# Fallacious Argument Classification in Political Debates

**Pierpaolo Goffredo**[1] , **Shohreh Haddadan**[2] , **Vorakit Vorakitphan**[1] ,
**Elena Cabrio**[1] and **Serena Villata**[1]
[1]Université Côte d'Azur, CNRS, Inria, I3S, France
[2]University of Luxembourg, Luxembourg
{pierpaolo.goffredo,elena.cabrio,serena.villata}@univ-cotedazur.fr, shohreh.haddadan@uni.lu,
vorakit.vorakitphan@inria.fr

## Abstract

Fallacies play a prominent role in argumentation since antiquity due to their contribution to argumentation in critical thinking education. Their role is even more crucial nowadays as contemporary argumentation technologies face challenging tasks as misleading and manipulative information detection in news articles and political discourse, and counter-narrative generation. Despite some work in this direction, the issue of classifying arguments as being *fallacious* largely remains a challenging and an unsolved task. Our contribution is twofold: first, we present a novel annotated resource of 31 political debates from the U.S. Presidential Campaigns, where we annotated six main categories of fallacious arguments (i.e., *ad hominem*, *appeal to authority*, *appeal to emotion*, *false cause*, *slogan*, *slippery slope*) leading to 1628 annotated fallacious arguments; second, we tackle this novel task of fallacious argument classification and we define a neural architecture based on transformers outperforming state-of-the-art results and standard baselines. Our results show the important role played by argument components and relations in this task.

## 1 Introduction

The notion of fallacy is inherently connected to argumentation. Standard dictionaries (such as the Oxford English Dictionary) record the varied meanings of a fallacy as an "invalid argument" or "faulty reasoning". Even if these meanings are related, they are also taken up by different disciplines with different emphasis: in logic the focus is on formally invalid arguments; in cognitive science on faulty, biased reasoning; in communication science on the deceptive and persuasive nature of fallacious discourse [Lewiński and Oswald, 2013]. In the pragma-dialectical theory of argumentation [Eemeren and Grootendorst, 1992; Eemeren, 2010], fallacies are defined as "derailments of strategic manoeuvring", meaning speech acts that violate the rules of a rational argumentative discussion for assumed persuasive gains. These derailments of strategic manoeuvring are particularly significant in political discourse, where informal fallacies are strategically employed by politicians to put forward their own positions [Mohammed and Lewinski, 2013; Zurloni and Anolli, 2013]. This deceptive strategic manoeuvring can lead to faulty and biased reasoning by the audience as well as to the subsequent formulation of further invalid arguments derived from those proposed by politicians. Therefore, the nefarious consequences of such misleading arguments has to be limited, being them comparable to those of propaganda and disinformation spread, with the goal to support critical thinking education.

In this paper, we address this issue by proposing a novel approach to automatically identify different categories of fallacious arguments in political debates. We first define and annotate the most prominent categories of fallacious arguments in political discourse on an existing dataset of U.S. presidential debates [Haddadan *et al.*, 2019], i.e., *ad hominem*, *appeal to authority*, *appeal to emotion*, *false cause*, *slogan*, *slippery slope*. We then train a neural classifier based on a transfomer-based model with an attention mechanism called Longformer [Beltagy *et al.*, 2020] to address the task in an automated way.

Our core contribution is twofold:

- We built a novel large linguistic resource of political debates where we annotated 1628 fallacious arguments. This is the largest existing dataset of political debates annotated with heterogeneous fallacious arguments (6 main categories, 14 sub-categories).
- We propose a new transformer-based model architecture, fine-tuned on argumentation features, and we address an extensive evaluation obtaining very promising results. We show that detecting argument components and relations in the debates is a necessary step to improve the model's result in classifying fallacious arguments.

The work we present in this paper is motivated by the lack of existing resources of fallacious argumentation in political discourse, and the need for effective methods to address this task. Despite the few existing approaches [Habernal at al. [2018a; 2018b]], classifying fallacious arguments largely remains an unsolved task. Our contribution advances the state of the art with a novel resource and an effective method.

## 2 Related Work

In the last years, there have been a few works aiming at automatically detecting fallacious content in argumentation, relying on Natural Language Processing methods. Habernal *et al.* [2017] developed the open-source software "Argotario", a gaming platform that serves both for educational purposes and as a crowd-sourcing data-acquisition platform to annotate fallacy types in everyday argumentation (*ad hominem*, *appeal to emotion*, *red herring*, *hasty generalization*, *irrelevant authority*). As a result of their study, they release an annotated dataset of fallacious arguments in both English and German. The German dataset contains 430 gold-labeled arguments. They experiment with fallacious argument type classification using Support Vector Machine and BiLSTM with German word vectors, achieving 50.9% accuracy and 42.1% macro-F1 on this dataset [Habernal *et al.*, 2018a]. In this work, each argument is considered to be entirely in one of the categories of "fallacy" or "non-fallacious".

