Identifying and Characterizing Ad hominem Fallacy Usage in The Wild

BTP-II report submitted to

Indian Institute of Technology Kharagpur

in partial fulfilment for the award of the degree of

Bachelor of Technology

in

Electronics & Electrical Communication Engineering

by

Utkarsh Patel

(18EC35034)

Under the supervision of

Prof. Mainack Mondal and Prof. Animesh Mukherjee

Department of Computer Science & Engineering

Indian Institute of Technology Kharagpur

Spring Semester, Academic Session 2021-22

April 25, 2022

DECLARATION

I certify that

(a) The work contained in this report has been done by me under the guidance of my supervisor.

(b) The work has not been submitted to any other Institute for any degree or diploma.

(c) I have conformed to the norms and guidelines given in the Ethical Code of Conduct of the Institute.

(d) Whenever I have used materials (data, theoretical analysis, figures, and text) from other sources, I have given due credit to them by citing them in the text of the thesis and giving their details in the references. Further, I have taken permission from the copyright owners of the sources, whenever necessary.

Date: April 25, 2022 (Utkarsh Patel)
Place: Kharagpur (18EC35034)

DEPARTMENT OF COMPUTER SCIENCE & ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "**Identifying and Characterizing Ad hominem Fallacy Usage in The Wild**" submitted by **Utkarsh Patel** (Roll No. 18EC35034) to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Bachelor of Technology in Electronics & Electrical Communication Engineering is a record of bona fide work carried out by him under my supervision and guidance during Spring Semester, Academic Session 2021-22.

Prof. Mainack Mondal and Prof. Animesh Mukherjee
Department of Computer Science & Engineering
Indian Institute of Technology Kharagpur
Kharagpur - 721302, India

April 25, 2022

Abstract

Today, participating in discussions on online forums is extremely commonplace and these discussions have started rendering a strong influence on the overall opinion of online users. Naturally, twisting the flow of the argument can have a strong impact on the minds of naïve users which in the long run might have socio-political ramifications, for example, winning an election or spreading targeted misinformation. Thus, these platforms are potentially highly vulnerable to malicious players who might act individually or as a cohort to breed fallacious arguments with a motive to sway public opinion. *Ad hominem* arguments are one of the most effective forms of such fallacies. Although a simple fallacy, it is effective enough to sway public debates in the offline world and can be used as a precursor to shutting down the voice of opposition by slander.

In this work, we take a first step in shedding light on the usage of ad hominem fallacies in the wild. First, we build a powerful ad hominem detector based on transformer architecture with high accuracy (F1 more than 83%, showing a significant improvement over prior work), even for datasets for which annotated instances constitute a very small fraction. We then used our detector on 265k arguments collected from the online debate forum – *CreateDebate*. Our crowdsourced surveys validate our in-the-wild predictions on CreateDebate data (94% match with manual annotation). Our analysis revealed that a surprising 30% of CreateDebate content contains ad hominem fallacy, and a cohort of highly active users post significantly more ad hominem to suppress opposing views. Then, our temporal analysis revealed that ad hominem argument usage increased significantly since the 2016 US Presidential election, not only for topics like Politics, but also for Science and Law. We conclude by discussing important implications of our work to detect and defend against ad hominem fallacies.

Acknowledgements

It has been a long journey through unknown lands, over sky-high peaks, and through deep and dark troughs. I am thankful for the many people I got to interact and who accompanied me on the way. First, I'd like to thank my supervisors Prof. Mainack Mondal and Prof. Animesh Mukherjee. The freedom you gave me to pursue my interests and follow my curiosity has made all the difference. Thank you for your sensible advice and patience, your support and for fostering a productive research environment.

I am also grateful for the support of my family and friends. Special thanks go to my mom and dad. I'm extremely lucky to have met individuals en route who took a risk on me and helped me grow. I would like to thank Mr. Soham Poddar and Mr. Punyajoy Saha for helping me setup and review the crowd-sourced surveys. To Mr. Sasi Bhushan for helping me with extracting the attention scores and creating heat maps with it. I am grateful for all the amazing people I had the chance to get to know at IIT Kharagpur.

Contents

List of Figures

List of Tables

Abbreviations

General notations

e.g.	**e**xemplum **g**ratia (*en*: for example)
et al.	**e**t **a**lia (*en*: and others)
i.e.	**i**d **e**st (*en*: that is)

Neural networks

CNN	**C**onvolutional **N**eural **N**etwork
GAN	**G**enerative **A**dversarial **N**etwork
BiLSTM	**Bi**directional **L**ong **S**hort-**T**erm **M**emory

Natural language processing

BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers

Graph theory

SCC	**S**trongly **C**onnected **C**omponents

Symbols

λ	Count of level-1 comments
ρ	Count of direct replies
$X(\lambda)$	Set of authors who posted at least λ level-1 comments
$Y(\rho)$	Set of authors who received at least ρ direct replies
$S(\lambda, \rho)$	Set of authors given as $X(\lambda) \cap Y(\rho)$
`[CLS]`	Token padded before a sentence in BERT
`[SEP]`	Token padded at the end of a sentence or in the middle of two sentences in BERT

Chapter 1

Introduction

Today online forums and social media sites facilitate easy collaborative opinion formation for billions of users surpassing geographical boundaries. However, perhaps quite naturally, this process of opinion formation also involves the process of imitating and participating in online arguments where multiple parties often present their conflicting views. The caveat here is that all the arguments presented in online debates are not always sound. They often contain *deceptive arguments in disguise* (Kennedy, 1993). Intuitively in online forums, the users present informal fallacies, necessitating the analysis of contents to identify them (as opposed to the formal ones, which can be examined using logical representations) (Sahai et al., 2021).

Amongst different fallacies, *ad hominem* is perhaps the most famous one in the offline world (Macagno, 2013; Schiappa and Nordin, 2013; Zalta, 2004; Woods, 2007). Ad hominem or *against the person* is a fallacious argument, based on feelings of bias (mostly irrelevant to the argumentation), rather than reality, reason, and rationale. However, despite a long history of dissecting and condemning ad hominem fallacies in the offline world, even online users are no stranger to the usage of ad hominem fallacies (Goodman, 2020; Redinger, 2020). Ad hominem arguments are often personal attacks on someone's character or motive rather than an attempt to address the reasoning that they presented. People tend to use ad hominem arguments because they want to appeal to others' emotions rather than reasoning.

Recently, there has been substantial research concerning investigating and countering hate speech, misinformation as well as cyberbullying within the user-generated content posted on social media (Mondal et al., 2017, 2018; Das et al., 2021; Mathew et al., 2020a,b). In the same vein, although relatively rare, some very recent works are exploring the detection of ad hominem fallacies in the wild using computational methods (Habernal et al., 2018; Sahai et al., 2021). However, these works focused more on the detection of ad hominem (and other) fallacies using automated methods in online forums. There is not much work shedding light on the lay of the land for ad hominem usage over time. We aim to bridge this gap.

In this work, we present a data-driven exploration of ad hominem arguments in the wild using CreateDebate[1], an online discussion forum as an experimental testbed. We used an in-house high-accuracy and high external validity ad hominem detector on a dataset containing more than 18k posts with 265k comments generated by 15k users of CreateDebate. Next, we analyzed the detected large-scale ad hominem arguments to shed light on in-the-wild ad hominem usage. Specifically, we have made three key contributions to this work.

First, we developed ad hominem detectors considering two scenarios—when the annotated data is abundant and when it is not, using an annotated dataset from previous work on Reddit Habernal et al. (2018). Our models significantly improved over the ad hominem detectors reported in prior work and achieved a macro-F1 score of 0.84. Furthermore, we evaluated the predictions of our detector on the CreateDebate dataset using a user study. Our user study demonstrated that results from our detector (trained on Reddit data) are also externally valid—achieving a 94% accuracy on the CreateDebate data.

Second, we leveraged this detector on our large-scale CreateDebate data and found that a significant 29.97% (i.e., almost one-third) of all the comments on CreateDebate are ad hominem. On further investigation, we uncovered that a community of highly active users posts a disproportionately high volume of ad hominem fallacies (more than 50% of their total comments).

