Identifying and Characterizing Logical Fallacies

MTP-2 report submitted to

Indian Institute of Technology Kharagpur

in partial fulfilment for the award of the degree of

Master of Technology (Dual Degree 5 Years)

in

Electronics & Electrical Communication Engineering

by

Utkarsh Patel

(18EC35034)

Supervisors: Prof. Mainack Mondal and Prof. Animesh Mukherjee
Co-supervisor: Prof. Amitalok Budkuley

Department of Electronics & Electrical Communication Engineering

Indian Institute of Technology Kharagpur

Spring Semester, Academic Session 2022-23

May 2, 2023

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.

Utkarsh Patel

Date: May 2, 2023 (Utkarsh Patel)
Place: Kharagpur (18EC35034)

DEPARTMENT OF ELECTRONICS & ELECTRICAL COMMUNICATION ENGINEERING

INDIAN INSTITUTE OF TECHNOLOGY KHARAGPUR

KHARAGPUR - 721302, INDIA

CERTIFICATE

This is to certify that the project report entitled "**Identifying and Characterizing Logical Fallacies**" submitted by **Utkarsh Patel** (Roll No. 18EC35034) to Indian Institute of Technology Kharagpur towards partial fulfilment of requirements for the award of degree of Master of Technology (Dual Degree 5 Years) 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 2022-23.

1/5/2023

Supervisors: Prof. Mainack Mondal and Prof. Animesh Mukherjee
Co-supervisor: Prof. Amitalok Budkuley
Indian Institute of Technology Kharagpur
Kharagpur - 721302, India

May 2, 2023

Abstract

Reasoning is a crucial facet of human intelligence. People debate on a huge range of topics on online platforms like Facebook, Reddit, etc. Debates can be lengthy, with users exchanging a wealth of information and opinions. However, conversations do not always go smoothly, and users sometimes engage in unsound argumentation techniques to prove a claim. These techniques are called fallacies. Fallacies are persuasive arguments that provide insufficient or incorrect evidence to support the claim. The use of fallacious arguments is extremely commonplace in offline and online discussions. These discussions often render a strong influence on the overall opinion of the people. Twisting the flow of the arguments can have a strong impact on the minds of naïve individuals, which in the long run might have socio-political ramifications, for example, winning an election or spreading misinformation.

In this work, we propose a new linguistic approach to detect logical fallacies by identifying distinct dependency paths that are specific to each type of fallacy. The proposed framework was tested using dependency parsing and fallacy detection experiments, which revealed that the dependency paths generated by the framework accurately represented the logical structure of each fallacy. Despite the presence of overlapping dependencies, the framework identified highly discriminative paths for each type of fallacy. The proposed method provides an alternative and effective approach to detecting fallacies by extracting unique dependency paths, which could potentially lead to the development of sophisticated models for identifying and detecting logical fallacies.

Acknowledgements

The path has been a challenging one, full of unfamiliar territories, towering mountains, and unfathomable depths. I am grateful for the numerous individuals with whom I had the opportunity to connect and who joined me on this expedition.

I would like to take this opportunity to express my sincere gratitude to my supervisors—Prof. Animesh Mukherjee and Prof. Mainack Mondal—who have played a crucial role in my master's thesis. I am greatly indebted to Prof. Animesh Mukherjee for his invaluable guidance, unwavering support, and insightful feedback throughout my research journey. His knowledge and expertise have been instrumental in shaping my ideas and research approach. I appreciate his encouragement and the freedom he gave me to explore different avenues to solve research problems. I would like to express my deepest appreciation and gratitude to Prof. Mainack Mondal, who played a crucial role in shaping my research by providing his valuable expertise and insights. His guidance has helped me gain a better and more comprehensive understanding of the subject matter. His guidance and mentorship have been invaluable throughout this journey and have taught me the importance of scientific thinking and the need for a sound justification of any research approach. His encouragement and support have helped me to develop my skills and confidence as a researcher.

I would like to express my gratitude to my co-supervisor, Prof. Amitalok Budkuley, for his involvement in this journey. Although his role may not have been significant in shaping the direction of my research, his support and guidance played an important role. I am grateful for his contribution to my academic journey.

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, Mr. Punyajoy Saha and Mr. Sasi Bhushan. I am grateful for all the amazing people I had the chance to get to know at IIT Kharagpur.

Contents

Declaration **ii**

Certificate **iii**

Abstract **v**

Acknowledgements **vi**

Contents **vii**

List of Figures **ix**

List of Tables **xi**

Abbreviations **xiii**

Symbols **xv**

1 Introduction **1**

2 Prior Works **5**

3 Data Collection **9**
- 3.1 CreateDebate Forum 9
  - 3.1.1 For-against vs. Perspective Debates 11
- 3.2 LOGIC Dataset 12
- 3.3 Empath Dictionary 14

4 Experiments **17**
- 4.1 Logical Fallacy Detection in CreateDebate Forum 17
- 4.2 Study on CreateDebate's User Network 18
- 4.3 Comprehensive Linguistic Study of Logical Fallacies in CreateDebate 23
  - 4.3.1 Dependency Parsing 23
  - 4.3.2 Parsing Tree-structures 25
  - 4.3.3 Identifying Most Discriminative Dependency Paths for a Given Logical Fallacy 25

5 Concluding Discussions **35**

Bibliography **37**

List of Figures

4.1 Distribution of Logical Fallacies in the Politics Forum of CreateDebate. . . 18
4.2 Count of fallacious users across for-against and perspective debates. 19
4.3 The average number of *allies* a user has when they post a logically fallacious comment in both for-against and perspective debates. 20
4.4 The average number of *enemies* a user has when they post a logically fallacious comment in both for-against and perspective debates. 20
4.5 Average degree centrality of fallacious users across for-against and perspective debates. 21
4.6 Network density for for-against and perspective user network. 22
4.7 Example of a tree structure that is formed by applying dependency parsing on a sample text. In this example, the verb *charge* is identified as the root of the tree, and *universities* is the nominal subject of the verb *charge*. Additionally, *fees* is identified as the direct object of the verb *charge*. The adjectives *deemed* and *huge* are also identified as the modifiers of *universities* and *fees*, respectively. 24
4.8 Variation in dependency path length by popularity. Shorter dependency paths are much more frequent in the Politics CreateDebate forum. 26
4.9 Jacquard overlap of top dependency paths (by frequency) between For-against and Perspective debates for Fallacy of Relevance and Faulty Generalization. 26

List of Tables

3.1 Basic statistics of our collected CreateDebate dataset. The first post in our dataset was posted on February 20, 2008 and the last post was updated on November 24, 2021 (Patel et al., 2023). 10
3.2 Distribution of For-Against debates and Perspective debates across different forums. 11
3.3 Types of logical fallacies along with their composition in the dataset, their descriptions and examples (Jin et al., 2022). 14

4.1 Hate categories with example of hate targets (Mondal et al., 2017). 23
4.2 Dependency paths from the tree structure illustrated in Figure 4.7 25
4.3 Most discriminative paths for logical fallacies. 30