Focusing only on the "Ad hominem" fallacy, [Habernal *et al.*, 2018b] created another annotated dataset from the Change My View subreddit on the Reddit platform, achieving a high inter-annotator agreement of 0.79 Cohen's $\kappa$. They investigate the importance of the context in distinguishing fallacious argumentation. They also prove insights into the triggers of the "Ad hominem" fallacy using a Self Attentive Embedding Neural Network Architecture. They obtain an accuracy of 0.810 using a Convolutional Neural Network architecture for the prediction of ad hominem arguments using a balanced dataset of negative/positive fallacies in 7,242 arguments.

[Sahai *et al.*, 2021] create a dataset of fallacious arguments, retrieving data using fallacy keywords extracted from Wikipedia on Reddit platform and then filter out false positives. They come up with 8 fallacy types recognized mostly in comments which correspond to some extent to the propaganda techniques annotated by [Da San Martino *et al.*, 2019]. They reach a moderate inter annotator agreement. Their dataset consists of 1708 fallacious arguments balanced over different fallacy types. They later examine different models to identify the occurrence of a fallacious argument and to classify the fallacy type both at comment and at token level.

Related to fallacies, [Da San Martino *et al.*, 2019] define an annotation scheme of 18 propaganda techniques and annotate 451 news articles with such labels, ending up with a dataset of 7,485 spans containing propaganda techniques. Such dataset is released in the context of the shared task NLP4IF'19[1] on fine-grained propaganda detection. The same authors propose a multi-granularity network architecture on top of the BERT contextualized embeddings to identify propagandist samples on different levels of granularity, e.g., document-level, paragraph-level, sentence-level. As a follow up, in 2020 another shared task on the same topic was proposed at SemEval (T11) [Da San Martino *et al.*, 2020] reducing the number of propaganda categories, and proposing a more restrictive evaluation scheme. As already introduced, some of the categories used to annotate propaganda partially overlap with the ones used to define fallacious arguments.

[1]https://propaganda.qcri.org/nlp4if-shared-task/

## 3 The ElecDeb60To16-fallacy Dataset

To investigate fallacious arguments in political debates, we extend the annotations of the ElecDeb60To16 dataset [Haddadan *et al.*, 2019], that collects televised debates of the presidential election campaigns in the U.S. from 1960 to 2016. To the best of our knowledge, it is the biggest available dataset of political debates annotated with both argument components (evidence, claim) and relations (support, attack). Given that we aim at investigating possible correlations in the occurrence of fallacious arguments within certain argument components or relations, this dataset is an optimal starting point.

### 3.1 List of Annotated Fallacies

Before starting the annotation process, an expert annotator carried out an exploratory study on the arguments put forward by the candidates in the *ElecDeb60To16* dataset, to verify which of the fallacy types – among those mentioned in the annotation scheme of [Da San Martino *et al.*, 2019] and in the categorization of [Walton, 1987] – were mainly present in political discourse. As a result, in this study, we decided to focus on the 6 types of fallacies which occur more frequently in political debates, meaning *Ad Hominem*, *Appeal to Emotion*, *Appeal to Authority*, *Slippery Slope*, *False Cause*, and *Slogans*. The first three types of fallacies are further divided into sub-categories. In the rest of this section, we provide a definition of each of these categories, and Table 1 shows some examples from the *ElecDeb60To16* dataset.

**Ad Hominem.** When the argument becomes an excessive attack on an arguer's position [Walton, 1987]. It covers the three sub-types defined in [Habernal *et al.*, 2018b], e.g., *general ad hominem* (an attack on the character of the opponent), *tu quoque ad hominem* (the "You did it first" attack) and *bias ad hominem* (an attack in which the arguer implies that the opponent is personally benefiting from his stance in the argument); and *Name-calling,Labeling*, i.e., when the arguer calls the opponent by an offensive label.

**Appeal to Emotion.** The unessential loading of the argument with emotional language to exploit the audience emotional instinct. Sub-categories: *appeal to pity*, *appeal to fear*, *loaded language* (i.e., increasing the intensity of a phrase by using emotionally loaded descriptive phrases - either positive or negative) and *flag waving*, which appeals to the emotion of a group of people by referring to their identity.

