukraine	China
7138_010	2014	Ukraine	Chile
7138_011	2014	Ukraine	Argentina
7138_012	2014	Ukraine	Australia
7138_013	2014	Ukraine	Republic Of Korea
7138_014	2014	Ukraine	Nigeria
7138_015	2014	Ukraine	Chad
7138_016	2014	Ukraine	Jordan
7138_017	2014	Ukraine	Luxembourg
7138_018	2014	Ukraine	Ukraine
7138_020	2014	Ukraine	Russian Federation

Table 9: Distribution of annotated speeches 7138 in terms of year, topic and speaker's country.

Speech ID	Year	Topic/Agenda Item	Country
7154_002	2014	Ukraine	Unite Nations
7154_004	2014	Ukraine	Russian Federation
7154_005	2014	Ukraine	Lithuania
7154_006	2014	Ukraine	USA
7154_007	2014	Ukraine	UK
7154_008	2014	Ukraine	France
7154_009	2014	Ukraine	Rwanda
7154_011	2014	Ukraine	Australia
7154_012	2014	Ukraine	China
7154_013	2014	Ukraine	Argentina
7154_014	2014	Ukraine	Republic Of Korea
7154_017	2014	Ukraine	Chile
7154_019	2014	Ukraine	Ukraine
7154_021	2014	Ukraine	Russian Federation
7154_023	2014	Ukraine	USA
7154_025	2014	Ukraine	Ukraine
7154_027	2014	Ukraine	Russian Federation

Table 10: Distribution of annotated speeches 7154 in terms of year, topic and speaker's country.

Speech ID	Year	Topic/Agenda Item	Country
7165_004	2014	Ukraine	UK
7165_005	2014	Ukraine	France
7165_006	2014	Ukraine	Rwanda
7165_007	2014	Ukraine	Argentina
7165_008	2014	Ukraine	USA
7165_009	2014	Ukraine	China
7165_010	2014	Ukraine	Republic of Korea
7165_011	2014	Ukraine	Luxembourg
7165_014	2014	Ukraine	Australia
7165_015	2014	Ukraine	Russian Federation
7165_016	2014	Ukraine	Lithuania
7165_019	2014	Ukraine	Ukraine
7165_021	2014	Ukraine	Russian Federation

Table 11: Distribution of annotated speeches 7165 in terms of year, topic and speaker's country.

Speech ID	Year	Topic/Agenda Item	Country
7219_004	2014	Ukraine	UK
7219_006	2014	Ukraine	USA
7219_007	2014	Ukraine	Lithuania
7219_008	2014	Ukraine	China
7219_009	2014	Ukraine	Australia
7219_010	2014	Ukraine	Luxembourg
7219_011	2014	Ukraine	France
7219_013	2014	Ukraine	Argentina
7219_014	2014	Ukraine	Nigeria
7219_015	2014	Ukraine	Russian Federation
7219_017	2014	Ukraine	Republic of Korea
7219_018	2014	Ukraine	Rwanda
7219_019	2014	Ukraine	Ukraine
7219_025	2014	Ukraine	Indonesia
7219_029	2014	Ukraine	Belgium

Table 12: Distribution of annotated speeches 7219 in terms of year, topic and speaker's country.

Speech ID	Year	Topic/Agenda Item	Country
7643_002	2016	United Nations peacekeeping operations	Egypt
7643_004	2016	United Nations peacekeeping operations	USA
7643_005	2016	United Nations peacekeeping operations	Egypt
7643_006	2016	United Nations peacekeeping operations	UK
7643_007	2016	United Nations peacekeeping operations	Russian Federation
7643_008	2016	United Nations peacekeeping operations	China
7643_009	2016	United Nations peacekeeping operations	France
7643_010	2016	United Nations peacekeeping operations	USA
7643_011	2016	United Nations peacekeeping operations	Venezuela
7643_012	2016	United Nations peacekeeping operations	Ukraine
7643_013	2016	United Nations peacekeeping operations	Malaysia
7643_014	2016	United Nations peacekeeping operations	Spain
7643_015	2016	United Nations peacekeeping operations	Senegal
7643_016	2016	United Nations peacekeeping operations	New Zealand
7643_017	2016	United Nations peacekeeping operations	Japan
7643_018	2016	United Nations peacekeeping operations	Uruguay

Table 13: Distribution of annotated speeches 7643 in terms of year, topic and speaker's country.

Speech ID	Year	Topic/Agenda Item	Country
7658_012	2016	Women and peace and security	UK
7658_013	2016	Women and peace and security	USA
7658_015	2016	Women and peace and security	China
7658_017	2016	Women and peace and security	New Zealand
7658_023	2016	Women and peace and security	France
7658_024	2016	Women and peace and security	Russian Federation
7658_026	2016	Women and peace and security	Egypt
7658_029	2016	Women and peace and security	Angola
7658_070	2016	Women and peace and security	Namibia

Table 14: Distribution of annotated speeches 7658 in terms of year, topic and speaker's country.

