Before Name-calling: Dynamics and Triggers of Ad Hominem Fallacies in Web Argumentation

Ivan Habernal[†] **Henning Wachsmuth**[‡] **Iryna Gurevych**[†] **Benno Stein**[‡]

[†] Ubiquitous Knowledge Processing Lab (UKP) and Research Training Group AIPHES
Department of Computer Science, Technische Universität Darmstadt, Germany
www.ukp.tu-darmstadt.de www.aiphes.tu-darmstadt.de

[‡] Faculty of Media, Bauhaus-Universität Weimar, Germany
<firstname>.<lastname>@uni-weimar.de

Abstract

Arguing without committing a fallacy is one of the main requirements of an ideal debate. But even when debating rules are strictly enforced and fallacious arguments punished, arguers often lapse into attacking the opponent by an *ad hominem* argument. As existing research lacks solid empirical investigation of the typology of ad hominem arguments as well as their potential causes, this paper fills this gap by (1) performing several large-scale annotation studies, (2) experimenting with various neural architectures and validating our working hypotheses, such as controversy or reasonableness, and (3) providing linguistic insights into triggers of ad-hominem using explainable neural network architectures.

1 Introduction

Human reasoning is lazy and biased but it perfectly serves its purpose in the argumentative context (Mercier and Sperber, 2017). When challenged by genuine back-and-forth argumentation, humans do better in both generating and evaluating arguments (Mercier and Sperber, 2011). The dialogical perspective on argumentation has been reflected in argumentation theory prominently by the pragma-dialectic model of argumentation (van Eemeren and Grootendorst, 1992). Not only sketches this theory an ideal normative model of argumentation but also distinguishes the wrong argumentative moves, *fallacies* (van Eemeren and Grootendorst, 1987). Among the plethora of prototypical fallacies, notwithstanding the controversy of most taxonomies (Boudry et al., 2015), *ad hominem* argument is perhaps the most famous one. Arguing *against the person* is considered faulty, yet is prevalent in online and offline discourse.[1]

Although the ad hominem fallacy has been known since Aristotle, surprisingly there are very few empirical works investigating its properties. While Sahlane (2012) analyzed ad hominem and other fallacies in several hundred newspaper editorials, others usually only rely on few examples, as observed by de Wijze (2002). As Macagno (2013) concludes, ad hominem arguments should be considered as multifaceted and complex strategies, involving not a simple argument, but several combined tactics. However, such research, to the best of our knowledge, does not exist. Very little is known not only about the feasibility of ad hominem theories in practical applications (the NLP perspective) but also about the dynamics and triggers of ad hominem (the theoretical counterpart).

This paper investigates the research gap at three levels of increasing discourse complexity: ad hominem in isolation, direct ad hominem without dialogical exchange, and ad hominem in large inter-personal discourse context. We asked the following research questions. First, what qualitative and quantative properties do ad hominem arguments have in Web debates and how does that reflect the common theoretical view (RQ1)? Second, how much of the debate context do we need for recognizing ad hominem by humans and machine learning systems (RQ2)? And finally, what are the actual triggers of ad hominem arguments and can we predict whether the discussion is going to end up with one (RQ3)?

We tackle these questions by leveraging Web-

[1] According to 'Godwin's law' known from the internet pop-culture (https://en.wikipedia.org/wiki/Godwin's_law), if a discussion goes on long enough, sooner or later someone will compare someone or something to Adolf Hitler.

based argumentation data (*Change my View* on Reddit), performing several large-scale annotation studies, and creating a new dataset. We experiment with various neural architectures and extrapolate the trained models to validate our working hypotheses. Furthermore, we propose a list of potential linguistic and rhetorical triggers of ad hominem based on interpreting parameters of trained neural models.[2] This article thus presents the first NLP work on multi-faceted ad hominem fallacies in genuine dialogical argumentation. We also release the data and the source code to the research community.[3]

2 Theoretical background and related work

The prevalent view on argumentation emphasizes its pragmatic goals, such as persuasion and group-based deliberation (van Eemeren et al., 2014), although numerous works have dealt with argument as product, that is, treating a single argument and its properties in isolation (Toulmin, 1958; Habernal and Gurevych, 2017). Yet the social role of argumentation and its alleged responsibility for the very skill of human reasoning explained from the evolutionary perspective (Mercier and Sperber, 2017) provide convincing reasons to treat argumentation as an inherently dialogical tool.

The observation that some arguments are in fact 'deceptions in disguise' was made already by Aristotle (Aristotle and Kennedy (translator), 1991), for which the term *fallacy* has been adopted. Leaving the controversial typology of fallacies aside (Hamblin, 1970; van Eemeren and Grootendorst, 1987; Boudry et al., 2015), the *ad hominem* argument is addressed in most theories. Ad hominem argumentation relies on the strategy of attacking the opponent and some feature of the opponent's character instead of the counter-arguments (Tindale, 2007). With few exceptions, the following five sub-types of ad hominem are prevalent in the literature: **abusive ad hominem** (a pure attack on the character of the opponent), **tu quoque ad hominem** (essentially analogous to the "He did it first" defense of a three-year-old in a sandbox), **circumstantial ad hominem** (the "practice what you preach" attack and accusation of hypocrisy), **bias ad hominem** (the attacked opponent has a hidden agenda), and **guilt by association** (associating the opponent with somebody with a low credibility) (Schiappa and Nordin, 2013; Macagno, 2013; Walton, 2007; Hansen, 2017; Woods, 2008).