Abbreviations

General notations

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

Natural language processing

NLP	**N**atural **L**anguage **P**rocessing
BERT	**B**idirectional **E**ncoder **R**epresentations from **T**ransformers

Dependency Parsing

ADJ	**Adj**ective
amod	**A**djectival **Mod**ifier
nsubj	**N**ominal **Subj**ect
dobj	**D**irect **Obj**ect

Statistics

KL divergence	**K**ullback–**L**eibler divergence

Symbols

G	Directed graph representing user network
V	Set of nodes in G
E	Set of edges in G
u, v	Sample nodes in G
$\mathcal{T}$	Tree-structure obtained after dependency parsing
$\mathcal{N}$	Set of nodes in $\mathcal{T}$ which are identified as nouns or pronouns
$\mathcal{E}$	Set of nodes in $\mathcal{T}$ whose text matches with an Empath seed word or a slur word
P, Q	Dummy probability distribution
$\mathcal{C}_i$	Class label for logical fallacies
$\mathcal{P}_i$	Probability distribution of dependency paths for $\mathcal{C}_i$
$\mathcal{U}$	Universal probability distribution of dependency paths
$\mathcal{S}_i$	Scoring function used in our framework
$\mathcal{A}, \mathcal{B}, \mathcal{C}$	Placeholders (masks) used in the logical form of a fallacy

1

Introduction

Opinions are a fundamental aspect of human nature, and people have them about almost everything. Some opinions are based on personal experiences, beliefs, values, and knowledge, while others may be influenced by cultural, societal, and environmental factors. However, it is also important to note that not all opinions are created equal. Some opinions may be based on misinformation, ignorance, or prejudice, and can lead to harmful or negative consequences. It is essential to critically evaluate opinions and consider evidence-based information before accepting or rejecting them. Online forums and social media platforms provide an easy and accessible way for people to share their opinions and engage in debates with others. However, the nature of online communication can also make it easier for deceptive arguments to be presented, often disguised as logical or rational arguments (Kennedy, 1993). Deceptive arguments can be used to manipulate or influence others' opinions and can lead to misunderstandings and conflicts. To avoid falling prey to deceptive arguments, it is important to develop critical thinking skills, evaluate sources of information carefully, and look for evidence-based arguments. It is also essential to engage in respectful and constructive discussions with others, listening to their viewpoints, and being open to changing our opinions based on new evidence.

The term "argument" here does not refer to a bitter dispute or heated exchange. An argument is a group of statements, one or more of which, the premises, support or provide evidence for another, the conclusion. The premises of an argument are those statements that together constitute the reasons for believing the conclusion to be true. Some premises are conclusions of previous arguments, while others may be statements of fact, personal observations, expert testimony, or expressions of common knowledge. Premises may also be found in the form of definitions, principles, or rules, which, together with other premises, are used in an attempt to support the truth of the conclusion (Damer, 2008).

Argumentation plays a critical part in our lives as it helps us make decisions and reason about the the world around us. Sanders et al. (1994) have shown that learning how to argue increases the ability to identify weak arguments and decreases the tendency to use verbal aggressiveness. For evaluation, the argument is reconstructed into what is called a standard logical form. Whether this extraction of the argument from its original context is done mentally or in writing, it is an important part of the process of effectively evaluating the argument.

A standard format that exhibits the logical structure of an argument is as follows:

```
Since (premise),
which is a conclusion supported by (subpremise),
and (premise),
which is a conclusion supported by (subpremise),
and (premise),
[and (implicit premise)]
and (rebuttal premise),
Therefore, (conclusion) (Damer, 2008).
```

Blair (2006) argue that any complex argument can be reconstructed to make this logical structure explicit. There is a very clear difference between an argument and a good argument. Making a claim supported by at least one other claim does create an argument, but the quality of the argument depends on the strength of the supporting evidence and the reasoning used to connect the claims. A good argument is one that presents strong and relevant evidence to support the claim and uses logical reasoning to connect the claims in a clear and convincing way. A good argument also takes into account counterarguments and acknowledges any weaknesses or limitations in the evidence. On the other hand, a weak argument may present insufficient or irrelevant evidence, use flawed or illogical reasoning, or ignore counterarguments and limitations. A weak argument may fail to persuade or convince others and may be subject to criticism or rejection. There are five criteria of a good argument. It must have—a well-formed structure, premises that are relevant to the truth of the conclusion, premises that are acceptable to a reasonable person, premises that together constitute sufficient grounds for the truth of the conclusion, and premises that provide an effective rebuttal to all anticipated criticisms of the argument. An argument that meets all of these conditions is a good one, and its conclusion should be accepted. If an argument fails to satisfy these conditions, it is probably not a good argument.

Human reasoning is often affected by cognitive biases and logical fallacies that can lead to flawed or inaccurate conclusions. Logical fallacies are errors in reasoning that can occur in arguments, often leading to misleading or false conclusions. A fallacy is a violation of one of the criteria of a good argument. Fallacies are weak arguments that seem convincing, however, their evidence does not prove or disprove the argument's conclusion. Fallacies are mistakes in reasoning that typically do not seem to be mistakes. Indeed, the word "fallacy" comes from "to deceive" or "deceitful" in Latin and Old French. Fallacious arguments usually have the deceptive appearance of being good arguments. Some common examples of logical fallacies include ad hominem attacks, strawman arguments, false dilemma, circular reasoning, and appeals to emotion. These fallacies can occur in everyday conversations, debates, and even in media reports. In daily life, fallacious arguments can be as harmless as "All tall people like cheese" (faulty generalization) or "She is the best because she is better than anyone else" (circular claim). However, logical fallacies are also intentionally used to spread misinformation, for instance "Today is so cold, so I don't believe in global warming" (faulty generalization) or "Global warming doesn't exist because the earth is not getting warmer" (circular claim) (Jin et al., 2022).

Recently, there has been substantial research on logical fallacy detection within user-generated content posted on online platforms (Patel et al., 2023; Habernal et al., 2018b; Sahai et al., 2021; Jin et al., 2022). Detecting logical fallacies is a challenging problem because it requires the model to recognize patterns of reasoning that may not be immediately apparent. Logical fallacies can be subtle, and they can be difficult to identify without a deep understanding of the context and the underlying principles of logic.(Blair, 2006). Moreover, different types of logical fallacies require different types of reasoning and knowledge to detect them. For example, detecting a strawman fallacy may require understanding the opponent's argument and recognizing how it differs from the distorted version presented in the fallacious argument. Detecting a false dilemma fallacy may require identifying other options or possibilities that have been ignored or overlooked. To detect logical fallacies, machine learning models must be trained on large datasets of examples of fallacious arguments and non-fallacious arguments. This requires a thorough understanding of the different types of logical fallacies and the ability to recognize them in various contexts. While significant progress has been made in developing machine learning models to detect logical fallacies, it remains an ongoing area of research, and there is still much to be learned about how to create accurate and effective models for detecting logical fallacies in different contexts.

