Mining Arguments in Political Domain with Transformer Language Models

Mariia Poiaganova
University of Potsdam, Germany
poiaganova@uni-potsdam.de

Abstract

This project is focused on argumentation mining framed through the tasks of argument detection (predict whether the utterance is an argument or not) and argument component identification (predict whether the argumentative utterance is a claim or a premise). We implement BERT and RoBERTa models and approach both tasks on the sentence-level. The second task, component identification, is also modelled on the argumentative discourse unit level. To train and test our models, we use a large-scale corpus of the US presidential debates data (Haddadan et al., 2019). Additionally, we study models' generalizability to the data within the same domain. For that, we collect and manually annotate a novel dataset of diplomatic speeches presented in the United Nations Security Council.

1 Introduction

Argumentation Mining is becoming an increasingly popular research area within Natural Language Processing. It contributes to the performance of such systems as automated decision making (Bench-Capon, 2009) and decision assistance (Rinott, 2015). Argumentation mining techniques enhance language understanding and verbal reasoning in chat-bots (Galitsky, 2019).

At its core, argument mining is the computational analysis of arguments. Arguments are complex structures, comprised of several units - argument components. The central component of an argument is a claim – a controversial standpoint that needs to be justified. It is justified through a premise – the second argument component, which provides reasons why the claim is true and should be perceived as such. Minimal argumentative structures are argumentative discourse units (hereinafter, ADUs) (Peldszus and Stede, 2013), and they can span across entire sentences or more fine-grained units, e.g., conjunct clauses.

Argument mining is traditionally a process, comprised of subsequent steps: 1) argumentation filtering – finding argumentative texts among the input data; 2) unit segmentation – finding the argumentative discourse units within the argumentative text and detecting their boundaries; 3) unit type classification – distinguishing between claims and premises; 4) relation identification – finding whether premises support or attack claims.

Apart from the tasks within the traditional argument mining pipeline, several promising research directions have been emerging. For example, many recent studies are dedicated to argument generation. (Alshomary et al., 2021a) study the potential of audience-aware argument generation and (Schiller et al., 2021) of aspect-controlled argument generation. A study by (Alshomary et al., 2021b) attempts at generating counterarguments. Another developing area is argument quality assessment (Wachsmuth et al., 2017).

In our study, we remain within the scope of traditional argument mining pipeline and focus on the two tasks. The first one is argumentation filtering, which we refer to Task 1, – classifying sentences as argumentative or not. The second task is the unit type classification, Task 2, – among argumentative sentences or discourse units, detect whether the input is a claim or a premise.

In-domain and Cross-domain Argument Mining. Another relevant direction of argumentation mining research is finding the training corpus optimal for generalizing to different domains. It is a challenging task because argument conceptualization proves to vary across different domains (Daxenberger et al., 2017). At the same time, creating large manually annotated datasets to adapt for every possible domain (e.g., legal documents,

academical papers, student essays etc.) is a time-consuming and intellectually demanding process, and, thus, a consensual training corpus has to be found. In our paper, we aim at investigating scenario when the training and testing corpora are of slightly different genres but are within the same domain. Globally, we focus on political speeches, but take two different types of those: presidential debates and diplomatic speeches in the United Nations Security Council (hereinafter, UNSC). Despite both are representing political discourse, presidential debates and diplomatic speeches exhibit certain functional and structural differences. Presidential debates aim at persuading the audience to take the speaker's stance towards a certain, typically, social or political issue. Debates play a significant role in promoting the candidate's program and shaping the politician's image. Diplomatic speeches, while also aimed at persuasion, express collective opinion, the nation's positions regarding current problems in the world and contribute to the nation's public image. Another difference lies in the fact that diplomatic speeches are prepared monologues, while presidential debates, even if also partially pre-written, allow for some spontaneity due to their dialogical nature. From the thematic angle, a single debate round would typically cover a broader range of topics than a diplomatic meeting, which usually focuses on a specific problem, especially in the setting of the United Nations Security Council gatherings. Thus, given such differences, we might expect different conceptualization of argumentation within these two types of political speeches, and our goal is to understand whether one type of political discourse data is enough to make robust predictions on the other type.