**Appeal to Authority.** When the arguer mentions the name of an authority or a group of people who agreed with her claim either without providing any relevant evidence, or by mentioning popular non-experts, or the acceptance of the claim by the majority.

**Slippery Slope.** It suggests that an unlikely exaggerated outcome may follow an act. The intermediate premises are usually omitted and a starting premise is usually used as the first step leading to an exaggerated claim.

**False Cause.** The misinterpretation of the correlation of two events for causation [Walton, 1987]. Politicians tend to apply this technique when they affiliate the cause of an improvement to their party, or the failure to their opponent's party.

| Fallacy Category | Sub-category | Sample |
|---|---|---|
| Ad Hominem | General | You were totally out of control. |
| | Bias ad hominem | But when I look at what you have proposed, you have what is called now the Trump loophole, because it would so advantage you and the business you do. |
| | Tu quoque | First of all, what my opponent wants you to forget is that he voted to authorize the use of force and now says it's the wrong war at the wrong time at the wrong place. |
| | Name-calling,Labeling | Such a nasty woman! |
| Appeal to Emotion | Appeal to fear | These terrorists are serious, they're deadly, and they know nothing except trying to kill. |
| | Appeal to pity | I think of the man who grabbed me by the shoulder once with tears in his eyes and said his daughter was dying of cancer and he thanked me for giving him a chance to spend some time with her without losing his job because of the Family and Medical Leave Act. |
| | Loaded Language | we pointed out how ridiculous this attempt was by the Environmental Protection Agency. |
| | Flag waving | Communism is the enemy of all religions; and we who do believe in God must join together. |
| Appeal to authority | Without evidence | Admiral Mullen suggests that Senator Obama's plan is dangerous for America. |
| | False authority | I don't think General Douglas MacArthur would like that too much. |
| | Popular opinion | Let me just tell you who the jury is. The people of Texas. There's only been one governor ever elected to back-to-back four-year terms, and that was me |
| Slippery Slope | | Now what do the Chinese Communists want? They don't want just Quemoy and Matsu; they don't want just Formosa; they want the world |
| False cause | | In a place like Chicago, where thousands of people have been killed, thousands over the last number of years, in fact, almost 4000 have been killed since Barack Obama became president, over almost 4000 people in Chicago have been killed. |
| Slogan | | Make America great again! |

Table 1: Fallacious argument categories and sub-categories with examples from the ElecDeb60To16 dataset.

**Slogan.** It is a brief and striking phrase used to provoke excitement of the audience, and is often accompanied by another type of fallacy called *argument by repetition*.

### 3.2 Annotation Phase

After carefully defining the annotation schema presented in the previous section and describing it into annotation guidelines, three annotators with background in computational linguistics carried out the annotation of the dataset. After a first annotation round on a data sample, the guidelines were updated to conciliate the disagreements among the annotators, in particular with respect to the boundaries of the spans to be annotated. Afterwards, 9 sections[2] from 5 different debates from different years were annotated to compute inter-annotator agreement, reported in Table 2 as moderate agreement. Annotations were then reconciled by an expert annotator before adding them to the released dataset.

The semantic annotation platform INCEpTION [Klie *et al.*, 2018] was used to perform the annotation process. Annotators were presented with raw data, meaning that we hide the existing annotations of argument components and relations present in the original ElecDeb60To16 dataset, to avoid annotation bias. The rest of the dataset was randomly and equally split among the three annotators, that carried out independently the annotation task[3]. As a result, 31 debates of the ElecDeb60To16 dataset are fully annotated with fallacies[4].

| **Fallacy Type** | **Observed Agr.** | **Krippendorff's $\alpha$** |
|---|---|---|
| Ad Hominem | 0.9961 | 0.5315 |
| Appeal to Authority | 0.9945 | 0.5806 |
| Appeal to Emotion | 0.9759 | 0.4640 |
| Slogans | 0.9989 | 0.5995 |

Table 2: IAA, three annotators, 9 sections from 5 different debates (only 4 types of fallacies were present in the annotated data sample.)

### 3.3 Statistics and Data Analysis

Table 3 reports on the number of annotated fallacious argument types per category. Appeal to Emotion is the most frequent fallacy in the ElecDeb60To16 dataset (1016 instances), while Slippery Slope is the less frequent one (57 instances).

We tokenized the annotated fallacious arguments to compute the average number of words in each sub-category. As expected, Slogans, Name Calling and Loaded Language cover the shortest spans, with respectively 8.31, 9.93 and 11.5 tokens on average.