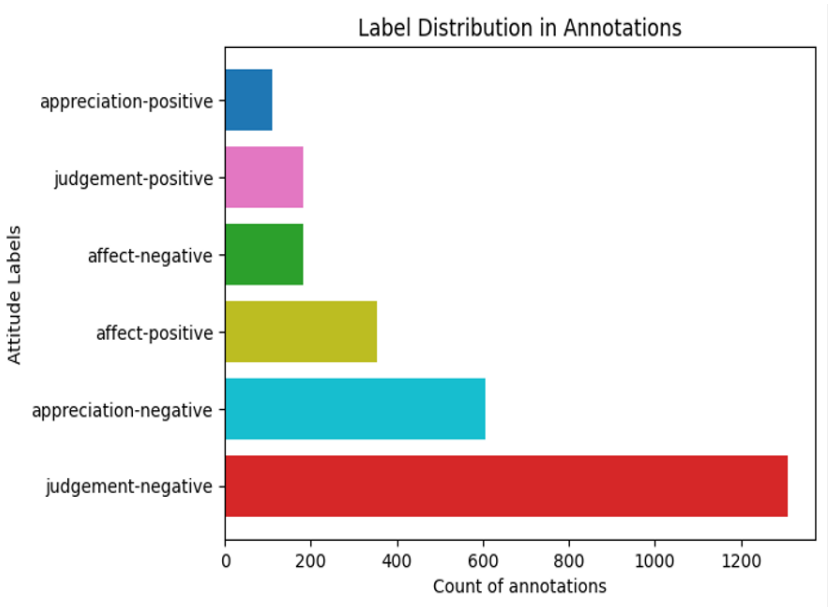

Figure 6: Distribution of *attitude* labels in the annotated speeches before removal of duplicate *EDUs*.

Subdomain	Positive Examples	Negative Examples
Affect	• "We congratulate you" • "I am absolutely confident that we will be able to do that" • "It is a special honour and a source of deep satisfaction" • "We wish to express our support and our thanks" • "We fervently hope"	• "I am afraid" • "Unfortunately" • "the hope that peace was finally taking root in the Middle East has fast evaporated" • "We are deeply disturbed" • "very disappointing"
Judgement	• "It is the job of this body to stand up for peace and to defend those in danger" • "There is no other option than peace" • "No one worked harder than the United States" • "We understand better than anyone else" • "We will work patiently and with perseverance" • "The Secretary-General and his Special Envoy have made tremendous efforts"	• "An unwanted exception" • "It would be naive" • "Security Council has failed so far to properly weigh" • "It is indeed difficult for the international community to keep pace" • "A clear threat to peace and security is posed"
Appreciation	• "*dramatic* improvement" • "*an extremely crucial* role" • "*encouraging* signs" • "*true* security" • "A *just and lasting solution*" • "A *fair, balanced and global political solution*" • "An *extraordinary* achievement" • "*Not entirely irrelevant*"	• "grave consequences' • "counter-productive consequences" • "the occupying Power" • "catastrophic loss of life" • "The draft is unbalanced and counterproductive and, quite frankly, out of touch with the reality in the region" • "Complex and treacherous international balances"

Table 15: Annotation Examples from *UNSC dataset* based on Anisimova (2023)

Label	**Before**	**After**
affect-negative	183	148
affect-positive	354	279
judgement-negative	1308	553
judgement-positive	182	177
appreciation-negative	607	338
appreciation-positive	109	109

Table 16: Difference of the annotation distribution of labels before and after removing duplicate *EDUs*.