Overall framework of the research project can be described through three research objectives. First, we are aim at solving argument filtering and component identification tasks with transformer language models using data from the political domain. Transfer learning has shown to exhibit state-of-the-art performance in multiple Natural Language Processing tasks, and, to the best of our knowledge, performance of such models on the two tasks in question has not been reported on the USElecDeb corpus (Haddadan et al., 2019). Second, we aim at understanding the most optimal textual span for the component identification task by comparing models' performance given the input on the sentence-level and ADU-level. Third, we investigate the generalizability potential of such models on the different type of data yet from the same domain. We do so by introducing a novel corpus of diplomatic speeches, manually annotated with respect to their argumentative structure.

2 Related Literature

Recent developments in argument mining in political discourse have resulted in several annotated corpora and the researchers achieved promising results on different tasks. For example, (Lippi and Torroni, 2016) collect an original dataset based on the 2015 UK political elections debates. They combine features from texts and speech to test the potential improvement of claim detection by expanding the feature set to the spoken language dimension.

Another example is the work of (Vissler et al., 2019). The authors build US2016, the corpus of transcriptions of television debates leading up to the 2016 US presidential elections, and reactions to the debates on Reddit.

Finally, (Haddadan et al., 2019) present a large-scale corpus of presidential debates held in the United States and annotate the units of speeches as claims and premises. They frame their experiments as two classification tasks: argument filtering and argument component classification. On the methodological side, they test various models, including Support Vector Machines (SVM) and a Feed-Forward Neural Network using linguistic features for data representation. They also implement FastText word embeddings (Joulin et al., 2016) fed into Long Short-Term Memory neural model (Hochreiter and Schmidhuber, 1997) and report their best results on this model. Average f1-scores for the argument filtering and component detection tasks achieve 0.84 and 0.67 respectively. Our research largely relies on their work, what is reflected in using USElecDeb corpus for training and in analogically framing our research tasks.

Regarding cross-domain experiments, several studies attempted at finding an optimal training corpus for accurate argument detection and classification. For example, (Schäfer et al., 2022) study claim detection in a cross-domain setting. They use four corpora, all varying in genres and sizes, and experiment with different train-test scenarios, depending on which corpora are used or left out.

They conclude that a large training sample, homogeneous claim proportions, and less formal language might be the characteristics of the "universal" training corpora.

3 Data

3.1 Presidential Debates Corpus

To train our models, we use USElecDeb16To60 v.01 (we later refer to it as USElecDeb) corpus introduced in (Haddadan et al., 2019). It is a collection of debate transcripts from past presidential and vice-presidential debates held in the United Stated. The speeches were originally retrieved from the website of the Commission on Presidential Debates[1]. The corpus includes debate transcripts dating from years 1960 to 2016, covering 39 debate transcripts in total.

The debates are annotated according to their argumentative structure and contain two types of labels – claims and premises. Three expert annotators independently conducted the annotations of debates, relying on the guidelines handbook created by other expert annotators. The observed agreement percentage and inter-annotator agreement at sentence-level for argumentative - non- argumentative sentences are 0.83% and $\varkappa = 0.57$ respectively. For the argument component they are 63% and $\varkappa = 0.4$. Table 1 shows the distribution of argument components in the final version of the corpus.

Total	Argument	Non-Argument	Claim	Premise
29.621	22.280	7.252	11.964	10.316

Table 1: Classes distribution in the dataset, displayed in the number of sentences

Table 2 shows the original splits of the corpora, as provided by the authors. We use these original sets to train and test our models.

	Train	Test	Validation
Task 1	14.044	8.455	7.033
Task 2	10.464	6.575	5.241

Table 2: Data splits statistics, displayed in number of sentences

Arguments in political debates. This subsection describes corpus annotations. Certain examples of claims and premises are provided. Claims are marked in **bold**, premises in *Italics* and the component boundaries are defined by [square brackets].

Claims. As defined in (Haddadan et al., 2019), applied to the political debates discourse, a claim can be:
- a policy advocated by a candidate;
- a candidate's stance towards a policy;
- a candidate's opinion regarding a certain issue;
- a candidate's personal judgement.

Examples below illustrate claims in the presidential debates.