The topic of fallacies, which might be considered as sub-topic of argumentation quality, has recently been investigated also in the NLP field. Existing works are, however, limited to the monological view (Wachsmuth et al., 2017; Habernal and Gurevych, 2016b,a; Stab and Gurevych, 2017) or they focus primarily on learning fallacy recognition by humans (Habernal et al., 2017, 2018a). Another related NLP sub-field includes abusive language and personal attacks in general. Wulczyn et al. (2017) investigated whether or not Wikipedia talk page comments are personal attacks and annotated 38k instances resulting in a highly skewed distribution (only 0.9% were actual attacks). Regarding the participants' perspective, Jain et al. (2014) examined principal roles in 80 discussions from the *Wikipedia: Article for Deletion* pages (focusing on stubbornness or ignoredness, among others) and found several typical roles, including 'rebels', 'voices', or 'idiots'. In contrast to our data under investigation (Change My View debates), Wikipedia talk pages do not adhere to strict argumentation rules with manual moderation and have a different pragmatic purpose.

Reddit as a source platform has also been used in other relevant works. Saleem et al. (2016) detected hateful speech on Reddit by exploiting particular sub-communities to automatically obtain training data. Wang et al. (2016) experimented with an unsupervised neural model to cluster social roles on sub-reddits dedicated to computer games. Zhang et al. (2017) proposed a set of nine comment-level dialogue act categories and annotated 9k threads with 100k comments and built a CRF classifier for dialogue act labeling. Unlike these works which were not related to argumentation, Tan et al. (2016) examined persuasion strategies on Change My View using word overlap features. In contrast to our work, they focused solely on the successful strategies with delta-awarded posts. Using the same dataset, Musi (2017) recently studied concession in argumentation.

[2] An attempt to address the plea for thinking about problems, cognitive science, and the details of human language (Manning, 2015).

[3] https://github.com/UKPLab/naacl2018-before-name-calling-habernal-et-al

3 Data

Change My View (CMV) is an online 'place to post an opinion you accept [...] in an effort to understand other perspectives on the issue', in other words an online platform for 'good-faith' argumentation hosted on Reddit.[4] A user posts a **submission** (also called **original post(er); OP**) and other participants provide arguments to change the OP's view, forming a typical tree-form Web discussion. A special feature of CMV is that the OP acknowledges convincing arguments by giving a **delta** point (Δ). Unlike the vast majority of internet discussion forums, CMV enforces obeying strict rules (such as no 'low effort' posts, or accusing of being unwilling to change view) whose violation results into deleting the comment by moderators. These formal requirements of an ideal debate with the notion of violating rules correspond to incorrect moves in critical discussion in the normative pragma-dialectic theory (van Eemeren and Grootendorst, 1987). *Thus, violating the rule of 'not being rude or hostile' is equivalent to committing ad hominem fallacy.* For our experiments, we scraped, in cooperation with Reddit, the complete CMV including the content of the deleted comments so we could fully reconstruct the fallacious discussions, relying on the rule violation labels provided by the moderators. The dataset contains ≈ 2M posts in 32k submissions, forming 780k unique threads.

We will set up the stage for further experiments by providing several quantitative statistics we performed on the dataset. Only 0.2% posts in CMV are ad hominem arguments. This contrasts with a typical online discussion: Coe et al. (2014) found 19.5% of comments under online news articles to be incivil. Most threads contain only a single ad hominem argument (3,396 threads; there are 3,866 ad hominem arguments in total in CMV); only 35 threads contain more than three ad hominem arguments. In 48.6% of threads containing a single ad hominem, the ad hominem argument is the very last comment. This corresponds to the popular belief that if one is out of arguments, they start attacking and the discussion is over. This trend is also shown in Figure 1 which displays the relative position of the first ad hominem argument in a thread. Replying to ad hominem with another ad hominem happens only in 15% of the cases; this speaks for the attempts of CMV participants

[4] https://www.reddit.com/r/changemyview/

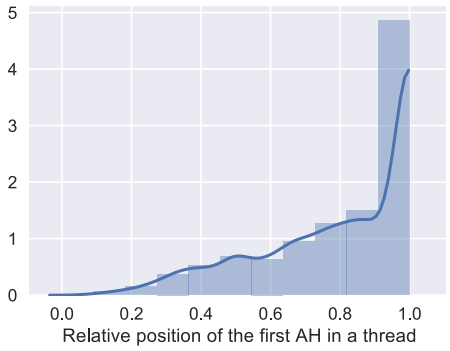

Figure 1: 'No discussion after ad hominem.' Distribution of the number of comments before the first ad hominem is committed proportional to the thread length.

to keep up with the standards of a rather rational discussion.

Regarding ad hominem authors, about 66% of them start attacking 'out of blue', without any previous interaction in the thread. On the other hand, 11% ad hominem authors write at least one 'normal' argument in the thread (we found one outlier who committed ad hominem after writing 57 normal arguments in the thread). Only in 20% cases, the ad hominem thread is an interplay between the original poster and another participant. It means that there are usually more people involved in an ad hominem thread. Unfortunately, sometimes the OP herself also commits ad hominem (12%).

We also investigated the relation between the presence of ad hominem arguments and the submission topic. While most submissions are accompanied by only one or two ad hominem arguments (75% of submissions), there are also extremes with over 50 ad hominem arguments. Manual analysis revealed that these extremes deal with religion, sexuality/gender, U.S. politics (mostly Trump), racism in the U.S., and veganism. We will elaborate on that later in Section 4.2.