Table 4 shows the distribution of fallacy categories over the debate years. Given that for some years there is more than one

[2]Debates in the ElecDeb60To16 are divided into sections, where each section starts with the moderator/panelist or an audience member asking a question on a new topic [Haddadan *et al.*, 2019].

[3]To avoid biased annotations, we ensured the fair conduct of the annotators by hiding them the candidate's identity.

[4]https://github.com/pierpaologoffredo/IJCAI2022

| Category | Sub-category | Freq. |
|---|---|---|
| Ad Hominem (188) | General | 79 |
| | Bias ad hominem | 48 |
| | Tu quoque | 30 |
| | Name-calling,Labeling | 31 |
| Appeal to Emotion (1016) | Appeal to fear | 87 |
| | Appeal to pity | 102 |
| | Loaded Language | 676 |
| | Flag waving | 151 |
| Appeal to authority (234) | Without evidence | 125 |
| | False authority | 43 |
| | Popular opinion | 66 |
| Slippery Slope | | 57 |
| False cause | | 69 |
| Slogan | | 64 |

Table 3: The number of annotated fallacies annotated per category.

debate, to carry out additional analysis of diachronic changes in debating strategies, we perform data normalization. More specifically, given that debates can have a different length, different number of speech turns and even a different number of candidates debating (e.g., in 1992), we normalize the length of the fallacious snippet with respect to the total length of the debates in one year. An example analysis (that could be of interest to scholars in political sciences) that can be made on the ElecDeb60To16-fallacy dataset is on the presence of fallacious arguments of type "Ad Hominem" in the different debate years (see Table 4), interesting showing that such strategy was often used by candidates in 2016. It is also interesting to underline that, in the debates of 2004, a higher percentage of fallacies from the Appeal to Fear sub-category (of Appeal to Emotion) was employed with respect to other debate years, that might be due to the fact that the dominating topic of 2004 debates was the Iraq war.

## 4 Fallacy Classification in Political Debates

In this section, we first describe the fallacious argument classification task, then we report on the experimental setting.

### 4.1 Fallacious Argument Classification

We cast this task as a sequence classification problem. First, we focus on a multi-class classification task to classify the fallacies observed in the debates. Then, we enhance our classifier with argumentation-based features (i.e., argument components and relations) within each fallacious argument. To this end, we test and adapt SOTA language models based on the transformer architecture as they are challenging baselines to compare with. In particular, the well-known Pre-trained Language Models (PLM) BERT [Devlin *et al.*, 2019] and RoBERTa [Liu *et al.*, 2019] are considered as baselines.

Each debate is composed of two parts: *i)* the portion of the debate containing the fallacious argument in the presidential debate, and *ii)* the fallacious argument snippet itself. Additional information are the year of the debate, the date, and the section (which starts with the moderator's question introducing a new topic to be discussed) (see Section 3 for details).

Our objective is not only to identify and classify the fallacies, but to do so taking in account the context in the debate in which they are put forward. BERT [Devlin *et al.*, 2019] shows some limitations in our setting, given that it allows a maximum sequence length of 512. Whilst a fallacy can fit in this length limit, this is not the case for the entire context of the fallacious argument. Each speech is significantly longer as it contains the arguments proposed by both candidates.

In this work, we use more advanced PLMs to tackle such lengthy speeches, i.e., Longformer [Beltagy *et al.*, 2020] and Transformer-XL [Dai *et al.*, 2019] which have the ability to capture long-input texts to perform the classification. Longformer is a transfomer-based model with an attention mechanism that scales linearly with sequence length, making it easy to process documents of thousands of tokens or longer. The attention mechanism of Longformer is a combination of a windowed local-context self-attention and the global attention of the context. The local attention of Longformer is primarily used to build contextual representations, while the global attention allows Longformer to build full sequence representations for prediction. The model is pre-trained with the Masked Language Modeling (MLM) approach, similarly to RoBERTa [Liu *et al.*, 2019]. Transformer-XL, instead, enables learning dependency beyond a fixed length without disrupting temporal coherence. Combining recurrence and relative positional encoding, it can model longer-term dependency than RNNs and vanilla Transformers.

In the different experimental settings, the features we considered for the fallacious argument classification task are the following: political discourse speech context, fallacious argument snippet, argument component and relation labels. The argument component feature refers to the two basic argument components, i.e., *premise* and *claim*, and the argument relation feature refers to bipolar argument relations, i.e., *attack* and *support*, plus the *equivalence* relation [Cabrio and Villata, 2018; Lawrence and Reed, 2019]. These features are extracted from the annotated argument components and relations in the ElecDeb60To16 dataset [Haddadan *et al.*, 2019].