(1) Nixon-Kennedy, October 13, 1960:
Nixon: **[I favor higher salaries for teachers]**.

(2) Nixon-Kennedy, October 13, 1960:
Nixon: **[Senator Kennedy's position and mine completely different on this]**. **[I favor the present depletion allowance]**. **[I favor it]** *[not because I want to make a lot of oil men rich]*, but because *[I want to make America rich]*.

(3) Clinton-Bush-Perot, October 11, 1992:
Clinton: **[I've worked hard to create good jobs and to educate people]**. *[My state now ranks first in the country in job growth this year, fourth in income growth, fourth in reduction of poverty, third in overall economic performance, according to a major news magazine]*.

Premises. In order to support or attack a claim, politicians provide reasons or justifications for their judgements. Premises can refer to precise events, include data on a certain issue, give an example of a formerly adopted policy and its consequences. Sometimes candidates can use examples from the past as in (4), where the candidate aims at proving the efficiency of his office, by describing some positive developments which took place during his presidency. Some language markers (e.g., "because", "for example", "since") might appear in premises.

(4) Obama-Romney, 2012:
Obama: **[America remains the one indispensable nation]. [And the world needs a strong America]**, and **[it is stronger now than when I came into office]**. *[Because we ended the war in Iraq, we were able to refocus our attention on not*

[1] https://www.debates.org/

only the terrorist threat, but also beginning a transition process in Afghanistan]. [It also allowed us to refocus on alliances and relationships that had been neglected for a decade].

3.2 Novel UN-UNSC Corpus

To study the question of models' generalizability within the same domain, we manually annotate a novel dataset, comprised of diplomatic speeches given at the United Nations Security Council. UNSC is an organ of the United Nations, responsible for maintaining international peace and security. The council meets whenever peace is threatened. We retrieve texts of the speeches delivered during such meetings from the Digital Library of the United Nations. The archive[2] includes all speeches presented in the Council since 1946, and it allows to select speeches according to their topic.

For our purposes, we select speeches covering the military conflict in Ukraine which initiated in 2014. Political conflict discourse is another natural application scenario for argumentation mining as it provides statements supporting or condemning the actions of the conflict's parties, along with assertions regarding expected future developments. We refer to a new corpus as "UC-UNSC", with UC standing for "Ukrainian Conflict".

We retrieve speeches spanning from years 2014 to 2018. Within every speech, we manually label argumentative units as claims or premises. The dataset contains 144 speeches in total given by representatives of 24 different countries. Tables 3 and 4 show speeches' distribution per year and per country.

Year	Speeches
2014	93
2015	27
2016	11
2017	7
2018	6

Table 3: UC-UNSC corpus statistics: number of speeches per year

Since 2014 was the year of the conflict onset, most of discussions were held during that year.

Being two main conflicting sides, Russia and Ukraine dominate in the number of speeches.

[2]https://digitallibrary.un.org/?ln=en

Country	Speeches
Russia	25
Ukraine	16
United States	15
United Kingdom	11
France	11
China	11
Lithuania	8
Australia	7
Rwanda	6
The Republic of Korea	6
Luxembourg	5
Argentina	4
Chile	4
Nigeria	3
Jordan	2
Sweden	1
Ethiopia	1
Angola	1
Belgium	1
New Zealand	1
Venezuela	1
Spain	1
Chad	1
Indonesia	1
UNSC Briefing	1

Table 4: UC-UNSC corpus statistics: number of speeches per country

Sentence- and component-level statistics of the dataset are represented in Tables 5 and 6. By component we mean a single independent argumentative unit. We consider a sentence argumentative if it contains at least one claim or premise component. Like in the corpus of (Haddadan et al., 2019), we observe that the number of claims prevails the number of premises, which is usually not typical for argument corpora, but is often the case for political discourse. It is because sometimes, especially in shorter talks, speakers do not provide premises to justify their claims.

Annotating argument components. Initially, we relied on the guidelines provided by the authors of USElecDeb to detect argumentative components in the political discourse. Yet, as the genres are slightly different, we met several types of

Total	Argument	Non-Argument	Claim	Premise
4.751	3.814	937	2.077	1.737

Table 5: Classes distribution in the dataset on the sentence-level, displayed in the number of sentences

Total	Claim	Premise
4.103	2.239	1.864

Table 6: Classes distribution in the dataset on the component-level, displayed in the number of ADUs

arguments specific to our data. The section below details argument annotations as applied to diplomatic speeches in the UNSC and provides several examples of those. All examples are marked in the same manner: claims are in **bold**, premises are in *italics*, brackets indicate the boundaries.