4 Experiments

The experimental part is divided into three parts according to the increasing level of discourse complexity. We first experiment with ad hominem in isolation in section 4.1, then with direct ad hominem replies to original posts without dialogical exchange in section 4.2, and finally with ad hominem in a larger inter-personal discourse context in section 4.3.

4.1 Ad hominem without context in CMV

The first experimental set-up examines ad hominem arguments in *Change my view* regardless of its dialogical context.

4.1.1 Data verification

Ad hominem arguments labeled by the CMV moderators come with no warranty. To verify their reliability, we conducted the following annotation studies. First, we needed to estimate parameters of crowdsourcing and its reliability. We sampled 100 random arguments from CMV without context: positive candidates were the reported ad hominem arguments, whereas negative candidates were sampled from comments that either violate other argumentation rules or have a delta label. To ensure the maximal content similarity of these two groups, for each positive instance the semantically closest negative instance was selected.[5] We then experimented with different numbers of Amazon Mechanical Turk workers and various thresholds of the MACE gold label estimator (Hovy et al., 2013); comparing two groups of six workers each and 0.9 threshold yielded almost perfect inter-annotator agreement (0.79 Cohen's κ). We then used this setting (six workers, 0.9 MACE threshold) to annotate another 452 random arguments sampled in the same way as above.

Crowdsourced 'gold' labels were then compared to the original CMV labels (balanced binary task: positive instances (ad hominem) and negative instances) reaching accuracy of 0.878. This means that the ad hominem labels from CMV moderators are quite reliable. Manual error analysis of disagreements revealed 11 missing ad hominem labels. These were not spotted by the moderators but were annotated as such by crowd workers.

4.1.2 Recognizing ad hominem arguments

We sampled a larger balanced set of positive instances (ad hominem) and negative instances using the same methodology as in section 4.1.1, resulting in 7,242 instances, and casted the task of recognition of ad hominem arguments as a binary supervised task. We trained two neural classifiers, namely a 2-stacked bi-directional LSTM network (Graves and Schmidhuber, 2005), and a convolutional network (Kim, 2014), and evaluated them using 10-fold cross validation. Thorough the paper we use pre-trained `word2vec` word embeddings (Mikolov et al., 2013). Detailed hyper-parameters are described in the attached source codes. As results in Table 1 show, the task of recognizing ad hominem arguments is feasible and almost achieves the human upper bound performance.

Model	Accuracy
Human upper bound estimate	0.878
2 Stacked Bi-LSTM	0.782
CNN	**0.810**

Table 1: Prediction of ad hominem arguments

4.1.3 Typology of ad hominem

While binary classification of ad hominem as presented above might be sufficient for the purpose of red-flagging arguments, theories provide us with a much finer granularity (recall the typology in section 2). To validate whether this typology is empirically relevant, we executed an annotation experiment to classify ad hominem arguments into the provided five types (plus 'other' if none applies). We sampled 200 ad hominem arguments from threads in which interlocution happens only between two persons and which end up with ad hominem. The Mechanical Turk workers were shown this last ad hominem argument as well as the preceding one. Each instance was annotated by 16 workers to achieve a stable distribution of labels as suggested by Aroyo and Welty (2015). While 41% arguments were categorized as *abusive*, other categories (*tu quoque*, *circumstantial*, and *guilt by association*) were found to be rather ambiguous with very subtle differences. In particular, we observed a very low percentage agreement on these categories and a label distribution spiked around two or more categories. After a manual inspection we concluded that (1) the theoretical typology does not account for longer ad hominem arguments that mix up different attacks and that (2) there are actual phenomena in ad hominem arguments not covered by theoretical categories. These observations reflect those of Macagno (2013, p. 399) about ad hominem moves as multifaceted strategies.

We thus propose a list of phenomena typical to ad hominem arguments in CMV based on our empirical study. For this purpose, we follow up with

[5] Similarity was computed using a cosine similarity of average embedding vectors multiplied by the argument length difference to minimize length-related artifacts. The sample was balanced with roughly 50% positive and 50% negative instances.

another annotation experiment on 400 arguments, with seven workers per instance.[6] The goal was to annotate a text span which made the argument an ad hominem; a single argument could contain several spans. We estimated the gold spans using MACE and performed a manual post-analysis by designing a typology of causes of ad hominem together with their frequency of occurrence. The results and examples are summarized in Table 2.

4.1.4 Results and interpretation

The data verification annotation study (section 4.1.1) has two direct consequences. First, the high κ score (0.79) answers RQ2: for recognizing ad hominem argument, no previous context is necessary. Second, we still found 5% overlooked ad hominem arguments in CMV thus a moderation-facilitating tool might come handy; this can be served by the well-performing CNN model (0.810 accuracy; section 4.1.2).

The existing theoretical typology of ad hominem arguments, as presented for example in most textbooks, provides only a very simplified view. On the one hand, some of the categories which we found in the empirical labeling study (section 4.1.3) do map to their corresponding counterparts (such as the vulgar insults). On the other hand, some ad hominem insults typical to online argumentation (illiteracy insults, condescension) are not present in studies on ad hominem. Hence, we claim that any potential typology of ad hominem arguments should be multinomial rather than categorical, as we found multiple different spans in a single argument.