[1] https://www.createdebate.com/

Third, we checked whether ad hominem arguments were always used in such a high volume, or was it just a recent trend? It appeared that the fraction of ad hominem arguments showed a sharp rise after 2016. This trend was prominent not only in the Politics subforum but transcended in subforums like Science and Law. We found a striking correlation—the users posting in the Politics subforum of CreateDebate have very high overlap in subforums like Science and Law. The rise of ad hominem arguments in Politics subforums seems to be triggered by the 2016 US Presidential election, which resulted in the users active in Political subforum posting insulting comments in other forums as well, hence, significantly increasing the ad hominem usage.

Ethical considerations: In this work, we collected and analyzed data from CreateDebate and also conducted an annotation survey for validating our classifier. However, since we were analyzing user-generated data in this work, we tried our best to conduct our study ethically and protect the privacy of the users in our dataset. Specifically, we leveraged the previous work by Eysenbach and Till (2001) to check the ethics of our work. We noted that CreateDebate is a moderate-sized forum with around 15k members, and no registration was necessary to view and collect the CreateDebate data, signifying it was an 'open' forum. Finally, the debate topics often revolved around general phenomena (e.g., election), signifying the potentially public nature of our collected dataset. Nonetheless, following the footsteps of previous work by Cook et al. (2018), we hashed usernames after data collection to protect the privacy of the users during our analysis. Along the same lines of ethical consideration, for our annotation study, we did not collect any personally identifiable data from our participants to protect their privacy. In the next section, we shall start with related research to put our work in context.

Chapter 2

Prior Work

We divide the related prior works into two important sub-parts—exploration on ad hominem argumentation, especially in the online forums like ChangeMyView[1] and CreateDebate, and usage of Generative Adversarial Networks (GANs) in NLP.

2.1 Investigation on ad hominem argumentation

Aristotle first identified that some arguments are indeed *deceptions in disguise* (Kennedy, 1993). He called them fallacies. The ad hominem arguments are addressed in most of the follow-up treatises of fallacies (Hamblin, 1970; Eemeren and Grootendorst, 1987; Boudry et al., 2015). Ad hominem arguments are used in a debate for simply attacking the opponents' traits instead of countering their arguments (Tindale, 2007). Naturally, ad hominem arguments are based on feelings of bias rather than reason, often involving personal attacks on someone's character or motive. Though arguing *against the person* is considered faulty, these arguments are used in online debate forums and social media sites (Habernal et al., 2018; Sahai et al., 2021). Ad hominem arguments are multifaceted and use complex strategies, involving not a simple argument, but a cumulation of several combined

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

tactics (Macagno, 2013). However, the majority of this research was often aimed at dissecting what constitutes ad hominem in philosophy, rather than evaluating its usage in the real world (Schiappa and Nordin, 2013; Macagno, 2013; Zalta, 2004; Woods, 2007).

Recently, the scenario started to change when researchers working on NLP aimed to identify different fallacies in online forums. To that end, Wulczyn et al. (2017) annotated 38k instances of Wikipedia talk page comments for detecting personal attacks on the forum and Jain et al. (2014) studied principal roles in discussions from the Wikipedia Article for Deletion pages, and extracted several typical roles like 'idiots', 'voices', 'rebels', etc. which might be considered signals for ad hominem fallacies.

Studies on ChangeMyView: More recent work exploited online discussion forums like Reddit, primarily the ChangeMyView subreddit, to detect naturally occurring hate-speech and ad hominem fallacies. Wei et al. (2016a) studied the impact of different sets of features on the identification of persuasive comments. Tan et al. (2016) developed a framework for analyzing persuasive arguments and malleable opinions. Habernal et al. (2018) investigated ad hominem argumentation at three levels of discourse complexity (arguments in isolation, in direct replies to original post without dialogical context and in a larger inter-personal discourse context). Sahai et al. (2021) extended this work and found the types of fallacies. Our work builds on this type of detection methods, yet extends them considerably.

Studies on CreateDebate: Abbott et al. (2016) developed Internet Argument Corpus (v2.0), a collection of corpora for research in political debate on Internet forums, which contains a sample from CreateDebate (3K posts) and includes topic annotations. Wei et al. (2016b) analyzed the disputation action in the online debate by labeling a set of disputing argument pairs extracted from CreateDebate and performing annotation studies. Trabelsi and Zaïane (2014) suggested a fine-grained probabilistic framework for improving the quality of opinion mining from online contention texts. Hasan and Ng (2014) exploited stance information for reason classification, proposing systems of varying complexity for modeling stances and reasons. Qiu (2015) modeled user posting behaviors and user opinions for viewpoint discovery and proposed an integrated model that jointly considers arguments, stances, and attributes. Qiu et al. (2015) predicted user stances on a variety of topics

and assembled user arguments, interactions, and attributes into a collaborative filtering framework that exploits recently introduced fast inference methods.

2.2 Text classification using deep learning for our detector

We explored a number of state-of-the-art transformer architectures like Google's BERT (Devlin et al., 2019), OpenAI's GPT-1 (Radford and Narasimhan, 2018) for our experiments. These architectures are trained over large-scale annotated datasets and only require fine-tuning for a targeted task to achieve high accuracy for various NLP applications. However, one major disadvantage of these architectures is that even fine-tuning them often requires at least thousands of annotated examples for the targeted tasks. However, in our use case, obtaining thousands of annotated ad hominem arguments (and an equal number of non-ad hominem arguments) is costly and might be difficult to obtain. To that end, we leveraged a recent technique—integrate these humongous pre-trained models with Generative Adversarial Networks (GANs) (Goodfellow et al., 2014). In GANs, a 'generator' is trained to produce samples resembling some data distribution. This training process 'adversarially' depends on a 'discriminator', which instead is trained to distinguish examples of the generator from the real instances. SS-GANs are an extension to GANs where the discriminator also assigns class labels to each example while discriminating whether or not it was naturally produced (Salimans et al., 2016). Croce et al. (2020) proposed 'GAN-BERT' that extends the fine-tuning of BERT-like architectures with unlabeled data in a generative adversarial setting using SS-GAN schema. Building on this related work helped us create an accurate yet explainable detector with limited data. Next, we will start with describing our approach to develop the classifier.

Present work: Our work builds a highly accurate detector, beating the models used by Habernal et al. (2018) in macro averaged F1-score for classification and establishes the external validity of the detector on CreateDebate data. Then using this detector, we measured the prevalence of ad hominem in the wild—we found that the amount of ad hominem in recent times increased manyfold, more than the figures hinted in any of

the previous works. Our highly accurate detector relied on the recent advances in text classification using transformer models as discussed above.

Chapter 3

Ad hominem detection

For the purposes of our experiment, we use the ad hominem argumentation in the ChangeMyView (CMV) dataset (Habernal et al., 2018) as the benchmark. ChangeMyView is a popular subreddit in which a user (called OP, original poster) posts an opinion and other users write comments to change the perspective of OP about the posted opinion. OP can acknowledge convincing arguments by giving *delta* points.

Unlike regular debate forums, strict rules are enforced on this subreddit content, violating which results in the deletion of the content by the moderators. The CMV dataset contains 7242 comments from this subreddit (3622 instances with the label 'ad hominem' and 3620 instances with the label 'none'). The dataset was created maintaining a balance of syntactic and semantic similarity between the instances of the two-class labels. We will use this dataset to build powerful classifiers for ad hominem detection.

Habernal et al. (2018) used CNN and 2 Stacked Bi-LSTM models for detecting ad hominem comments. They reported 10-fold cross validation results. We used the BERT model (case-insensitive, base) and carried out the same 10-fold cross validation experiments. The results are presented in Table 3.1.

Model	**Accuracy**	**Macro-F1**
CNN	0.808	0.807
2 Stacked Bi-LSTM	0.781	0.781
BERT	**0.834**	**0.834**

TABLE 3.1: 10-fold cross validation results of the models on CMV dataset.