Claims. In the diplomatic speeches, where the thematic focus is a military conflict, claims might be the nation's interpretations and evaluations of the current situation. In the example (5), the representative of France first claims that Russia did not follow a protocol it agreed upon, and then provides a list of reasons why this is the case, to support their first judgement.

(5) France, 2014:

On the other hand, **[the Russian side has complied with none of the 17 April commitments]**. *[There has been no condemnation of the separatist actions that have spawned new violence and no call for public buildings to be evacuated]. [There has been no appeal to the pro-Russian militants to exercise restraint and end their attacks on munitions depots and on their compatriots...]*

Another possible form of claims in the diplomatic speeches is the nation's stance towards the conflicting party's policies and activities, as in example (6).

(6) The Republic of Korea, 2014:

[We strongly condemn the detention of military monitors of the Organization for Security and Cooperation in Europe (OSCE), as well as of Ukrainian staff by illegal armed groups].

In addition, claims might be represented as the nation's expectations towards further development of the conflict. Such claims can appear in a standalone manner and not necessarily be supported by premises, as in (7). At the same time, they can appear in a group of consequent claims, advocating for the same main idea, but covering it from different angles, as in (8).

(7) The Republic of Korea, 2014:

[All provocative actions and hostile rhetoric aimed at destabilizing Ukraine must cease immediately].

(8) Lithuania, 2014:

[The safety of the international observers deployed across Ukraine must be guaranteed by all of the parties]. **[We take this opportunity to reiterate our strong condemnation of the kidnapping of a team of military inspectors deployed under the OSCE 2011 Vienna Document]**. **[We reiterate our call on Russia to continue using all of its influence on the pro-Russian separatists to free, unconditionally and without delay, the seven monitors from OSCE participating States, whom they have been detaining in Sloviansk for one week now, as well as the Ukrainian personnel accompanying them]**.

We observe that claims can come along with certain linguistic means. They can take the form of complement clauses as "we think", "we believe" or be in the form of discourse connectives as "thus", "as a result", "this is why". In diplomatic speeches of our corpus, the presence of such modal verbs as "must" and "should" might indicate the country's position towards a situation and the expectations towards other party's policies. Utterances starting with "we condemn", "we call for", "we reiterate our position" are likely to be claims as well.

Premises. The most common type of premises typical for the conflict-related diplomatic speeches include references to some events or documents. Such statements can feature dates, participants, describe precise actions, consequences, including, e.g., number of victims, number and types of weapons used. Examples (9), (10) and (11) are representative of premises. More rarely, like in the example (12), premises can be preceded by an interrogative structure, initiating the reasoning

and justifications on a certain issue.

(9) United States, 2014:

Since 17 April, **[the Government of Ukraine has acted in good faith and with admirable restraint to fulfil its commitments]**. *[The Kyiv city hall and its surrounding area are now clear of all Maidan barricades and protestors]. [Over the Easter holiday, Ukraine voluntarily suspended its counter-terrorism initiative, choosing to de-escalate despite its fundamental right to provide security on its own territory and for its own people]. [Unlike the separatists, Ukraine has cooperated fully with the OSCE special monitoring mission and allowed its observers to operate in regions about which Moscow had voiced concerns regarding the treatment of ethnic Russians].*

(10) United Kingdom, 2014:

[The situation in eastern Ukraine has continued to deteriorate]. *[Armed groups stormed the Prosecutor's office in Donetsk yesterday, further increasing the number of Government buildings occupied since the 17 April Geneva agreement]. [We remain seriously concerned about the kidnapping and continued detention of the Organization for Security and Cooperation in Europe's Vienna Document inspectors. . .]*

(11) Ukraine, 2015:

As of today, **[the Russian Federation is continuing its military aggression in the Donetsk and Luhansk regions of Ukraine]** *[by sending military units into our territory, delivering heavy weapons to the local terrorist groups, training, equipping and financing mercenaries and waging an information war].*

(12) Ukraine, 2015:

On mobilization, yes, **[we are in the process of reforming our army, which was fully destroyed in recent years]**. Why are we doing that? Because of the facts expressed today - *[the enlargement of the Russian military presence in in Donbas, Ukraine, with thousands of Russian nationals and sophisticated weaponry]. [We have a right to defend ourselves]*. **[That is why we are doing so, ourselves]**.

