Sentiment Analysis and Argumentation Mining in United Nations Security Council Speeches

Juliane Hanel, Atreya Shankar
{hanel,shankar}@uni-potsdam.de
PM: Mining Sentiments and Arguments, WiSe 2019/20
Prof. Dr. Manfred Stede
Applied Computational Linguistics
University of Potsdam

April 5, 2020

Abstract

The United Nations Security Council (UNSC) corpus from Schönfeld et al. (2019) is a novel political speech corpus containing ∼65,000 textual speech records from ∼5,000 security council meetings over the years of 1995-2017. Due to its large size and wide temporal distribution, it could be a very useful corpus for various Natural Language Processing (NLP) tasks in the political domain. One limitation of the current version of this corpus is its lack of extensive handwritten annotations; which diminish its utility for downstream supervised NLP tasks. In order to address this limitation, our project aims to evaluate and provide machine-driven sentiment and argumentation annotations for this corpus. We utilize state-of-the-art sentiment analysis and argumentation mining tools for this purpose, which include `VADER`, `AFINN`, and `TextBlob` for sentiment analysis and a political-domain fine-tuned version of the `ALBERT` language model for argumentation mining. Our automatic annotations, while not being as reliable as human annotations, nevertheless provide an initial foothold for future human annotations to follow-through.

Contents

1 Introduction

The United Nations Security Council (UNSC) corpus, detailed in Schönfeld et al. (2019), is a novel political speech corpus containing 65,393 textual speech records from 4,460 security council meetings over the years of 1995-2017. Due to its large size and wide temporal distribution, it could be a very useful corpus for various Natural Language Processing (NLP) tasks in the political domain.

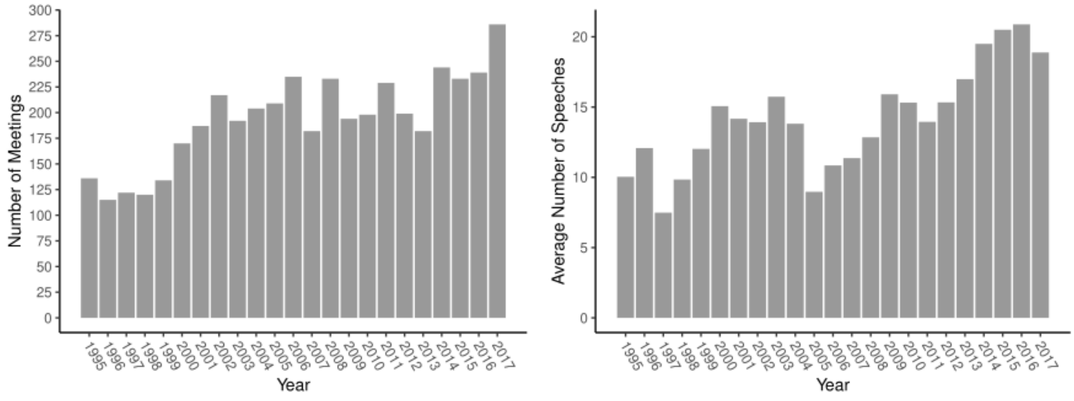
One limitation of the current version of this corpus is its lack of extensive hand-written annotations (Schönfeld et al., 2019); which diminish its utility for down-stream supervised NLP tasks. In order to address this limitation, our project aims to evaluate and provide machine-driven sentiment and argumentation annotations for this corpus. We utilize state-of-the-art sentiment analysis and argumentation mining tools for this purpose, which include `VADER`, `AFINN`, and `TextBlob` for sentiment analysis and a political-domain fine-tuned version of the `ALBERT` language model for argumentation mining. Our automatic annotations, while not being as reliable as human annotations, nevertheless provide an initial foothold for future human annotations to follow-through.

In section 2, we describe and define various background concepts that are used for the methodologies in this paper. In section 3, we describe the various methodologies implemented in this study. In sections 4 and 5, we describe our results and discuss their implications. Lastly, we conclude this study in section 6 and provide some recommendations for future work in section 7.

2 Background Concepts

2.1 United Nations Security Council (UNSC) Corpus

As the UNSC corpus was already briefly described in the introduction, we will proceed to describe some key statistics regarding this corpus. As mentioned, this corpus contains textual data from 65,393 speeches which were collected from 1995-2017. In Figure 1, we can observe how the number of speeches and meetings vary temporally. We can, for example, notice how there was a peak in the number of speeches per meeting in 2003; which corresponds to the onset of the Iraq war.



(a) Numbers of meetings per year

(b) Average number of speeches per meeting

Fig. 1. Temporal distribution of the number of speeches and meetings in the UNSC corpus (Schönfeld et al., 2019)

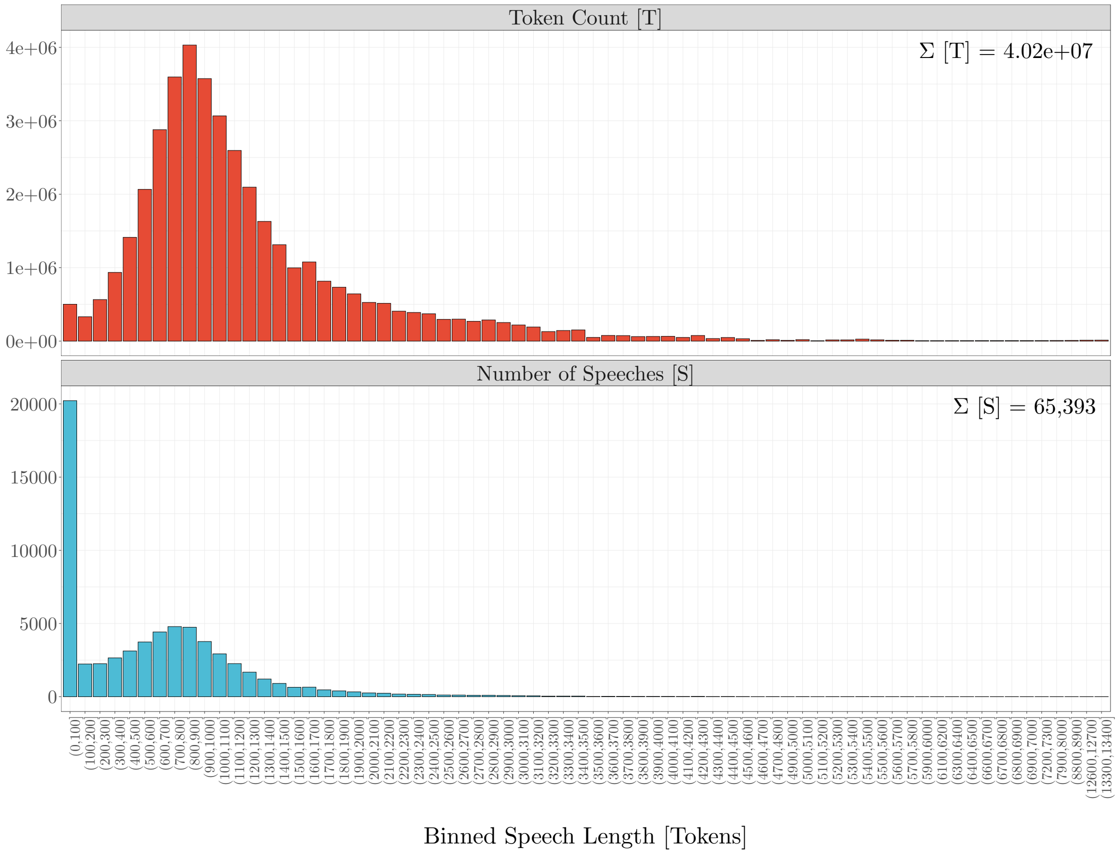


Fig. 2. Token count (top; red) and number of speeches (bottom; blue) by binned speech length in the UNSC corpus

To get a better idea of the token statistics of the UNSC speeches, we conducted a simple tokenization and binning procedure on the corpus. Here, we removed certain redundant parts of each speech, such as the initial tokens describing which speaker is speaking. Then we tokenized the left-over speeches using the `nltk` python module. After this, we binned all of the speeches by their token lengths into intervals of 100 tokens, and plotted the number of tokens and speeches in each binned category as shown in Figure 2.

Here, we can observe that the number of speeches peaks at the binned speech length category of `(0,100]`, while the number of tokens peaks at the binned speech length category of `(800,900]` tokens. This tells us that most of the UNSC speeches are generally short and have a length less than or equal to 100 tokens, but the majority of tokens are found in longer length speeches between 800 and 900 tokens. This information will come into further use in later parts of this paper.

2.2 Sentiment Analysis

According to Liu (2012):

> "Sentiment analysis and opinion mining is the field of study that analyzes people's opinions, sentiments, evaluations, attitudes, and emotions from written language. It is one of the most active research areas in natural language processing and is also widely studied in data mining, Web mining, and text mining."

Sentiment analysis can be used to detect various sentiment polarities, for example detecting if a given utterance is positive/negative or subjective/objective (Liu, 2012). In such cases, the sentiment value of an utterance typically takes on a real value from -1 to $+1$, where each extremum indicates one of the polarities of interest. Furthermore, sentiment analysis can be conducted on various granularities or scales (Liu, 2012). These can range from a sentence level to a document level and can include various user-defined aspects. For example, given a certain online product review, sentiment analysis could decipher the sentiment polarity towards aspects such as service, product quality and delivery time.

In this paper, we will generally be dealing with sentiment analyses of entire UNSC speeches. This would therefore correspond to sentiment analysis on a document level. In the next subsections, we will proceed to describe some commonly used automatic sentiment analysis tools.

2.2.1 VADER

The Valence Aware Dictionary for Sentiment Reasoning (VADER) is a simple rule-based model for general sentiment analysis of social media text (Hutto and Gilbert, 2015). VADER utilizes a generalized valence-based and human-curated sentiment lexicon to score utterances for their sentiment polarity. Since it was created from handcrafted rules, the VADER sentiment analysis system does not have training data. However, it has been evaluated on various sentiment datasets and shows high classification accuracy; as well as a strong correlation to human-annotator performance (Hutto and Gilbert, 2015).

VADER has different scoring regimes for sentiment polarities of utterances. Table 1 gives a summary of the various VADER sentiment scores.

Scores	Description
Positive Negative Neutral	All three of these scores refer to the proportion of the input text that is positive, negative or neutral. Ultimately, the sum of the positive, negative and neutral scores must add up to 1.
Compound	The compound score is computed by adding the polarity scores of each word in the sentiment lexicon, adjusted according to the rules, and then normalized between -1 (most extreme negative) and $+1$ (most extreme positive). One can refer to the compound score as a normalized and weighted composite sentiment score.

Table 1. Tabular summary of VADER sentiment scores

One limitation of VADER is that it was constructed to perform on "microblog-like" contexts found typically on social media platforms such as *Twitter*. However, Hutto and Gilbert (2015) do state that VADER is a robust tool and can perform well on diverse domains. For this reason, we consider VADER as a feasible sentiment analysis tool for the UNSC corpus.

2.2.2 AFINN

AFINN is a widely used multilingual lexicon containing words and their valence, manually rated between -5 and +5. For sentence-wise rating, the valence of the words is simply aggregated. The latest version of the English AFINN lexicon[1] contains more than 3,300 words. The author, Finn Nielsen, has created a Python wrapper called `afinn` that grants programmers easy access to the lexicon (Nielsen, 2011). Despite its simplicity, AFINN performs very well on various tasks in benchmark testing (Ribeiro et al., 2016). Due to its extensive and relatively unbiased contents, we consider AFINN to be a valuable addition to VADER. The compound scores VADER provides give an idea of the document-level sentiment in the UNSC speeches. For sentence-wise rating, we will rely on AFINN.

