Is This The Real Life? Is This Just Fallacy? – Identifying and Characterizing Ad hominem Fallacy Usage in The Wild

Project-I (EC47007) 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

Autumn Semester, 2021-22

November 21, 2021

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: November 21, 2021 (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 "**Is This The Real Life? Is This Just Fallacy? – 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 Autumn Semester, 2021-22.

Prof. Mainack Mondal and Prof. Animesh Mukherjee

Department of Computer Science & Engineering

Indian Institute of Technology Kharagpur

Kharagpur - 721302, India

November 21, 2021

Abstract

Today online forums and social media sites facilitate easy collaborative opinion formation for billions of users surpassing geographical boundaries. However, intuitively, this same dynamics of opinion formation also let large-scale usage of fallacies like *Ad-hominem* (personal attack in response to a logical question) creep into online debates by malicious actors. Ad-hominem, although a simple fallacy, is effective enough to sway public debates in offline world and can be used as a precursor to shutting the voice of opposition by slander.

In this work, we take a first step in shedding light on the use of Ad-hominem fallacies in the wild. First, we create BERT-like explainable models to detect Ad-hominem and provide linguistic insights into triggers of hurling Ad-hominem fallacies in the wild, even for datasets with as small as 100 annotated examples only. Our models achieved macro-F1 score of 0.84 on fully annotated datasets and 0.74 on datasets with only 4% annotated examples. We then applied our models on an online debate forum – Create Debate (114k comments from 7k users) and conversations on popular Facebook pages (65k comments from 43k users). We performed several crowd-sourced surveys to validate our in-the-wild predictions and observed an average (across two datasets) macro-F1 score of 0.94. Our investigations reveal that Ad-hominem fallacies are routinely used in online debates and are more frequently leveraged by cohorts of popular/influential users to win an argument. We conclude by pointing out the important implications of our work.

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

Ad hominem arguments are based on feelings of bias (mostly irrelevant to the argumentation), rather than realities, reason, and rationale. They 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 other's emotions rather than their reasoning.

1.1 Overview

To understand triggers and dynamics of ad hominem argumentation and how different they are for a debate portal and a social media site, we scraped political debates from Create Debate and posts and comments from popular political figures and news agencies in US from Facebook. We performed network studies on the users, classified the comments using BERT (Devlin et al., 2019) fine-tuned on annotated Change My View (Reddit) dataset (Habernal et al., 2018) and validated in-the-wild predictions via several crowd-sourced surveys.

1.2 Research Questions

1. Why people tend to use ad hominem arguments while debating **(RQ1)**?

2. How much annotated examples are required for identifying ad hominem argumentation by state-of-the-art NLP models **(RQ2)**?

3. How robust are the Transformer models in generating key-phrases that make the comment ad hominem **(RQ3)**?

RQ1 attempts to investigate the triggers of ad hominem argumentation and the qualitative and quantitative properties of Web forums and social media sites that encourages these phenomenon. RQ2 attempts to investigate the approaches that would require minimal annotated examples to identify ad hominem argumentation with very high confidence. RQ3 attempts to investigate how reliable and robust the Transformer models, e.g. BERT, are for generating the key-phrases that make the comments ad hominem.

1.3 Key Observations

For RQ1, we found out that Create Debate and Facebook are very different platforms with respect to how users interact and debate. Hyperactive user interactions are very prominent on Create Debate, but not so on Facebook.

For RQ2, we found out that using SS-GAN schema (Salimans et al., 2016) applied over BERT drastically reduces the requirement of a large annotated dataset for end-to-end training.

For RQ3, we found out that BERT performs exceptionally well in generating key-phrases that are the building blocks for ad hominem argumentation, achieving precision up to 0.97.

Chapter 2

Prior Work

2.1 Ad hominem Argumentation

It was Aristotle who first observed that some arguments are in fact 'deceptions in disguise'. With few exceptions, the ad hominem argumentations are categorized in following subtypes (Schiappa and Nordin, 2013; Macagno, 2013):

1. Abusive ad hominem (*a pure assault on the personality of the rival*)
2. Tu quoque ad hominem (*an assault on somebody for an apparent shortcoming by the way they have communicated their perspective*)
3. Circumstantial ad hominem (*"practice what you preach" type attacks*)
4. Bias ad hominem (*assuming the attacked opponent has a hidden agenda*)
5. Guilt-by-association (*associating the adversary with someone with a low believability*)

2.2 Semi-Supervised Learning using GANs

State-of-the-art transformer architectures like BERT (Devlin et al., 2019) are trained over large scale annotated datasets and are fine-tuned over a targeted task to achieve state-of-the-art results in various NLP tasks. One bottleneck is that these achievements are

obtained when thousands of annotated examples are available for the targeted tasks. One possible solution is to integrate these humongous pre-trained models with 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 is instead trained to distinguish samples of the generator from the real instances. SS-GANs (Salimans et al., 2016) are an extension to GANs where the Discriminator also assigns a class label to each example while discriminating whether or not it was naturally produced.

A significant drop is observed on the performance when BERT is fine-tuned with less than 200 annotated instances. However, the SS-GAN schema applied over BERT, i.e., GAN-BERT, decreases the requirement for annotated examples; even with less than 200 annotated examples it is feasible to acquire results tantamount with a fully supervised setting. (Croce et al., 2020).

Chapter 3

Data

3.1 Create Debate

Create Debate[1] is an online debate forum that calls itself 'tool that democratizes the decision-making process through online debate'. It allows the users to view debates of various types (for-against, perspective, and challenge debates) and topics (politics, entertainment, law, etc.) and sort the debates by time, state (open/closed) and other features like most-heated, most-votes, most-arguments, etc. We scraped 10k political debates containing 114k arguments and counter-arguments.