4.2 Triggers of first level ad hominem

In the following section, we increase the complexity of the studied discourse by taking the original post into account.

4.2.1 Annotation study

We already showed that ad hominem arguments are usually preceded by a discussion between the interlocutors. However, 897 submissions (original posts; OPs) have at least one intermediate ad hominem (in other words, the original post is directly attacked). We were thus interested in what triggers these first-level ad hominem arguments. We hypothesize two causes: (1) the *controversy* of the OP, similarly to some related works on news comments (Coe et al., 2014) and (2) the *reasonableness* of the OP (whether the topic is reasonable to argue about). We model both features on a three-point scale.[7]

We sampled two groups of OPs: those which had some ad hominem arguments in any of its threads but no delta (ad hominem group) and those without ad hominem but some deltas (Delta group). In total, 1,800 balanced instances were annotated by five workers and the resulting value was averaged for each item.[8]

Statistical analysis of the annotated 1,800 OPs revealed that ad hominem arguments are associated with more controversial OPs (mean controversy 1.23) while delta-awarded arguments with less controversial OPs (mean controversy 1.06; K-S test;[9] statistics 0.13, P-value: 7.97×10^{-7}). On the other hand, reasonableness does not seem to play such a role. The difference between ad hominem in reasonable OPs (mean 1.20) and delta in reasonable OPs (mean 1.11) is not that statistically strong; (K-S test statistics: 0.07, P-value: 0.02).

4.2.2 Regression model for predicting controversy and reasonableness

We further built a regression model for predicting controversy and reasonableness of the OPs. Along with Bi-LSTM and CNN networks (same models as in 4.1.2) we also developed a neural model that integrates CNN with topic distribution (CNN+LDA). The motivation for a topic-incorporating model was based on our earlier observations presented in section 3. In particular, we trained an LDA topic model (k = 50) (Blei et al., 2003) on the heldout OPs and during training/testing, we merged the estimated topic distribution vector with the output layer after convolu-

[6] Here we decided on seven workers per item by relying on other span annotation experiments done in a similar setup (Habernal et al., 2018b).

[7] Controversy: 1 = 'not really controversial', 2 = 'somehow controversial', 3 = 'very controversial' and reasonableness: 1 = 'quite stupid', 2 = 'neutral', 3 = 'quite reasonable'. Examples of (a) not really controversial: *"I Don't Think Monty Python is Funny"*, (b) very controversial: *"Blacks are generally intellectual inferior to the other major races"*, (c) quite stupid: *"Burritos are better than sandwiches"*, and (d) quite reasonable: *"Nations whose leadership is based upon religion are fundamentally backwards"*.

[8] A pilot crowd sourcing annotation with 5 + 5 workers showed a fair reliability for controversy (Spearman's ρ 0.804) and medium reliability for reasonableness (Spearman's ρ 0.646).

[9] Kolmogorov-Smirnov (K-S) test is a non-parametric test without any assumptions about the underlying probability distribution.

Type	(%)	Example spans
Vulgar insult	31.3	"Your just an asshole", "you dumb fuck", etc.
Illiteracy insult	13.0	"Reading comprehension is your friend", "If you can't grasp the concept, I can't help you"
Condescension	6.5	"little buddy", "sir", "boy", "Again, how old are you?"
Ridiculing and sarcasm	6.5	"Thank you so much for all your pretentious explanations", "Can you also use Google?"
'Idiot'-insults	6.5	"Ever have discussions with narcissistic idiots on the internet? They are so tiring"
Accusation of stupidity	4.3	"You have no capability to understand why", "You're obviously just Nobody with enough brains to operate a computer could possibly believe something this stupid"
Lack of argumentation skills	4.3	"You're making the claims, it's your job to prove it. Don't you know how debating works?", "You're trash at debating."
Accusation of trolling	3.9	"You're just a dishonest troll", "You're using troll tactics"
Accusation of ignorance	3.5	"Please dont waste peoples time pretending to know what you're talking about", "Do you even know what you're saying?"
"You didn't read what I wrote"	3.0	"Read what I posted before acting like a pompous ass", "Did you even read this?"
"What you say is idiotic"	2.6	"To say that people intrinsically understand portion size is idiotic.", "Your second paragraph is fairly idiotic"
Accusation of lying	2.6	"Possible lie any harder?", "You are just a liar."
"You don't face the facts and ignore the obvious"	1.7	"Willful ignorance is not something I can combat", "How can you explain that? You can't because it will hurt your feelings to face reality"
Accusation of ad hominem or other fallacies	1.7	"You started with a fallacy and then deflected.", "You still refuse to acknowledge that you used a strawman argument against me"
Other	8.3	"Wow. Someone sounds like a bit of an anti-semite", "You're too dishonest to actually quote the verse because you know it's bullshit"

Table 2: What makes an argument ad hominem: results of the empirical study of labeling spans in 400 ad hominem arguments.

Controversy (Spearman's ρ)	
Human upper bounds	0.804
Bi-LSTM	0.539
CNN	0.559
CNN-LDA	**0.569**
Reasonableness (Spearman's ρ)	
Human upper bounds	0.646
Bi-LSTM	0.332
CNN	0.320
CNN-LDA	**0.385**

Table 3: Results of predicting controversy and reasonableness of the original post.

tion and pooling. We performed 10-fold cross validation on the 1,800 annotated OPs and got reasonable performance for controversy prediction (ρ 0.569) and medium performance for reasonableness prediction (ρ 0.385), respectively; both using the CNN+LDA model (see Table 3).