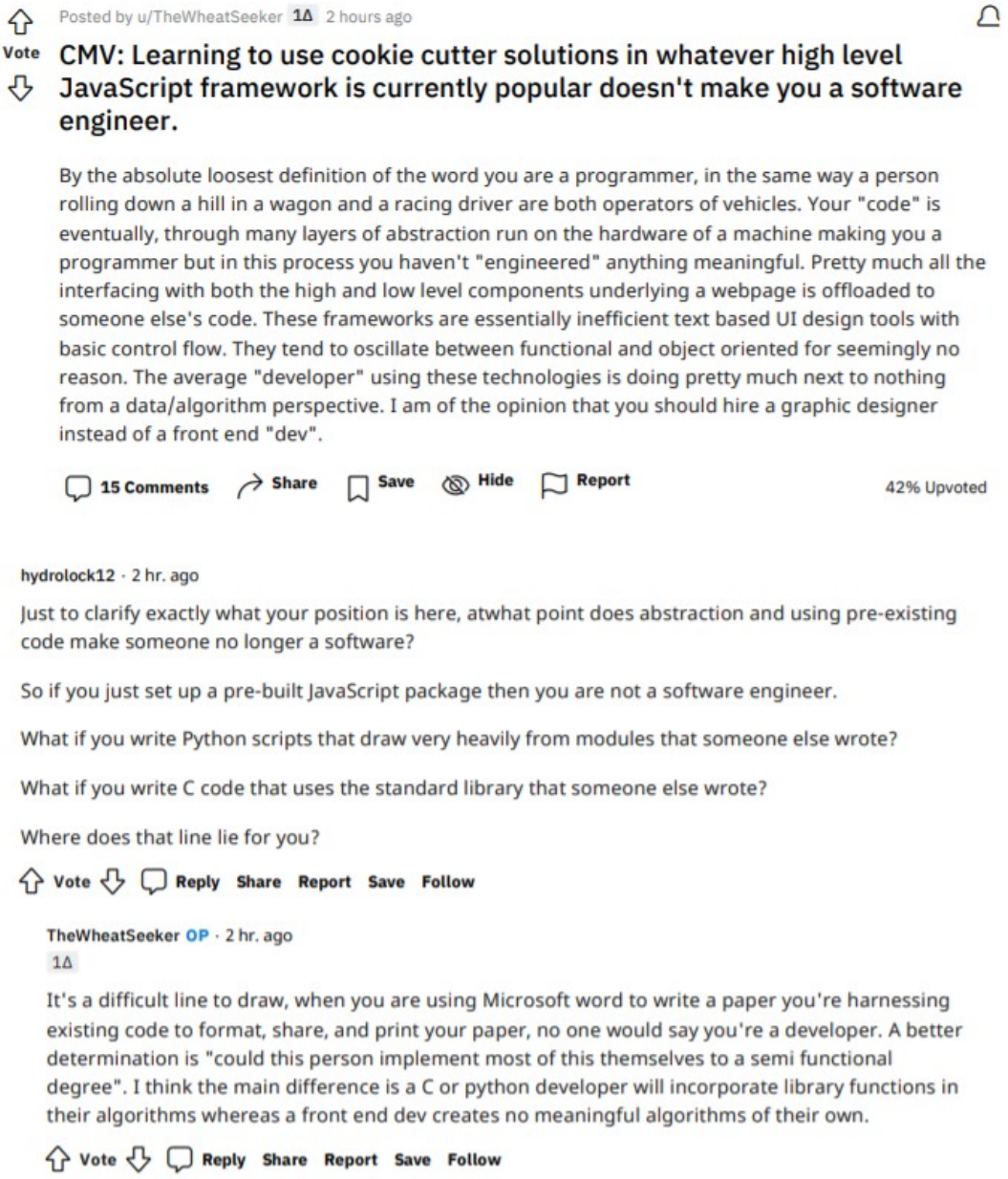


FIGURE 3.1: Example snapshot showing a post and a few comments for the Change-MyView forum.

As noted in Table 3.1, we achieved a 2.6% improvement in accuracy over the baselines. We further investigate the segments that could be potentially responsible for flagging an argument as ad hominem. We cast an explanation of triggers and dynamics of ad hominem argumentation as a supervised learning problem and draw theoretical insights by a retrospective interpretation of the learned model. As `[CLS]` token is the aggregate

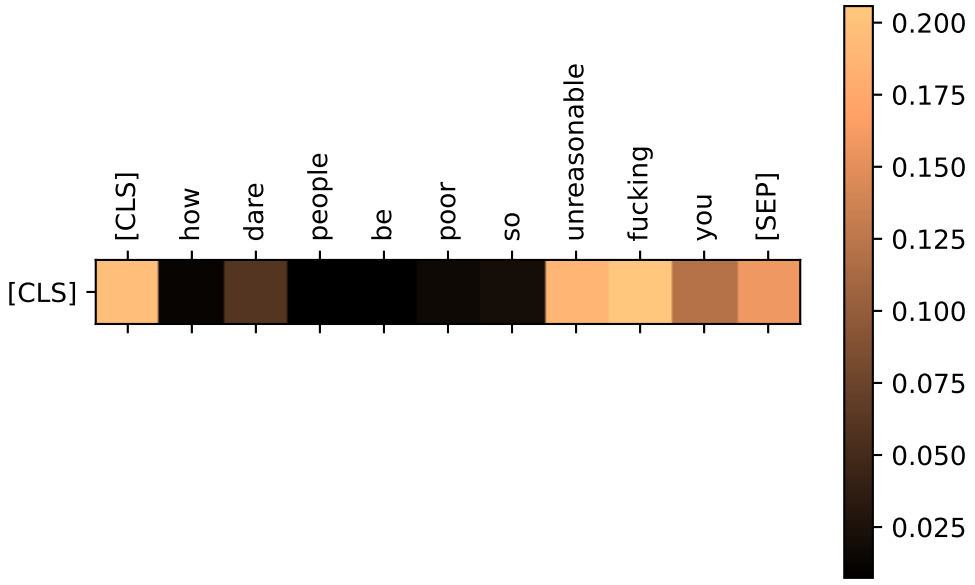


FIGURE 3.2: Visualization of attention scores for the `[CLS]` token for an example comment.

representation of the input sequence for classification tasks, we use attention scores for `[CLS]` token to detect key tokens influencing the classification. A sample visualization is shown in Figure 3.2. We greedily select top three tokens (excluding `[CLS]` and `[SEP]` tokens) on the basis of attention scores so that the trigrams centered at those tokens do not overlap and highlight these trigrams in the visualization. A sample visualization is shown in Figure 3.4. We analyzed these highlighted trigrams for the comments which were flagged as ad hominem and observed that the BERT model is not only beating the baselines in terms of accuracy, it can also be easily interpreted to extract linguistic insights into potential triggers for ad hominem argumentation.

One bottleneck for such studies is the cost of generating an annotated dataset. Annotating ad hominem arguments is a very costly task as they are difficult to comprehensively and objectively define (Sheng et al., 2020). Hence, we simulate a situation where we assume that labeled instances constitute a very small fraction of the dataset. While doing training on different folds, we intentionally drop the class labels of a given fraction of instances and

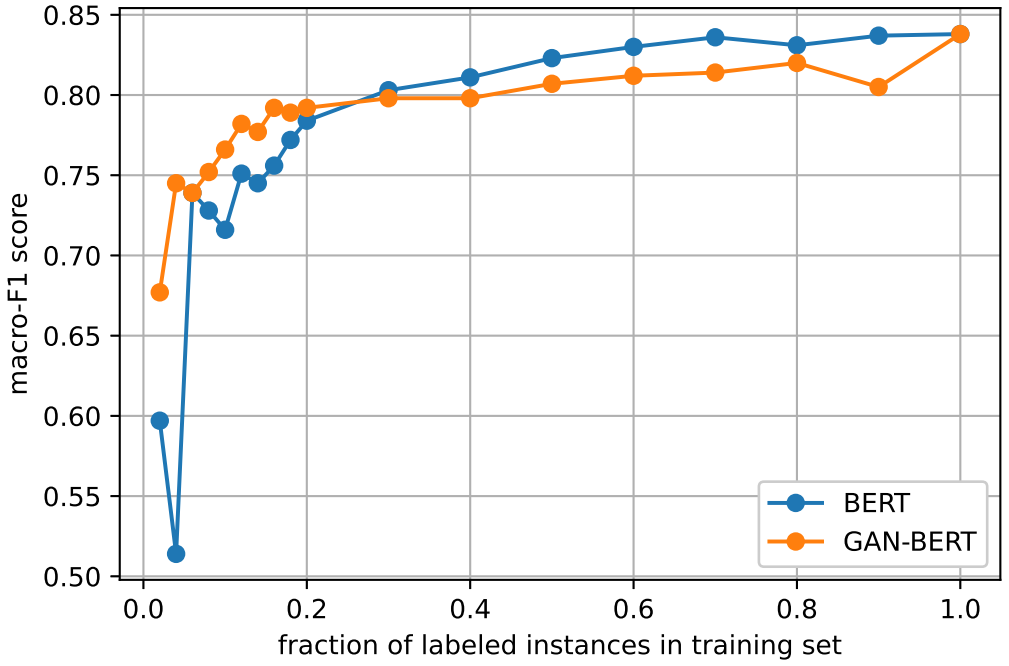


FIGURE 3.3: Macro-F1 scores for ad hominem classification for different fraction of labeled instances in the training set.