We also observe several linguistic indicators of premises. We agree with the list provided by the authors of the USElecDeb and state that such means as "because", "as" and "for example" might indicate premises. We expand this list with respect to our data and add the following markers: "this is the reason why. . . ", "the report shows. . . ", "as reported/stated. . . ". We note that, however, even though certain lexical means might signal premises, their presence does not necessarily mean that the unit is a premise and, it is more important to consider the paragraph as a whole rather than simply rely on linguistic markers.

4 Problem Setting

We approach both tasks as binary classification problems. In the Task 1, we classify utterances as argumentative or non-argumentative, and in the Task 2, we classify utterances as claims or premises. We first perform classification on the sentence level. The sentence is considered argumentative if it includes at least one argumentative unit – a claim or a premise. In fact, both in the presidential debates' dataset and in our corpus, some sentences contained a claim and a premise at the same time. For such cases, following the practice of (Haddadan et al., 2019), we label the sentence according to its longer component, the length is computed in the number of symbols. Sentence-level approach is rather simplistic and suggests the need for a more fine-grained segmentation. Thus, we additionally approach task 2 on the ADU-level. For this, we use the original labelled sequences from both corpora.

To evaluate the models, we use f1-measure and compute it per each class along with the weighted average score.

5 Experimental Setting

We fine-tune different implementations of BERT and RoBERTa for our tasks. For both Task 1 and Task 2, we experiment with *bert-base-uncased* model. For the Task 1, we make the use of *roberta-argument*, which is *roberta-base* fine-tuned on approximately 25.000 sentences labelled as argumentative or not. (Stab et al., 2018). The corpus contains texts on different controversial topics, including abortion, school uniforms, death penalty, marijuana legalization, nuclear energy, cloning, gun control, and minimum wage. For the Task 2 experiments, distinguishing between claims and premises, we fine-tune stan-

Task	Class	Majority Baseline	LSTM	BERT	RoBERTa
Task 1	Argument	0.810	0.902	0.907	0.912
	None	0.000	0.524	0.617	0.613
	Average	0.551	0.818	0.842	0.846
Task 2	Claim	0.680	0.658	0.723	0.711
	Premise	0.000	0.654	0.625	0.675
	Average	0.350	0.656	0.675	0.693

Table 7: f1-scores for Task 1 and Task 2, classification of argumentative and non-argumentative sentences. Test set: USElecDeb

Task	Class	Majority Baseline	BERT	RoBERTa
Task 2	Claim	0.715	0.761	0.758
	Premise	0.000	0.682	0.690
	Average	0.398	0.726	0.728

Table 8: f1-scores for Task 2, classification of claims and premises on ADU-level. Test set: USElecDeb

dard *roberta-base*. Both BERT and RoBERTa implementations include 12 encoder layers, and each layer is a transformer layer with 12 attention heads. It results in 144 separate attention mechanisms. At the top of the last hidden state, there is a feed-forward layer with softmax activation function. The total number of parameters for both BERT and RoBERTa is 110M.

As advised in the original paper (Devlin et al., 2019), we fine-tune the models on two to four epochs, and observe that two epochs are optimal for all the settings. To facilitate our results' reproduction, we provide exemplar notebook in the GitHub repository along with details on model and hyperparameter settings for each experiment. The models are implemented using bert-for-sequence-classification – a simple open-source framework for fine-tuning BERT-like models on the sequence classification tasks[3].

Every model is trained and evaluated with train and validation sets, while test sets are left out only for prediction. For both tasks, we train the models on ElecDeb corpus and test them, first, on the original test set provided by the authors and, second, on the full UC-UNSC dataset. For the Task 2 on the ADU-level, we manually split the data into train, validation and test sets and stratify them in order to keep the same class distributions.