We then used the trained model and extrapolated on all held-out OPs (1,267 ad hominem and 10,861 delta OPs, respectively). The analysis again showed that ad hominem arguments tend to be found under more controversial OPs whereas delta arguments in the less controversial ones (K-S test statistics: 0.14, P-value: 1×10^{-18}). For reasonableness, the rather low performance of the predictor does not allow us draw any conclusions on the extrapolated data.

4.2.3 Results and interpretation

Controversy of the original post is immediately heating up the debate participants and correlates with a higher number of direct ad hominem responses. This corresponds to observations made in comments in newswire where 'weightier' topics tended to stir incivility (Coe et al., 2014). On the other hand, 'stupidity' (or 'reasonableness') does not seem to play any significant role. The CNN+LDA model for predicting controversy (ρ 0.569) might come handy for signaling potentially 'heated' discussions.

4.3 Before calling names

In this section, we focus on the dialogical aspect of CMV debates and dynamics of ad hominem fallacies. Although ad hominem arguments appear in many forms (Section 4.1.3), we treat all ad hominem arguments equal in the following ex-

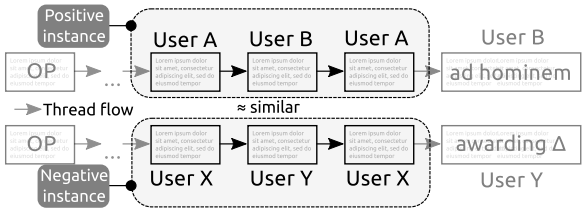

Figure 2: Sampling instances for learning triggers of ad hominem.

periments.

4.3.1 Data sampling

So far we explored what makes an ad hominem argument and whether debated topic influences the number of intermediate attacks. However, possible causes of the argumentative dynamics that ends up with ad hominem remain an open question, which has been addressed in neither argumentation theory nor in cognitive psychology, to the best of our knowledge. We thus cast an explanation of triggers and dynamics of ad hominem discussions as a supervised machine learning problem and draw theoretical insights by a retrospective interpretation of the learned models.

We sample positive instances by taking three contextual arguments preceding the ad hominem argument from threads which are an interplay between two persons. Negative samples are drawn similarly from threads in which the argument is awarded with Δ as shown in Figure 2.[10] Each instance consists of the three concatenated arguments delimited by a special OOV token. This resulted in 2,582 balanced training instances.

4.3.2 Neural models

The alleged lack of interpretability of neural networks has motivated several lines of approaches, such as layer-wise relevance propagation (Arras et al., 2017) or representation erasure (Li et al., 2016), both on sentiment analysis. As our task at hand deals with multi-party discourse that presumably involves temporal relations important for the learned representation, we opted for a state-of-the-art self-attentive LSTM model. In particular, we re-implemented the Structured Self-Attentive Embedding Neural Network (SSAE-NN) (Lin et al., 2017) which learns an embedding matrix representation of the input using attention weights. To make the attention even more interpretable, we replaced the final non-linear MLP layers with a single linear classifier (softmax). By summing over one dimension of the attention embedding matrix, each word from the input sequence gets associated with a single attention weight that gives us insights into the classifier's 'features' (still indirectly, as the true representation is a matrix; see the original paper).[11] The learning objective is to recognize whether the thread ends up in ad hominem or delta. We trained the model in 10-fold cross-validation and although our goal is not to achieve the best performance but rather to gain insight, we also tested a CNN model (accuracy 0.7095) which performed slightly worse than the SSAE-NN model (accuracy 0.7208).

4.3.3 Results and interpretation

During testing the model, we projected attention weights to the original texts as heat maps and manually analyzed 191 true positives (ad hominem threads recognized correctly), as well as 77 false positives (ad hominem threads misclassified as delta) and 84 false negatives (delta as ad hominem), in total about 120k tokens. The full output is available in the supplementary materials, we use IDs as a reference in the following text.

In the following analysis, we solely relied on the weights of words or phrases learned by the attention model, see an example in Figure 3. Based on our observations, we summarize several linguistic and argumentative phenomena with examples most likely responsible for ad hominem threads in Table 4.

The identified phenomena have few interesting properties in common. First, they all are topic-independent rhetorical devices (except for the loaded keywords at the bottom). Second, many of them deal with meta-level argumentation, i.e., arguing about argumentation (such as missing support or fallacy accusations). Third, most of them do not contain profanity (in contrast to the actual ad hominem arguments of which a third are vulgar insults; cf. Table 2). And finally, all of them should be easy to avoid.

Misleading 'features' False positives revealed properties that misled the network to classify delta threads as ad hominem threads.

[10] To ensure as much content similarity as possible, we used same similarity sampling as in section 4.1.1.

[11] We also experimented with regularizing the attention matrix as the authors proposed, but it resulted in worse performance.