This work is focused on examining logical fallacies that occur in real-world contexts using a data-driven approach using CreateDebate[1], an online discussion forum as an experimental testbed. We refer the reader to (Patel et al., 2023) for additional details about this dataset. We used a structure-aware transformer architecture (Jin et al., 2022) for logical fallacy detection on the CreateDebate dataset. Since CreateDebate facilitates two types of debates, namely For-Against debates and Perspective debates, we also examined the network properties of the user network for each type of debate. We also conducted an analysis of the linguistic structure of personal attacks through dependency parsing. Our findings revealed that each logical fallacy can be distinguished by its distinct use of links in the dependency graph of the text.

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

2

Prior Works

Aristotle first identified that some arguments are indeed *deceptions in disguise* (Kennedy, 1993). When evaluating new ideas or attempting to resolve differences of opinion, humans often use argumentation. An argument consists of a claim or conclusion that needs to be validated, premises or evidence that support the claim, and an inference relation between the evidence and conclusion that either validates or disproves the conclusion. A fallacy, on the other hand, is a flawed argument where either the inference relation or the premises are incorrect.

A fallacy is a violation of one of the criteria of a good argument. Fallacies, then, stem from one or more of the following:

- A structural flaw in the argument
- A premise that is irrelevant to the conclusion

- A premise that fails to meets the standards of acceptability
- A set of premises that together are insufficient to establish the argument's conclusion
- A failure to give an effective rebuttal to the anticipated criticisms of the argument

An argument that does not meet one or more of these criteria is considered fallacious. Arguments that fail to convince others to reach the intended conclusion not only fail to meet one or more of the five criteria of a good argument, but they may also fail to meet a criterion in a variety of ways that share similar characteristics with other violations of the same criterion. These common violations have been given specific names due to their prevalence, and logicians have organized these fallacies into categories based on shared properties of the mistakes (Damer, 2008).

Fallacies can be classified into two categories: formal and informal. Formal fallacies are invalid logical formulas that can be easily represented, such as *denying the antecedent*, which is an incorrect application of modus tollens. In contrast, informal fallacies are easier to understand and describe without resorting to logical representations, although many can still be represented as invalid arguments (Hansen, 2015).

An initial effort for creating an extensive dataset of fallacies was made in Habernal et al. (2017). A platform for educational games was developed by the authors, aimed at improving players' debating skills. Players can earn points by using valid arguments and adding new fallacies to the platform while attempting to deceive other participants with invalid arguments. A subsequent study (Habernal et al., 2018a) reported a dataset of approximately 300 arguments generated through the platform, indicating the necessity of exploring alternative approaches for constructing more extensive datasets of fallacies.

The issue of ad hominem fallacies in conversations was tackled in Habernal et al. (2018b). The authors utilized the subreddit *ChangeMyView*[1], which is a forum for polite discussions, where users can share their opinions even if they may be flawed, in order to obtain a better understanding of other perspectives on the issue. The fallacy dataset comprises of comments that were removed by the moderators for violating the forum's rule of being respectful and non-hostile, thereby committing an ad hominem fallacy. Patel et al. (2023)

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

measured the prevalence of ad hominem in the CreateDebate forum. They found that the amount of ad hominem in recent times increased manyfold, more than the figures hinted in any of the previous works. Their analysis revealed for the first time, the extremely worrying prevalence of ad hominem in the wild—one-third of the posts in the CreateDebate forum were ad hominems and a small cohort of highly active users hurl the largest number of ad hominems. Quite interestingly, hurling ad hominems accelerated at time periods closer to the 2016 US Presidential election. Overall, political debates are found to be at the core of increased ad hominem usage with its effects transcending to other topics like religion, science and law.

A methodology was designed by Sahai et al. (2021) for aligning informal fallacies mentioned on Reddit within the pragma-dialectic theory of argumentation and for mining and labeling fallacies in online discussions. It was found that the additional conversational context plays an important role in predicting fallacious arguments. A dataset of $2,449$ samples of 13 logical fallacy types was collected, and extensive experiments were conducted using 12 existing language models. In addition, a structure-aware classifier was designed, which outperformed the best language model in the task of logical fallacy detection (Jin et al., 2022).

3
Data Collection

3.1 CreateDebate Forum

CreateDebate is a website for social networking debates that has been in operation since 2008. Its primary goal is to aid groups of individuals in navigating topics, perspectives, and beliefs by providing a platform centered on concepts, discourse, and democratic principles. At CreateDebate, conversations frequently center around striving for agreement and comprehension in order to arrive at more informed and effective conclusions.

On CreateDebate, users can create posts and moderate the content themselves. This platform allows users to express their opinions and perspectives on a variety of topics, which are organized into 14 different forums, such as Politics, Entertainment, Science, etc. The majority of the content on CreateDebate is public, and users can interact with each other's posts by writing comments, supporting or disputing them, or providing clarification. However, due to the weak moderation on the site, there may be instances of logical fallacies in the discussions. This makes CreateDebate a valuable tool for investigating the prevalence of logical fallacies over time. By analyzing the conversations and discourse on CreateDebate, researchers can gain insight into how logical fallacies are used and how they can be addressed to improve the quality of online discourse.

For our study, we utilized an automated approach to gather the entire publicly accessible CreateDebate dataset, spanning across all the different topics on the platform since its launch. However, to keep our presentation concise and focused, we will only showcase the key findings from the most popular forums, namely Politics, Religion, World News, Science, Law, and Technology. It's worth noting that the outcomes from the remaining topical forums remained consistent with the presented results.

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 3.1: Basic statistics of our collected CreateDebate dataset. The first post in our dataset was posted on February 20, 2008 and the last post was updated on November 24, 2021 (Patel et al., 2023).

We present the general statistics of the dataset in Table 3.1. Overall, the six forums comprise a total of 18, 227 posts and 264,814 comments posted by 14,961 individual users over a period of 14 years. As part of our research, we confirmed that all of the posts included in our analysis were written in the English language. Leveraging this vast dataset of online discussions that spanned over several years, we employed logical fallacy detection techniques on the posts and comments on the CreateDebate platform. This approach can provide valuable insights into how online debates and discussions can turn aggressive and hurtful, affecting the overall quality of discourse on online platforms.

3.1.1 For-against vs. Perspective Debates

CreateDebate hosts two distinct types of debates, namely For-Against debates and Perspective debates. These debates allow users to express their views and opinions on various topics and engage with other users who may hold different perspectives. In For-Against debates, users take a stand on a particular topic and argue in favor or against it. On the other hand, Perspective debates enable users to share their viewpoints on a topic without necessarily taking a stance for or against it. By providing these different debate formats, CreateDebate seeks to foster a healthy and inclusive discussion environment where users can exchange ideas and learn from each other.

Topic	**For-Against Posts %**	**Perspective Posts %**	**For-Against Comments %**	**Perspective Comments %**
Politics	48%	52%	68%	32%
Religion	60%	40%	76%	24%
World News	58%	42%	73%	27%
Science	55%	45%	74%	26%
Law	66%	34%	82%	18%
Technology	54%	46%	79%	21%

Table 3.2: Distribution of For-Against debates and Perspective debates across different forums.