[3]https://pypi.org/project/bert-for-sequence-classification/

6 Results and Discussion

6.1 Sentence-level Classification: Task 1 and Task 2

Our first results depict sentence-level classification of argumentative vs. non-argumentative sentences as well as sentence-level classification of claims vs. premises, both trained and tested on the USElecDeb corpus. Table 7 shows the achieved scores. We compare the BERT and RoBERTa results against majority baseline - a model which would always classify every instance as a class with the higher distribution, i.e., argumentative class for the Task 1 and claim class for the Task 2. Since LSTM implementation was the best reported result on the two tasks for the corpus in question, we also compare the models against our previous implementation of LSTM [4].

For the Task 1, the results show that both BERT and RoBERTa-Argument significantly outperform majority baseline and outperform LSTM for up to 3%. They show particularly prominent improvement (up to 9%) over LSTM in guessing the non-argumentative class. We expected RoBERTa implementation to show better performance compared to BERT since it was pre-trained on the argumentative corpus. Despite that, the difference between BERT and RoBERTa-Argument seems to be only marginal.

For the Task 2, both models again achieve higher scores compared to LSTM. On this task,

[4]https://github.com/a-moi/Argument-Mining-Presidential-Debates

Task	Class	Majority Baseline	BERT	RoBERTa
Task 1	Argument	0.891	0.906	0.912
	None	0.000	0.484	0.412
	Average	0.715	0.823	0.813
Task 2	Claim	0.705	0.761	0.758
	Premise	0.000	0.693	0.717
	Average	0.384	0.730	0.739

Table 9: f1-scores for tasks 1 and 2, classification on a sentence-level. Test set: UC-UNSC

Task	Class	Majority Baseline	BERT	RoBERTa
Task 2	Claim	0.710	0.723	0.714
	Premise	0.000	0.707	0.725
	Average	0.390	0.716	0.719

Table 10: f1-scores for task 2, classification on ADU-level. Test set: UC-UNSC

RoBERTa implementation is slightly better than BERT, especially when predicting premises, the score increase is 5% compared to BERT.

6.2 ADU-level Classification: Task 2

Table 8 represents results of the classification of claims and premises on ADU-level. The weighted average scores are higher compared to the results of the same task on the sentence-level even when taking into account slightly different distributions of classes in the two test samples (majority baseline for claim class is 0.680 on the sentence-level task and 0.715 on the ADU-level task). Both claim and premise classes are better distinguished on the ADU-level. These results might be the initial indication that sentences are too large for argument type classification and the future approaches should adopt a more fine-grained fragmentation. Comparing BERT and RoBERTa, we conclude that the models exhibit nearly the same performance.

6.3 Testing on the UC-UNSC Corpus

We save all the previously reported models and test them on the novel UC-UNSC Corpus instead of the test set from USElecDeb Corpus. Table 9 shows results for Task 1 and Task 2, classification on a sentence-level.

For the Task 1, weighted average results on the diplomatic speeches from the UNSC are lower compared to the "full in-domain" results. The differences are especially significant for the non-argumentative class, where the f1-score falls from 0.617 to 0.484 on BERT and from 0.613 to 0.412 on RoBERTa-Argument. Note that, also, even though the weighted average drops just slightly, the majority baseline again points out to different class distributions in the two sets. Thus, in the USElecDeb test set, the overall performance is almost 30% higher compared to the average baseline score, while in the UC-UNSC test set, the average score is only 11% higher. This suggests that functional and linguistic differences in speeches' characteristics effect the models' generalizability. It also might indicate that the task of finding a "universal" training corpus for argument classification is complex, since even within close genres of the same domain the results vary drastically. However, such low results might also be due to the limitations in proposed annotations. This is a pilot study of a newly developed corpus, and for the moment, only one annotator participated in the process. Argumentation structure annotation task is known to be complex even for experts, and inter-annotator agreement is typically moderate. Thus, we intend to involve more annotators in the future and verify our results on the updated corpus annotations.

For the Task 2 on the sentence-level, argument component identification, which is traditionally considered to be more complex than argument filtering, we get results higher than those in USElecDeb test set, with RoBERTa outperforming the majority baseline for almost 36%, compared to 34% on the original test set. From the angle of corpora interpretations, this might indicate that claims and premises keep their conceptualizations across two genres, while the non-argumentative

sentences exhibit different characteristics (as was seen due to lower results on the Task 1).