1.3.18 Comment 18 - Ad hominem

super stupid has just said that the mule team has explained to it that women have a 63 percent dis approval of trump super stupid did the mule team tell you that or was it abc clear up your confused world for us

1.3.18 Comment 18 - Ad hominem

super stupid has just said that the mule team has explained to it that women have a 63 percent dis approval of trump super stupid did the mule team tell you that or was it abc clear up your confused world for us

FIGURE 3.4: An example visualization of highlighted triggers generated using the attention scores of `[CLS]` token for an ad hominem comment using BERT *(cyan)* and GAN-BERT *(orange)*.

then fine-tune the BERT model on only the labeled instances and evaluate its performance. This is repeated for different fractions of labeled instances while training. We observe that as we increase the fraction of unlabeled instances, the macro-F1 score for classification degrades. This is the major disadvantage of the transformer architectures, like BERT, that they require at least thousands of annotated examples to achieve state-of-the-art results on the targeted tasks. However, with very less annotated examples, they fall apart.

We experimented with GAN-BERT to leverage the unlabeled instances in the training set in a generative adversarial setting, which we simply cannot use while working with BERT. As the results in Figure 3.3 show, the requirement of labeled instances can be drastically reduced (up to only 50-100 labeled instances) for obtaining similar performance as that of BERT. When low fraction of instances in training set are annotated, GAN-BERT beats BERT. One of the possible reasons could be that as the number of annotated instances is low, GAN-BERT leverages its GAN architecture to use the unlabeled instances to train its discriminator. As this utility is not present in BERT, its performance is not very good. Naturally, when larger fraction of annotated instances are available in the training set BERT beats GAN-BERT since the utility of the GAN architecture is diminished. Figure 3.4 compares the visualizations obtained by using BERT (when 100% training instances are labeled) and GAN-BERT (when 15% training instances are labeled). We observe that even with much less annotated data, GAN-BERT is able to detect triggers similar to that of BERT. In the rest of the paper we will report results using the BERT-based model, the GAN-BERT models produced similar results.

Chapter 4

Detecting ad hominem fallacies in the wild

Now that we created an accurate and explainable ad hominem detector using the Reddit data, we aimed to test the usage of ad hominem fallacies in discussion forums from the wild. To that end, we chose CreateDebate as our experimental testbed.

4.1 Collecting CreateDebate dataset

CreateDebate is a social networking debate website, launched in 2008. It was built around ideas, discussion, and democracy to help groups of people to sort through issues, viewpoints, and opinions. The discussions in CreateDebate often aim towards reaching consensus and understanding to make better decisions. CreateDebate is very similar to Reddit with a notable exception— its moderation policy is different, only the debate creator can be the debate moderator[1].

Each CreateDebate post is created by a user who also acts as its moderator. The forum allows users to write their perspectives as comments on the posts. Other users can support a comment, dispute it or clarify it as replies. The site, like Reddit, doesn't limit the depth of comment nesting. CreateDebate forum is divided into 14 topical forums–Politics,

[1] https://www.createdebate.com/about/faq

Entertainment, World News, Religion, Law, Science, Technology, Sports, Comedy, Business, Travel, Shopping, Health, and NSFW. The majority of the content in CreateDebate is public for all the forums. We found CreateDebate to be suitable for our investigation since it is a popular discussion-based forum in the wild with loose moderation. In effect, CreateDebate provides us an opportunity to measure the prominence of ad hominem fallacy usage over time.

Topic	**# Posts**	**# Comments**	**# Users**
Politics	10,434	119,850	7,686
Religion	2,841	77,418	4,563
World News	2,008	27,418	3,622
Science	1,276	20,691	2,837
Law	759	11,016	1,436
Technology	909	8,421	2,674
Total	18,227	264,814	14,961

TABLE 4.1: Basic statistics of our collected CreateDebate dataset. The first post in our dataset was posted on Feb 20, 2008 and the last post was updated on Nov 24, 2021.

In this work, we programmatically collected the complete publicly available CreateDebate dataset for all topical CreateDebate forums from the inception of the CreateDebate service. However, for brevity, we will present primarily results from the top six CreateDebate forums—Politics, Religion, World News, Science, Law, and Technology. Results from all other topical forums remained the same. We present the general statistics of the dataset in Table 4.1—in totality, these six forums contain 18,227 posts with 264,814 comments made by 14,961 users uploaded over 14 years. We verified that all posts we collected from CreateDebate were in English. We leveraged this large-scale discussion dataset (posted over the years) collected from CreateDebate to detect ad hominem from online discussions. However, we faced a crucial question—is our detector, trained over the Reddit CMV dataset, extendable to the CreateDebate dataset? We explored this question next.

4.2 Validating our ad hominem detector on CreateDebate discussions

After collecting the CreateDebate data, we faced a dilemma—our ad hominem detector was fine-tuned on the Reddit CMV dataset (as noted in the previous section), however,

CreateDebate is a very different forum with potentially different userbase and linguistic styles (including syntactic and semantic differences with Reddit). Thus, in this section, we will report a real-world data-driven survey that establishes the validity of our detector on the CreateDebate dataset.

Study setup: Our goal was to test the accuracy of our model output on the CreateDebate dataset. To that end, we ran our BERT-based detector on this dataset and randomly sampled 50 comments which were classified as ad hominem by our detector and another 50 comments which were non-ad hominem. Next, we created a simple online survey. We used Prolific[2], a crowdsourcing platform, to recruit participants for our survey. We recruited 18+ years old US nationals who were fluent in English and had a 100% approval rating on the platform and had participated in at least 200 previous studies. In this survey we presented a set of comments (from our sample of 100 random CreateDebate comments) along with a link to access the original CreateDebate discussion and its replies to each participant. Then we asked the participants to mark each of those comments as ad hominem or non ad hominem. For each comment, we additionally presented (in case the participant deem it to be an ad hominem) top three phrases identified by our model (with highest weights) and enquired if these key phrases indeed make this content ad hominem (the participants could also add their own phrases in a free form text field). This part of our survey was aimed to validate the explainability of our model. In total 15 participants gave three responses for each of the 100 comments (each participant gave responses for a batch of 20 comments); comparing the annotators across 5 batches yielded substantial inter-annotator agreement (0.73 Fliess' κ). The average time per participant was 8 minutes and we compensated them with $1. The survey instrument is provided in the Appendix A.

Results: We found that for 94% of CreateDebate comments, the labels given by participants were the same as the predicted label by our model, signifying the high validity of our model output even on the CreateDebate dataset. Furthermore, for 94.3% of ad hominem comments, the key phrases identified by our model (with the highest attention scores) exactly matched with the participant-identified phrases. This shows the power of the generalizability of our model.

[2] https://www.prolific.co/

Chapter 5

Characterizing ad hominem fallacy usage in CreateDebate discussions

We used our (almost) accurate and explainable detector on the CreateDebate dataset and characterized the ad hominem fallacies. We will start by exploring the volume of ad hominem fallacies in the wild.

5.1 Usage of ad hominem fallacies in CreateDebate

Topic	**Ad hominem comments**		**Ad hominem users**	
	(#)	(%)	(#)	(%)
Politics	42,718	35.64	2,686	34.95
Religion	20,194	26.08	1,712	37.52
World News	6,701	24.44	1,191	32.88
Science	5,437	26.28	996	35.11
Law	2,642	23.98	482	33.57
Technology	1,401	16.64	676	25.28
Total	79,093	29.97	4,965	33.19

TABLE 5.1: Basic statistics of ad hominem content in our collected CreateDebate dataset (users who have 50% or more comments as ad hominem arguments are referred here as ad hominem users).