2.2.3 TextBlob

`TextBlob` is a python library used for processing textual data. This library also incorporates a successful lexical sentiment analysis tool, which is inherited from the `pattern` python library. TextBlob's default sentiment analyzer, known as `PatternAnalyzer`, works in a similar way compared to VADER and assigns sentiment scores to relevant words using a manually annotated sentiment lexicon (Smedt and Daelemans, 2012). It then averages the lexical sentiment scores to create an overall sentiment score. Table 2 gives a summary of the two types of scores that this tool can output.

Scores	Description
Polarity	This score takes on real-values from –1 to +1, where –1 indicates negative while +1 indicates positive sentiment polarity.
Subjectivity	This score takes on real-values from 0 to +1, where 0 indicates an objective utterance while +1 indicates a subjective utterance.

Table 2. Tabular summary of TextBlob `PatternAnalyzer` sentiment scores

2.3 Argumentation Mining

According to Van Eemeren and Grootendorst (2004):

> "Argumentation is a verbal, social, and rational activity aimed at convincing a reasonable critic of the acceptability of a standpoint by putting forward a constellation of propositions justifying or refuting the proposition expressed in the standpoint."

Argumentation mining can be defined as the process of decomposing text into Argumentative Discourse Units (ADUs) and connecting these units into a comprehensive argumentation tree, where an ADU is defined as the *"span of text that plays a single role for the argument being analyzed, and is demarcated by neighboring text spans that play a different role, or none at all"* (Stede and Schneider, 2018). We will expound further on corpus-specific ADUs and argumentation trees in section 2.3.1. Stede and Schneider (2018) further provide an exhaustive overview for the process of argumentation mining (not necessarily in numerical order):

1. Identify argumentative text (or a portion of a text)

[1] https://github.com/fnielsen/afinn/blob/master/afinn/data/AFINN-en-165.txt

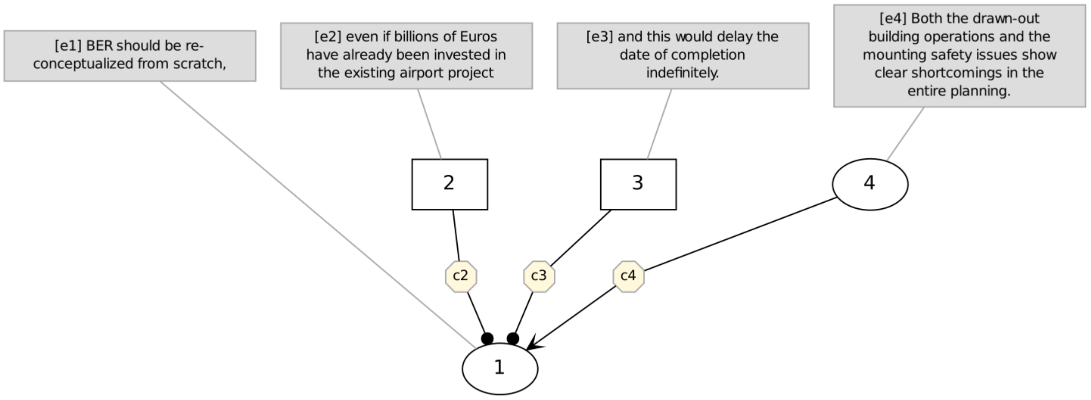

[e1] BER should be re-conceptualized from scratch,
[e2] even if billions of Euros have already been invested in the existing airport project
[e3] and this would delay the date of completion indefinitely.
[e4] Both the drawn-out building operations and the mounting safety issues show clear shortcomings in the entire planning.
2
3
4
c2
c3
c4
1

Fig. 3. Sample argumentation tree from argumentative microtext corpus detailed in Peldszus and Stede (2015); bottom-most node represents the central claim; arrow-head represents supporting ADU; circle represents attacking ADU

2. Segment the text into ADUs
3. Identify the central claim
4. Identify the role/function of ADUs
5. Identify relations between ADUs
6. Build the overall structural representation or argumentation tree
7. Identify the type and the quality of the argumentation

It is worth noting that there exist significant differences in the granularities of ADUs and their corresponding argumentation trees depending on the corpus of reference. Some corpora, such as the argumentative microtext corpus detailed in Peldszus and Stede (2015), go into deep detail regarding the relationships between ADUs; specifying central claims, support/attack segments and even describing subtypes of these ADUs, as seen in Figure 3. Other corpora, such as the US Election Debate corpus detailed in Haddadan et al. (2019), go into less detail and simply identify ADUs as claims or premises; and only specify basic hierarchies to build an argumentation tree.

Overall, argumentation mining provides the ability to automatically extract important information on the argumentative components of text. This process is extremely valuable for downstream NLP tasks which require distilled information; such as text summarization and question answering.

2.3.1 Copora

As mentioned previously, the granularities of ADUs and the complexity of argumentation trees are usually dependent on the corpus of reference. Table 3 shows a summary of three high-quality annotated argumentation corpora that are available to the public. Since the ultimate aim of this study is to apply an argumentation classifier on the UNSC corpus, it would make sense to select a well-suited argumentation corpus in order to train a relevant argumentation classifier.

Annotated Argumentation Corpus	Data Instances	Seq. Length Statistics [Tokens][†]	ADU Granularity and Argumentation Tree Complexity
Argumentative microtext corpus (Peldszus and Stede, 2015)	112 texts	$\overline{X} = 78.2$ $\sigma = 21.5$	Fine-grained ADUs with support/attack nature and subtypes, complex argumentation trees
Persuasive essay corpus (Stab and Gurevych, 2017)	402 essays	$\overline{X} = 366.0$ $\sigma = 62.9$	Fine-grained ADUs with support/attack nature, complex argumentation trees
US election debate corpus (Haddadan et al., 2019)	6,559 speech turns	$\overline{X} = 110.4$ $\sigma = 151.6$	Coarse ADUs as either claims or premises, simple argumentation trees

[†] $\overline{X}$ and σ represent the mean and standard deviations of sequence token lengths

Table 3. Tabular summary of three prospective annotated argumentation corpora

Referring back to Figure 2, we can observe that the UNSC corpus has a wide distribution of speech token lengths. In order to be similar to the UNSC corpus, a well-suited annotated argumentation corpus should have the following properties:

1. Relatively long sequence lengths similar to the UNSC corpus; which has a mean sequence length of ∼600 tokens.

2. Sufficiently large number of training data instances with negative (non-argumentative) examples to robustly train an argumentation classifier.

3. If possible, the corpus should exist in the political domain in order to maximize domain similarity.

4. If possible, utilize newly published corpus to contribute to research findings.

Given these requirements, points 2, 3, and 4 are satisfied best by the US Election Debate (USED) corpus. Point 1 is best satisfied by the persuasive essay corpus. However, as a mitigating factor towards point 1, the USED corpus has a very large standard deviation to mean ratio in its sequence length; which implies that there are many data instances which are also longer in length.

2.3.2 Models

In order to gain information about state-of-the-art argumentation classification models, we conducted a survey based on recent publications and summarized key information on three models in Table 4. Since these models were published in or before 2019, we can observe that they were trained/evaluated on the persuasive essay and argumentative microtext corpora, and not the USED corpus due to it only being released in mid-2019.

Furthermore, we can observe that both Potash et al. (2016) and Kuribayashi et al. (2019) train their argumentation classifiers on the assumption that ADU

Model	Corpus[†]	Task	Language Encoding	Best Performance [Macro-F_1][‡]
Joint Pointer Network (Potash et al., 2016)	PEC MTC	Classify ADU types and linkages given pre-defined spans	Bag-of-Words GloVe	PEC: 0.801 MTC: 0.777
BLCC tagger and LSTM-ER (Eger et al., 2017)	PEC	Sequence tagging for ADU spans, types and linkages	GloVe	Exact: 0.449 Half-span: 0.505
Span-based BiLSTM (Kuribayashi et al., 2019)	PEC MTC	Classify ADU types and linkages given pre-defined spans	GloVe ELMo	PEC: 0.818 MTC: 0.782

[†] PEC: Persuasive Essay Corpus, MTC: Argumentative Microtext Corpus
[‡] Where necessary, F_1 scores were averaged amongst joint tasks for brevity

Table 4. Tabular summary of three state-of-the-art argumentation classification models

spans in text are pre-defined. This, while being a reasonable assumption for partially annotated text, is not a sound assumption for non-annotated text where there exist no pre-defined ADU spans, as is the case for the UNSC corpus.

This limitation was in fact noted by Eger et al. (2017), which perhaps inspired them to create a more ground-up methodology of conducting sequence tagging in order to classify ADU spans, types and linkages altogether as a joint task. This technique of sequence tagging to form a ground-up framework is notable and will be revisited in section 3.2.3 of this study.

Another limitation of all three models is their language encoding techniques. Recent developments in NLP have shown that transformer-based language models, such as BERT, far outperform other language encodings such as Bag-of-Words, GloVe and ELMo (Devlin et al., 2018). Exploring new BERT-style language encodings could be an interesting avenue to further our research.

2.4 Transformer-Based Language Models

Since the introduction of the transformer architecture in the iconic "Attention is all you need" paper by Vaswani et al. (2017), we have seen the transformer architecture being used extensively in various NLP tasks, particularly in unsupervised language encoding. Transformers show significant advantages over traditional recurrent encoder-decoder frameworks due to their self-attention mechanism; which has shown to both increase performance, and efficiency through improved parallelization (Vaswani et al., 2017).

Since previous argumentation models shown in Table 4 did not leverage on state-of-the-art transformer architectures, it could be a good addition for us to conduct our research using such transformer architectures.

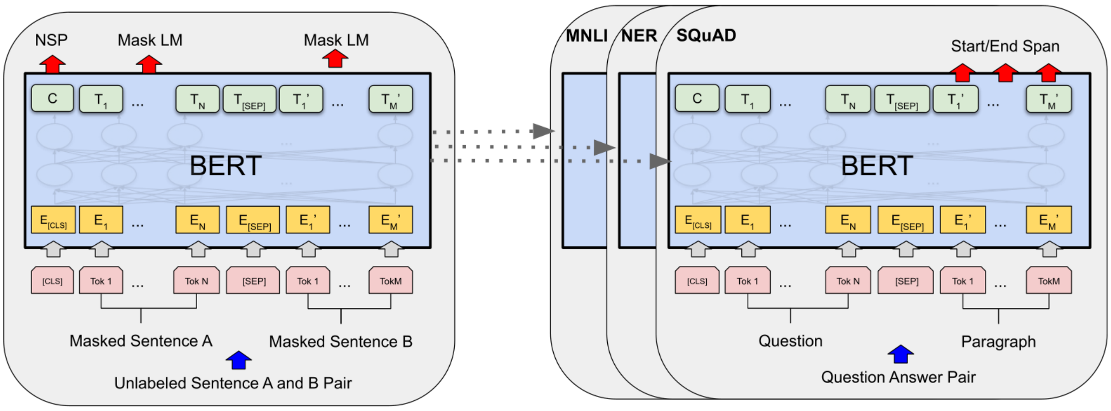


Fig. 4. Schematic of BERT pre-training (left) and fine-tuning (right) for downstream NLP tasks (Devlin et al., 2018)

2.4.1 BERT

BERT, which abbreviates Bidirectional Encoder Representations from Transformers, is a novel context-aware language model trained using transformers (Devlin et al., 2018). After pre-training on unsupervised language modeling tasks, BERT can be fine-tuned for downstream NLP tasks such as question-answering and natural entity recognition (Devlin et al., 2018). A schematic for this process can be seen in Figure 4. Fine-tuning BERT on various NLP tasks has shown significant performance improvements over preceding baselines; which is a large factor for recent academic interest and publications regarding the model.

Architecture

In Devlin et al. (2018), two variants of BERT known as BERT$_{\text{BASE}}$ ($L = 12$, $H = 768$, $A = 12$, $P = 110$ million) and BERT$_{\text{LARGE}}$ ($L = 24$, $H = 1024$, $A = 16$, $P = 340$ million) were tested; where L refers to the number of transformer blocks, H refers to the hidden size, A refers to the number of self-attention heads and P refers to the total number of parameters. The models were trained on 4 and 16 Cloud TPUs respectively, training for a total of 4 days.