Pattern	Attitude Label	Example
ADV+ ADJ	affect-negative	(1) Lithuania is **deeply concerned** [affect-negative] about the sharp deterioration of the situation in eastern Ukraine. (UNSC_2014_SPV.7154_spch005)
	appreciation-negative	(2) [...] and the situation is **extremely dangerous** [appreciation-negative]. (UNSC_2014_SPV.7154_spch004)
ADJ + (ADJ) + NOUN (1 or more)	appreciation-negative	(3) Involving them in the dialogue would prevent a further escalation of the **current serious crisis** [appreciation-negative], [...] (UNSC_2014_SPV.7219_spch015)
	appreciation-positive	(4) [...] all they want and ask for is a **peaceful country** [appreciation-positive]. (UNSC_2014_SPV.7154_spch009)
	judgement-negative	(5) [...] **serious crimes** [judgement-negative] against humanity, [...] (UNSC_2016_SPV.7658_spch029)
	judgement-positive	(6) I also take note of the **very important comment** [judgement-positive] [...] (UNSC_2016_SPV.7643_spch010)
it v-link ADJ that (Su and Hunston, 2019)	affect-negative	(7) **It is unfortunate that** [affect-negative] much time has been lost. (UNSC_2014_SPV.7138_spch020)
	judgement-positive	(8) **It is true that** [judgement-positive] the people in Washington, DC, [...] (UNSC_2014_SPV.7154_spch021)
it v-link ADJ to-info (Su and Hunston, 2019)	affect-negative	(9) **It is unacceptable to** [affect-negative] target international observers [...] (UNSC_2014_SPV.7165_spch010)
	appreciation-negative	(10) [...] and **it is essential to** [appreciation-positive] provide them with the necessary means [...] (UNSC_2016_SPV.7658_spch023).
NOUN NEG	affect-negative	(11) Argentina continues to follow the situation with **concern** [affect-negative], in particular in the east of Ukraine. (UNSC_2014_SPV.7154_spch013)
	appreciation-negative	(12)It has become painfully evident that the Ukraine crisis [appreciation-negative] will continue to deepen [...] (UNSC_2014_SPV.7154_spch002)
	judgement-negative	(13) [...] that the **terrorists** [judgement-negative] have at least two SA-ll Buk missile systems. (UNSC_2014_SPV.7219_spch019)
PRONOUN/NOUN + (ADV + ADJ) + VERB	affect-positive	(14) **We thank** [affect-positive] Under-Secretary-General Feltman for his message [...] (UNSC_2014_SPV.7219_spch009)
	judgement-positive	(15) **Our Government respects** [judgement-positive] freedom of expression and the right to peaceful the assembly, [...] (UNSC_2014_SPV.7165_spch019)

Table 17: Shared patterns between *attitude* labels.

Pattern	Attitude Label	Example
PRONOUN + V-link + (ADV/NEG ADV)+ ADJ	affect-negative	(1) **We are truly shocked** [affect-negative] by the number of victims among the civilian population in eastern Ukraine. (UNSC_2014_SPV.7219_spch007)
	affect-positive	(2) **I am sincerely thankful** [affect- positive] to the members of the Security Council [. . .] (UNSC_2014_SPV.7154_spch025)
PRONOUN + (ADV) + VERB	affect-negative	(3) **We deplore the** [affect-negative] loss of nearly 300 innocent civilian lives aboard that commercial airliner. (UNSC_2014_SPV.7219_spch017)
	affect-positive	(4) **We fully support** those efforts [. . .] (UNSC_2014_SPV.7154_spch005)
PRONOUN + V-modal + (ADV) + like + to-info	affect-negative	(5) **I should just like to** [affect-negative] say two things. (UNSC_2014_SPV.7154_spch027)
	affect-positive	(6) **I would also like to thank** [affect- positive] Under-Secretary-General Feltman for his briefing, [. . .] (UNSC_2014_SPV.7165_spch019)

Table 18: Shared patterns in *affect*.

Pattern	Attitude Label	Example
ADV	affect-negative	(1) Unfortunately [affect-negative], we are deeply concerned [. . .] (UNSC_2014_SPV.7165_spch019).
NOUN + VERB + ADJ + CONJ + ADJ	affect-negative	(2) China is shocked and grieved [affect-negative] by the downing of Malaysia Airlines Flight MH-17 in eastern Ukraine. (UNSC_2014_SPV.7219_spch008).
DET + NOUN + (ADV) + VERB	affect-negative	(3) My delegation strongly condemns [affect-negative] the kidnapping of unarmed OSCE military verification mission observers and Ukrainian security personnel. (UNSC_2014_SPV.7165_spch016)

Table 19: Shared patterns in negative *affect*.

Pattern	Attitude Label	Example
PRONOUN + VERB + (to-info)	affect-positive	(1) **We want to insist** [affect-positive] on that point. (UNSC_2016_SPV.7643_spch011)
PRONOUN + VERB + ADJ + that-info	affect-positive	(2) **I am proud that** [affect- positive] all future United Kingdom-hosted peacebuilding events will ensure that women's voices are heard. (UNSC_2016_SPV.7658_spch012)
PRONOUN + VERB (wish) + to + thank	affect-positive	(3) **I wish to thank you** [affect-positive], Madam President, for convening this meeting. (UNSC_2014_SPV.7154_spch005)

Table 20: Shared patterns in positive *affect*.

Pattern	Attitude Label	Example
VERB + ADV + ADJ	appreciation-positive	(1) That point **is highly relevant** [appreciation-positive] for Africa, [. . .] (UNSC_2016_SPV.7658_spch029)

Table 21: Shared patterns in *appreciation*.

Pattern	Attitude Label	Example
to+ VERB+ that-info	judgement-positive	(1) [. . .] the United Nations should rise **to playing that role** [judgement-positive]. (UNSC_2014_SPV.7138_spch008)

Table 22: Shared patterns in *judgement*.