**Baselines.** For the tested architecture with BERT and RoBERTa baselines, we use the same transformer model to produce logits (*L*) regarding the snippet-level with the default pre-trained model *bert-base-uncased, roberta-base*, learning rate of 5e-3, and $\alpha$ of 0.5.

**Proposed Architecture.** Our approach is based on the Longformer model empowered with the argumentation features, and the context of the fallacious argument in the debate. Figure 1 visualises our neural architecture for fallacious argument classification. Each debate is processed into four components: the dialogue context, the fallacious argument snippet, the argument component, and argument relation. Each component is then extracted in the embedded vectors using the PLM of interest. Each embedding has its own transformer-based classifier to finally obtain a logit (*L*). All transformer models apply Adam optimizer, dropout 0.1, and CrossEntropy as a loss function. The loss is produced per classifier i.e., fallacy-snippet ($loss_{snippet}$), speech ($loss_{speech}$), argument component ($loss_{ArgComp}$), and argument relation ($loss_{ArgRel}$). We then join the *loss* of each classifier to have a joint-loss learning with $\alpha = 0.5$ [Vorakitphan *et al.*, 2021]. We ar-

| Year of Debate | Number of Debates | Ad Hominem | Appeal to Authority | Appeal to Emotion | False Cause | Slippery Slope | Slogans | Average per debate | Total |
|---|---|---|---|---|---|---|---|---|---|
| **1960** (Kennedy-Nixon) | 4 | 10 | 24 | 95 | 12 | 12 | 1 | 38.5 | 154 |
| **1976** (Carter-Ford) | 3 | 5 | 8 | 42 | 4 | 4 | 4 | 22.3 | 67 |
| **1980** (Carter-Reagan) | 2 | 5 | 12 | 77 | 2 | 3 | 5 | 52 | 104 |
| **1984** (Mondale-Reagan) | 2 | 3 | 13 | 35 | 3 | 3 | 3 | 30 | 60 |
| **1988** (Bush-Dukakis) | 1 | 4 | 19 | 31 | 2 | 3 | 4 | 63 | 63 |
| **1992** (Bush-Clinton-Perot) | 2 | 11 | 19 | 74 | 8 | 3 | 2 | 58.5 | 117 |
| **1996** (Clinton-Dole) | 2 | 10 | 24 | 93 | 6 | 2 | 10 | 72.5 | 145 |
| **2000** (Bush-Gore) | 2 | 8 | 25 | 140 | 5 | 8 | 11 | 98.5 | 197 |
| **2004** (Bush-Kerry) | 4 | 32 | 38 | 135 | 13 | 10 | 4 | 58 | 232 |
| **2008** (Mccain-Obama) | 3 | 7 | 21 | 67 | 4 | 1 | 2 | 34 | 102 |
| **2012** (Obama-Romney) | 1 | 0 | 2 | 16 | 1 | 1 | 2 | 22 | 22 |
| **2016** (Clinton-Trump) | 3 | 93 | 29 | 211 | 9 | 7 | 16 | 121.6 | 365 |
| **Total** | 31 | 188 | 234 | 1016 | 69 | 57 | 64 | 52.5 | 1628 |

Table 4: Distribution of annotated fallacious argument spans among different debate years.

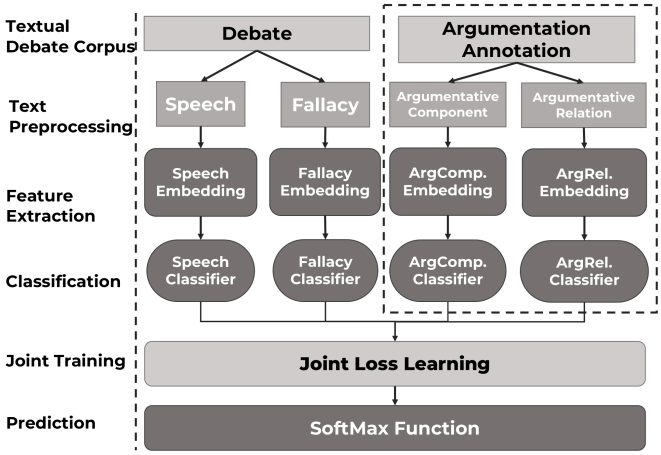


Figure 1: Pipeline for the task of fallacious argument classification.

range these alignments of $L$ to calculate the average loss as a joint loss ($loss_{joint\_loss}$) from each *loss* element. The function used before back-propagation is $loss_{joint\_loss} = \alpha * \frac{(loss\_speech+loss\_sentence+loss\_comp+loss\_rel)}{N\_loss}$ where $N_{loss}$ stands for the number of *loss* elements taken into the model.