There are 6892 authors active on the forum. We analyzed the distribution of authors on the basis of: number of arguments (all level) and level-1 arguments they wrote, and number of direct replies they received. Out of 6892, only 5815 authors initiate discussions in at least one thread. The statistics are shown in Figure 3.1. We observe that number of authors who wrote n comments decreases monotonically as we increase n, just what one would expect from a regular debate forum. The same statistics are observed for level-1 comments and direct replies. However, one thing that differentiates Create Debate from some other regular debate forum is that most of the contents are posted by only a small percentage of user base. From Table 3.1, it can be observed that only top 10 (0.15%) posted about 26% level-1 comments for the entire corpus.

[1]Debate Forum | Online Debate Community | CreateDebate: https://www.createdebate.com

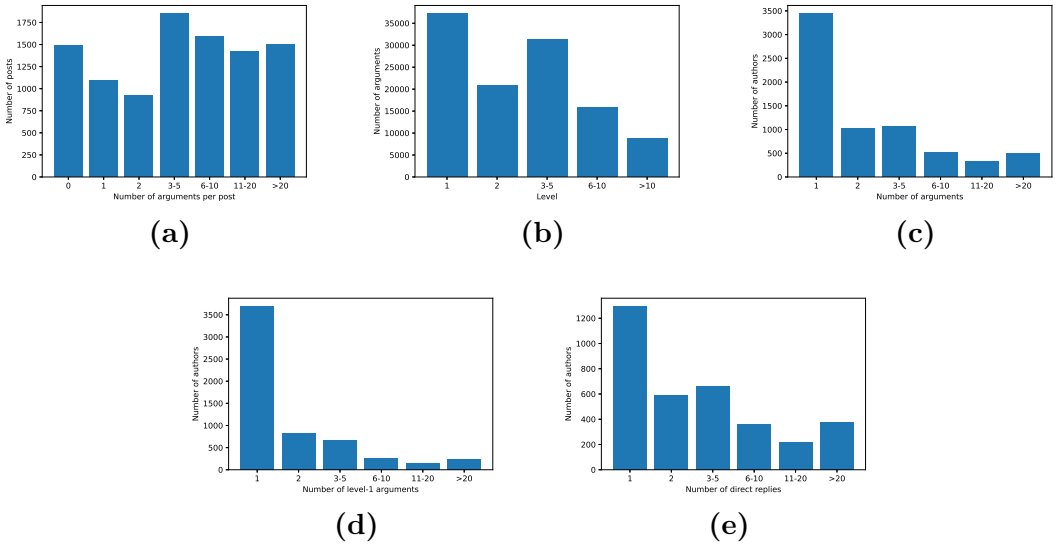


FIGURE 3.1: Distribution of **(a)** arguments per post, **(b)** arguments on each level, **(c)** authors w.r.t. comment frequency, **(d)** authors w.r.t. level-1 comment frequency, **(e)** authors w.r.t. direct reply frequency for Create Debate corpus

Username	**Number of level-1 comments posted**
AlofRI	1445
outlaw60	1388
excon	1078
brontoraptor	1077
PhxDemocrat	1030
Chinaman	845
FromWithin	770
Grenache	603
PrayerFails	521
HighFalutin	506

TABLE 3.1: Top 10 users on the basis of level-1 comment count in Create Debate corpus

Table 3.1 and Table 3.2 have 6 common usernames. These are the users who initiate discussion in majority of the threads and also receive large number of direct replies, indicating a possibility of link farming on Create Debate forum.

Let $X(\lambda)$ be the set of authors who wrote at least λ level-1 comments, and $Y(\rho)$ be the set of authors who received at least ρ direct replies. We consider the set $S(\lambda, \rho) = X(\lambda) \cap Y(\rho)$ and try to understand dynamics of this set by varying parameters λ and ρ. For the given set $S(\lambda, \rho)$, we start by creating a directed graph showing support and dispute between the authors in S. The weights of the edge between nodes A and B in support and dispute

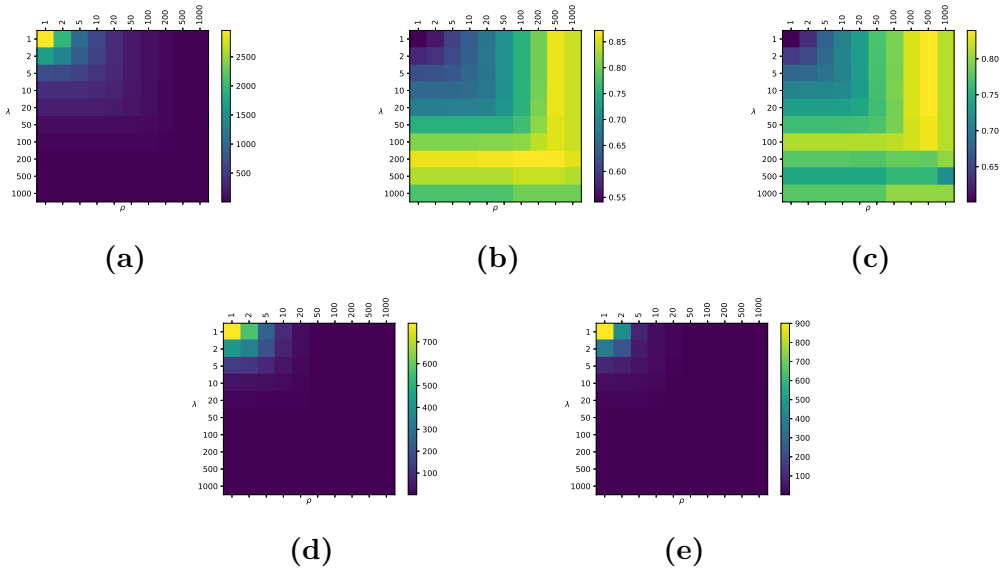


FIGURE 3.2: Variations in **(a)** number of authors in $S(\lambda, \rho)$; reciprocity in **(b)** support graph and **(c)** dispute graph; number of strongly-connected components in **(d)** support graph and **(e)** dispute graph for different λ and ρ for Create Debate corpus