The prediction pattern remains consistent also in the Task 2 on the ADU-level, where both BERT and RoBERTa yield scores very close to those of the Task 2 on the sentence-level. The average scores are 0.716 and 0.719 for the two models respectively, compared to 0.726 and 0.728 on the original test set. We also observe that the claim detection drops on the UC-UNSC corpus, while the premise detection slightly increases.

6.4 Linguistic Error Analysis

We inspect model's incorrect predictions on the UC-UNSC dataset on both classification tasks on the sentence-level. To this end, on each task, we manually retrieve thirty false positives (sentences incorrectly predicted as argumentative on task 1 and as claims on task 2) and thirty false negatives (sentences incorrectly predicted as non-argumentative on task 1 and as premises on task 2), all randomly selected. We use predictions of BERT implementation for the task 1 error analysis and RoBERTa for the task 2, as those were more accurate in their respective tasks.

Inspecting false negative sentences in the task 1, we observe that five out of thirty examples are sentences the meaning of which can only be captured inside the context (e.g, "Enough is enough."), and outside of the speech, such sentences would not necessarily be considered argumentative. Another interesting observation is a large number of interrogative sentences among those incorrectly classified as non-arguments. The proportion of such sentences in the whole false negative sample is 0.26, compared to 0.02 in the entire test set. Arguments stated in the interrogative form are typically rare and very context dependent, hence, given not enough examples in the training set, the model might have not captured the pattern. In the false positive sample, 12 out of 30 examples are sentences mentioning precise events and their respective dates, related legal documents, numbers of victims, etc. Outside of context, these could easily be premises, but were not labelled as such as they were reported outside of an argument and did not lead to any conclusion, i.e., claim.

Regarding errors in the task 2, four out of thirty incorrectly classified claims were utterances featuring "premise"-like units while being a claim by their communicative function. For example, the sentence "We strongly condemn the kidnapping near the town of Sloviansk on 25 April of a team of military inspectors deployed under the 2011 Vienna document of the Organization for Security and Cooperation in Europe (OSCE)" states the country representative's stance towards an event – their condemnation of it. But the textual unit detailing the circumstances of the kidnapping contains linguistic markers typical to premises. As for the premises classified as claims, about half of the sample examples represent hard annotation cases. Those are the cases when a sentence could theoretically serve as both a claim and a premise, but the final judgement was largely depending on the context. A sentence "Rwanda remains of the view that military action will only worsen the already tense situation." could be a claim as it represents a county's opinion towards the future course of events. However, in the speech, it is followed by "We encourage all the parties involved, particularly in the current situation, to exercise full restraint.", which makes us interpret the second sentence as a claim and the first one as a premise, meaning that Rwanda believes that that countries should exercise full restraint because the opposite would only worsen the already tense situations. Also, out of sixty false positive and false negative combined instances, seven sentences contain both a premise and a claim. This reflects the limitation of sentence-level classification approach, and such double-labeled sentences might indeed confuse the model.

7 Conclusion

In this project, we addressed the problem of argument mining and looked at it through two classification tasks. The first task implied argument detection on a sentence-level, while the second task was dedicated to classification of argument components - claims and premises, also on a sentence-level. We made the use of BERT and RoBERTa implementations on both tasks and achieved, to the best of our knowledge, the highest results on the USElecDeb corpus. To mitigate negative effects of sentence-level component identification task (caused by the fact that sometimes both components might be in a sentence),we repeated the second task experiments on the argumentative discourse unit level and did so by taking the original labelled fragments. We observed the first evidence that such more fine-grained fragmentation might

be beneficial for the argumentative unit classification.

Additionally, we studied the question of models' generalizability to a corpus of different genre yet within the same political domain. These cross-corpora experiments suggest that argument detection is a complicated task when it comes to predicting, even on the same-domain data. At the same time, component detection models have shown to generalize well.

In the future, we aim at continuing the UC-UNSC corpus development. We plan to increase the number of annotators and, thus, to provide more reliable annotations. We also intend to expand the annotations towards argument component relations and create labels on whether a premise supports or attacks a claim. Such annotations would allow for more advanced argumentation representations and modelling.