Pre-training Tasks

BERT was trained with two pre-training objectives; namely Masked Language Model (MLM) and Next Sentence Prediction (NSP). Under the MLM task, some percentage of input tokens are masked at random and the model is trained to predict the identities of the masked tokens. This task incorporates a global context of the input text, since all of the context would be necessary to predict the masked token(s). Under NSP, some consecutive sentences are shuffled at random and BERT learns to predict whether sentence B indeed follows sentence A. Through this process, BERT incorporates sequential context for natural language understanding and this pre-training task has been shown to help with downstream NLP tasks such as question-answering.

Special Tokens

BERT incorporates certain special tokens in order to perform the aforementioned pre-training tasks; namely `[CLS]`, `[SEP]`, `<pad>` and `<unk>`. The

`[CLS]` token is found at the start of every sequence and its position in the model output represents the aggregate representation of the entire sequence; which is why this token's position is used for sequence classification tasks. The `[SEP]` token is a separator token which is used to separate sentence pairs for the NSP pre-training task. Devlin et al. (2018) also suggests that the `[SEP]` token should be used at the end of sequences; although there are some discrepancies regarding this in their paper. For simplicity, we assume that the `[SEP]` token should additionally be placed at the end of input sequences. The `<pad>` and `<unk>` tokens represent padding and unknown tokens respectively. BERT additionally enforces a maximum input sequence length of 512 tokens.

WordPiece Tokenization

BERT uses the WordPiece Model (WPM) for tokenization of its input text. The WPM is a novel and powerful tokenization model that tokenizes words into sub-word units (Wu et al., 2016). This is a powerful approach as it mitigates the large vocabularies and out-of-vocabulary (OOV) issues of purely word-based tokenizers, as well as the small vocabularies and semantic information loss of purely character-based tokenizers. Following is an example of the WPM model's output, which has been adapted from Wu et al. (2016):

Input Sequence: Jet makers feud over seat widths

WordPiece Output: _J et _makers _fe ud _over _seat _width s

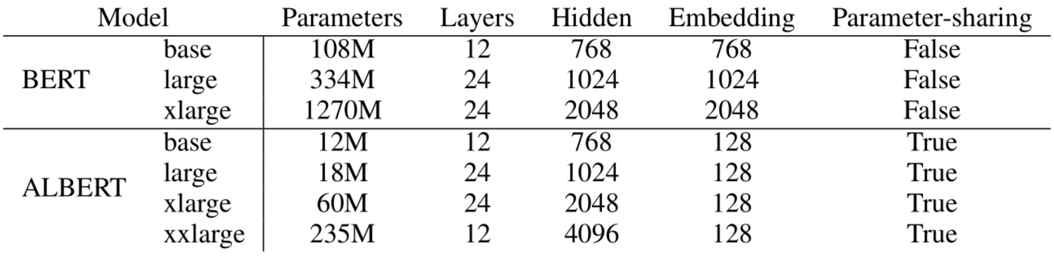
"_" is a special character used to mark the beginning of a word. A token which does not start with "_" is a segmented sub-word of a main word; for example "et" being a sub-word of "Jet" in the example above. It is worth noting here that a combination of using the WPM and special tokens in BERT result in input sequences becoming much longer than they originally were. As a result, some input sequences in corpora might exceed BERT's 512 token hard upper-limit and might need to be discarded from analysis.

Fine-tuning

For pre-training BERT, large unsupervised corpora such as BookCorpus (800 million words) and English Wikipedia (2,500 million words) are used (Devlin et al., 2018). Fine-tuning follows after pre-training and involves slightly adjusting all the parameters in BERT in order to perform on downstream NLP tasks and corresponding supervised corpora. Compared to pre-training, fine-tuning is relatively fast and computationally inexpensive (Devlin et al., 2018). Fine-tuning BERT on downstream NLP tasks has shown significant improvements in state-of-the-art performances, which include a 7.7% point absolute improvement of the GLUE score for the General Language Understanding Evaluation (GLUE) task.

2.4.2 ALBERT

A major criticism of the aforementioned BERT language model is its sheer size in terms of parameters; which can ultimately lead to GPU/TPU memory limitations, long training times and unexpected model degradation. To address these



Model		Parameters	Layers	Hidden	Embedding	Parameter-sharing
BERT	base	108M	12	768	768	False
	large	334M	24	1024	1024	False
	xlarge	1270M	24	2048	2048	False
ALBERT	base	12M	12	768	128	True
	large	18M	24	1024	128	True
	xlarge	60M	24	2048	128	True
	xxlarge	235M	12	4096	128	True

Fig. 5. Tabular excerpt from Lan et al. (2019) on BERT vs. ALBERT architecture

issues, Lan et al. (2019) propose a "lite" version of BERT (a.k.a. ALBERT) which utilizes factorized embedding parameterization and cross-layer parameter sharing to reduce memory consumption and increase training speed. Figure 5 shows how this optimization results in a drop in the number of parameters in the ALBERT model(s); for example the ALBERT$_{\text{BASE}}$ model can be reduced to 12 million parameters, making it a suitable candidate to train even on a single GPU.

To further improve performance, Lan et al. (2019) replace the NSP task in BERT's unsupervised pre-training with a sentence-order prediction (SOP) task. This involves the same task as per NSP for positive examples. However, negative examples involve two consecutive sentences being swapped in terms of their order. According to Lan et al. (2019), the SOP task pushes ALBERT to learn more fine-grained distinctions on discourse-level coherence properties. This modification has shown to improve downstream NLP task performance for multi-sentence problems. Additionally, ALBERT uses the SentencePiece tokenization model (SPM) from Kudo and Richardson (2018), which incorporates the Unigram Language Model (ULM) from Kudo (2018). The SPM is similar to BERT's WPM, but is purportedly faster and more versatile compared to the WPM.

Despite these key differences, ALBERT is still very similar to BERT in terms of its general architecture and building blocks, which can be seen as well in Figure 5. Additionally, it is worth noting that the 512 token hard upper-limit for input sequences also applies for ALBERT. Ultimately, since ALBERT has shown to both perform and scale better than its BERT counterpart (with particular focus on discourse-level coherence for its SOP task), we consider ALBERT as a viable model for use in our argumentation mining task. This application will be further expounded on in section 3.2.3 of our study.

3 Methodologies

3.1 Sentiment Analysis

3.1.1 Data Preprocessing

For conducting the sentiment analysis, preprocessing of the data was necessary. Since the 65,393 speeches contained in the corpus were provided as separate text files, we decided to combine the speeches for each year, allowing for an analysis of the sentiment development over time. We resolved several issues that arose in the process, e.g. duplicate speech IDs or missing meta data.

The yearly overview used for the analysis included the speech ID, date, speech text, cleaned text, topic, country, name of the speaker, participant type, subjectivity score, sentiment score, number of positive sentences, and number of negative sentences for each speech.

In order to perform inter- and intracountry comparisons, we calculated the yearly average for each feature (i.e. sentiment, subjectivity, positive and negative sentences) for each country. Using this average data, we also calculated the standard deviation of the scores for each country over time.

We also created additional overviews for analyzing the the core nations of the UNSC, the sentiment of the core nations related to the Iraq War. Lastly, we created two more overviews for a detailed country-level sentiment analysis during two specific resolutions related to the invasion of Iraq.

After creating these data overviews in Python, we stored each one locally as `csv` file since computation on each access of the data would have been too costly. For the purpose of easy access, we prepared several helper functions which load the data into the Python data structure `Pandas DataFrame`. We chose the DataFrame for data storage and manipulation since it efficiently handles large data and is much more flexible than standard data structures.

3.1.2 Sentiment Polarity and Subjectivity

As mentioned in section 2.2.1, we relied on commonly used automatic sentiment analysis tools for our analysis. As for the processing of the speeches for this analysis, we chose to clean the speeches, which means that we removed stop words (common words with no inherent meaning) and transformed the text into lowercase. This was done for the sentiment analysis using VADER and the subjectivity analysis using TextBlob. For the calculation of the amount of positive and negative sentences per speech with AFINN, we also tokenized the words in the speeches. Tokenization yielded no improved results for sentiment and subjectivity analysis. For text manipulation, we used the Python `nltk` library.

As for the visualizations of the analysis results, we relied on Python's `matplotlib` plotting library, expanded with the `Seaborn` library. All analyses and data handling was performed with `Python 3.7.4`. Our final product for sentiment analysis is a human-readable `json` file mapping UNSC speech IDs to sentiment and subjectivity scores, available in our public GitHub repository[2].

3.2 Argumentation Mining

3.2.1 US Election Debate (USED) Corpus

As mentioned in section 2.3.1, we decided that the USED corpus would be the best choice for us to train an argumentation classifier. Regarding corpus-specific ADUs, Haddadan et al. (2019), the authors of the USED corpus, define a claim (in the political domain) as *"a policy advocated by a party or a candidate to be undertaken which needs to be justified in order to be accepted by the audience."* Similarly, they define a premise as an *"assertion made by the debaters for supporting their claims (i.e., reasons or justifications)."*

The USED corpus is annotated at a character-span level, with spans being annotated either as claims or premises. It follows naturally that character spans

[2] https://github.com/atreyasha/sentiment-argument-mining

that are not annotated can be considered as non-argumentative text. Furthermore, the USED corpus also has been annotated for linkages between argumentative spans. This typically comes in the form of N consecutive premises being recorded after a claim; which indicate that those N premises belong to the preceding claim. Linkage types, such as attack or support relations, are not annotated in this corpus.

As mentioned in section 2.3.2, we would like to conduct argumentation mining using a bottom-up approach by using sequence tagging; since this process would also lend itself to quick application on the UNSC corpus. Given the time and resource restrictions in this study, we decided to focus solely on classifying tokens into argumentation candidates; such as claims, premises or neither. We omit the task of linking spans from claims and premises and instead recommend this task for further studies.

It is also worth noting that Haddadan et al. (2019) performed a similar task as ours on the USED corpus. However, they trained a classifier to identify argumentative sentences and argumentative sentence types. Performing such a task at a sentence-level removes many of the fine-grained character spans which do not necessarily span entire sentences. In fact, multiple character spans can be included within a sentence. As a result, we will be unable to compare our results directly with theirs since we perform a more fine-grained token-level classification task.