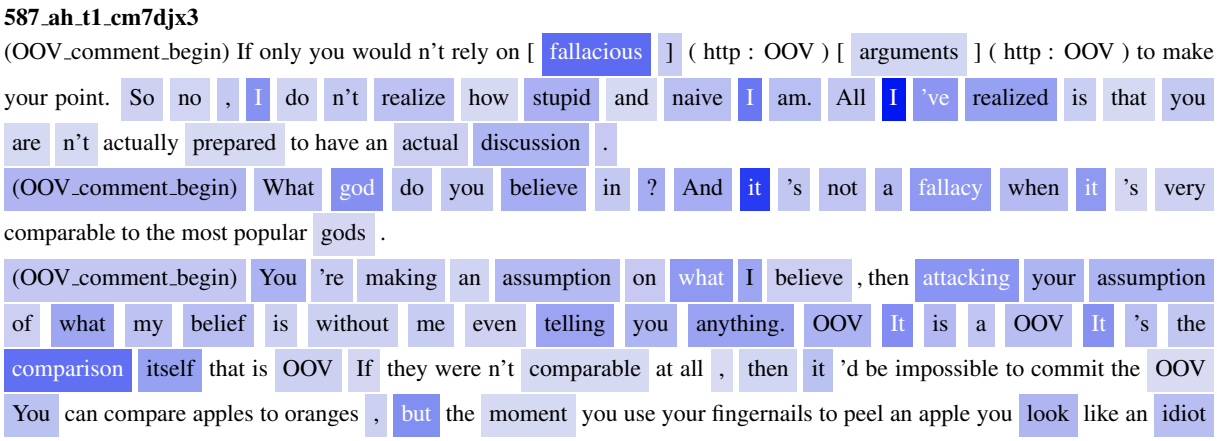

Figure 3: An example of reconstructed word weight heat map extracted from the attention matrix for a thread which ends up in ad hominem; three previous arguments are shown (see Figure 2 for sampling details).

Phenomena	**Examples**
Introducing vulgar intensifiers or interrogatives	**388(-1)** *"Where the fuck is your idea to ..."*, **712(-2)** *"no shortage of fucking gun"*, **1277(-1)** *"This is fucking CMV"*, **428(-2)** *"I'm fucking trans!"*, **2018(-2)** *"an arrogant fuck"*, **1277(-2)** *"What the fuck are you smoking?"*
Direct imperatives	**1003(-3)** *"You should get more mad about it"*, **857(-2)** *"You need to do a lot better than that."*, **233(-2)** *"So now delete your post"*, **749(-1)** *"google this fact as well"*, **1276(-1)** *"Just look back at the reasons why ..."*
Accusing of believing in or using propaganda	**522(-1)** *"It's right wing propaganda?"*, **1003(-1)** *"If you're not outraged, you're not paying attention to our propaganda that says the opposite of literally thousands of published research articles"*
Accusation of fallacies or bad argumentation practice	**238(-3)** *"your snide remarks and poor argumentation skills"*, **1117(-2)** *"you're circle jerking A vs. B"*, **263(-3)** *"You're grasping at straws"*, **78(-3)** *"You sure like changing words and statements to make your argument appear more cogent, don't you?"*, **210(-1)** *"Your arguments range from ... to ..."*, **1085(-3)** *"It's only a fallacy"*, **144(-1)** *"You haven't presented any evidence or argument that disagrees with anything I've said."*, **587(-3)** *"If only you wouldn't rely on fallacious arguments"*
Reinterpreting opponent's positions	**982(-1)** *"The fact that you obviously think ... reveals ..."*, **982(-2)** *"What makes you think I see myself ... ?"*, **1060(-3)** *"That kind of thinking is ..."*, **760(-1)** *"If I'm understanding you correctly"*, **405(-1)** *"... deluded yourself into believing factually incorrect things"* **587(-1)** *"You're making an assumption on what I believe, then attacking your assumption of what my belief is without me even telling you anything."*
Accusation of not reading the other party's arguments	**586(-1)** *"... me without even reading my ..."*, **240(-1)** *"You are just reading it wrong."*, **310(-1)** *"Oh, you're not actually reading my ..."*
Pointing at missing or unsupported evidence and facts	**1238(-2)** *"unsupported bullshit as before"*, **1121(-3)** *"you can't chose your facts"*, **931(-1)** *"If that's your only argument ..."*, **486(-2)** *"unsubstantiated statement"*, **486(-1)** *"unsupported claims"*, **71(-2)** *"factually correct"*, **915(-1)** *"But for the sake of argument, your points are pitifully .."*, **388(-3)** *"Please provide statistics ... It's silly to debate statistics without actual numbers."*
UPPERCASE	**1238(-3)** *"NO ONE CLAIMED THAT ... ARE NOT ... AGAINST ..."*
Sarcasm	**78(-2)** *"But I'm sure you know best"*, **310(-1)** *"Have a nice day."*, **1276(-1)** *"Good luck with that"*
Mentions of trolling	**701(-2)** *"Then you are giving trolls the victory then?"*
Loaded keywords	*Nazi, rape, racist*

Table 4: Phenomena resulting into ad hominem learned by the SSAE-NN model. The first number is the instance ID (available in the supplementary material), the minus number in parentheses is the position of the argument before the ad hominem.

- These include **topic words** (such as *racism, blacks, slave, abortion*) which reflects the implicit bias in the data.
- Actual interest mixed with indifference in **sarcasm** is also problematic (**185(-2)** *"That's a very interesting ..."*).
- Another problematic phenomena is also **expressed disagreement** (**678(-2)** *"overheated rhetoric"*, **203(-2)** *"But I suppose this argument is ..."*, **230(-2)** *"But I don't think it's quite ..."*, **938(-1)** *"I disagree too, however ..."*).

False negatives were caused basically by presence of many 'informative' **content words** (**980** *unemployment, quarterly publication, inflation data*, **474** *actual publications, this experiment, biological ailments, medical doctorate*, **1214** *graduate degree, education, health insurance*) and **misinterpreted sarcasm** (**285(-1)** *"Also this is a cute analogy"*).