Username	**Number of direct replies received**
excon	5371
outlaw60	4039
AlofRI	2694
Cartman	2513
brontoraptor	2365
BurritoLunch	2345
FromWithin	1577
Amarel	1297
PrayerFails	1200
JustIgnoreMe	1069

TABLE 3.2: Top 10 users on the basis of direct reply count in Create Debate corpus

networks represents how many times author A agreed/disagreed with author B via direct reply. The results are shown in Figure 3.2. We observe that number of authors, $|S(\lambda, \rho)|$, decreases if either λ or ρ is increased, however reciprocity in support and dispute graphs increases with an increase in the parameters λ and ρ. This signifies that the hyperactive user interactions are very prominent in Create Debate forum.

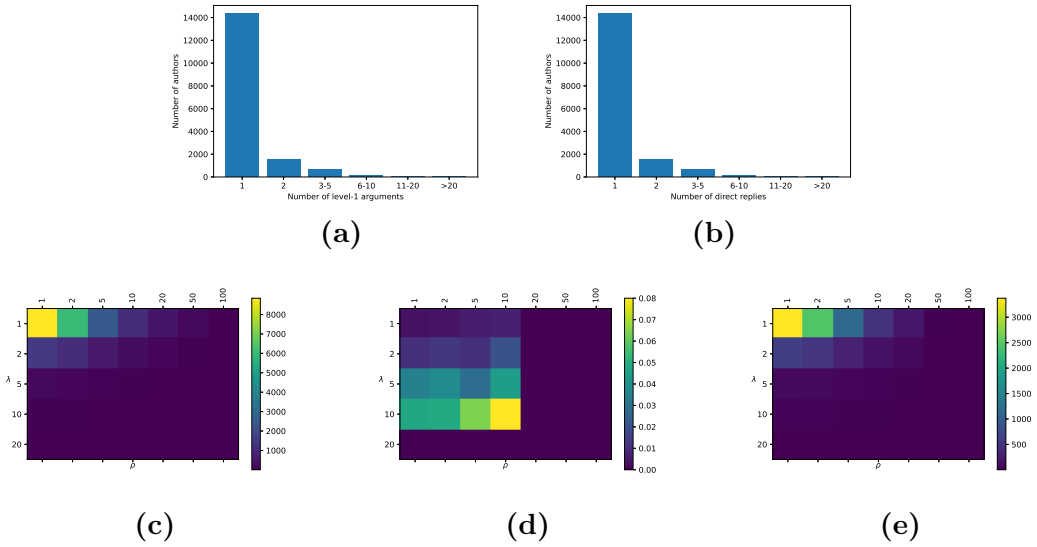


FIGURE 3.3: Distribution of **(a)** authors w.r.t. level-1 comment frequency, **(b)** authors w.r.t. direct reply count; variations in **(c)** number of authors in $S(\lambda, \rho)$, **(d)** reciprocity among users, and **(e)** number of strongly-connected components for different λ and ρ for Facebook corpus

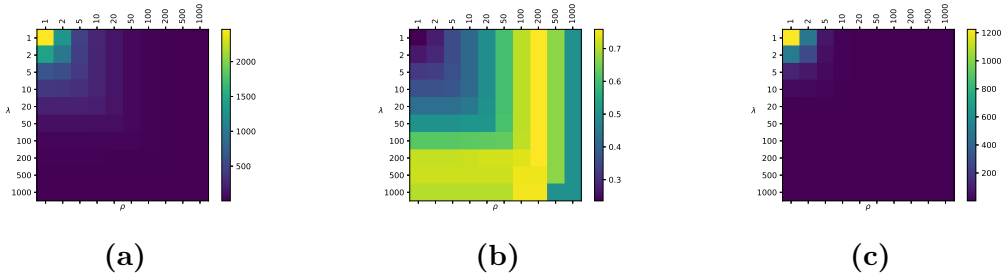


FIGURE 3.4: Variations in **(a)** number of authors in $S(\lambda, \rho)$, **(b)** reciprocity among users, and **(c)** number of strongly-connected components for different λ and ρ for truncated Create Debate corpus

3.2 Facebook

We scraped 65k comments by 43k users from the Facebook pages of five popular political figures and news agencies in the US: Barack Obama, Donald Trump, Joe Biden, Fox News and Brietbart. We analyzed the distribution of authors w.r.t. level-1 comment frequency and direct replies count. We also investigated possibility of link farming for this corpus as well (Figure 3.3). We observed one striking dissimilarity between Create Debate and Facebook in terms of reciprocity among the users. For Create Debate, we observed that as the λ and ρ parameters are increased, the authors in $S(\lambda, \rho)$ tend to show a very high

reciprocity. However, the same is not true for Facebook. Beyond a certain value of λ and ρ, a sudden drop in reciprocity is observed, which indicates that the hyperactive users on Facebook don't often interact with one another. We were not sure whether this behavior is the representative of how the users interact on Facebook, or it is just an anomaly due to very limited entries per user in our Facebook dataset. To resolve this doubt, we performed the same link farming analysis on a truncated version of Create Debate, having similar statistics as our Facebook dataset. We observed that even after this down-scaling, similar reciprocity statistics are observed for Create Debate (Figure 3.4), as if user dynamics are scale-invarient. We conclude that Facebook is a very different platform as compared to Create Debate with respect to how users interact and debate.