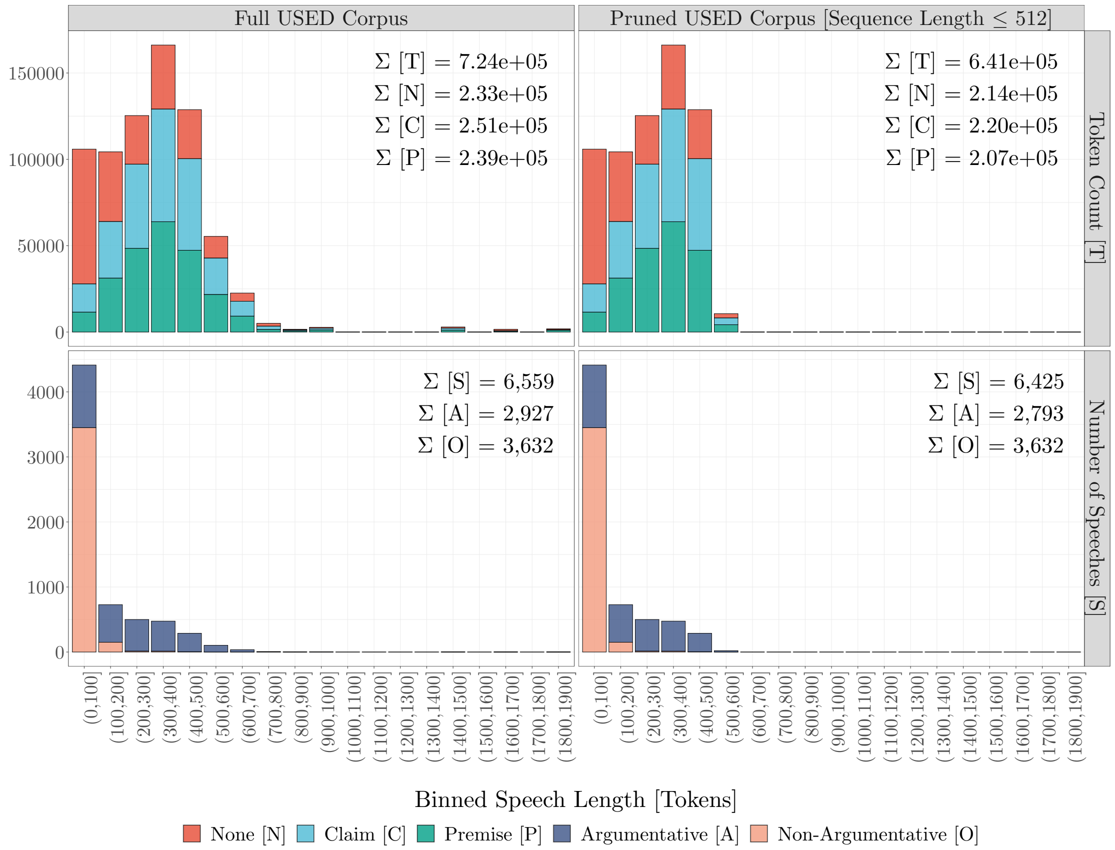


Fig. 6. Full USED corpus (left-column) and pruned USED corpus (right-column) by binned speech length with token count (top-row) and number of speeches (bottom-row)

3.2.2 Data Preprocessing

Following a similar approach from Eger et al. (2017), we start by preparing the USED corpus for a sequence tagging task. Firstly, we converted character span annotations in the USED corpus to token-level annotations. Each token in the USED corpus gets mapped to one of three labels; specifically the “None” (N), “Claim” (C) or “Premise” (P) label. Speeches that had at least one of either C or P tokens are considered argumentative (A), while speeches containing only N tokens are considered non-argumentative (O). The left column of Figure 6 shows the distribution of these classes of tokens as well as the number of speeches/types in the full USED corpus. It is worth noting that we omit the Beginning-Inside-Outside (BIO) tagging scheme as recommended in Eger et al. (2017); to keep our methodology simple and prevent label imbalances in the already small USED corpus. We also observe that sequences longer than 100 tokens tend to be more argumentative or have more C and P labels, indicating that longer sequences are of greater interest for argumentation mining in the USED corpus.

After mapping tokens to the aforementioned argumentation classes, we attempt to remove USED corpus-specific bias. In particular, we remove the initial token(s) in each speech where the identity of the person giving the speech is stated. Next, we use the SPM model to segment words into lowercased subwords; in accordance with the pre-packaged uncased ALBERT tokenizer. Finally we add BERT’s special tokens; specifically a [CLS] and [SEP] token at the start and end of the sequence respectively. As mentioned in 2.4.1, this process increases the token lengths of

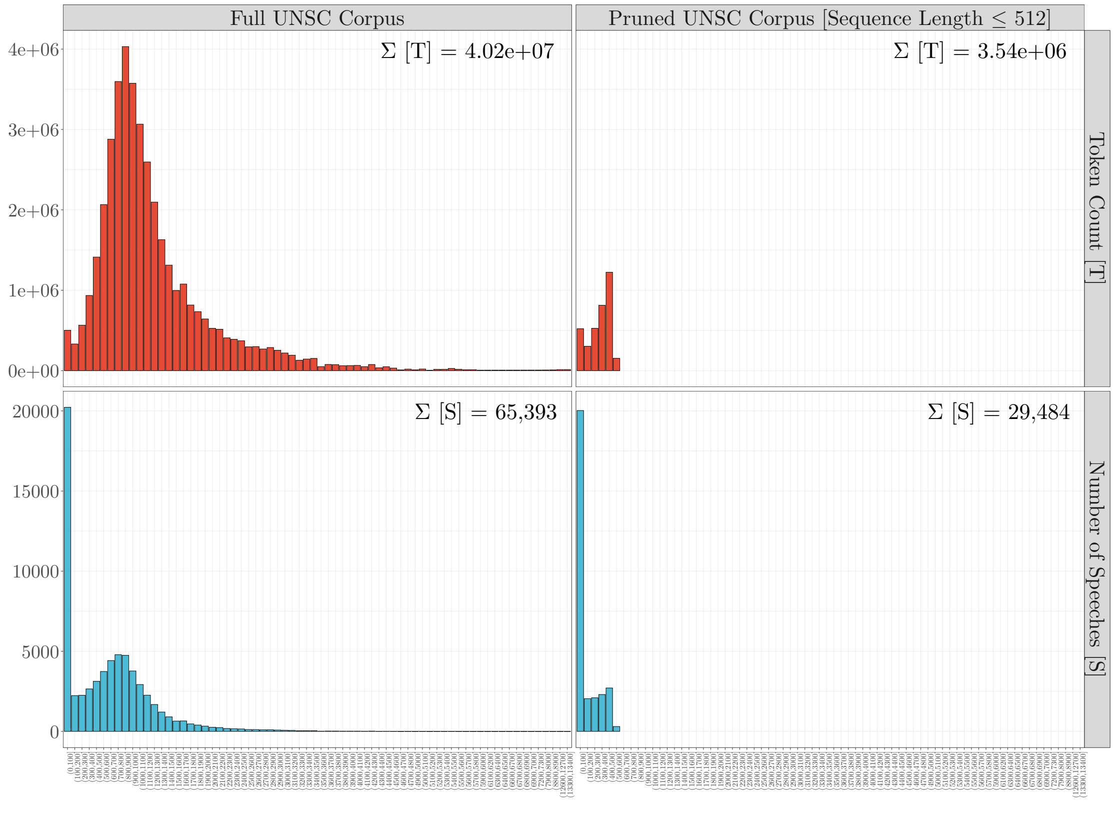


Fig. 7. Full UNSC corpus (left-column) and pruned UNSC corpus (right-column) by binned speech length with token count (top-row) and number of speeches (bottom-row)

speeches; with some speeches exceeding the maximum sequence length of 512 tokens. To address this, we prune the USED corpus and remove sequences longer than 512 tokens. The effect of this pruning process can be seen in the right column of Figure 6, with a minimal loss of 11% of tokens and 2% of speeches in the USED corpus.

The same data preprocessing procedure (naturally minus argumentation labelling) is then performed on the UNSC corpus to ensure that both these corpora can be treated similarly by the argumentation classifier model. For domain debiasing in the UNSC corpus, we also remove initial tokens which state the identity of the speaker. Figure 7 shows the effect of pruning on the UNSC corpus, with a much greater loss of 91% of tokens and 55% of speeches. Unfortunately, this is a necessary measure and is one of the major limitations of using the ALBERT model for such long sequences.

Finally, after pruning the corpora and having a clean set of input sequences, we insert BERT's `<pad>` tokens to cap up all sequences to 512 tokens, as this is necessary for the static computation graph that was implemented in our code. For evaluation purposes, we randomly split the USED corpus into $\{\text{training} \cup \text{validation}\}$ and test sets with a 70:30 ratio. We then split the $\{\text{training} \cup \text{validation}\}$ set into training and validation sets with a 85:15 ratio.

3.2.3 Model Fine-Tuning and Evaluation

As mentioned in section 2.4.2, we decided that the ALBERT language model would be a good choice for us to fine-tune for our argumentation mining task. With the USED and UNSC corpora preprocessed, we needed to finalize the ALBERT model for fine-tuning. Out of the box, the ALBERT model is an encoder model. In order to use it for a supervised argumentation sequence tagging task, we would need to extend ALBERT with custom decoders such that the output of the model would be in the appropriate dimensions. Table 5 provides a summary of three models that we fine-tuned in this study.

For the fine-tuning process, we assume a warmup-cooldown learning rate profile, where the learning rate rises linearly over the warmup epochs until a maximum learning rate and then exponentially decays till an end learning rate at a fixed up-

Model	Description	Parameters
TD_Dense	5 stacked time-distributed dense layers on top of the ALBERT model; with Rectified Linear Unit (ReLU) and softmax activations	11,659,334
1D_CNN	5 stacked 1-dimensional convolutional layers on top of the ALBERT model; with ReLU and softmax activations	12,790,598
Stacked_LSTM	3 stacked Long Short-Term Memory (LSTM) layers on top of the ALBERT model; with ReLU and softmax activations	14,709,446

Table 5. Tabular summary of three end-to-end ALBERT model types with custom decoders

per limit of 100 epochs. We utilize the Adam optimizer for fine-tuning all our models. To determine the best model and corresponding hyperparameters, we conduct a grid-search over model-type, warmup-epochs, maximum learning rate and end learning rate.

It is worth noting here that we fix our fine-tuning batch-size at 10 samples to prevent GPU out-of-memory (OOM) issues. This is mainly because the ALBERT model has a high memory consumption that is exacerbated by its $O(N^2)$ space complexity due to the self-attention mechanism, where N is the input sequence length. Because of this, our model fine-tuning process suffers from noisy gradients. We recommend solutions for this issue in section 7 of our study.

For the evaluation process, we employ early stopping with a patience value of 5 epochs; measuring the cross-entropy loss on the validation set. At the end of fine-tuning each model, we compute the model's performance on the test dataset and record its Macro-F_1 score over N, C and P argument labels. The model with the best test F_1 score is deemed as the best performing model. We provide our best fine-tuned model in our GitHub repository[2].

For all the aforementioned processes, we utilized the `bert-for-tf2` python library, which is written using Google's `TensorFlow` API. In terms of hardware, we utilized a single NVIDIA GeForce GTX 1080 Ti GPU with 12 GB RAM provided by the University of Potsdam.

3.2.4 Prediction on UNSC

After following the aforementioned fine-tuning and model evaluation process, we arrive at our best performing model. In order to use this model to predict argumentation token-level argumentation candidates in the UNSC corpus, we simply pipe the preprocessed UNSC corpus into the model and acquire the model's outputs.

Due to the static configuration of our model, the output will have a sequence length of 512 tokens. We then remove the output sequence positons which correspond to the BERT `[CLS]`, `[SEP]` and `<pad>` special tokens; leaving behind only the important sub-word tokens that have been classified into N, C or P argumentation candidates. As a final product, we compile our UNSC predictions into a human-readable `json` file in our GitHub repository[2], which maps UNSC speech IDs to token-level argumentation predictions.

4 Results

4.1 Sentiment Analysis

Given the abundance of UNSC session data available, we decided to narrow the scope of our analysis using different approaches. First, we examined to most popular topics in order the determine whether trends can be seen. Secondly, we took a look at the big picture and analyzed the sentiment development for each country in the time frame 1995-2017. Finally, we decided to analyze the core nations of the UNSC, since the corpus provides consistent data for the entire time frame. In addition to this, we analyzed the sentiment in Iraq War related sessions.