From Table 3.2, it is noticeable that the majority of topical forums on CreateDebate exhibit a fairly even distribution of both For-Against and Perspective type posts. However, it appears that For-Against debates tend to attract a higher number of comments compared to Perspective debates. This indicates that users are more likely to engage in discussions where there is a clear stance taken on a particular topic.

3.2 LOGIC Dataset

Fallacy	**Description**	**Example**
Faulty Generalization (18.01%)	An informal fallacy wherein a conclusion is drawn about all or many instances of a phenomenon on the basis of one or a few instances of that phenomenon. is an example of jumping to conclusions.	"I met a tall man who loved to eat cheese. Now I believe that all tall people like cheese."
Ad Hominem (12.33%)	An irrelevant attack towards the person or some aspect of the person who is making the argument, instead of addressing the argument or position directly.	"What can our new math teacher know? Have you seen how fat she is?"
Ad Populum (9.47%)	A fallacious argument which is based on affirming that something is real or better because the majority thinks so.	"Everyone should like coffee: 95% of teachers do!"
False Causality (8.82%)	A statement that jumps to a conclusion implying a causal relationship without supporting evidence.	"Every time I wash my car, it rains. Me washing my car has a definite effect on the weather."
Circular Claim (6.98%)	A fallacy where the end of an argument comes back to the beginning without having proven itself.	"J.K. Rowling is a wonderful writer because she writes so well."

Appeal to Emotion (6.82%)	Manipulation of the recipient's emotions in order to win an argument.	"It is an outrage that the school wants to remove the vending machines. This is taking our freedom away!"
Fallacy of Relevance (6.61%)	Also known as red herring, this fallacy occurs when the speaker attempts to divert attention from the primary argument by offering a point that does not suffice as counterpoint/supporting evidence (even if it is true).	"Why are you worried about poverty? Look how many children we abort every day."
Deductive Fallacy (6.21%)	An error in the logical structure of an argument.	"It is possible to fake the moon landing through special effects. Therefore, the moon landing was a fake using special effects."
Intentional Fallacy (5.84%)	A custom category for when an argument has some element that shows intent of a speaker to win an argument without actual supporting evidence.	"No one has ever been able to prove that extraterrestrials exist, so they must not be real."
Fallacy of Extension (5.76%)	An arugment that attacks an exaggerated or caricatured version of your opponent's position.	"Their support of the discussion of sexual orientation issues is dangerous: they advocate for the exposure of children to sexually explicit materials, which is wrong."

False Dilemma (5.76%)	A claim presenting only two options or sides when there are many options or sides.	“You’re either for the war or against the troops.”
Fallacy of Credibility (5.39%)	An appeal is made to some form of ethics, authority, or credibility.	“My professor, who has a Ph.D. in Astronomy, once told me that ghosts are real. Therefore, ghosts are real.”
Equivocation (2.00%)	An argument which uses a key term or phrase in an ambiguous way, with one meaning in one portion of the argument and then another meaning in another portion of the argument.	“I don’t see how you can say you’re an ethical person. It’s so hard to get you to do anything; your work ethic is so bad”

Table 3.3: Types of logical fallacies along with their composition in the dataset, their descriptions and examples (Jin et al., 2022).

The LOGIC dataset is a collection of typical examples of logical fallacies sourced from diverse online educational resources that are designed to either teach or evaluate students’ comprehension of logical fallacies. These resources may include websites, textbooks, and other materials that deal with the subject matter. The examples were carefully selected based on their relevance and frequency of occurrence in various contexts. The dataset was created to aid in the development of machine learning models for the automatic detection of logical fallacies. In total, the LOGIC dataset comprises 2,449 examples of logical fallacies that are categorized into 13 different types. For more information regarding LOGIC dataset, we refer the reader to (Jin et al., 2022).

3.3 Empath Dictionary

Fast et al. (2016) introduced *Empath*, which allows researchers to understand the emotional and semantic content of large-scale text data. Empath works by analyzing text data at the level of individual words and phrases, and grouping them into categories based on their emotional and semantic associations. These categories are based on a large set of human-labeled seed words, which are used to train a machine learning model to identify similar words and phrases in text data.

The authors demonstrate the effectiveness of Empath in a number of experiments, including analyzing the emotional content of tweets, detecting patterns in online discussions, and predicting the success of startup pitches based on the language used in pitch videos. They argue that Empath is a more flexible and accurate tool than existing techniques, as it allows researchers to create custom categories of words and phrases based on their specific research questions.

Empath has 194 built-in pre-validated topical and emotional categories. Each category contains an average of 83 seed words. To focus on detecting logical fallacies, we selected 61 relevant categories and also included a high-quality dictionary of slur words[1] that we created in-house.

[1] A slur word is a derogatory or insulting term used to refer to a person or group of people based on their race, ethnicity, gender, sexual orientation, religion, or other personal characteristics. Slur words can be very offensive and hurtful to the targeted individuals or communities. Many societies and cultures have social norms and laws against using slur words as they can cause harm and perpetuate discrimination and prejudice.

4

Experiments

The experimentation phase is comprised of three distinct components. Initially, we use the LOGIC dataset to fine-tune the BERT [base, uncased] model (Devlin et al., 2019) and then apply it to the CreateDebate dataset for inference. This allows us to categorize each comment in CreateDebate into one of the logical fallacy types. Following that, we examine the user network of the CreateDebate forum to investigate the differences in user behavior between For-against debates and Perspective debates. Lastly, we conduct a comprehensive linguistic analysis for each logical fallacy type.

4.1 Logical Fallacy Detection in CreateDebate Forum

The initial experimental setup involves an analysis of logical fallacies in various topical forums in CreateDebate. For the purposes of our experiment, we use the LOGIC dataset

(Section §3.2) for fine-tuning BERT [base, uncased] model. The fine-tuned model is then used to classify every comment across different topical forums in CreateDebate.

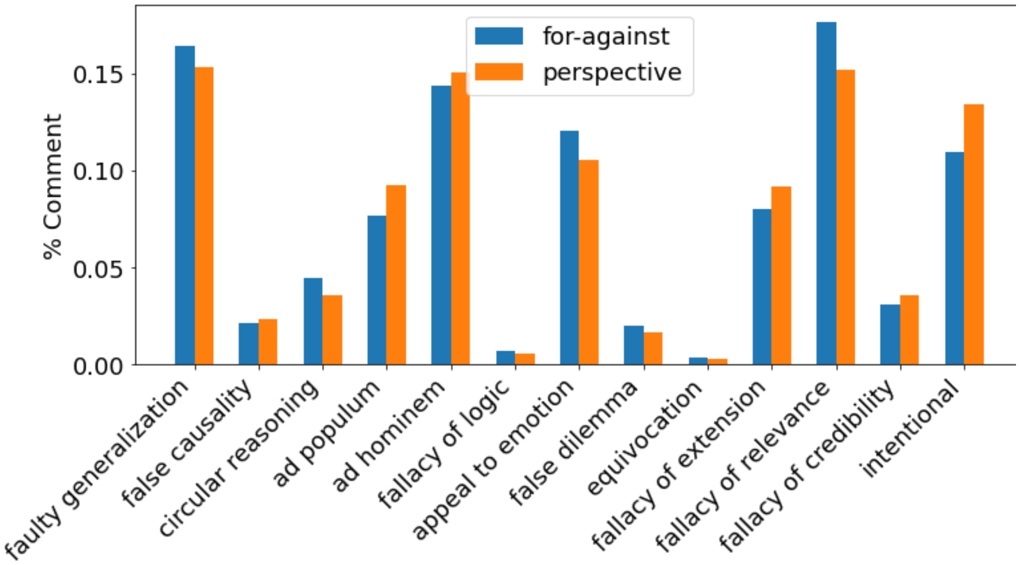

FIGURE 4.1: Distribution of Logical Fallacies in the Politics Forum of CreateDebate.