## 4.2 Experimental Setup

We apply different experimental settings. We first evaluate which model performs better on the classification of the main categories of fallacies. The transformer-based PLMs used in this setting are: BERT, RoBERTa, Longformer and Transformer-XL. The implementation is based on the huggingface transformer[5] using PyTorch (version 1.7.0). The selected learning rate is 5e-5, dropout 0.1 and batch size 1 for Transformer-XL, and 8 for the rest. For each model, the max size length used for the number of tokens changes: in BERT and RoBERTa models we used 128 for the fallacious argument snippet, in Transformer-XL[6] the fallacious argument snippet is set to 128 and the context speech text to 8192, and finally, in the Longformer we used 128 and 4096 respectively for the fallacious argument snippet and the speech context.

We also performed additional experiments with the best performing model (Longformer + $loss_{joint\_loss}$) to *i)* classify the 14 fallacious argument sub-categories, i.e., General Ad hominem, Bias ad hominem, Tu quoque, Name-calling Labeling, Appeal to fear, Appeal to pity, Loaded Language, Flag waving, Without evidence, False authority, Popular opinion, Slippery Slope, False cause, Slogan; and *ii)* classify the main categories, enriching the dataset with the argument component and relation features in an ablation test setting. In all the experimental settings, 80% of the dataset has been used for training and the remaining 20% for testing. We performed the train_and_test_split by sklearn to create the training and test sets with a random seed for the label distribution. To average the results, we performed experiments 3 times.

# 5 Evaluation

Table 5 shows the results we obtained with our transformer-based neural architecture for fallacious argument classification. Comparing the different models and approaches, we can observe that the highest F1-score is achieved by the Longformer with $loss_{joint\_loss}$ method. Results are very promising, in particular with the use of argumentation features, and outperform existing approaches like [Habernal *et al.*, 2018a; Habernal *et al.*, 2018b]. These results lead us to select this model as the proposed architecture and to experiment with it on the task of fallacious argument classification on sub-categories. Table 6 reports obtained results on this second task, showing a good performance on some labels, i.e., Flag waving, Slogan, Loaded language, and Without Evidence, which unsurprisingly are the most represented in our data.

**Ablation Test.** Furthermore, we performed an ablation test on the multi-class classification setting to show the impact of the argumentative features (both components and relations) on the classification of the main fallacious argument categories. Table 7 reports, for each of the main fallacious argument categories, the results obtained without considering the argumentation features, those obtained considering only the argument component and the argument relation features respectively, and finally the results obtained by combining both

[5]https://huggingface.co/docs/transformers/index

[6]Given that the high memory demand increased exponentially, we have been forced to set a max length.

| Model | Dataset | $loss_{joint\_loss}$ | Argum. Features | Precision | Recall | Macro avg F1-Score |
|---|---|---|---|---|---|---|
| BERT | Fallacy Main Category | No | None | 0,62 | 0,55 | 0,55 |
| RoBERTa | Fallacy Main Category | No | None | 0,58 | 0,56 | 0,53 |
| Longformer | Fallacy Main Category | No | None | 0,64 | 0,6 | 0,57 |
| Longformer | Fallacy Main Category | Yes | None | 0,66 | 0,61 | **0,61** |
| Transformer-XL | Fallacy Main Category | No | None | 0,61 | 0,45 | 0,47 |
| Transformer-XL | Fallacy Main Category | Yes | None | 0,61 | 0,51 | 0,53 |
| Longformer | Fallacy Sub-category | Yes | None | 0,44 | 0,45 | 0,42 |
| Longformer | Fallacy Main Category | Yes | Component Label | 0,88 | 0,81 | **0,83** |
| Longformer | Fallacy Main Category | Yes | Relation Label | 0,87 | 0,81 | **0,83** |
| Longformer | Fallacy Main Category | Yes | Comp + Rel Labels | 0,84 | 0,85 | **0,84** |

Table 5: Results of the multi-class sequence tagging task, on the average of three runs.