We simply run our BERT-based detector on CreateDebate data to find the answer to the question—do CreateDebate users leverage ad hominem fallacy? We present the answer in Table 5.1. Surprisingly, the percentage of ad hominem comments in the CreateDebate

Topic	Examples
Politics	Now Socialist have ya not paid any attention to the Gubernor of Michigan ? Damn Socialist are you as stupid as you seem to be when you engage your 1 brain cell
Religion	you don't have the right to believe in a sacrifice that never happen, why don't you find an ass to wipe you're a massive sack of shite with half of half a mind you're asinine you think that black is white you're more dishonest then a bag of kykes
World News	The Fat Boy Dicktater is launching missiles and is it for you to say the next missile does not carry an ICBM and how would you know ? Are you using a statement to prove a point you no nothing about ?
Science	Don't understand even the basics of how chemotherapy works or are anti-science I see. Oh Jesus Christ you are just soooooooooooooooooooo stupid. UV radiation has got nothing to do with chemotherapy you brainless Nazi retard. Chemotherapy uses gamma radiation to target cancer cells. It also makes patients extremely sick and causes their hair to fall out. You're an idiot and every single word you type is stupid.
Law	Amy anyone can speak in terms you whine about without your sensitive ears hearing what was said. Keep crying u college educated fool because you cannot stop free speech in any forum !
Technology	Excuse me you imbecile, I graduated from Bismarck State College with a bachelors in Farm and Ranch Management. I was known on campus as "The Great Debater", and successfully won 18 arguments. So far I have been gentlemanly, but if you keep up with this funny business you will make me unleash my inner demons and go full throttle debate god.

TABLE 5.2: Examples of ad hominem arguments across different topical subforums for CreateDebate.

forum is alarmingly high, especially for CreateDebate topical forums related to Politics (35.64%). In fact, a large number of users are using these ad hominem fallacies—34.95% for Politics, demonstrating, ad hominem fallacies are used rampantly in the wild. These numbers contrast with the Reddit CMV forum where Habernal et al. (2018) found only 0.02% posts to be ad hominem. Even for a regular online discussion, only 19.5% of comments under online news articles were found to be incivil (Coe et al., 2014), much lower than the reported fraction of ad hominems. We show some examples of topical ad hominems posted on CreateDebate in Table 5.2.

Username	# TLC	Username	# DR
UserA	1,445	**UserC**	5,371
UserB	1,388	**UserB**	4,039
UserC	1,078	**UserA**	2,694
UserD	1,077	UserK	2,513
UserE	1,030	**UserD**	2,365
UserF	845	UserL	2,345
UserG	770	**UserG**	1,577
UserH	603	UserM	1,297
UserI	521	**UserI**	1,200
UserJ	506	UserN	1,069

TABLE 5.3: Users who posted most of the top-level comments (TLC) *(left)* and who received most of the direct replies (DR) *(right)* on the Politics subforum (usernames are censored for privacy concerns and common users in both sides are bold-faced).

Now, we ask an obvious question—why is this percentage alarmingly high in contrast to CMV's 0.02%, even though there is a mechanism to *report* a comment on CreateDebate?

To answer this question, we focus on ad hominem posts from 'Politics' subforum (owing to its 35.64% ad hominem content).

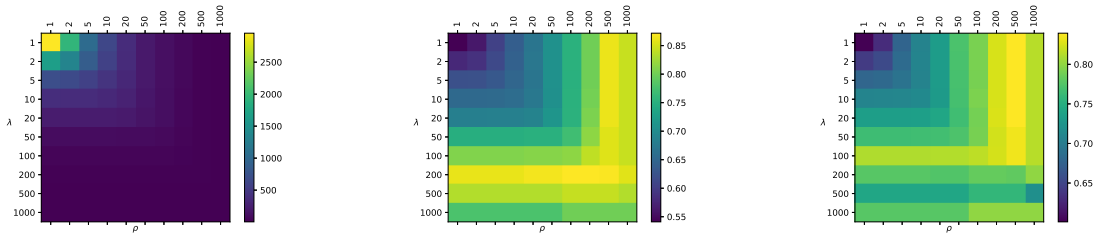
As the statistics in Table 5.3 show, the Politics subforum of CreateDebate follows a heavy tail distribution, where 10 most active users posted about 26% of the top-level comments, it also shows the users who received the most direct replies to their posts on the subforum. Interestingly, both sides have 6 usernames in common, signifying the possible influence of only a handful of key players. Thus, we ask—do these active users have any role in elevating the fraction of ad hominem arguments in the CreateDebate forum?

5.2 Correlation between activity and ad hominem

We wanted to investigate if a handful of users are colluding to upload disproportionally more ad hominem content. To that end, we define $X(\lambda)$ as the set of authors who posted at least λ top-level comments, and $Y(\rho)$ as the set of authors who received at least ρ direct replies to their comments. We define the set $S(\lambda, \rho)$ as $X(\lambda) \cap Y(\rho)$ and try to understand the activity of users corresponding to different levels of λ and ρ.

Highly active users act as a community while posting: We start by creating a directed graph showing support and dispute between the authors in $S(\lambda, \rho)$. The weights of the edge between nodes A and B in support and dispute networks represents how many times author A agreed/disagreed with author B via direct reply. As the results in Figure 5.1 show, the number of authors in the set decreases if either λ or ρ is increased, yet the reciprocity in support and dispute networks increases with an increase in λ and ρ. Our finding implies that the influential actors, who also happened to write most posts and receive most replies (high λ and ρ), participate in the debates not as an individual but as small-sized communities/cohorts of highly active users.

Community of highly active users post a disproportionately high volume of ad hominem posts: We started with the hypothesis that the alarmingly high ad hominem



(a) (b) (c)

FIGURE 5.1: Variation in (a) number of authors in $S(\lambda, \rho)$; reciprocity in (b) support network and (c) dispute network for different λ and ρ for CreateDebate political debates.

content present in the CreateDebate forum is mostly generated by small communities of highly active users. These users always post together and reply to each other. To check whether our hypothesis is true, we grouped the users participating in political debates on the CreateDebate forum based on their total comment count. Next, we test the fraction of ad hominem comments in the content generated by each group. The result is shown in Table 5.4. The set of authors who wrote more than 2000 comments constituted only 0.1% of the users, yet they post around 26.5% of total content with a whooping 55.2% of their content flagged as ad hominem. This group's activity is in stark contrast with the users who post less than 10 comments—they posted only 11.8% comments and a meager 16.8% of those comments were ad hominem. This finding confirms our hypothesis and identifies an intriguing pattern of ad hominem posting—these illogical personal-level attacks are often used by a highly active community or cohort of users, presumably to shut down the voice of less active opponents.

Group	**% Users**	**% Comments**	**% Ad hominem**
≤10	87.8	11.8	16.8
11–50	8.1	10.5	24.6
51–100	3.6	31.0	26.2
101–2000	0.4	20.2	26.1
≥ 2000	0.1	26.5	55.2
Total	100.0	100.0	35.64

TABLE 5.4: Prevalence of ad hominem within different groups for CreateDebate politcal debates.

Summary: CreateDebate dataset contains a surprisingly high volume of ad hominem—29.97%, which is much higher than the figures reported in earlier works on ad hominem content in other forums. Using the Politics subforum, we further showed that these ad

Characteristics	**Class $\mathcal{C}_1$**	**Class $\mathcal{C}_2$**	**MWU p-value**
# posts	49.26 ± 294.71	2.02 ± 2.25	0
reward points	169.41 ± 1070.35	6.89 ± 14.31	0
efficiency	88.31 ± 13.07	91.58 ± 13.55	1.86×10^{-142}
# allies	1.84 ± 9.11	0.14 ± 0.83	2.76×10^{-249}
# enemies	0.66 ± 6.21	0.03 ± 0.24	3.67×10^{-203}
# hostiles	0.59 ± 2.99	0.03 ± 0.20	0
reciprocity (SN)	0.64 ± 0.06	0.58 ± 0.02	3.18×10^{-264}
reciprocity (DN)	0.69 ± 0.05	0.64 ± 0.03	1.22×10^{-263}

TABLE 5.5: Average and standard deviation of different characteristics along with p value computed using Mann–Whitney U test of the distribution of the two classes $\mathcal{C}_1$ and $\mathcal{C}_2$. *SN* denotes support network and *DN* denotes dispute network.

hominem comments are largely facilitated by highly active CreateDebate users who collude among themselves to post together on the same discussion-reply threads and reply to each other[1]. Finally, we investigate—does this huge prevalence of ad hominem has any correlation with time, or was the ad hominem always prevalent in online forums?