The distribution of comments across various logical fallacy types in CreateDebate's Politics forum is displayed in Figure 4.1. The figure suggests that the distribution of comments is relatively even between For-against and Perspective debates. In the interest of brevity and clarity, we shall narrow our focus to the five most prevalent logical fallacies—Faulty generalization, ad hominem, fallacy of relevance, intentional fallacy, and appeal to emotion.

In the pursuit of knowledge, we may often stray
Towards the complex, with methods that seem to sway
But in the spirit of Occam's Razor, we stay
Simple, and let comment count guide our way.

4.2 Study on CreateDebate's User Network

Our initial step involved constructing directed graphs for both For-against and Perspective debates. In the directed graphs constructed for the study, the users are represented as nodes, while the interactions among them are indicated by the edges. These edges are assigned a weight corresponding to the frequency of direct replies by, say, user A to a post or comment authored by user B. If a user has posted a comment or a post in either for-against or perspective debates, then they will be included in the respective directed graphs for these debates. Otherwise, they will not be included.

CreateDebate provides profile pages for its users, which contain information such as the user's reward points, efficiency in debating, number of debates participated in, number of comments posted, and their activity on the forum (i.e., when they joined and when they were last online). Additionally, these pages allow us to view the user's allies, enemies, and hostiles. We can utilize these user pages as an alternative approach to construct directed graphs for analyzing network patterns. For this study, we will be using both of the approaches while performing network studies.

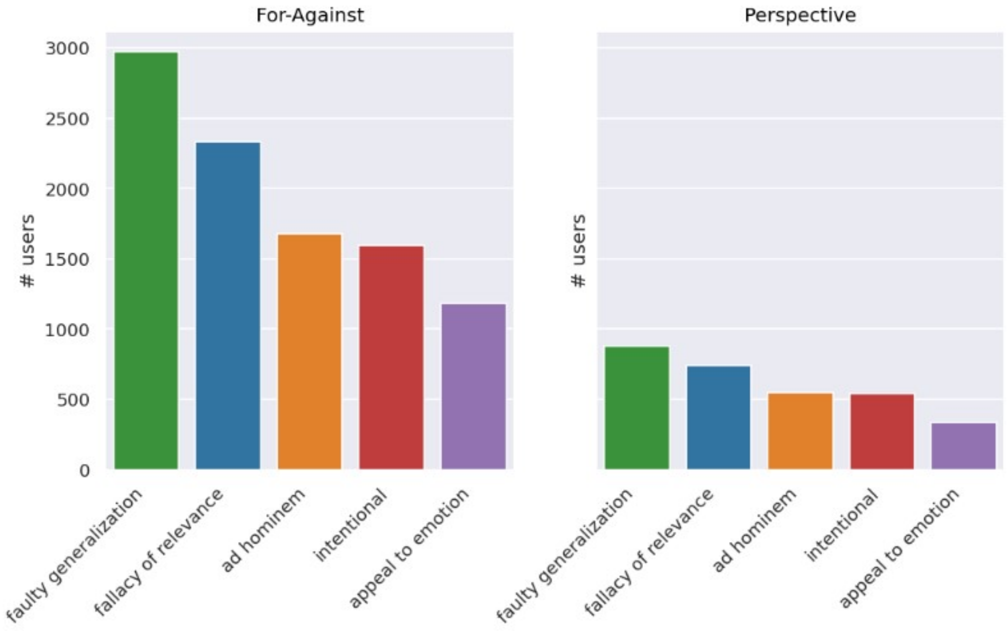

Figure 4.2: Count of fallacious users across for-against and perspective debates.

Upon closer observation of Figure 4.2, it becomes apparent that the number of users who participate in for-against debates is nearly three times more than the number of users involved in perspective debates. This difference in participation suggests a greater level of engagement and interest in for-against debates as compared to perspective debates, which could be attributed to a variety of factors such as personal preferences, beliefs, or the nature of the topic being discussed.

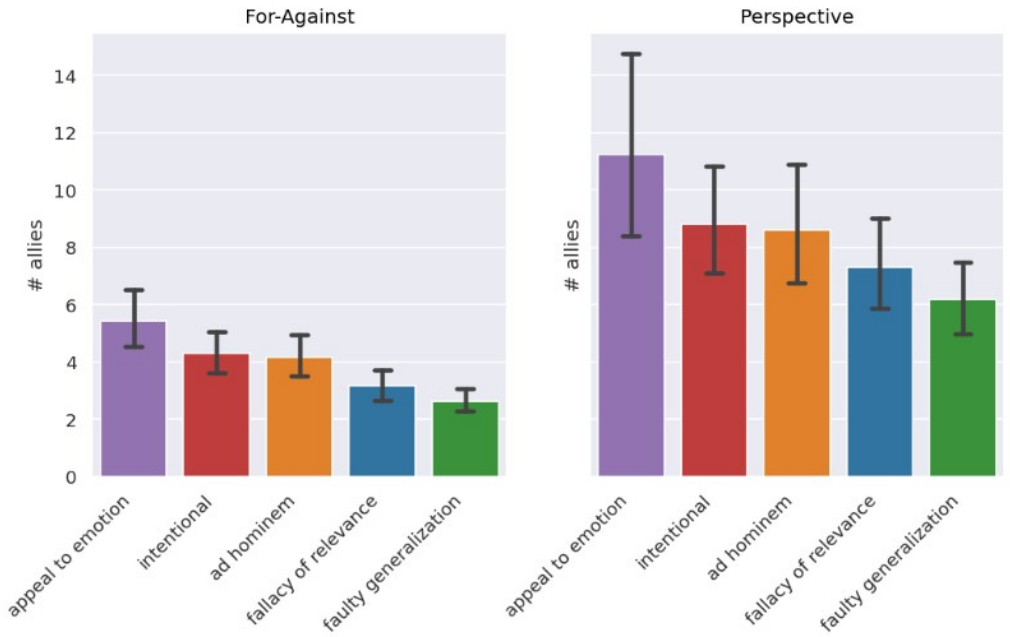

FIGURE 4.3: The average number of *allies* a user has when they post a logically fallacious comment in both for-against and perspective debates.

What's fascinating in Figure 4.3 is that even though there are fewer users posting fallacious comments in perspective debates, they have a larger number of allies in the overall forum network. This implies that these communities are closely interconnected and cooperative.