5 Conclusion

In this article, we investigated ad hominem argumentation on three levels of discourse complexity. We looked into qualitative and quantative properties of ad hominem arguments, crowdsourced labeled data, experimented with models for prediction (0.810 accuracy; 4.1.2), and proposed an updated typology of ad hominem properties (4.1.3). We then looked into the dynamics of argumentation to examine the relation between the quality of the original post and immediate ad hominem arguments (4.2). Finally, we exploited the learned representation of Self-Attentive Embedding Neural Network to search for features triggering ad hominem in one-to-one discussions. We found several categories of rhetorical devices as well as misleading features (4.3.3).

There are several points that deserve further investigation. First, we have ignored meta-information of the debate participants, such as their overall activity (i.e., whether they are spammers or trolls). Second, the proposed typology of ad hominem causes has not yet been post-verified empirically. Third, we expect that personality traits of the participants (BIG5) may also play a significant role in the argumentative exchange. We leave these points for future work.

We believe that our findings will help gain better understanding of, and hopefully keep restraining from, ad hominem fallacies in good-faith discussions.

Acknowledgments

This work has been supported by the ArguAna Project GU 798/20-1 (DFG), and by the DFG-funded research training group "Adaptive Preparation of Information form Heterogeneous Sources" (AIPHES, GRK 1994/1).

References

Aristotle and George Kennedy (translator). 1991. *On Rhetoric: A Theory of Civil Discourse*. Oxford University Press, USA.

Lora Aroyo and Chris Welty. 2015. Truth is a lie: Crowd truth and the seven myths of human annotation. *AI Magazine* 36(1):15–24. https://doi.org/10.1609/aimag.v36i1.2564.

Leila Arras, Grégoire Montavon, Klaus-Robert Müller, and Wojciech Samek. 2017. Explaining Recurrent Neural Network Predictions in Sentiment Analysis. In *Proceedings of the 8th Workshop on Computational Approaches to Subjectivity, Sentiment and Social Media Analysis*. Association for Computational Linguistics, Copenhagen, Denmark, pages 159–168. http://www.aclweb.org/anthology/W17-5221.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *Journal of Machine Learning Research* 3:993–1022. http://dl.acm.org/citation.cfm?id=944919.944937.

Maarten Boudry, Fabio Paglieri, and Massimo Pigliucci. 2015. The Fake, the Flimsy, and the Fallacious: Demarcating Arguments in Real Life. *Argumentation* 29(4):431–456. https://doi.org/10.1007/s10503-015-9359-1.

Kevin Coe, Kate Kenski, and Stephen A. Rains. 2014. Online and Uncivil? Patterns and Determinants of Incivility in Newspaper Website Comments. *Journal of Communication* 64(4):658–679. https://doi.org/10.1111/jcom.12104.

Stephen de Wijze. 2002. Complexity, Relevance and Character: Problems with Teaching the "Ad Hominem" Fallacy. *Educational Philosophy and Theory* 35(1):31–56. https://doi.org/10.1111/1469-5812.00004.

Alex Graves and Jürgen Schmidhuber. 2005. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. *Neural Networks* 18(5):602–610. https://doi.org/10.1016/j.neunet.2005.06.042.

Ivan Habernal and Iryna Gurevych. 2016a. What makes a convincing argument? Empirical analysis and detecting attributes of convincingness

in Web argumentation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Austin, Texas, pages 1214–1223. http://aclweb.org/anthology/D16-1129.

Ivan Habernal and Iryna Gurevych. 2016b. Which argument is more convincing? Analyzing and predicting convincingness of Web arguments using bidirectional LSTM. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Association for Computational Linguistics, Berlin, Germany, pages 1589–1599. http://www.aclweb.org/anthology/P16-1150.

Ivan Habernal and Iryna Gurevych. 2017. Argumentation Mining in User-Generated Web Discourse. *Computational Linguistics* 43(1):125–179. https://doi.org/10.1162/COLI_a_00276.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational Argumentation Meets Serious Games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Copenhagen, Denmark, pages 7–12. http://www.aclweb.org/anthology/D17-2002.

Ivan Habernal, Patrick Pauli, and Iryna Gurevych. 2018a. Adapting Serious Game for Fallacious Argumentation to German: Pitfalls, Insights, and Best Practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. European Language Resources Association (ELRA), Miyazaki, Japan, page (to appear).

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 2018b. The Argument Reasoning Comprehension Task: Identification and Reconstruction of Implicit Warrants. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, New Orleans, LA, page (to appear). https://arxiv.org/abs/1708.01425.

Charles L. Hamblin. 1970. *Fallacies*. Methuen, London, UK.

Hans Hansen. 2017. Fallacies. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*, Metaphysics Research Lab, Stanford University. http://plato.stanford.edu/entries/fallacies/.

Dirk Hovy, Taylor Berg-Kirkpatrick, Ashish Vaswani, and Eduard Hovy. 2013. Learning whom to trust with MACE. In *Proceedings of NAACL-HLT 2013*. Association for Computational Linguistics, Atlanta, Georgia, pages 1120–1130. http://www.aclweb.org/anthology/N13-1132.