5.3 Differences in characteristics of users who are posting ad hominem with those who aren't

CreateDebate maintains profile pages for its users. The profile pages contain the reward points of the users, their efficiency while debating, number of debates they participated in, number of comments they posted, when they joined the forum and when were they last online. We can also see the *allies*, *enemies* and *hostiles* of any given user. The proper definitions of these terms can be found on CreateDebate FAQ page[2].

We collected these characteristics for all the users of CreateDebate and partitioned the users into two classes—those who have posted ad hominem comment at least once on the forum ($\mathcal{C}_1$) and those who haven't ($\mathcal{C}_2$), and then computed the average and standard deviation of these characteristics for these two classes. We also considered the average and standard deviation of reciprocity observed by constructing support and dispute networks.

The results are shown in Table 5.5. It can be observed that the characteristics distribution of the two classes are statistically very different from each other (as noted with very low

[1]Our observations also hold for other topics; these results are not shown for brevity.
[2]https://www.createdebate.com/about/faq

p value for Mann-Whitney U test). Hence, we can use these characteristics to predict whether any given user is prone to ad hominem contributions. It is also possible to predict which users will engage in this behavior ahead of time by carefully tracking the temporal variation in these characteristics.

Chapter 6

Understanding temporal variations in ad hominem usage

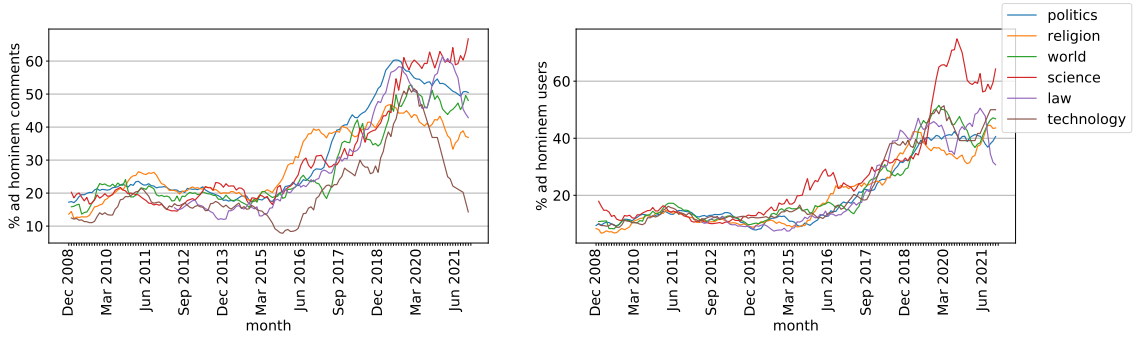


FIGURE 6.1: Variation in percentage of ad hominem posts posted per month on CreateDebate across different topics *(left)* and percentage of users for whom 50% or more posts are ad hominem per month *(right)* (averaged over 1 year).

CreateDebate was launched in 2008 as a tool 'that democratizes the decision-making process through online debate'. But as we have observed, the percentage of ad hominem content on the website is alarmingly high when compared to any other regular debate forum. So, what went wrong? To answer this question, we perform the first temporal analysis of ad hominem usage for the CreateDebate forum. Our first task involved generating temporal snapshots to capture the month-wise activity on the site. As our CreateDebate dataset spans from February 2008 to November 2021, it would be very clumsy to do the day-wise activity analysis, and a year-wise scheme will have only 14 data points. Hence, we chose

to study month-wise activities. The variation in the percentage of comments which were flagged as ad hominem and the percentage of users who were posting such comments is shown in Figure 6.1 for each month in the period. It can be observed that the plots for all the topical sub-forums follow a similar trend—initially they are stationary, then they show a steep rise and then they fall. In order to gain insights of what exactly triggered this sharp rise and fall, we performed change point detection experiments to quantitatively partition the corpus into three sub-corpora—the stationary $\mathcal{H}_1$, the rise $\mathcal{H}_2$ and the fall $\mathcal{H}_3$. We used variation in number of comments posted, percentage of comments which were flagged as ad hominem and percentage of users who were posting such comments for each month across all topical sub-forums as input to the change point detection algorithm which uses dynamic programming to find the optimal partition using RBF kernels as cost function. The cutoffs for the partitions as predicted by the algorithm are March 2017 and September 2019. Hence, the timeline for the sub-corpora are—$\mathcal{H}_1$ (February 2008 – February 2017), $\mathcal{H}_2$ (March 2017 – September 2019) and $\mathcal{H}_3$ (October 2019 – November 2021).

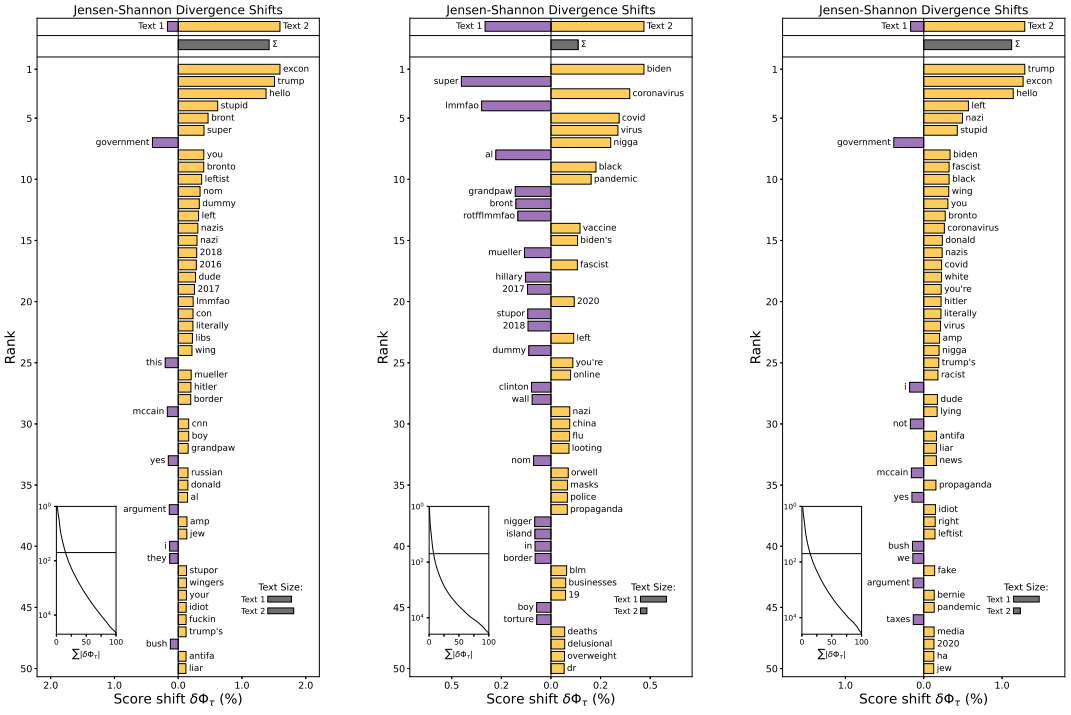


FIGURE 6.2: Word shift graphs for comparing sub-corpora (a) $\mathcal{H}_1$ and $\mathcal{H}_2$, (b) $\mathcal{H}_2$ and $\mathcal{H}_3$, and (c) $\mathcal{H}_1$ and $\mathcal{H}_3$.

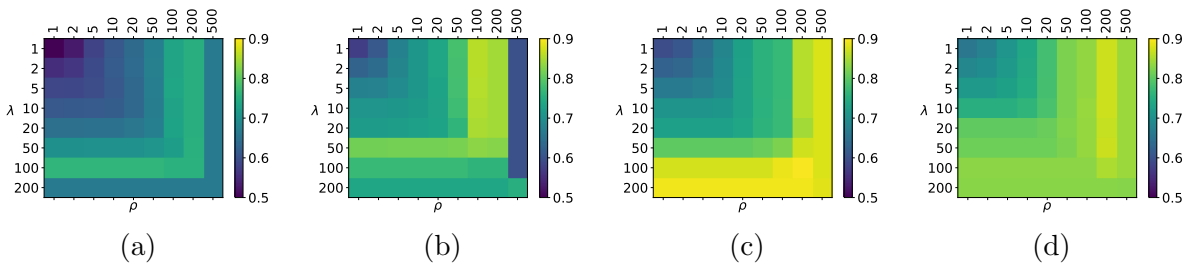


FIGURE 6.3: Reciprocity for CreateDebate politics subforum (a) support network and (b) dispute network constructed using $\mathcal{H}_1$, and (c) support network and (d) dispute network constructed using $\mathcal{H}_2$. λ and ρ were also varied.