Chapter 4

Experiments

The experimental part is divided into five parts. We first experiment with ad hominem detection in fully annotated datasets in Section 4.1, then with the datasets having very less number of annotated examples in Section 4.2, then we use the trained models to annotate comments scraped from the wild (Create Debate and Facebook) in Section 4.3, investigate what makes arguments ad hominem in Section 4.4 and finally we validate our models' predictions via crowd-sourced surveys in Section 4.5.

4.1 Identifying Ad hominem Arguments in Fully Annotated Datasets

The first experimental set-up examines ad hominem arguments in the fully annotated CMV dataset. Habernal et al. (2018) used two neural classifiers, namely a 2-stacked BiLSTM, and a CNN for the task.

Out of the 7242 comments in the CMV dataset, of which 3622 were labeled ad-hominem (+) and 3620 were labeled not ad hominem (-), we created a training set of 5242 (2622+, 2620-) comments, while the rest 2000 (1000+, 1000-) comments were put in the test set. We fine-tuned the BERT model (Devlin et al., 2019) on the training set and evaluated the model performance on the test set. Same was done for the models used by Habernal et al.

Model	Accuracy	Precision (AH)	Recall (AH)	F1 (AH)	Precision (None)	Recall (None)	F1 (None)	Macro F1
CNN	0.789	0.746	0.876	0.806	**0.850**	0.701	0.768	0.787
2 Stacked Bi-LSTM	0.726	0.671	**0.885**	0.764	0.831	0.567	0.674	0.726
BERT	**0.839**	**0.837**	0.840	**0.839**	0.840	**0.837**	**0.838**	**0.838**

TABLE 4.1: Prediction of ad hominem arguments against baselines for fully annotated datasets

Labeled fraction x	**Unlabeled fraction** y										**Best** y
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	
0.1	0.512	0.749	0.755	0.773	0.776	0.755	0.770	0.765	0.760	0.764	0.4
0.2	0.756	0.806	0.803	0.782	0.788	0.795	0.796	0.783	0.800		0.1
0.3	0.800	0.786	0.805	0.802	0.808	0.799	0.797	0.813			0.7
0.4	0.789	0.816	0.807	0.803	0.802	0.805	0.818				0.6
0.5	0.811	0.806	0.801	0.826	0.804	0.794					0.3
0.6	0.817	0.819	0.814	0.807	0.816						0.1
0.7	0.813	0.815	0.823	0.808							0.2
0.8	0.823	0.824	0.826								0.2
0.9	0.822	0.820									0.0
1.0	0.826										0.0

TABLE 4.2: Macro-F1 score computed over different pairs of labeled examples fraction x and unlabeled examples fraction y for GAN-BERT

(2018). The results are shown in Table 4.1. We achieved a 5.1% absolute improvement in macro-F1 score against the baselines.

4.2 Identifying Ad hominem Arguments in Sparsely Annotated Datasets

For creating sparsely annotated datasets, we intentionally drop the labels of a given fraction of comments in the training set. We fine-tune BERT on only labeled comments in the training set and evaluate its performance on the test set as created in Section 4.1. We observed that BERT performed poorly on these datasets with very less number of annotated training examples. We experimented with GAN-BERT (Croce et al., 2020) so as to leverage the unlabeled training examples, in a generative adversarial setting, which we simply ignore in the case of BERT. Results shows that the requirement of labeled examples can be drastically reduced (up to only 50-100 labeled examples), still obtaining good performances in the classification task. Table 4.2 and Table 4.3 show the computation of optimal fraction of unlabeled examples to be used for training GAN-BERT for a given fraction of labeled examples. Table 4.4 and Table 4.5 show the comparison between BERT and GAN-BERT as we vary fraction of labeled examples.

Labeled fraction x	**Unlabeled fraction** y										**Best** y
	0.0	**0.1**	**0.2**	**0.3**	**0.4**	**0.5**	**0.6**	**0.7**	**0.8**	**0.9**	
0.02	0.564	0.690	0.682	0.690	0.672	0.694	0.684	0.698	0.681	0.711	0.9
0.04	0.604	0.716	0.717	0.725	0.735	0.717	0.744	0.740	0.731	0.746	0.9
0.06	0.507	0.746	0.726	0.738	0.741	0.737	0.734	0.724	0.744	0.749	0.9
0.08	0.651	0.735	0.741	0.729	0.764	0.763	0.772	0.754	0.774	0.779	0.9

TABLE 4.3: Macro-F1 score computed over different pairs of very low labeled examples fraction x and unlabeled examples fraction y for GAN-BERT