Siddharth Jain, Archna Bhatia, Angelique Rein, and Eduard Hovy. 2014. A Corpus of Participant Roles in Contentious Discussions. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*. European Language Resources Association (ELRA), Reykjavik, Iceland, pages 1751–1756.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Doha, Qatar, pages 1746–1751. http://www.aclweb.org/anthology/D14-1181.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016. Understanding Neural Networks through Representation Erasure. In *arXiv preprint*. http://arxiv.org/abs/1612.08220.

Zhouhan Lin, Minwei Feng, Cicero Nogueira dos Santos, Mo Yu, Bing Xiang, Bowen Zhou, and Yoshua Bengio. 2017. A Structured Self-attentive Sentence Embedding. In *Proceedings of the 5th International Conference on Learning Representations (ICLR)*. Toulon, France, pages 1–15. http://arxiv.org/abs/1703.03130.

Fabrizio Macagno. 2013. Strategies of character attack. *Argumentation* 27(4):369–401. https://doi.org/10.1007/s10503-013-9291-1.

Christopher D. Manning. 2015. Computational Linguistics and Deep Learning. *Computational Linguistics* 41(4):701–707. https://doi.org/10.1162/COLI_a_00239.

Hugo Mercier and Dan Sperber. 2011. Why do humans reason? Arguments for an argumentative theory. *The Behavioral and Brain Sciences* 34(2):57–74; discussion 74–111. https://doi.org/10.1017/S0140525X10000968.

Hugo Mercier and Dan Sperber. 2017. *The Enigma of Reason*. Harvard University Press, Cambridge, MA, USA.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, Curran Associates, Inc., pages 3111–3119.

Elena Musi. 2017. How did you change my view? A corpus-based study of concessions' argumentative role. *Discourse Studies* page (in press). https://doi.org/10.1177/1461445617734955.

Ahmed Sahlane. 2012. Argumentation and fallacy in the justification of the 2003 War on Iraq. *Argumentation* 26(4):459–488. https://doi.org/10.1007/s10503-012-9265-8.

Haji Mohammad Saleem, Kelly P Dillon, Susan Benesch, and Derek Ruths. 2016. A Web of Hate: Tackling Hateful Speech in Online Social Spaces. In Guy De Pauw, Ben Verhoeven, Bart Desmet, and Els Lefever, editors, *Proceedings of the First Workshop on Text Analytics for Cybersecurity and Online Safety*. European Language Resources Association (ELRA), Portorož, Slovenia, pages 1–9.

Edward Schiappa and John P. Nordin. 2013. *Argumentation: Keeping Faith with Reason*. Pearson UK, 1st edition.

Christian Stab and Iryna Gurevych. 2017. Recognizing Insufficiently Supported Arguments in Argumentative Essays. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017)*. Association for Computational Linguistics, pages 980–990. http://www.aclweb.org/anthology/E17-1092.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of the 25th International Conference on World Wide Web*. ACM, Montreal, CA, page (to appear). http://arxiv.org/abs/1602.01103.

Christopher W. Tindale. 2007. *Fallacies and Argument Appraisal*. Cambridge University Press, New York, NY, USA, critical reasoning and argumentation edition.

Stephen E. Toulmin. 1958. *The Uses of Argument*. Cambridge University Press.

Frans H. van Eemeren, Bart Garssen, Erik C. W. Krabbe, A. Francisca Snoeck Henkemans, Bart Verheij, and Jean H. M. Wagemans. 2014. *Handbook of Argumentation Theory*. Springer, Berlin/Heidelberg.

Frans H. van Eemeren and Rob Grootendorst. 1987. Fallacies in pragma-dialectical perspective. *Argumentation* 1(3):283–301. https://doi.org/10.1007/BF00136779.

Frans H. van Eemeren and Rob Grootendorst. 1992. *Argumentation, communication, and fallacies: a pragma-dialectical perspective*. Lawrence Erlbaum Associates, Inc.

Henning Wachsmuth, Nona Naderi, Ivan Habernal, Yufang Hou, Graeme Hirst, Iryna Gurevych, and Benno Stein. 2017. Argumentation Quality Assessment: Theory vs. Practice. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*. Association for Computational Linguistics, Vancouver, Canada, pages 250–255. http://aclweb.org/anthology/P17-2039.

Douglas Walton. 2007. *Media Argumentation: Dialect, Persuasion and Rhetoric*. Cambridge University Press.

Alex Wang, William L. Hamilton, and Jure Leskovec. 2016. Learning Linguistic Descriptors of User Roles in Online Communities. In *Proceedings of 2016 EMNLP Workshop on Natural Language Processing and Computational Social Science*. Association for Computational Linguistics, Austin, Texas, pages 76–85. http://aclweb.org/anthology/W16-5610.

John Woods. 2008. Lightening up on the Ad Hominem. *Informal Logic* 27(1):109. https://doi.org/10.22329/il.v27i1.467.

Ellery Wulczyn, Nithum Thain, and Lucas Dixon. 2017. Ex Machina: Personal Attacks Seen at Scale. In *Proceedings of the 26th International Conference on World Wide Web*. International World Wide Web Conferences Steering Committee, Perth, Australia, pages 1391–1399. https://doi.org/10.1145/3038912.3052591.

Amy Zhang, Bryan Culbertson, and Praveen Paritosh. 2017. Characterizing Online Discussion Using Coarse Discourse Sequences. In *Proceedings of the Eleventh International AAAI Conference on Web and Social Media (ICWSM 2017)*. AAAI Press, Montreal, Canada, pages 357–366.