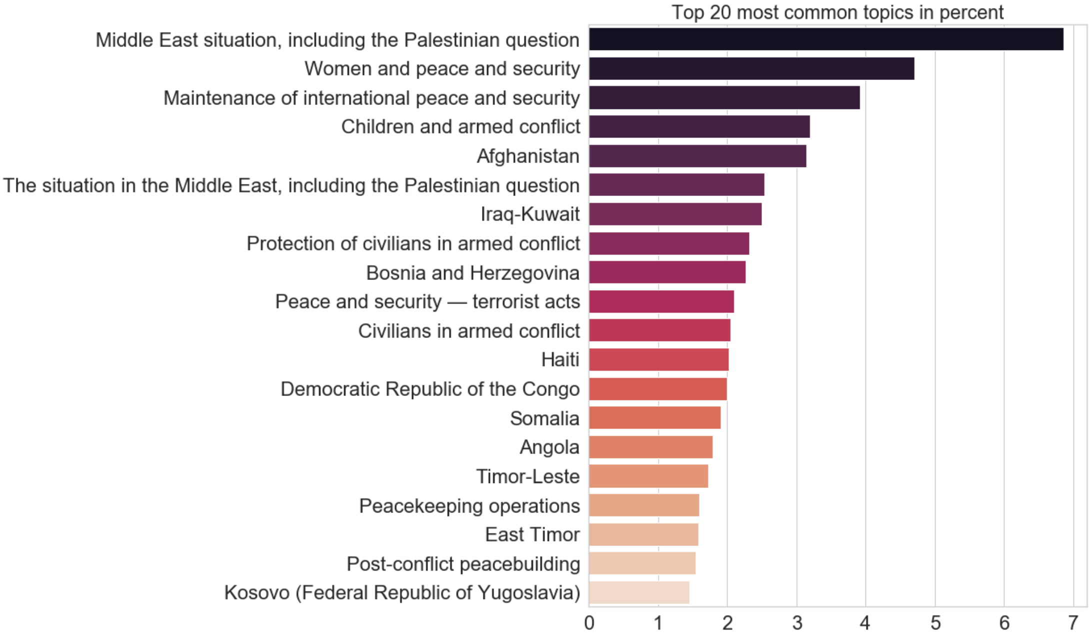


Fig. 8. The 20 most common topics in the UNSC

4.1.1 Topics

In order to get a grasp of the contents of the sessions, we examined the most common topics for the provided time frame and each year respectively. Figure 8 visualizes the 20 most common topics for the UNSC speeches. It can be observed that there is no uniform naming convention for the topics. The most discussed topic is *Middle East situation, including the Palestinian question* and the topic on rank six is *The situation in the Middle East, including the Palestinian question.* In other analyses, the topic name *The situation in the Middle East* occurs as well. It can be assumed that the different topic names refer to one and the same topic. Interestingly, the Middle East topic occurs in the most discussed topics for the first time in 2000 and remains in the top 5 most discussed topics for the entirety of 2000-2017, despite the inconsistent naming convention.

A shift in topics can be observed after the year 2001. The most common topics diverged from humanitarian ones (e.g. Africa, Angola, East Timor) to more fundamental ones (e.g. Peace and Security - terrorist acts, Women and Peace and Security, Post-conflict peacebuilding). Figure 9 and Figure 10 visualize the differences in topics for 1996 and 2003.

4.1.2 Sentiment in the UNSC

The sentiment analysis in the UNSC was conducted on 182 individual members of the council, sovereign nations and independent members were included. As could be expected, the participation varies among the members. For some countries, only scarce data is available, e.g. Belize, which only took part in four years of council meetings. Other countries attended the council meetings more frequently, with many countries appearing in all 23 years of the analysis. It is also worth noting that some countries' names changed over time, e.g. Zaire, which is now known as Democratic Republic of the Congo, and some countries do not exist

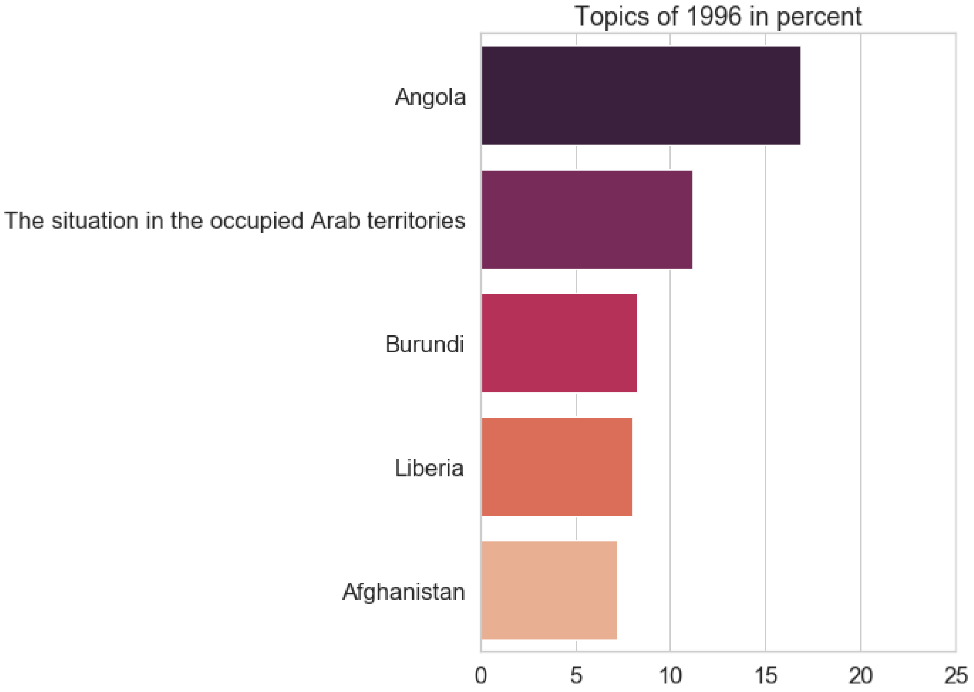


Fig. 9. Most common topics in 1996

anymore, e.g. the former Yugoslavia.

The visualizations of the sentiment and subjectivity analysis for Belize (Figure 11) and Germany (Figure 12) suggest a correlation between sentiment and subjectivity scores. A higher polarity in sentiment occurs with an increase in subjectivity in the analysis for Belize, while the lower sentiment scores come with an increase in objectivity for Germany. For Germany, especially the year 2003 stands out. 2003 was the beginning of the Iraq War, of which Germany was a vocal opponent (The Guardian, 2002b).

4.1.3 Sentiment in the Core Nations

After completing the topic and sentiment analysis for the UNSC, we decided to narrow the scope of our investigations. A comparative and exploratory analysis yields the best results if the data is not skewed, i.e. equally distributed. For this reason, we laid our focus for the sentiment analysis on the core nations of the UNSC, which are the People's Republic of China, the French Republic, the Russian Federation, the United Kingdom of Great Britain and Northern Ireland, and the United States of America.

These core nation are granted a permanent member status of the council as per the Charter of the United Nations, a treaty from 1945[3]. A permanent membership entails a *Power of Veto*, which allows voting against UNSC resolutions and thus preventing their adoption. The core nations were a suitable subject for our analysis since they participated in most of the UNSC meetings. Not only could we rely on data for every year provided in the corpus, we could also assume that important resolutions would be included in the analysis.

The sentiment analysis for the UNSC core nations shows a general downward

[3] https://www.un.org/en/charter-united-nations/

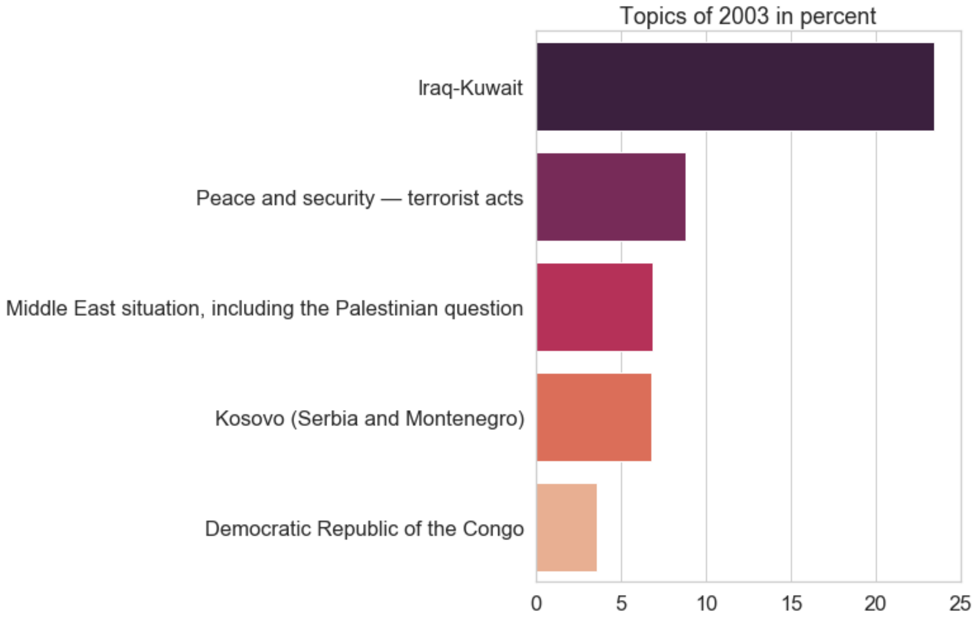


Fig. 10. Most common topics in 2003

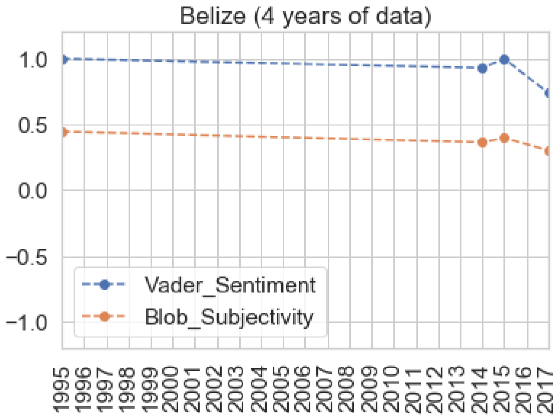


Fig. 11. Sentiment Analysis for Belize

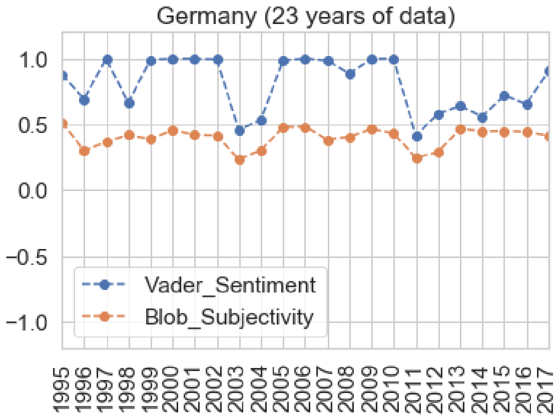


Fig. 12. Sentiment Analysis for Germany

trend in sentiment polarity (Figure 13). Significant negative peaks were revealed for historical events, e.g. the year 2002 for the Russian Federation and the year 2003 for the United States of America, both possibly related to the imminent Iraq War.

The visualization of the subjectivity analysis (Figure 14) shows a mostly consistent development of subjectivity in the core nations but suggests a slight downward trend, meaning a rise in objectivity, over the investigated time frame. This is consistent with the observation made in the sentiment analysis conducted on all members of the UNSC. The negative peaks in sentiment mentioned for Russia and the USA, 2002 and 2003 respectively, also appear as negative, more objective, peaks in the subjectivity analysis.