Model	**Accuracy**	**Precision (AH)**	**Recall (AH)**	**F1 (AH)**	**Precision (None)**	**Recall (None)**	**F1 (None)**	**Macro F1**
BERT (10%)	0.721	0.796	0.594	0.680	0.676	0.848	0.752	0.716
BERT (20%)	0.785	0.786	0.782	0.784	0.783	0.787	0.785	0.784
BERT (30%)	0.804	0.812	0.790	0.801	0.796	0.817	0.806	0.803
BERT (40%)	0.811	0.794	0.840	0.816	0.830	0.782	0.805	0.811
BERT (50%)	0.824	0.801	0.860	0.829	0.848	0.787	0.817	0.823
BERT (60%)	0.831	0.818	0.850	0.834	0.844	0.811	0.827	0.830
BERT (70%)	0.836	0.835	0.838	0.836	0.837	0.834	0.836	0.836
BERT (80%)	0.832	0.838	0.822	0.830	0.825	0.841	0.833	0.831
BERT (90%)	0.837	0.845	0.826	0.835	0.830	0.848	0.839	0.837
BERT (100%)	0.839	0.837	0.840	0.839	0.840	0.837	0.838	0.838
GAN-BERT (10%)	0.777	0.767	0.796	0.781	0.788	0.758	0.773	0.777
GAN-BERT (20%)	0.792	0.802	0.775	0.788	0.782	0.809	0.795	0.792
GAN-BERT (30%)	0.798	0.796	0.802	0.799	0.800	0.794	0.797	0.798
GAN-BERT (40%)	0.798	0.788	0.815	0.801	0.808	0.781	0.794	0.798
GAN-BERT (50%)	0.807	0.813	0.797	0.805	0.801	0.817	0.809	0.807
GAN-BERT (60%)	0.813	0.818	0.804	0.811	0.807	0.821	0.814	0.812
GAN-BERT (70%)	0.814	0.807	0.824	0.815	0.820	0.803	0.812	0.814
GAN-BERT (80%)	0.820	0.825	0.811	0.818	0.814	0.828	0.821	0.820
GAN-BERT (90%)	0.818	0.805	0.839	0.822	0.831	0.797	0.814	0.805
GAN-BERT (100%)	0.839	0.837	0.840	0.839	0.840	0.837	0.838	0.838

TABLE 4.4: Prediction of ad hominem arguments for different fractions of annotated training examples

Model	**Accuracy**	**Precision (AH)**	**Recall (AH)**	**F1 (AH)**	**Precision (None)**	**Recall (None)**	**F1 (None)**	**Macro F1**
BERT (2%)	0.600	0.585	0.682	0.630	0.619	0.517	0.563	0.597
BERT (4%)	0.584	0.548	0.964	0.699	0.850	0.204	0.329	0.514
BERT (6%)	0.739	0.742	0.732	0.737	0.736	0.746	0.741	0.739
BERT (8%)	0.729	0.751	0.684	0.716	0.710	0.773	0.740	0.728
BERT (10%)	0.721	0.796	0.594	0.680	0.676	0.848	0.752	0.716
BERT (12%)	0.752	0.766	0.725	0.745	0.739	0.778	0.758	0.751
BERT (14%)	0.745	0.737	0.762	0.749	0.754	0.728	0.741	0.745
BERT (16%)	0.756	0.739	0.792	0.764	0.776	0.720	0.747	0.756
BERT (18%)	0.772	0.762	0.792	0.776	0.783	0.752	0.767	0.772
BERT (20%)	0.785	0.786	0.782	0.784	0.783	0.787	0.785	0.784
GAN-BERT (2%)	0.678	0.688	0.650	0.668	0.668	0.705	0.686	0.677
GAN-BERT (4%)	0.746	0.758	0.721	0.739	0.734	0.770	0.752	0.745
GAN-BERT (6%)	0.739	0.734	0.750	0.742	0.744	0.728	0.736	0.739
GAN-BERT (8%)	0.752	0.736	0.786	0.760	0.770	0.718	0.743	0.752
GAN-BERT (10%)	0.767	0.739	0.826	0.780	0.803	0.708	0.752	0.766
GAN-BERT (12%)	0.783	0.791	0.768	0.779	0.775	0.797	0.786	0.782
GAN-BERT (14%)	0.777	0.784	0.764	0.774	0.770	0.790	0.780	0.777
GAN-BERT (16%)	0.793	0.800	0.780	0.790	0.785	0.805	0.795	0.792
GAN-BERT (18%)	0.790	0.805	0.764	0.784	0.775	0.815	0.795	0.789
GAN-BERT (20%)	0.792	0.802	0.775	0.788	0.782	0.809	0.795	0.792

TABLE 4.5: Prediction of ad hominem arguments for very low fractions of annotated training examples

As the Generator in GAN-BERT samples noise from Gaussian distribution, the results of a given pair of labeled and unlabeled fraction (x, y) tend to vary if we run the experiment again. Hence, the results are averaged over 3 iterations for GAN-BERT. From Table 4.2 and Table 4.3, it is observed that for a given fraction of labeled examples, say x, the macro-F1 score obtained when the unlabeled fraction is $(1 - x)$ is very close to the maximum

macro-F1 score obtained for labeled fraction x. Hence, for computing the results for Table 4.4 and Table 4.5, GAN-BERT (x%) is trained by using all of the unlabeled data.

When low fraction of training data is used as labeled examples, GAN-BERT beats BERT. One of the possible reasons is: as number of labeled examples is low, GAN-BERT leverages its GAN architecture to use unlabeled examples to train its Discriminator. As this feature is not present in BERT, hence its performance is not very good. However, when high fraction of training data is used labeled examples, BERT beats GAN-BERT. Possible reason could be: GAN-BERT uses a Generator to produce fake examples from Gaussian noise, hence its Discriminator (which is the classifier) has one more extra class for classification i.e., fake instance. It may be due to unnecessary complexity that GAN-BERT loses to BERT for large labeled dataset.