| | Precision | Recall | F1-score |
|---|---|---|---|
| Ad hominem | 0,47 | 0,60 | 0,52 |
| Appeal to fear | 0,47 | 0,41 | 0,43 |
| Appeal to pity | 0,60 | 0,47 | 0,51 |
| Appeal to popular opinion | 0,68 | 0,49 | 0,50 |
| Circumstantial Ad hominem | 0,27 | 0,30 | 0,28 |
| FalseAuthority | 0,00 | 0,00 | 0,00 |
| Falsecause | 0,19 | 0,44 | 0,27 |
| Flagwaving | 0,62 | 0,70 | **0,65** |
| LoadedLanguage | 0,85 | 0,82 | **0,83** |
| Name-Calling,Labeling | 0,33 | 0,22 | 0,27 |
| Slipperyslope | 0,45 | 0,31 | 0,32 |
| Slogan | 0,69 | 0,69 | **0,68** |
| Tuquoque | 0,00 | 0,00 | 0,00 |
| WithoutEvidence | 0,48 | 0,78 | **0,57** |
| accuracy | | | 0,63 |
| macro_avg | 0,44 | 0,45 | **0,42** |
| weighted_avg | 0,62 | 0,63 | 0,61 |

Table 6: Results of the multi-class sequence tagging task on the different categories, on the average of three runs.

| | **Original dataset** F1 | **Arg. Comp.** F1 | **Arg. Rel.** F1 | **Arg. Comp. & Rel.** F1 |
|---|---|---|---|---|
| Ad Hominem | 0,56 | 0,85 | 0,81 | 0,81 |
| Appeal to Auth. | 0,65 | 0,85 | 0,84 | 0,91 |
| Appeal to Em. | 0,85 | 0,93 | 0,93 | 0,94 |
| False Cause | 0,43 | 0,80 | 0,82 | 0,80 |
| Slippery slope | 0,50 | 0,78 | 0,79 | 0,84 |
| Slogans | 0,67 | 0,76 | 0,88 | 0,77 |
| accuracy | 0,75 | 0,88 | 0,89 | 0,89 |
| macro_avg | **0,61** | **0,83** | **0,83** | **0,84** |
| weighted_avg | 0,74 | 0,88 | 0,89 | 0,89 |

Table 7: Ablation test with argumentative features in details.

of them[7]. We can observe that adding the argumentation features to the neural model allows it to achieve even more satisfactory results. This is due to the additional context given by the argumentative components to the fallacious argument, in addition to the pure speech context extracted from the debate. The approach used to fine-tune the architecture proposed with these additional elements is based on the further addition of the component label loss and component relation label loss in the $loss_{joint\_loss}$ function (cf. Section 4.1).

[7]E.g., Slogans are mostly stated in claims, while False Cause and Slippery Slope include both the premises and claim of an argument.

| **Fallacy snippet** | **True** | **Pred.** |
|---|---|---|
| It is time for a change. | Slogans | App.Em. |
| I think if you raise taxes during a recession, you head to depression. | App.Em. | Slip.Sl. |
| Bill Clinton, as President, has provided that kind of leadership. We are more secure and stronger today because of Bill Clinton's handling of foreign policy. | FalseC. | AdHom. |

Table 8: Examples of wrong predictions by our system in the experimental setting using argumentative features.

**Error Analysis.** Table 8 shows some miss-classified examples requiring the injection of external knowledge (e.g., Wikipedia articles, archival articles) to be correctly classified. Most of the errors made by our best model are due to the fact that Slogans, Appeal to Emotions and Ad Hominem aim to manipulate the perceived sentiment of the audience, therefore share a similar vocabulary and an appeal to empathy. The reason lies also in the subtle notion of fallacious argument itself, which is sometimes hard to pin down to a single category of fallacy. A possible solution consists in addressing a new annotation of the dataset to assign one (or more) secondary fallacy category to the identified fallacious arguments.

## 6 Concluding Remarks

Fallacies remain a controversial issue in argumentation, as the argumentation schemes to identify such bad or invalid types of reasoning is somehow ineffective when applied to real world scenarios, like political debates [Boudry *et al.*, 2015]. The problem is that they abstract away from the specific content and dialectical context of the fallacy. We plan to empirically investigate the connection between the argumentative content and the context of the fallacy in our dataset. Furthermore, almost every known type of fallacy is a close neighbor to sound arguments in a debate. We will study how to generate sound arguments out of the identified fallacies and their context. Finally, the investigation of how to counter the formal invalidity of these fallacious arguments through newly generated counter-arguments remains a challenging follow up of this work.

## Acknowledgements

This work was partly supported by the French government, through the 3IA Côte d'Azur Investments in the Future project managed by the National Research Agency (ANR) with the reference number ANR-19-P3IA-0002. This work was partly supported also by EU Horizon 2020 project AI4Media, under contract no. 951911 (https://ai4media.eu/). Shohreh Haddadan hereby acknowledges that this research is supported by the Luxembourg National Research Fund (FNR) (10929115).