4.1.4 Sentiment regarding the Iraq War

As the topic for an in-depth analysis, we chose the Iraq War. Until today, the Iraq War and the UNSC's involvement in it is a highly polarizing topic. Kofi Annan,

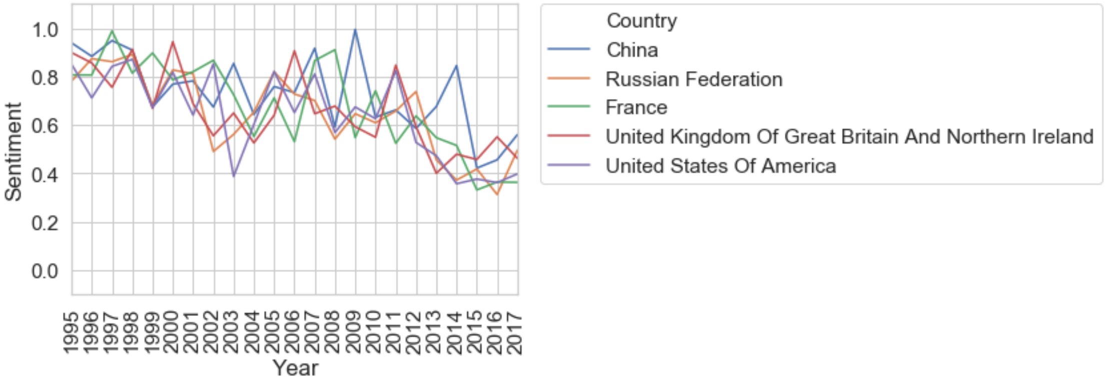


Fig. 13. Sentiment development in the UNSC core nations

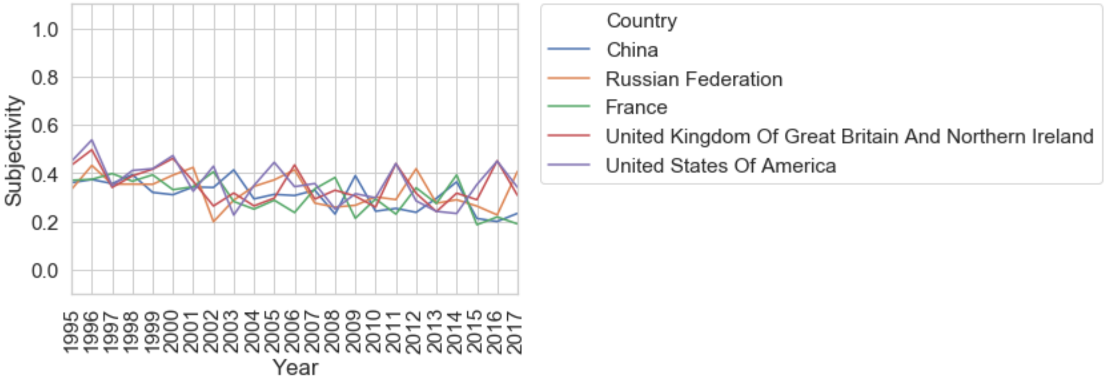


Fig. 14. Subjectivity development in the UNSC core nations

the United Nations Secretary-General, stated that “the US-led invasion of Iraq was an illegal act that contravened the UN charter” (BBC, 2004). The possibility of an invasion arose from the speculation that Iraq was in possession of weapons of mass destruction, after the UNSC council unanimously adopted Resolution 1441 which gave Iraq ”a final opportunity to comply with its disarmament obligations” (UNSC, 2002). This issue also divided the core nations of the Security Council.

The United States of America were a strong proponent of pursuing military action against Iraq, with or without approval of the UNSC (The Guardian, 2002c). The United Kingdom supported this policy. France on the other hand was strictly against the war. The former president Jacques Chirac went as far as saying he would make use of his veto right to prevent the adoption of a resolution that would allow for a military intervention in Iraq (New York Times, 2003). China was more reserved in taking a position on this matter, yet stated that their position was “extremely close to that of France” (CNN, 2003). Russia also did not have a consistent policy on this matter, first being against an invasion and later on becoming more neutral and even approving of the USA’s plans (Golan, 2004). On March 19, 2003, an U.S.-led coalition invaded Iraq.

General Sentiment In order to examine the general sentiment of the core nations regarding the Iraq War, we extracted all speeches with a topic labelled “Iraq” or “Iraq-Kuwait”.

For presenting the results of the analysis, we chose two countries with opposing opinions on the invasion of Iraq, France (Figure 15) and the United States of

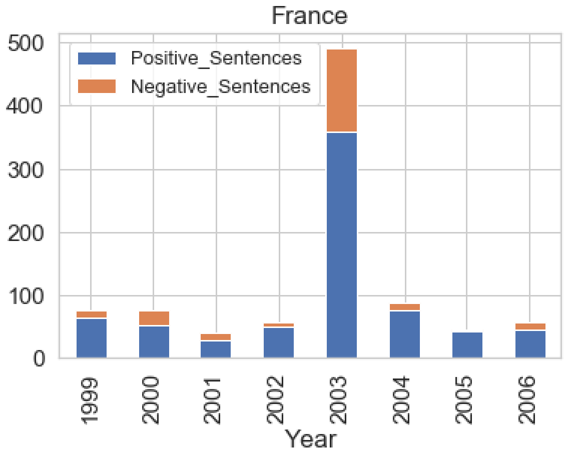


Fig. 15. Polarity distribution regarding Iraq for France

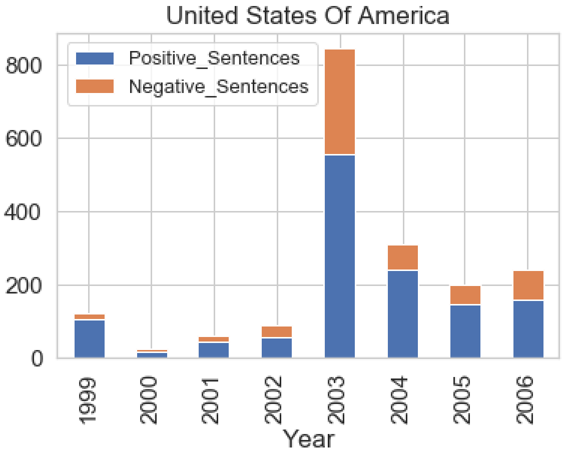


Fig. 16. Polarity distribution regarding Iraq for the USA

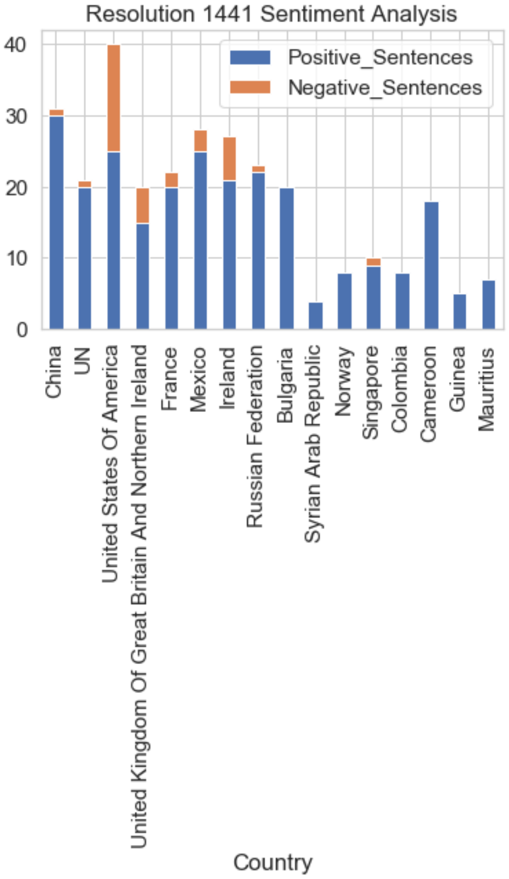


Fig. 17. Sentiment regarding Resolution 1441

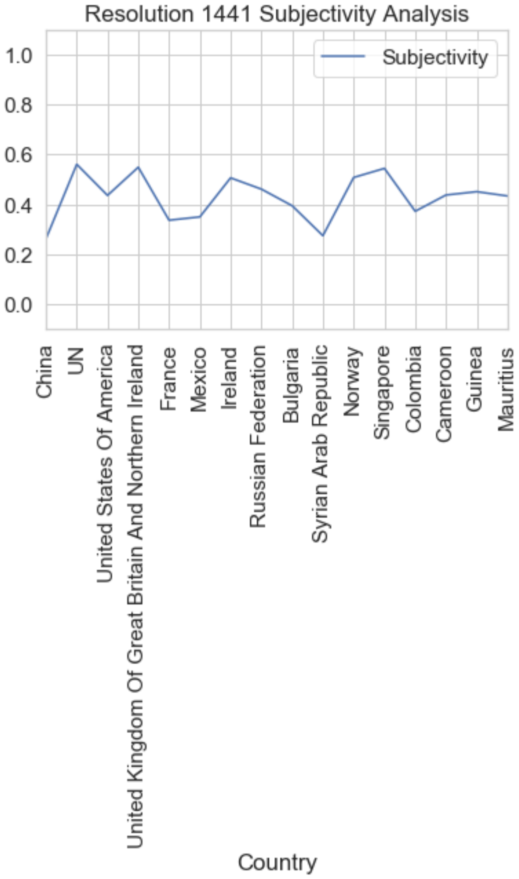


Fig. 18. Subjectivity regarding Resolution 1441

America (Figure 16). For both countries, the amount of sentences with polarity with regard to Iraq is the highest for the year 2003, due to the Iraq War. Before and after 2003, France contributed on average less than 100 sentences on this topic, an increase after 2003 cannot be observed. This is different for the USA. Before 2003, the analysis reveals years were the country contributed fewer sentences than France, with 1999 being the the year with the most contributions and 2000 being the year with the fewest. After 2003, the USA consistently contributed more than 200 negative and positive sentences per year, which is significantly more than all other core nations.

In 2003, the representatives of the USA contributed more than 800 sentences with polarity, with roughly a third of them being negative. For France, the country representatives uttered almost 500 relevant sentences with roughly a quarter of them being negative.

In-Depth: Iraq War related Resolutions In order to give a better insight into Iraq War related resolutions, we focused on two sessions that were highly significant in the year leading to the invasion of Iraq. First, we examined Resolution 1441, a resolution that was adopted in the 4644th session of the UNSC on 18 November 2002. As mentioned before, it granted Iraq ”a final opportunity to comply with its disarmament obligations” (UNSC, 2002) and subsequently lead to the invasion of Iraq, after the country allegedly failed to comply with said disarmament obligations.

In the sentiment analysis visualizations (Figure 17) it can be observed that the USA contributed more to the session than the other countries. The highest negative/positive sentence ratio can be observed for the USA and the UK, along with Ireland. While the former two countries were in favor of the “consequences” announced in Resolution 1441, Ireland had a neutral stance on this policy.

The subjectivity scores (Figure 18) are roughly distributed between 0.3 and

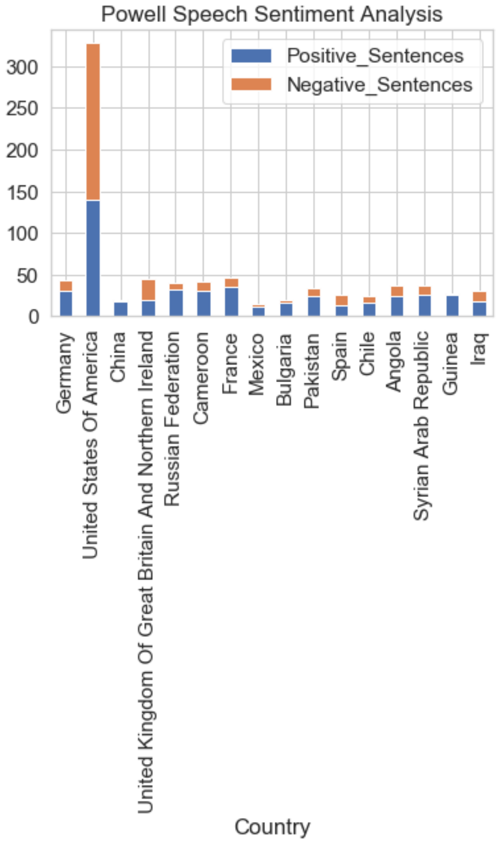


Fig. 19. Sentiment regarding Powell’s speech

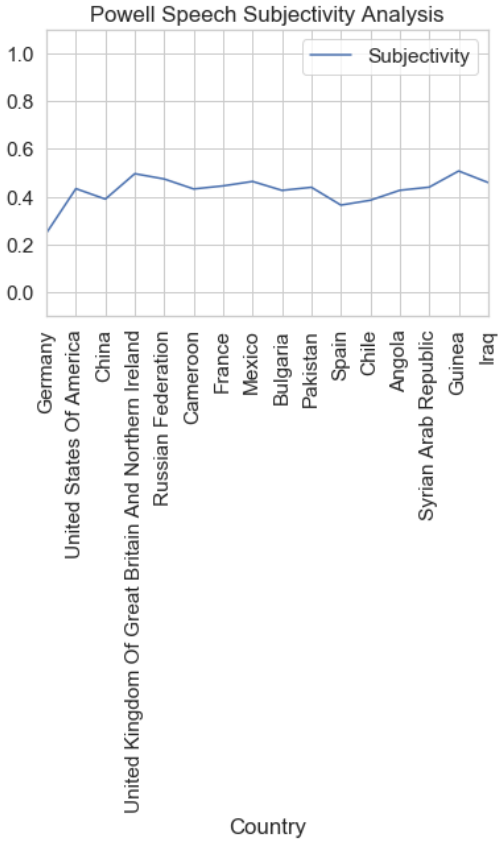


Fig. 20. Subjectivity regarding Powell’s speech

0.6, which indicates a varying level of objectivity with regard to this resolution.