4.3 Identifying Ad hominem Arguments in the Wild

Fine-tuned on annotated Change My View dataset, BERT was used to classify 114k comments in Create Debate corpus. 35% comments were classified as being ad hominem. We grouped the authors from the corpus on the basis of their comment count:

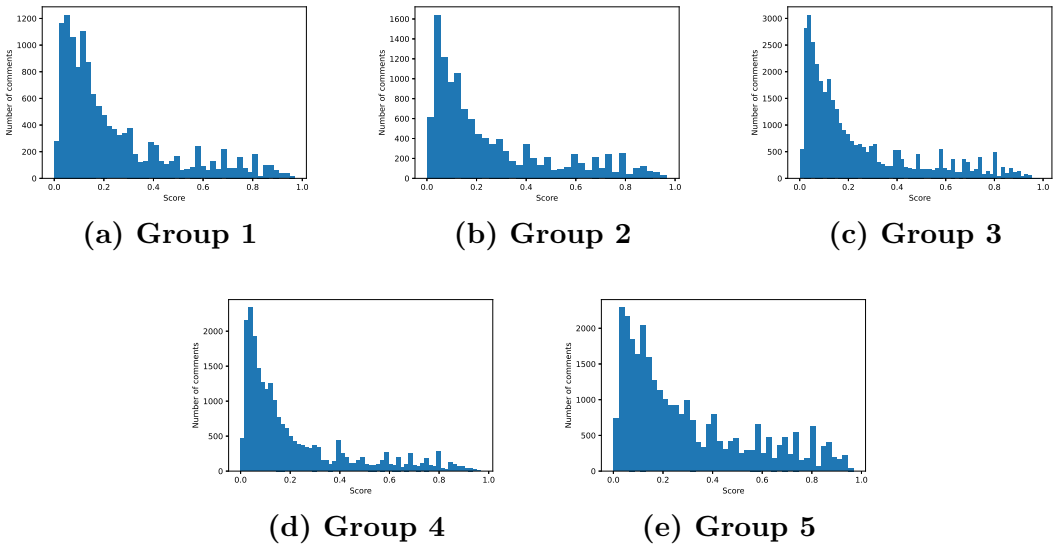
- Group 1: less than or equal to 10 comments
- Group 2: 11-50 comments
- Group 3: 51-500 comments
- Group 4: 501-2000 comments
- Group 5: greater than 2000 comments

Each group was analyzed using BERT and Perspective API[1]. Comments were analyzed against five attributes using Perspective API: *toxicity*, *identity attack*, *insult*, *threat* and *sexually explicit*. The group statistics are shown in Table 4.6. From Figure 4.1 – 4.5, we observed that for attributes like *toxicity* and *insult*, the distribution of scores for Group

[1]Perspective API: https://www.perspectiveapi.com/

Group #	Fraction of comments	Fraction of authors	Fraction of ad hominem comments
1	0.118	0.878	0.168
2	0.105	0.081	0.246
3	0.310	0.036	0.262
4	0.203	0.004	0.261
5	0.265	0.001	0.552

TABLE 4.6: Group statistics for Create Debate corpus



(a) Group 1 **(b) Group 2** **(c) Group 3**

(d) Group 4 **(e) Group 5**

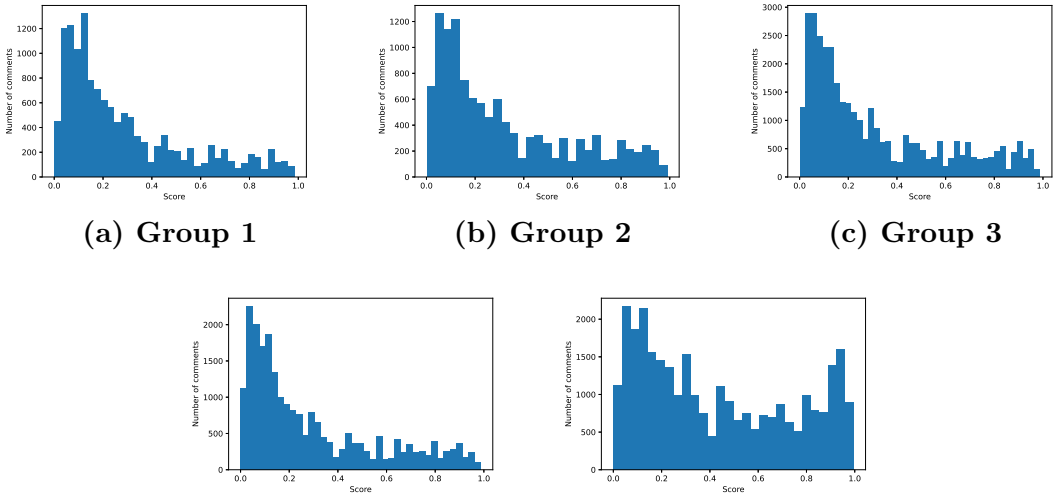
FIGURE 4.1: Perspective API scores for attribute *Identity Attack*

5 is very different from previous groups. However, there is no significant change in the distribution for other attributes like *identity attack*, *threat* and *sexually explicit*.

4.4 What Makes Arguments Ad hominem?

So far we detected ad hominem comments in fully annotated datasets, sparsely annotated datasets and in the wild. However, possible causes of the argumentative dynamics that ends up with ad hominem argument remain an open question. 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.

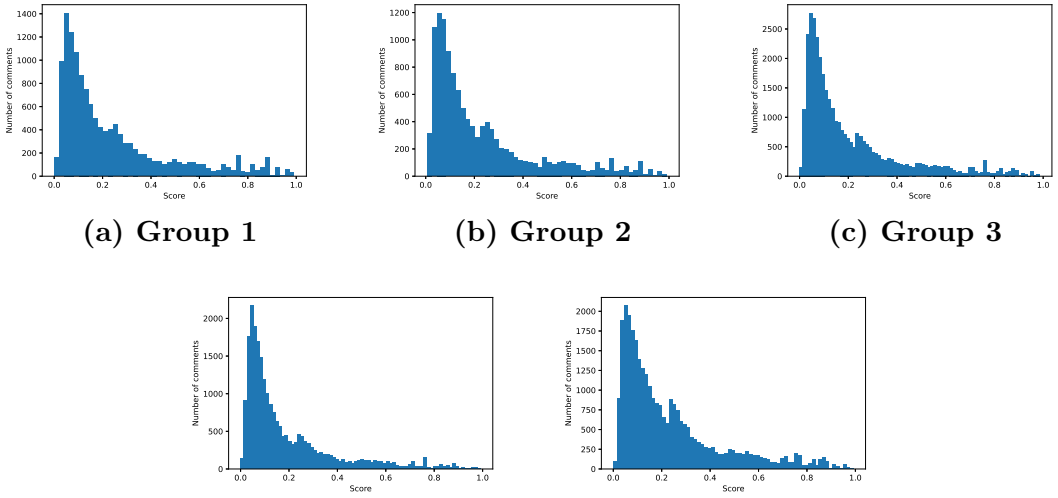
BERT is an encoder-based architecture. BERT-base has 12 encoder blocks (layers) and 12 attention heads. Consider the input sentence to be: `how dare people be poor so`