## References

[Beltagy *et al.*, 2020] Iz Beltagy, Matthew E. Peters, and Arman Cohan. Longformer: The long-document transformer. *CoRR*, abs/2004.05150, 2020.

[Boudry *et al.*, 2015] Maarten Boudry, Fabio Paglieri, and Massimo Pigliucci. The fake, the flimsy, and the fallacious: demarcating arguments in real life. *ARGUMENTATION*, 29(4):431–456, 2015.

[Cabrio and Villata, 2018] Elena Cabrio and Serena Villata. Five years of argument mining: a data-driven analysis. In *Proceedings of IJCAI 2018*, pages 5427–5433, 2018.

[Da San Martino *et al.*, 2019] Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, Rostislav Petrov, and Preslav Nakov. Fine-grained analysis of propaganda in news article. In *Proceedings of EMNLP-IJCNLP 2019*, pages 5636–5646. ACL, 2019.

[Da San Martino *et al.*, 2020] Giovanni Da San Martino, Alberto Barrón-Cedeño, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. SemEval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of SemEval 2020*, 2020.

[Dai *et al.*, 2019] Zihang Dai, Zhilin Yang, Yiming Yang, Jaime G. Carbonell, Quoc V. Le, and Ruslan Salakhutdinov. Transformer-xl: Attentive language models beyond a fixed-length context. *CoRR*, abs/1901.02860, 2019.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT 2019*, pages 4171–4186, 2019.

[Eemeren and Grootendorst, 1992] Frans H. Van Eemeren and Rob Grootendorst. *Argumentation, Communication, and Fallacies a Pragma-Dialectical Perspective*. Routledge, 1992.

[Eemeren, 2010] Frans H. Van Eemeren. *Strategic Maneuvering in Argumentative Discourse. Extending the Pragma-Dialectical Theory of Argumentation*. Amsterdam-Philadelphia: John Benjamins, 2010.

[Habernal *et al.*, 2017] Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. Argotario: Computational Argumentation Meets Serious Games. In *Proceedings of EMNLP 2017 (System Demonstrations)*, pages 7–12. ACL, 2017.

[Habernal *et al.*, 2018a] Ivan Habernal, Patrick Pauli, and Iryna Gurevych. Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices. In *Proceedings of LREC 2018*. ELRA, 2018.

[Habernal *et al.*, 2018b] Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. Before Name-Calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation. In *Proceedings of NAACL 2018*, pages 386–396. ACL, 2018.

[Haddadan *et al.*, 2019] Shohreh Haddadan, Elena Cabrio, and Serena Villata. Yes, we can! mining arguments in 50 years of us presidential campaign debates. In *Proceedings of ACL 2019*, pages 4684–4690, 2019.

[Klie *et al.*, 2018] Jan-Christoph Klie, Michael Bugert, Beto Boullosa, Richard Eckart de Castilho, and Iryna Gurevych. The inception platform: Machine-assisted and knowledge-oriented interactive annotation. In *Proceedings of ACL 2018 (System Demonstrations)*, pages 5–9. Association for Computational Linguistics, June 2018.

[Lawrence and Reed, 2019] John Lawrence and Chris Reed. Argument mining: A survey. *Comput. Linguistics*, 45(4):765–818, 2019.

[Lewiński and Oswald, 2013] Marcin Lewiński and Steve Oswald. When and how do we deal with straw men? a normative and cognitive pragmatic account. *Journal of Pragmatics*, 59:164–177, 2013.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[Mohammed and Lewinski, 2013] Dima Mohammed and Marcin Lewinski, editors. *Argumentation in political deliberation*. John Benjamins Publishing, 2013.

[Sahai *et al.*, 2021] Saumya Sahai, Oana Balalau, and Roxana Horincar. Breaking Down the Invisible Wall of Informal Fallacies in Online Discussions. In *Proceedings of ACL 2021*, pages 644–657, Online, August 2021. ACL.

[Vorakitphan *et al.*, 2021] Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. "Don't discuss": Investigating Semantic and Argumentative Features for Supervised Propagandist Message Detection and Classification. In *Proceedings of RANLP 2021*, 2021.

[Walton, 1987] Douglas Walton. *Informal Fallacies : Towards a Theory of Argument of Criticisms*. John Benjamins Publishing Company, Philadelphia, 1987.

[Zurloni and Anolli, 2013] Valentino Zurloni and Luigi Anolli. Fallacies as argumentative devices in political debates. In Isabella Poggi, Francesca D'Errico, Laura Vincze, and Alessandro Vinciarelli, editors, *Multimodal Communication in Political Speech. Shaping Minds and Social Action*, pages 245–257. Springer, 2013.