Secondly, we analyzed session 4701 from February 5, 2003 in which United States Secretary of State Colin Powell discussed the urgency of attacking Iraq as it was allegedly in possession of a mobile production facility for biological weapons (CIA, 2007). This speech was controversial in the UNSC and later on it became evident that it was not based on facts but solely on speculations. Neither the USA nor the UK were able to provide evidence for their claims (The Guardian, 2002a). Powell himself said later: "I regret it. I will always regret it. It was a terrible mistake on all our parts and on the intelligence community." (Harvard Gazette, 2015).

It can be observed that Powell (i.e. the USA) uses a very negative vocabulary (Figure 19). He contributes significantly more than the other participants (330 sentences) and more than half of his speech has a negative polarity. A similar ratio of positive to negative sentences can be observed for the UK, although they do not participate more than the other countries involved. This aligns with the previously discussed division of the UNSC regarding the invasion of Iraq and the fact that the USA and the UK were proponents of this approach.

The subjectivity scores (Figure 20) concerning the Powell session seem rather homogeneous in comparison to the ones regarding Resolution 1441, except for Germany, which has an unusually low subjectivity score. As discussed before, Germany was a strong opponent of the Iraq War.

4.2 Argumentation Mining

4.2.1 Model Fine-Tuning and Evaluation

As mentioned in section 3.2.3, we fine-tuned three different model types using a grid-search technique. We ran the grid-search fine-tuning (or training) for 1.5 days on a single GPU, evaluating our results against the USED test set. We compared these models with a baseline majority-class classifier; which predicted all tokens in the test set as claim tokens which form the majority class.

Model	Training Epochs‡	Test F_1	Test F_1 [N]	Test F_1 [C]	Test F_1 [P]
Baseline†	–	0.173	0.000	0.518	0.000
TD_Dense	17	**0.693**	**0.763**	**0.689**	0.627
1D_CNN	16	0.689	0.758	0.659	**0.651**
Stacked_LSTM	21	0.633	0.679	0.624	0.596

†Baseline model is a majority classifier for the C token, which has the highest frequency in the test set
‡Training epochs include five extra patience epochs; best model was saved from the lowest validation loss epoch

Table 6. Tabular summary of model performance on the USED test set; bold implies best performance for given category

Statistic	Macro-Average	None [N]	Claim [C]	Premise [P]
Precision	0.703	0.826	0.633	0.651
Recall	0.691	0.710	0.757	0.605

Table 7. Tabular summary of precision-recall statistics for TD_Dense model on the test set

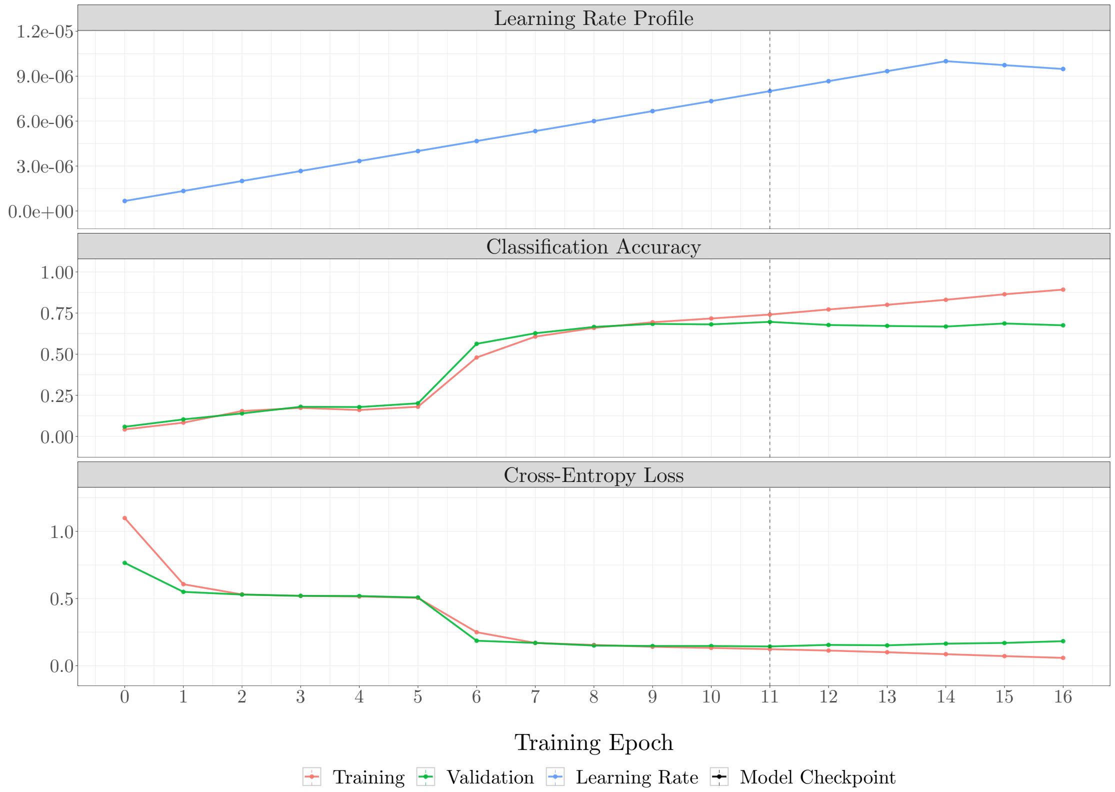


Fig. 21. TD_Dense model performance summary statistics by training epochs; black vertical line demarcates the 12th training epoch where validation cross-entropy loss was minimized

Our final model results are summarized in Table 6. While the 1D_CNN model performed best in classifying premise tokens, the TD_Dense model performed best on the overall USED test set with a Macro-F_1 score of 69.3%. Table 7 shows the precision-recall scores for the TD_Dense model on the USED test set. Overall, we can observe a higher performance for N tokens compared to C and P tokens; which is in some ways intuitive since N tokens aggregate together in non-argumentative speeches, while C and P tokens are distributed in tight spans within argumentative speeches. This might make the detection of C and P tokens more nuanced compared to the N tokens. Next, we can observe the performance of the TD_Dense model with respect to training epochs in Figure 21. This figure further emphasizes the point that fine-tuning is fast and computationally inexpensive; with the lowest validation cross-entropy loss being achieved in just 12 training epochs.

4.2.2 Prediction on UNSC

After identifying the best model as the TD_Dense model, we then used this model to predict token-level argumentation labels of the preprocessed (pruned) UNSC corpus. Figure 22 shows the token distribution of the predictions of our best classifier on the UNSC corpus. Similar to the token distributions in the USED corpus as shown in Figure 6, we observe that the proportions of claim and premise tokens, as well as argumentative speeches, increase as the speech sequence lengths increase. However, unlike the USED corpus' balanced token/speech distributions in Figure 6; our fine-tuned model's predictions appear to show a strong skew towards N tokens and non-argumentative speeches.

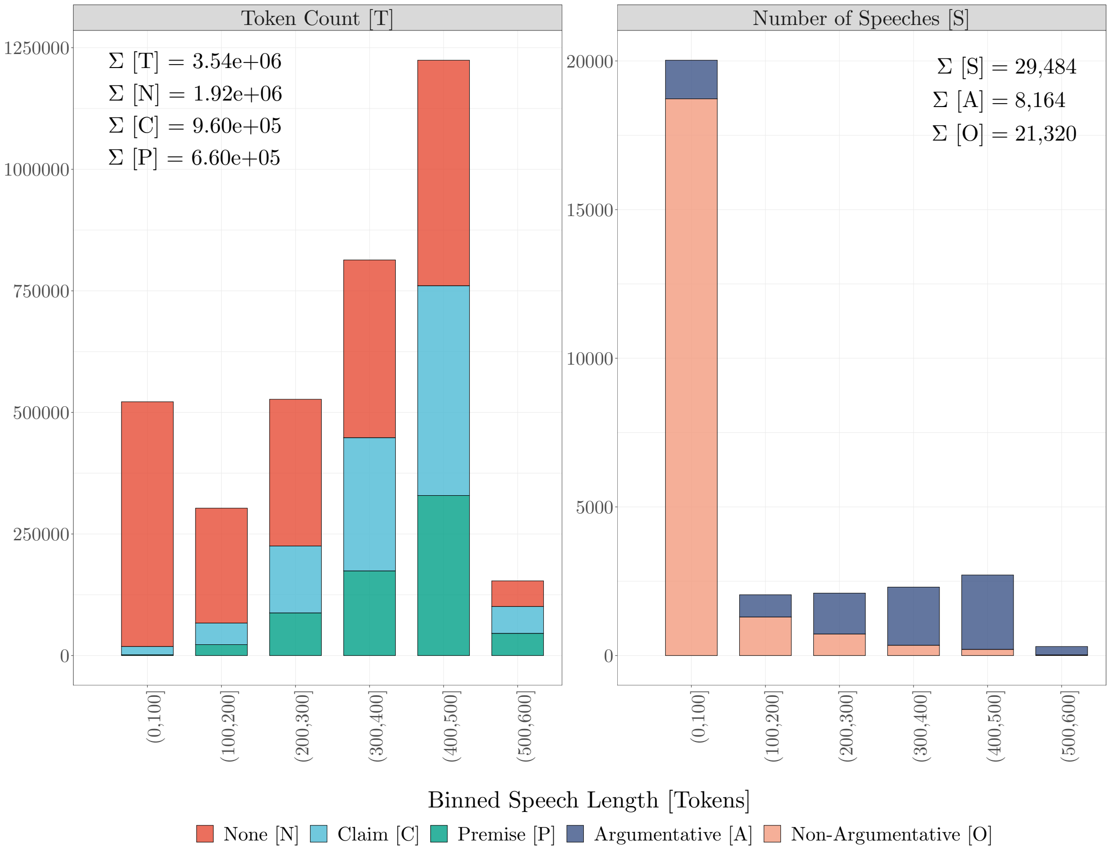


Fig. 22. Token type predictions from TD_Dense classifier on pruned UNSC corpus; non-argumentative (O) speeches have zero C and P tokens while argumentative (A) speeches contain at least one of either

There could be many reasons for this; one of which could be a bias towards more informal argumentation structures in the USED corpus which might conflict with the formalized argumentation structures in the USED corpus. In section 5.2.2, we will perform some manual semantic analyses of positive and negative UNSC predictions to sample, albeit in an extremely limited manner, performance on the UNSC corpus.

5 Discussion

5.1 Sentiment Analysis

The results of the sentiment analysis seem to reveal a correlation of negative sentiment polarity and objectivity. The trend can be observed best in active countries for which we have a lot of data, e.g. the UNSC core nations and Germany. In order to confirm the validity of this correlation assumption, we would have to find viable means of evaluation. For this project, we manually evaluated a small number of speeches in order to get a grasp of the accuracy of our computed subjectivity and sentiment scores. The evaluation is not a simple task, as we are handling almost 65,000 speeches, and even evaluating 1% of the data manually would be a labor-intensive tasks.