(a) **Group 1** (b) **Group 2** (c) **Group 3**

(d) **Group 4** (e) **Group 5**

FIGURE 4.2: Perspective API scores for attribute *Insult*



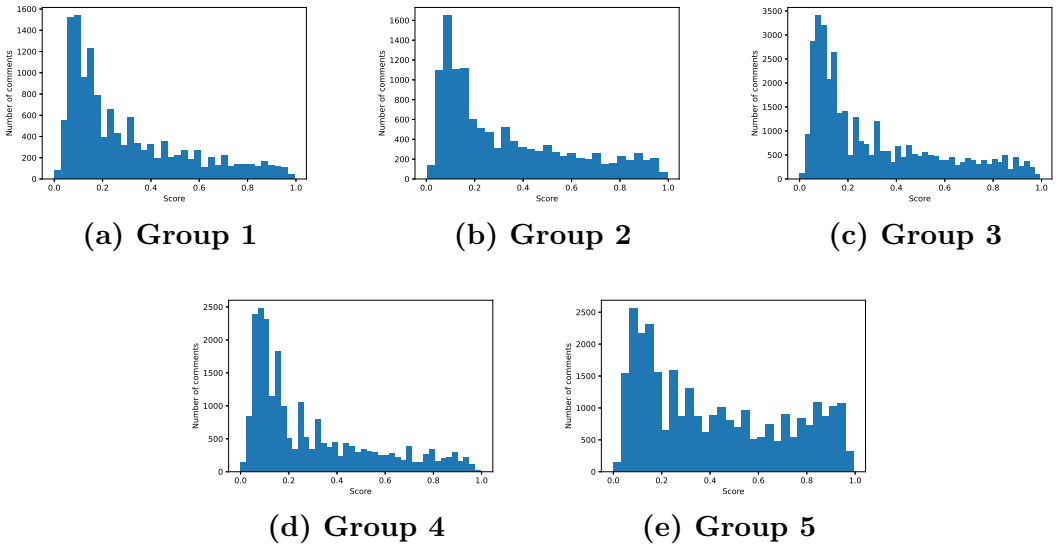
(a) **Group 1** (b) **Group 2** (c) **Group 3**

(d) **Group 4** (e) **Group 5**

FIGURE 4.3: Perspective API scores for attribute *Threat*

`unreasonable fucking you`. This input will be tokenized as: `[[CLS], how, dare, people, be, poor, so, unreasonable, fucking, you, [SEP]]`. Let's denote the size of tokenized input as m. Then, the attention scores returned by BERT will be a tensor, say A, of dimension $(12, 12, m, m)$, where $A[p, q, i, j]$ denotes the attention given to token T_i by token T_j in p^{th} encoder block and q^{th} attention head.

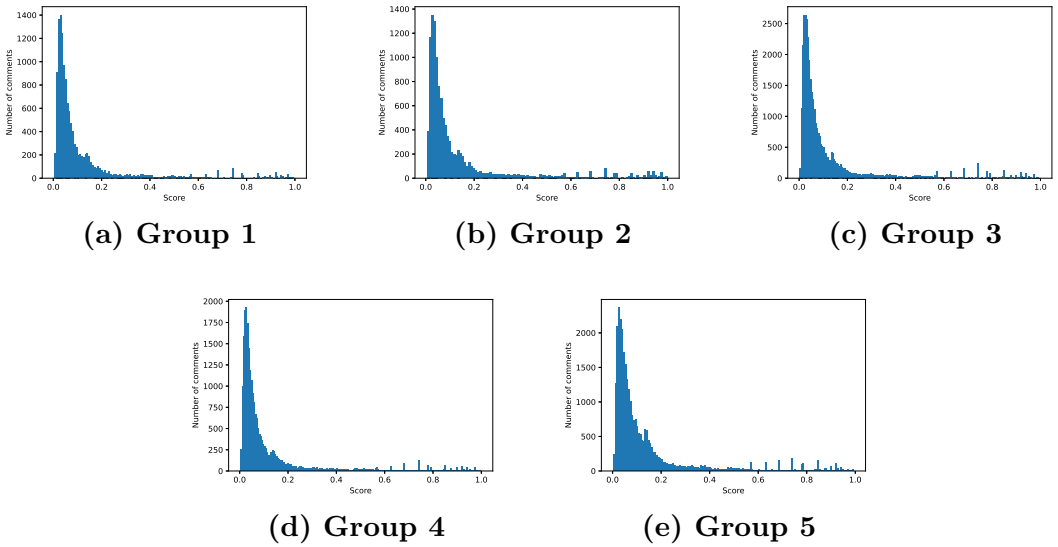
As `[CLS]` token is the aggregate representation of the input sequence for classification