References

Milad Alshomary and Henning Wachsmuth. Toward audience-aware argument generation. Patterns (New York, N.Y.) vol. 2,6 100253. 11 Jun. 2021, doi:10.1016/j.patter.2021.100253

Milad Alshomary, Shahbaz Syed, Martin Potthast and Henning Wachsmuth. "Argument Undermining: Counter-Argument Generation by Attacking Weak Premises." ArXiv abs/2105.11752 (2021): n. pag.

Trevor Bench-Capon, Katie Atkinson, and Peter McBurney. 2009. Altruism and Agents: An Argumentation Based Approach to Designing Agent Decision Mechanisms. In Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems - Volume 2, AAMAS 2009, pages 1073–1080.

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.

Boris Galitsky. Enabling a Bot with Understanding Argumentation and Providing Arguments. Developing Enterprise Chatbots, 2019, p. 465-532

Johannes Daxenberger, Steffen Eger, Ivan Habernal, Christian Stab, and Iryna Gurevych. 2017. What is the Essence of a Claim? Cross-Domain Claim Identification. In Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pages 2055–2066, Copenhagen, Denmark. Association for Computational Linguistics.

Shohreh Haddadan, Elena Cabrio, and Serena Villata. 2019. Yes, we can! Mining Arguments in 50 Years of US Presidential Campaign Debates. Proceedings Of The 57Th Annual Meeting Of The Association For Computational Linguistics. https://doi.org/10.18653/v1/p19-1463

Sepp Hochreiter, and Jürgen Schmidhuber. 1997. Long Short-Term Memory. Neural Computation, 9(8), 1735-1780. https://doi.org/10.1162/neco.1997.9.8.1735

Armand Joulin, Edouard Grave, Piotr Bojanowski, Matthijs Douze, Hervé Jégou and Tomás Mikolov. 2016. FastText.zip: Compressing text classification models. arXiv preprint arXiv:1612.03651.

Marco Lippi and Paolo Torroni. 2016. Argument mining from speech: Detecting claims in political debates. In Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence, February 12-17, 2016, Phoenix, Arizona, USA, pages 2979–2985.

Andreas Peldszus and Manfred Stede. 2013. From argument diagrams to argumentation mining in texts: A survey. International Journal of Cognitive Informatics and Natural Intelligence (IJCINI), 7(1):1–31.

Ruty Rinott, Lena Dankin, Carlos Alzate Perez, M. Mitesh Khapra, Ehud Aharoni, and Noam Slonim. Show Me Your Evidence — An Automatic Method for Context Dependent Evidence Detection. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, pages 440–450, 2015.

Robin Schaefer, René Knaebel, and Manfred Stede. 2022. On Selecting Training Corpora for Cross-Domain Claim Detection. In Proceedings of the 9th Workshop on Argument Mining, pages 181–186, Online and in Gyeongju, Republic of Korea. International Conference on Computational Linguistics.

Benjamin Schiller, Johannes Daxenberger, and Iryna Gurevych. 2021. Aspect-Controlled Neural Argument Generation. In Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages 380–396, Online. Association for Computational Linguistics.

Christian Stab and Iryna Gurevych. 2014. Annotating argument components and relations in persuasive essays. In COLING 2014, 25th International Conference on Computational Linguistics, Proceedings of the Conference: Technical Papers, August 23-29, 2014, Dublin, Ireland, pages 1501–1510. ACL.

Christian Stab, Tristan Miller, Benjamin Schiller, Pranav Rai, and Iryna Gurevych. 2018. Cross-topic Argument Mining from Heterogeneous Sources. In Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing,

pages 3664–3674, Brussels, Belgium. Association for Computational Linguistics.

Jacky Visser, Barbara Konat, Rory Duthie, Marcin Koszowy, Katarzyna Budzynska, and Chris Reed. 2019. Argumentation in the 2016 us presidential elections: annotated corpora of television debates and social media reaction. Language Resources and Evaluation.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational Argumentation Quality Assessment in Natural Language. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers, pages 176–187, Valencia, Spain. Association for Computational Linguistics.