While performing the manual evaluation, we realized that the sentiment scoring yielded very positive results for rather neutral speeches. We also observed that the subjectivity scores were mostly located around the 0.5 mark, indicating the speech is somewhere in between objective and subjective. We assume that this is due to the fact that the United Nations Security Council speeches use unusual language. The English is very formal and lacks colloquial words and intensifiers. This kind of formal language, combined with the specific defense related topics discussed, might not be captured well by the two sentiment frameworks we used for the analysis. For the subjectivity scores, we came to the conclusion that due to the spokesperson speaking on behalf of an entire country, the typical indicators of subjective speech are not necessarily present, even if the spokesperson expresses their own opinion.

After close inspection of speeches from an extensive set of countries, we also noticed patterns that could have possibly lead to the high subjectivity polarity in some cases. It is part of the rules of etiquette to thank the previous speaker and in some cases elaborate on the positive work they did relating to a topic. At the end of speeches, the spokesperson or a moderator introduces the new person in a similar fashion. These sections of affirmation are repetitive and not related to the topics themselves, yet they are included in the scoring since they are part of the speeches. This skews the scores and raises the sentiment polarity to a higher level. Furthermore, the speeches are oftentimes translated from the spokesperson's mother tongue to English, which can deprive the speeches of unique facets the speaker wanted to express since they can get lost in translation.

One last aspect we noticed was that the speeches are stripped of environmental factors, e.g. interruptions that happened or hesitation on the part of the speaker. This would add valuable information to both levels of our analysis.

Despite the issues we noticed during the evaluation, we believe our sentiment analysis provides interesting insights into the UNSC speech data. While the results should be taken with a grain of salt, they still depict the differing sentiments among participants. Especially the insights into the Iraq War related sessions allow for an interesting digression into this controversial topic.

Said insights revealed for example that the opponents of this military intervention had a lower ratio of uttered negative vs. positive sentences than the proponents did. Even after a simple visual exploration of the plots, it is evident that the USA were the driving force in this issue, as they contributed the most to the sessions related to Iraq. Especially in the second session we analyzed, the session that opened with Colin Powell's speech, it can be seen that the two proponents of the invasion of Iraq, the USA and the UK, had a very negative vocabulary. While the UK did not contribute as much to the topic as the USA, more than 50% of their contributions were negative. This degree of negativity cannot be observed in countries that held other opinions on the Iraq War.

5.2 Argumentation Mining

5.2.1 Model Performance Comparison

Since the USED corpus is a relatively new political-domain argumentation corpus, we were unable to find directly comparable results from other studies. However, we can make some approximate comparisons from related studies. As mentioned in 3.2.1, Haddadan et al. (2019) performed argumentation classification of the

USED corpus; albeit at a coarser sentence-level instead of the token-level. Their best model achieved 84.3% F_1 score for argumentative sentence identification and 67.3% F_1 score on claim/premise sentence classification. The better results could have stemmed from an intrinsically simpler sentence-level classification task.

Eger et al. (2017) conducted a similar methodology as ours on the PEC; with their best model achieving a 75.6% F_1 score for the argumentation tagging task. This result is definitely a positive one, however we would question the robustness of such a model due to likely symbolic overfitting on the small training PEC containing only 402 essays.

5.2.2 Prediction on UNSC

Due to time and resource limitations, we only manually review two predictions from our fine-tuned classifier on the UNSC corpus. We use the same coloring scheme for N, C and P tokens as per Figures 6 and 22. In the positive example, we can observe clear and expansive segmentation with the claim and premise being in appropriate locations with a discourse connective "but" between them. In the negative example, we can observe much more fragmentation of token spans; with the true premise after "because" being (mostly) wrongly labelled as a claim. The fragmentation of token spans is not entirely surprising, since such fragmented argumentation spans also exist in the USED corpus. The true premise being wrongly predicted as a claim is however a limitation of the classifier in this example.

Colour Scheme:

None (N) Claim (C) Premise (P)

Positive Example: UNSC_2004_SPV.5007_spch019

_we _have , _indeed , _a _broad _range _of _tools , _developed _in _accordance
_with _chapter _viii _of _the _charter , _to _facilitate _cooperation . _but
_we _need _fresh _ideas _in _order _to _improve _such _cooperation _and
_to _make _sure _that _stability _can _be _achieved _as _a _result _of
_cooperation _and _interaction .

Negative Example: UNSC_2009_SPV.6075_spch042

_just _very _briefly , _i _think _i _can _only _endorse _what _alain _le _roy
_has _just _said . _we _must _make _sure _that _we _commit _ourselves
_fully _to _actively _participating _in _this _process , _because _we _all
_see _that _the _outcome _of _such _a _good _dialogue _will _be _positive
_for _our _missions .

6 Conclusions

In this project, we analyzed the UNSC corpus using two different approaches, sentiment analysis and argumentation mining. We aimed to provide automatic annotations for this novel political speech corpus, as it currently lacks hand-written annotations which would greatly facilitate the work of NLP researchers, as well as political scientists. We introduced the reader to the background concepts of our analyses and explained our methodologies. The detailed description of our results and their discussion shows that we succeeded in fulfilling our goals. We pointed to factors future researchers should bare in mind, for example false-positive classification of speeches. Our final results and automatic annotations can be found in our GitHub repository[2].

While we do not provide definite and flawless annotations for the UNSC corpus, we succeeded in providing a well documented starting point for both sentiment and argumentation mining. We believe that the corpus can provide invaluable insights into the work and development of the United Nations Security Council and hope to have facilitated future research on this fascinating topic.

7 Recommendations

Given the resource and time limitations of our project, we were not able to perform extensive evaluations of our analyses. As this is fundamental for good scientific practice and the interpretation of results, we strongly recommend future researchers to employ an infrastructure for the evaluation of the produced annotations, e.g. by using crowdsourcing platforms.

As for the sentiment analysis, crowdsourcing manual annotations on word, n-gram or sentence level might also be an option worth considering. A partly annotated corpus would allow for machine learning analyses of the UNSC data and open up new perspectives. Furthermore, we think closer examination of specific subtopics might yield interesting results. Investigation of a subset of the corpus allows for a more fine-grained analysis and shows diverging opinions among countries better than a simple country-level analysis. Developing a strategy for identification and exclusion of non-relevant opening and closing words of speeches is also an open task that should be pursued.

For argumentation mining, we can make multiple recommendations for building upon our methodologies. On a data level, we could improve our training/validation and test data splits to create more homogeneous datasets. We could also attempt tagging tokens using the BIO scheme, as recommended by Eger et al. (2017), and check if that improves performance. In regards to fine-tuning, we recommend a transition from `TensorFlow` to `PyTorch` API's since there are more diverse collections of NLP-oriented deep-learning libraries which build upon the latter. Given limited hardware such as a single GPU, we would recommend memory-conserving training techniques such as gradient accumulation or checkpointing; which could allow for larger global batch-sizes and therefore less noisy gradients. We could also recommend multi-task training for the USED corpus with the second joint task being argumentation span linking using a graph neural network. In regards to actual application of models, we could recommend a thorough post-processing threshold analysis; such that the classifier could perform better on claim or premise tokens by intentional classification threshold design. Finally, we also propose using different pre-trained language models which could handle multiple sentences better; which include XLNet, RoBERTa or Google's recently released Reformer (or efficient transformer) model.

References

BBC (2004). Middle East - Iraq war illegal, says Annan. http://news.bbc.co.uk/2/hi/middle_east/3661134.stm.

CIA (2007). Iraqi Mobile Biological Warfare Agent Production Plants. https://www.cia.gov/library/reports/general-reports-1/iraqi_mobile_plants/index.html.

CNN (2003). China adds voice to Iraq war doubts. http://edition.cnn.com/2003/WORLD/asiapcf/east/01/23/sprj.irq.china/.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

Eger, S., Daxenberger, J., and Gurevych, I. (2017). Neural end-to-end learning for computational argumentation mining. *arXiv preprint arXiv:1704.06104*.

Golan, G. (2004). Russia and the Iraq War: was Putin's policy a failure? *Communist and Post-Communist Studies*, 37(4):429–459.

Groza, A. and Popa, O. (2016). Mining arguments from cancer documents using natural language processing and ontologies. pages 77–84.

Haddadan, S., Cabrio, E., and Villata, S. (2019). Yes, we can! mining arguments in 50 years of US presidential campaign debates. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4684–4690, Florence, Italy. Association for Computational Linguistics.

Harvard Gazette (2015). An inside view from Powell, complete with regrets. https://news.harvard.edu/gazette/story/2015/11/an-inside-view-from-powell-complete-with-regrets/.

Hutto, C. and Gilbert, E. (2015). Vader: A parsimonious rule-based model for sentiment analysis of social media text.

Kudo, T. (2018). Subword regularization: Improving neural network translation models with multiple subword candidates. *arXiv preprint arXiv:1804.10959*.

Kudo, T. and Richardson, J. (2018). Sentencepiece: A simple and language independent subword tokenizer and detokenizer for neural text processing. *arXiv preprint arXiv:1808.06226*.

Kuribayashi, T., Ouchi, H., Inoue, N., Reisert, P., Miyoshi, T., Suzuki, J., and Inui, K. (2019). An empirical study of span representations in argumentation structure parsing. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4691–4698.

Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., and Soricut, R. (2019). Albert: A lite bert for self-supervised learning of language representations.

Liu, B. (2012). Sentiment analysis and opinion mining. *Synthesis lectures on human language technologies*, 5(1):1–167.

New York Times (2003). France to Veto Resolution On Iraq War, Chirac Says.

Nielsen, F. Å. (2011). Afinn.

Peldszus, A. and Stede, M. (2015). An annotated corpus of argumentative microtexts. In *Argumentation and Reasoned Action: Proceedings of the 1st European Conference on Argumentation, Lisbon*, volume 2, pages 801–816.

Potash, P., Romanov, A., and Rumshisky, A. (2016). Here's my point: Joint pointer architecture for argument mining.

Ribeiro, F. N., Araújo, M., Gonçalves, P., André Gonçalves, M., and Benevenuto, F. (2016). SentiBench - a benchmark comparison of state-of-the-practice sentiment analysis methods. *EPJ Data Science*, 5(1):23.

Schönfeld, M., Eckhard, S., Patz, R., and van Meegdenburg, H. (2019). The un security council debates 1995-2017.

Sennrich, R., Haddow, B., and Birch, A. (2015). Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*.

Smedt, T. D. and Daelemans, W. (2012). Pattern for python. *Journal of Machine Learning Research*, 13(Jun):2063–2067.

Stab, C. and Gurevych, I. (2017). Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Stede, M. and Schneider, J. (2018). Argumentation mining. *Synthesis Lectures on Human Language Technologies*, 11(2):1–191.

The Guardian (2002a). Blix urges US and UK to hand over Iraq evidence. `https://www.theguardian.com/world/2002/dec/20/iraq.foreignpolicy`.

The Guardian (2002b). German leader says no to Iraq war. `https://www.theguardian.com/world/2002/aug/06/iraq.johnhooper`.

The Guardian (2002c). US will attack without approval. `https://www.theguardian.com/world/2002/nov/11/iraq.usa`.

UNSC (2002). Resolution 1441. `https://www.un.org/Depts/unmovic/documents/1441.pdf`.

Van Eemeren, F. and Grootendorst, R. (2004). *A systematic theory of argumentation: The pragma-dialectical approach*. Cambridge University Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis.

Wu, Y., Schuster, M., Chen, Z., Le, Q. V., Norouzi, M., Macherey, W., Krikun, M., Cao, Y., Gao, Q., Macherey, K., et al. (2016). Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.