(a) Group 1 **(b) Group 2** **(c) Group 3**

(d) Group 4 **(e) Group 5**

FIGURE 4.4: Perspective API scores for attribute *Toxicity*

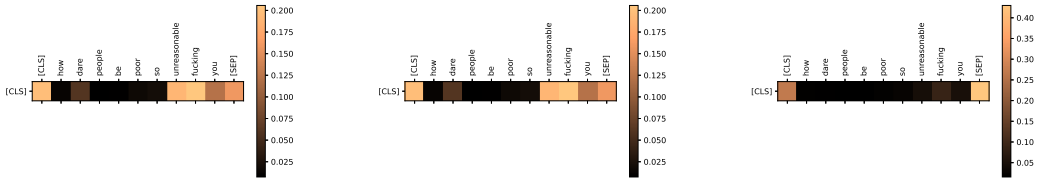


(a) Group 1 **(b) Group 2** **(c) Group 3**

(d) Group 4 **(e) Group 5**

FIGURE 4.5: Perspective API scores for attribute *Sexually Explicit*

tasks (Devlin et al., 2019), hence we use attention scores for `[CLS]` token to detect key tokens influencing the classification. A sample visualization is shown in Figure 4.6. We sampled 100 random comments each from Create Debate corpus and Facebook dataset. The sampling is done in a way that ensure a balance between the comments which are ad hominem with those which are not. For each comment, we extracted the attention vector for the `[CLS]` token averaged over all the encoder blocks and attention heads, selected



(a) **(b)** **(c)**

FIGURE 4.6: Attention scores for `[CLS]` token for **(a)** layer-11-head-8, **(b)** layer-11-head-averaged, and **(c)** layer-averaged-head-averaged

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

(a)

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

(b)

FIGURE 4.7: An example of reconstructed word weight heat map extracted from the attention vector of `[CLS]` token which ends up in ad hominem from **(a)** BERT and **(b)** GAN-BERT (Create Debate Batch-3 Comment-18)

the top 3 trigrams on the basis of the score of the center word, and projected them as heat maps. A sample heat map is shown in Figure 4.7. We analyzed these heat maps[2] for the comments which were predicted to be ad hominem. We observed that the BERT model was not only beating the baselines in terms of macro-F1 score of classification, it was also very good in picking the phrases which are the candidate triggers for ad hominem argumentation. Similar heat maps were produced and analyzed for GAN-BERT.

[2]The heat maps for 200 comments used in crowd-sourced surveys can be viewed here: https://github.com/utkarsh512/hatespeech

Dataset	Accuracy	Precision	Recall	F1-score
Create Debate	0.940	0.940	0.940	0.940
Facebook	0.810	0.900	0.763	0.826

TABLE 4.7: Evaluating BERT on in-the-wild predictions using crowd-sourced labels

Dataset	Exact matches	Exact matches or contained
Create Debate	0.943	0.977
Facebook	0.896	0.933

TABLE 4.8: Precision for the key-phrases generated using BERT over crowd-sourced key-phrases

4.5 Validating Models for In-The-Wild Predictions

We performed several crowd-sourced surveys to validate the labels and the heat maps produced by BERT. We used Prolific[3] platform for conducting the surveys. We recruited 18+ years old US nationals who were fluent in English and had a 100% approval rating and participated in at least 200 previous studies on Prolific. These ensured the accounts were sufficiently well-used for us to detect ad hominem comments in the wild. We sampled 100 comments from Create Debate dataset such that half of them were predicted as ad hominem by BERT. We created 5 batches with 20 comments (half of them being ad hominem) each. Each batch was annotated by 3 participants. We asked the participants to labels the comments as ad hominem or not. They were also asked to select key-phrases which they think made the comment ad hominem, if they are labeling it so. They had two choices: (a) choose some of 3 key-phrases which BERT generated and/or write their own keyphrases, or (b) choose 'None' option if they labeled it as 'not ad hominem'.

For each comment, we assumed that the labels provided by the majority of the Prolific participants is the gold label for that comment and evaluated the labels predicted by BERT using them. We performed the same studies with Facebook dataset. The results are shown in Table 4.7 and Table 4.8.

[3]Prolific: https://prolific.co/

Chapter 5

Conclusion

Throughout this study, we investigated approaches to identify and characterize ad hominem arguments in Web argumentation with very high confidence. We scraped political debates from Create Debate and posts and comments from popular political figures and news agencies in US from Facebook. We performed network studies on the users, classified the comments using BERT fine-tuned on annotated Change My View (Reddit) dataset and validated in-the-wild predictions via several crowd-sourced surveys.

We looked into the user dynamics for Create Debate forum and Facebook and found out that these two platforms are very different in the ways users interact and debate by performing link-farming analysis. We found out that using SS-GAN schema (Salimans et al., 2016) applied over BERT drastically reduces the requirement of a large annotated dataset for end-to-end training. We observed that BERT performed exceptionally well in generating key-phrases that are the building blocks for ad hominem argumentation, achieving precision up to 0.97.

There are a few focuses that merit further examination. First, we have overlooked metainformation of the members, such as their overall activity. Second, we expect that personal characteristics of the members (Top 10) may also play a significant role in the factious trade. Third, we expect that the temporal analysis of the Create Debate forum might give a better understanding of how the use of ad hominem arguments in Web debates has evolved through time.

We leave these focuses for future work. We accept that our discoveries will assist gain with bettering comprehension of, and hopefully keep restraining from, ad hominem argumentation in good-faith discussions.

Bibliography

Croce, D., Castellucci, G., and Basili, R. (2020). GAN-BERT: Generative adversarial learning for robust text classification with a bunch of labeled examples. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2114–2119, Online. Association for Computational Linguistics.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

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, page 2672–2680.

Habernal, I., Wachsmuth, H., Gurevych, I., and Stein, B. (2018). Before name-calling: Dynamics and triggers of ad hominem fallacies in web argumentation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 386–396, New Orleans, Louisiana. Association for Computational Linguistics.

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

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. Curran Associates, Inc.

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