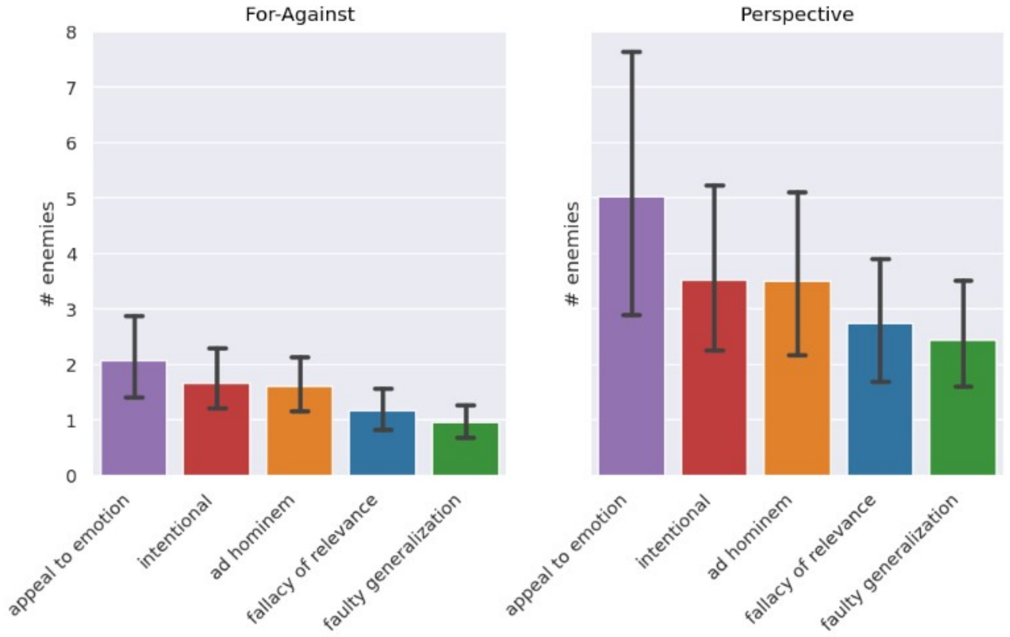

FIGURE 4.4: The average number of *enemies* a user has when they post a logically fallacious comment in both for-against and perspective debates.

Figure 4.4 suggests that the fallacious users in perspective debates have a higher number of enemies, suggesting that their comments are more likely to be challenged by other users. However, these users also have a larger number of allies, indicating a close-knit and supportive community. The discrepancy in the number of allies and enemies between for-against and perspective debates underscores the distinct dynamics of each type of debate.

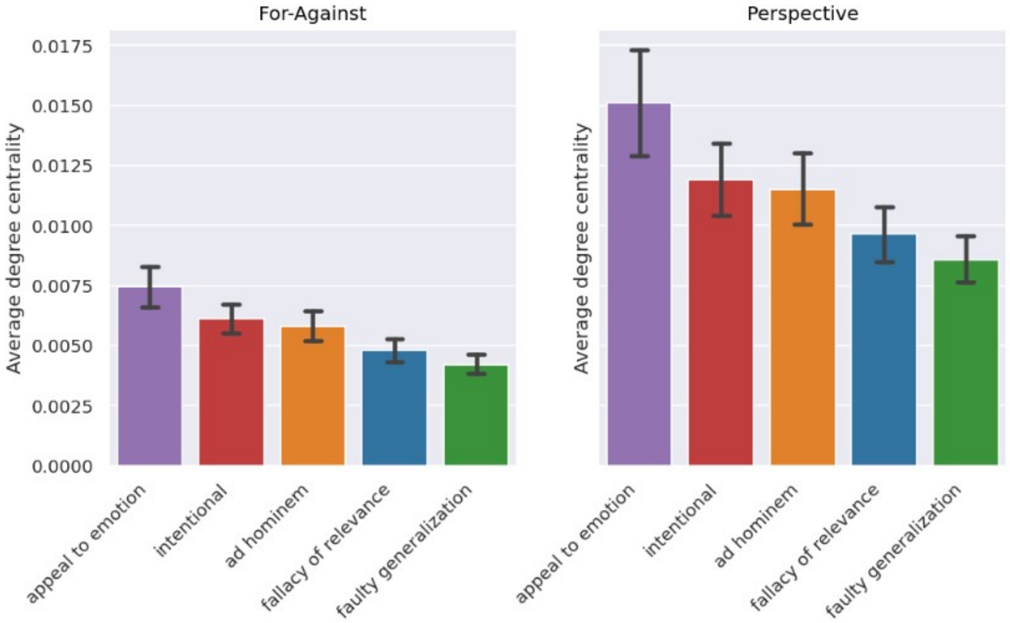

Figure 4.5: Average degree centrality of fallacious users across for-against and perspective debates.

From Figure 4.5, it can be inferred that the users who make fallacious comments in perspective debates have a higher degree centrality[1] in the overall network when compared to those in for-against debates. This indicates that despite there being fewer users posting fallacious comments in perspective debates, they have more connections with other users in the forum. This higher degree centrality could be due to several reasons. It's possible that these users have been active in the forum for a long time, or they have built a strong rapport with other users. It could also be because the community around perspective debates is tightly knit, as we saw earlier. Overall, this figure highlights an interesting pattern in the network of the CreateDebate forum, which could be useful in understanding the behavior of users and the dynamics of different types of debates.

[1] For a graph $G \equiv (V, E)$, the degree centrality of a node $v \in V$ is the number of its direct neighbors, normalized by $|E| - 1$

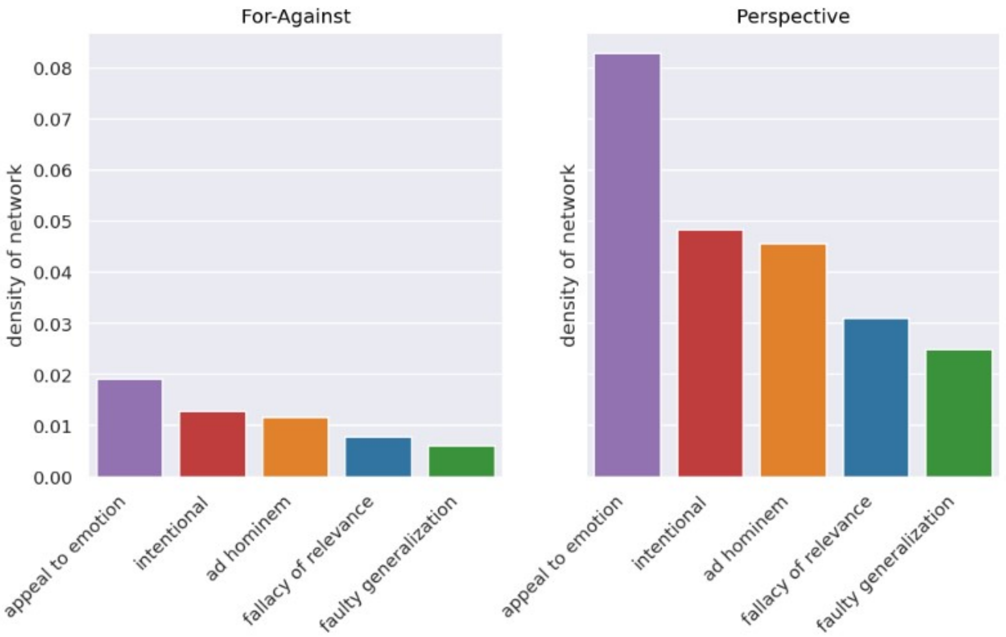

Figure 4.6: Network density for for-against and perspective user network.