We then compared these sub-corpora by generating word-shift graphs using Jensen-Shannon divergence, which are shown in Figure 6.2. It can be observed that the use of terms that act as triggers to ad hominem argumentation is very prominent in $\mathcal{H}_2$ when compared to $\mathcal{H}_1$. These triggers are also present in $\mathcal{H}_3$, but they are not as dominant as in $\mathcal{H}_2$. We constructed the support and the dispute networks for $\mathcal{H}_1$ and $\mathcal{H}_2$ (see Figure 6.3) and observed that the reciprocity has increased significantly between users who wrote at least 100 top-level comments or received at least 200 direct replies, for support as well as dispute networks.

It is very interesting to note that the timelines for 2016 US Presidential election and the Covid-19 outbreak are very close to the predicted cutoffs of the partitions. We observed that CreateDebate forum was heavily used for political debates during the 2016 US Presidential election. Our hypothesis is that usage of ad hominem argumentation was accelerated after the 2016 US Presidential debates – forum becoming highly polar, people choosing sides. As it has been observed throughout history, the use of illogical arguments is very prevalent when people are discussing Politics, but to win debates, the use of ad hominem comments skyrocketed on the forum. However, due to the Covid-19 pandemic, the activity of users on the forum declined significantly, thus curbing the ad hominem usage.

This is a plausible explanation about rising ad hominem arguments in political debates. But what about other topical forums like Religion? Why is the ad hominem content increasing for these topics? To understand this complex phenomenon, we partitioned the religious debates using the above scheme and peeked into what users are talking about. We

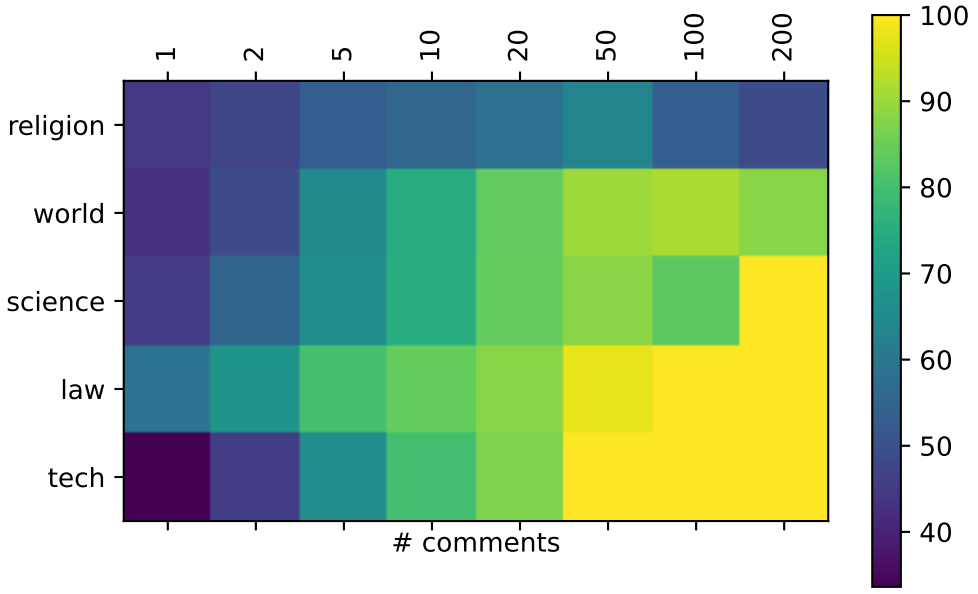


FIGURE 6.4: Percentage overlap of user base for different topic categories with Politics. #comments denotes the threshold for filtering users with different comment counts.

observed that the religious debates that were published before the 2016 US Presidential debates have a negligible political angle, however, those published after the Presidential debates were highly convoluted with Politics. As the results in Figure 6.4 show, each topical sub-forum on the CreateDebate site has a huge user overlap with Politics, especially the highly active users. For some categories like Science and Law, the overlap approaches 100%. This explains why the increase in ad hominem comments and users posting them across different categories show similar trends as Politics (see Figure 6.1). So, it seems plausible to believe that political discussions are the root cause of the alarmingly high ad hominem content on CreateDebate, which skyrocketed apparently after the 2016 US Presidential debates started.

Chapter 7

Concluding Discussions

In this work, we shed a data-driven light for the first time about ad hominem usage in the wild by leveraging the CreateDebate data posted by tens of thousands of users over a period of more than a decade. We reported creating detectors with high accuracy whose judgment matched that of users for 94% of the cases. Using this detector, we uncover that around *one-third* of all content on CreateDebate is simply ad hominem, which is extremely surprising. We deep-dived to find that a cohort of highly active users is responsible for this high fraction of ad hominems. Moreover, users particularly influenced by the political discourse resorted to this fallacy. While almost all data-driven studies suffer from intrinsic bias, we strongly believe this work is still valuable for understanding the ecosystem of cyber aggression as well as designing novel and more respectful debating platforms. In this final section, we will discuss the limitations as well as key implications of our work.

7.1 Limitations

Our work has a couple of limitations. First, our detector is bound by the annotation quality and volume of the CMV dataset. We believe we could have achieved higher accuracy with more data. However, even our detector achieved a significantly high accuracy compared to the prior works and we established that this detector is also valid on CreateDebate—a different forum and dataset, underpinning the efficacy of the detector. Second, our results

might or might not be generalizable beyond Reddit, CreateDebate and in general online forums to share opinions and debate. Even then it has a lot of societal importance since as prior works show such debate forums in themselves are extremely important to shape the Internet and shape public opinion (Proferes et al., 2021) Third, our findings are often correlated and not causality. However, we believe these correlations show underlying large-scale behavioral patterns which facilitate ad hominems and in effect a toxic culture. Thus, in spite of this limitation, our work is still useful as a first attempt to shed light on ad hominem usage patterns.

7.2 Implications

We identify three key implications of our work for platform designers as well as platform regulators.

Defending against logical fallacies is becoming important: One very surprising, and concerning finding is that our dataset from CreateDebate, a popular opinion-forming and debating forming forum is filled with ad hominem fallacies. This finding highlights the importance of understanding and defending against logical fallacies which perpetuates toxic culture and attempts to stop any opposing arguments using irrelevant personal attacks. Thus, to counter cyber-aggression and bring back respect to online spaces the platforms should acknowledge this issue and actively design defenses against such fallacies. The key focus today is on defending against hate speech and misinformation. However, where hate speech and misinformation lie to the users, ad hominem fallacy, increasing at an alarming speed shows a drift to stop any opposing voice.

It is possible to detect and defend against these fallacies via automated means: Our work demonstrated that we can leverage current extremely powerful techniques like BERT and GAN-BERT to create highly accurate ad hominem detectors, even when only a handful of annotated posts are available. Thus our work can also be interpreted as a very strong proof of concept about using automated means (e.g., classifiers) by the platforms to protect against such fallacies along with hate speech and misinformation.

Users need to be nudged to reduce the usage of these fallacies: Finally, the results of this work strongly hints at the need of nudging users to reduce logical fallacies. As our results show, highly active user cohorts in CreateDebate are using ad hominem in more than 50% of their comments. Hence, community members can help report users showing such behavior. Furthermore, the substantially high ad hominem usage was rooted in the political climate of 2016 and now is spreading through other topics and forums, polluting the online space. The high volume of the affected population possibly even hint that many of these users might not even realize that they are utilizing fallacious arguments. Thus the current platforms should focus on nudging the users against the usage of potential ad hominems even before they upload a fallacious post. Using the models discussed in this paper, online debate forums and social media sites can nudge users when they are posting ad hominem comments by providing them the ad hominem triggers. Moderators can also decrease exposure of such comments by pushing them at the bottom of the thread, and flag them. Overall, we strongly believe that our findings will help policymakers and platform developers help detect and defend against ad hominem fallacies in online opinion influencing forums.

Appendix A

Survey instrument

This section contains the survey instrument that we used during annotation studies.

A.1 Instructions

Identifying Personal Attacks in Comment Chains

An ad hominem argument (or argumentum ad hominem in Latin) is used to counter another argument. However, it's based on feelings of prejudice (often irrelevant to the argument), rather than facts, reason, and logic. An ad hominem argument is often a personal attack on someone's character or motive rather than an attempt to address the reasoning that they presented.