Figure 4.6 suggests that in the perspective debate network, there are more connections or relationships between users, while in the for-against debate network, there are relatively fewer connections. The density[2] of the network is measured by the number of connections between users divided by the maximum possible number of connections. A denser network suggests a higher level of interaction and communication among users. The difference in network density between the two types of debates may indicate differences in the type and level of engagement, as well as the social dynamics within each debate.

To summarize, CreateDebate forum has almost three times more fallacious users engaged in for-against debates than in perspective debates, indicating a greater interest in the former type. However, even though there are fewer fallacious comments in perspective debates, users making these comments have more allies and enemies, indicating a close-knit and supportive community that challenges these comments. Fallacious users in perspective debates also have a higher degree centrality in the overall network than those in for-against debates, which could be attributed to various reasons such as their rapport with other users or a tightly knit community. The network of the perspective debate is denser than for-against debates, indicating more communication and interaction among users. This

[2] For a graph $G \equiv (V, E)$, its density is given as $\frac{|E|}{|V|(|V|-1)}$.

disparity in network density may suggest differences in the level and type of engagement and social dynamics between the two debate types.

4.3 Comprehensive Linguistic Study of Logical Fallacies in CreateDebate

Mondal et al. (2017) created a set of categories for hate speech with examples of targeted groups in order to gain a better understanding of how hate speech is expressed on social media sites such as Twitter and Whisper. By categorizing and identifying specific examples of hate speech, the authors were able to quantify the extent to which hate speech is present on these platforms, which could help inform strategies for combating this issue. The paper provides a comprehensive analysis of the prevalence and characteristics of hate speech on these two platforms, shedding light on the need for continued efforts to address this problem in social media.

Categories	**Example of hate targets**
Race	nigga, nigger, black people, white people
Behavior	insecure people, slow people, sensitive people
Physical	obese people, short people, beautiful people
Sexual orientation	gay people, straight people
Class	ghetto people, rich people
Gender	pregnant people, cunt, sexist people
Ethnicity	chinese people, indian people, paki
Disability	retard, bipolar people
Religion	religious people, jewish people
Other	drunk people, shallow people

Table 4.1: Hate categories with example of hate targets (Mondal et al., 2017).

Using the hate categories and corresponding hate targets listed in Table 4.1, we filtered 61 topical and emotional categories relevant for logical fallacy detection from the Empath dictionary (Section §3.3).

4.3.1 Dependency Parsing

Dependency parsing is an important natural language processing technique used to analyze the grammatical structure of a sentence and determine the relationships between words.

In this technique, each word in a sentence is identified and then analyzed to understand how it relates to other words in the sentence. This analysis creates a tree-like structure that represents the grammatical relationships between words, with the main verb in the sentence typically at the root of the tree. Dependency parsing has a wide range of applications in NLP, including machine translation, sentiment analysis, and named entity recognition. Additionally, this technique can be used to extract valuable information from text data, such as relationships between entities or events described in a document. This is achieved by analyzing the grammatical structure of the text and identifying patterns and relationships between words, which can help in understanding the meaning of the text.

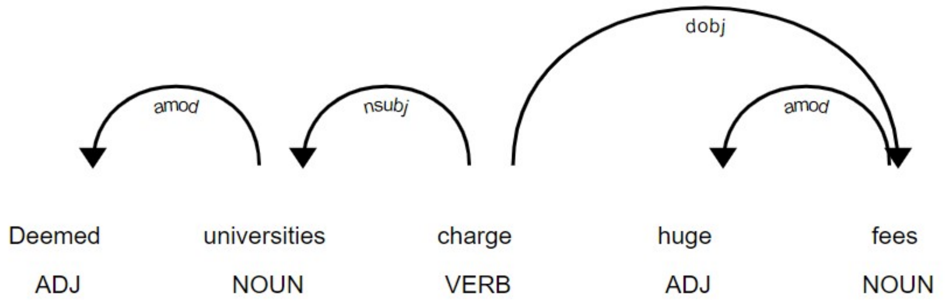

Figure 4.7: Example of a tree structure that is formed by applying dependency parsing on a sample text. In this example, the verb *charge* is identified as the root of the tree, and *universities* is the nominal subject of the verb *charge*. Additionally, *fees* is identified as the direct object of the verb *charge*. The adjectives *deemed* and *huge* are also identified as the modifiers of *universities* and *fees*, respectively.

To conduct our experiment, we utilize dependency parsing on every comment that has been posted on the Politics forum of the CreateDebate website. The process of lemmatization is then applied to the texts present in every node of the tree-structure. Lemmatization is a natural language processing technique that involves reducing words to their base or root form, which is known as the lemma. For example, the lemma of the word “running” is “run”, the lemma of “jumped” is “jump”, and so on. This technique helps to group together different forms of the same word and reduces the complexity of the text. This makes it easier to analyze and extract meaningful information from the text data. Dependency parsing helps us analyze the grammatical structure of each sentence, and we store the resulting tree structure to further explore and study the language used in the comments.

By parsing the comments, we can extract useful information about the relationships between different words, which can help us understand the sentiment and themes present in the political discussions on CreateDebate. This analysis of the comments' grammatical structure can aid in identifying patterns of behavior and attitudes towards political topics.

4.3.2 Parsing Tree-structures

For a given tree-structure $\mathcal{T}$ obtained after dependency parsing, let $\mathcal{N}$ denote the set of nodes which are identified as nouns or pronouns, and let $\mathcal{E}$ denote the set of nodes whose text matches with an Empath seed word or a slur word after lemmatization. We traverse the tree and generate dependency path between two nodes u and $v, \forall u \in \mathcal{N}, \forall v \in \mathcal{E}$, if the path exists. For example, in Figure 4.7, let's assume that the word *huge* belongs to Empath set $\mathcal{E}$, then we would get two dependency paths:

$\mathcal{N}$	$\mathcal{E}$	**Dependency Path**
fees	*huge*	$\textit{fees} \xleftarrow{amod} \textit{huge}$
universities	*huge*	$\textit{universities} \xrightarrow{nsubj} \textit{charge} \xleftarrow{dobj} \textit{fees} \xleftarrow{amod} \textit{huge}$

Table 4.2: Dependency paths from the tree structure illustrated in Figure 4.7

For the purposes of our experiment, a greater emphasis is placed on the interdependence between two words in a given path, as opposed to the individual words themselves.