Sometimes, people utilize ad-hominem argument (fallacy) because they want to appeal to other's emotions rather than their reasoning (since they are based on personal attack). Ad-hominem is often used in toxic conversations or comment chains in the internet.

A.2 Examples

Let's review several ad hominem examples. Unfortunately, they're prevalent in the courtroom and in politics, so we'll begin there. To no surprise, ad hominem arguments also occur in any sort of daily interaction, so we'll review a few more everyday examples, too.

The more you read about examples of ad hominem arguments, the more you'll be able to spot them and, if need be, defend yourself against such arguments.

Next, gave five examples of Ad hominems identified from prior work in four situations—In the Court, In the Political Debates, Used in the Media, In Everyday Conversations.

A.3 Task

In this task, you will be shown 20 comments, one comment per page. For each comment, you will be asked whether the given comment is ad hominem argument or not. For additional context, each comment is provided with an URL of the full conversation (post and comments). You will also be asked to select some keywords from the comments shown, which you think, best describes your judgment (ad hominem or otherwise).

Note – Devices you can use to take this study: Desktop and Tablet

For each of the 20 comments show the following

- Show the comment excerpt (with a link to the conversation for added context)
- Do you think this is an ad-hominem comment? ◯ Yes ◯ No
- If participant chose ad hominem Select the phrases from the comments, which you think, makes it an ad hominem comment. If some other phrase makes it ad hominem, please enter that in 'Other' option. ☐ word 1 ☐ word 2 ☐ word 3 ☐ other ____

Bibliography

Abbott, R., Ecker, B., Anand, P., and Walker, M. (2016). Internet argument corpus 2.0: An SQL schema for dialogic social media and the corpora to go with it. In *(LREC'16)*, pages 4445–4452.

Boudry, M., Paglieri, F., and Pigliucci, M. (2015). The fake, the flimsy, and the fallacious: Demarcating arguments in real life. *Argumentation*, 29(4):10–1007.

Coe, K., Kenski, K., and Rains, S. A. (2014). Online and uncivil? patterns and determinants of incivility in newspaper website comments. *Journal of Communication*, 64:658–679.

Cook, N., Ayers, S., and Horsch, A. (2018). Maternal post traumatic stress disorder during the perinatal period and child outcomes: A systematic review. *J Affect Disord .*, 2018(225):18–31.

Croce, D., Castellucci, G., and Basili, R. (2020). GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *ACL '2020*, pages 2114–2119. Association for Computational Linguistics.

Das, M., Saha, P., Dutt, R., Goyal, P., Mukherjee, A., and Mathew, B. (2021). You too brutus! trapping hateful users in social media: Challenges, solutions & insights.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL'2019*, pages 4171–4186. Association for Computational Linguistics.

Eemeren, F. H. V. and Grootendorst, R. (1987). Fallacies in pragma-dialectical perspective. *Argumentation*, 1(3):283–301.

Eysenbach, G. and Till, J. E. (2001). Ethical issues in qualitative research on internet communities. *BMJ*, 323(7321):1103–1105.

Goodfellow, I. J., Pouget-Abadie, J., Mirza, M., Xu, B., Warde-Farley, D., Ozair, S., Courville, A., and Bengio, Y. (2014). Generative adversarial nets. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, pages 2672–2680.

Goodman, J. C. (2020). Four years of argument ad hominem. https://www.independent.org/news/article.asp?id=13359. Accessed: 2021-01-14.

Habernal, I., Wachsmuth, H., Gurevych, I., and Stein, B. (2018). Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *NAACL'18*, pages 386–396. Association for Computational Linguistics.

Hamblin, C. L. (1970). Fallacies. *Tijdschrift Voor Filosofie*, 33(1):183–188.

Hasan, K. S. and Ng, V. (2014). Why are you taking this stance? identifying and classifying reasons in ideological debates. In *EMNLP'14*, pages 751–762, Doha, Qatar. Association for Computational Linguistics.

Jain, S., Bhatia, A., Rein, A., and Hovy, E. (2014). A corpus of participant roles in contentious discussions. In Chair), N. C. C., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., and Piperidis, S., editors, *LREC'14*, Reykjavik, Iceland. European Language Resources Association (ELRA).

Kennedy, G. A. (1993). Aristotle "on rhetoric": A theory of civic discourse. *Philosophy and Rhetoric*, 26(4):322–327.

Macagno, F. (2013). Strategies of character attack. *Argumentation*, 27:369–401.

Mathew, B., Illendula, A., Saha, P., Sarkar, S., Goyal, P., and Mukherjee, A. (2020a). Hate begets hate: A temporal study of hate speech.

Mathew, B., Saha, P., Yimam, S. M., Biemann, C., Goyal, P., and Mukherjee, A. (2020b). Hatexplain: A benchmark dataset for explainable hate speech detection.

Mondal, M., Silva, L., Correa, D., and Benevenuto, F. (2018). Characterizing usage of explicit hate expressions in social media. *New Review of Hypermedia and Multimedia*, 24(2):110–130.

Mondal, M., Silva, L. A., and Benevenuto, F. (2017). A Measurement Study of Hate Speech in Social Media . In *Proceedings of the 25th ACM Conference on Hypertext and Social Media (HT'17)*, Prague, Czech Republic.

Proferes, N., Jones, N., Gilbert, S., Fiesler, C., and Zimmer, M. (2021). Studying reddit: A systematic overview of disciplines, approaches, methods, and ethics. *Social Media + Society*, 7(2):20563051211019004.

Qiu, M. (2015). *Mining user viewpoints in online discussions*. PhD thesis, School of Information Systems, Singapore Management University.

Qiu, M., Sim, Y., Smith, N. A., and Jiang, J. (2015). Modeling user arguments, interactions, and attributes for stance prediction in online debate forums. In *SDM*.

Radford, A. and Narasimhan, K. (2018). Improving language understanding by generative pre-training.

Redinger, C. (2020). Opinion: It's not debatable, ad hominem attacks destroy constructive conversation. https://unothegateway.com/opinion-its-not-debatable-ad-hominem-attacks-destroy-constructive-conversation/. Accessed: 2021-01-14.

Sahai, S., Balalau, O., and Horincar, R. (2021). Breaking down the invisible wall of informal fallacies in online discussions. In *ACL'21*, pages 644–657, Online. Association for Computational Linguistics.

Salimans, T., Goodfellow, I., Zaremba, W., Cheung, V., Radford, A., Chen, X., and Chen, X. (2016). Improved techniques for training gans. In Lee, D., Sugiyama, M., Luxburg, U., Guyon, I., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 29.

Schiappa, E. and Nordin, J. P. (2013). *Argumentation: Keeping Faith with Reason*. Pearson UK.

Sheng, E., Chang, K., Natarajan, P., and Peng, N. (2020). "nice try, kiddo": Ad hominems in dialogue systems. *CoRR*, abs/2010.12820.

Tan, C., Niculae, V., Danescu-Niculescu-Mizil, C., and Lee, L. (2016). Winning arguments. *Proceedings of the 25th International Conference on World Wide Web*.

Tindale, C. W. (2007). *Fallacies and Argument Appraisal*. Cambridge University Press.

Trabelsi, A. and Zaïane, O. R. (2014). Finding arguing expressions of divergent viewpoints in online debates. In *LASM'14*, pages 35–43, Gothenburg, Sweden. Association for Computational Linguistics.

Wei, Z., Liu, Y., and Li, Y. (2016a). Is this post persuasive? ranking argumentative comments in online forum. In *ACL'16*, pages 195–200, Berlin, Germany. Association for Computational Linguistics.

Wei, Z., Xia, Y., Li, C., Liu, Y., Stallbohm, Z., Li, Y., and Jin, Y. (2016b). A preliminary study of disputation behavior in online debating forum. In *ArgMining'16*, pages 166–171, Berlin, Germany. Association for Computational Linguistics.

Woods, J. (2007). Lightening up on the ad hominem. *Informal Logic*, 27:109–134.

Wulczyn, E., Thain, N., and Dixon, L. (2017). Ex machina: Personal attacks seen at scale. In *WWW'17*, pages 1391–1399. International World Wide Web Conferences Steering Committee.

Zalta, E. N. (2004). *The Stanford Encyclopedia of Philosophy*. Stanford, CA: The Metaphysics Research Lab.