Following the above steps, for every comment in the Politics forum, we parse its tree structure obtained after dependency parsing and generate a list of dependency paths associated with that comment.

4.3.3 Identifying Most Discriminative Dependency Paths for a Given Logical Fallacy

In the preceding section, we obtained all feasible dependency paths, with certain restrictions, from the dependency tree for each comment. Nevertheless, not all dependency paths are specific to a particular logical fallacy category. Figure 4.9 demonstrates that there is significant overlap of dependency paths among various categories.

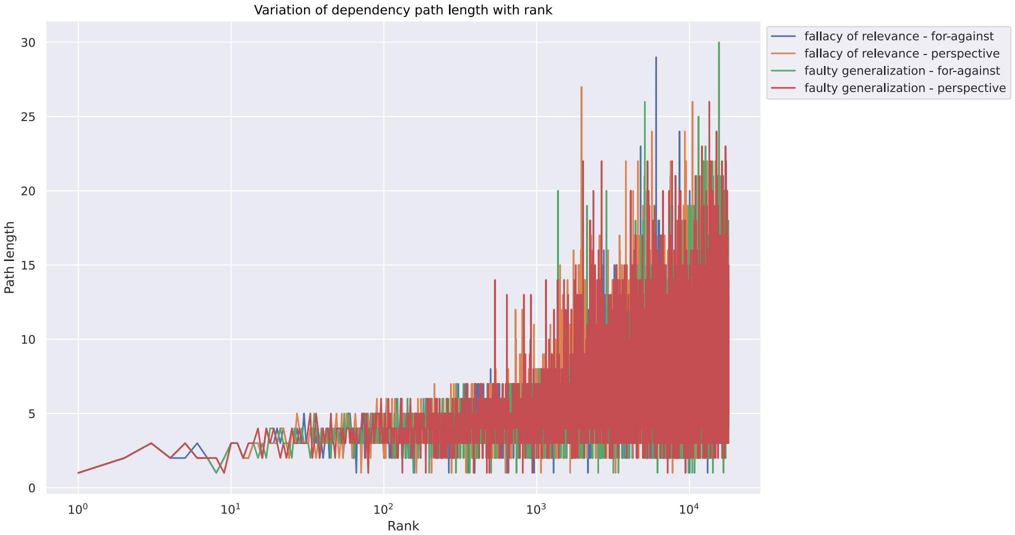

FIGURE 4.8: Variation in dependency path length by popularity. Shorter dependency paths are much more frequent in the Politics CreateDebate forum.

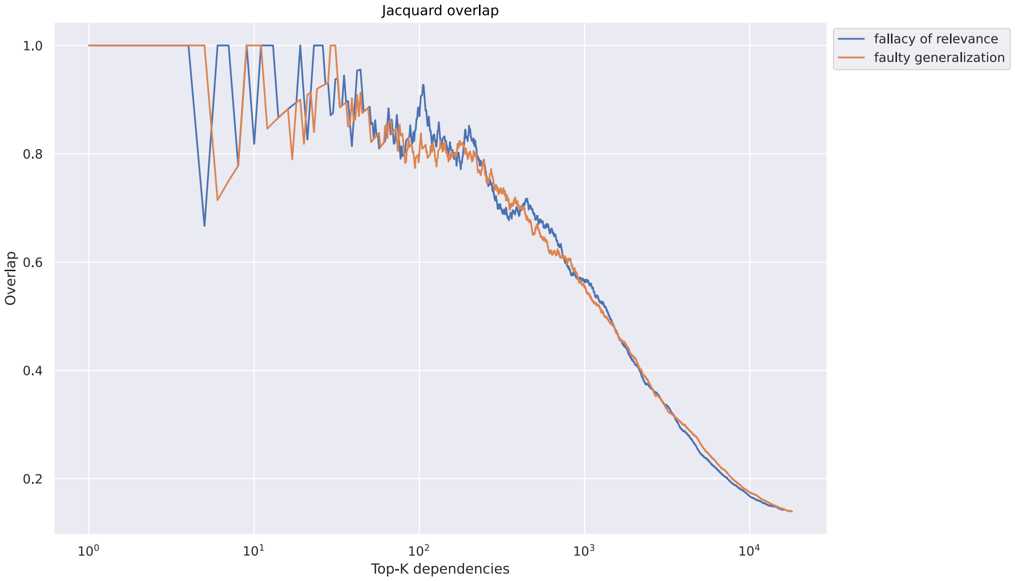

FIGURE 4.9: Jacquard overlap of top dependency paths (by frequency) between For-against and Perspective debates for Fallacy of Relevance and Faulty Generalization.

Certain types of dependencies can be more unique to a given logical fallacy because they capture the underlying syntactic and semantic relationships between words that are specific to that fallacy. For example, in the case of the "ad hominem" fallacy, the most discriminating dependencies might be those that involve a noun or pronoun referring to the opponent or the speaker themselves, as these are often the targets of personal attacks in ad hominem arguments. Similarly, in the case of the "straw man" fallacy, the most discriminating dependencies might involve a verb that indicates misrepresentation or distortion of an opponent's argument, as this is a key characteristic of straw man arguments.

By focusing on the dependencies that are most unique to a given fallacy, we can develop more accurate and effective models for identifying and classifying instances of that fallacy in natural language text. This can be particularly useful in applications such as automated fact-checking, where it is important to identify and flag instances of logical fallacies in news articles or social media posts.

Therefore, we have created a framework inspired by KL divergence to identify the most distinctive dependency paths of a specific category.

KL divergence is commonly used to compare two probability distributions, P and Q. It measures the amount of additional information needed to encode samples from one distribution using a code optimized for another distribution.

$$KL(P||Q) = \sum_x s(x) = \sum_x P(x) \log\left(\frac{P(x)}{Q(x)}\right)$$

In the above equation, higher (more positive) the value of $s(x)$, more discramative x is for the distribution P when compared to Q.

For the purposes of this experiment, let there be n logical fallacy classes and all the classes be mapped to a unique non-negative integer. For example, *faulty generalization* can be class $\mathcal{C}_0$, *fallacy of relevance* can be class $\mathcal{C}_1$ and so on. Let $\mathcal{P}_i$ denote the probability distribution of dependency paths in for class $\mathcal{C}_i$. We also define an universal probability distribution $\mathcal{U}$, given as,

$$\mathcal{U}(x) = \frac{1}{n}\sum_{i}\mathcal{P}_i(x)$$

We then define the score of a dependency path x in the logical fallacy class $\mathcal{C}_i$ as

$$\mathcal{S}_i(x) = \mathcal{P}_i(x)\log\left(\frac{\mathcal{P}_i(x)}{\mathcal{U}(x)}\right)$$

For each logical fallacy class $\mathcal{C}_i$, we select the top-5 dependency paths on the basis of the score $\mathcal{S}_i(x)$.

Fallacy Name	**Logical Form**	**Most Discriminative Dependency Paths**
Fallacy of Relevance	It is claimed that $\mathcal{A}$ implies $\mathcal{B}$, whereas $\mathcal{A}$ is unrelated to $\mathcal{B}$.	1. *nominal subject → clausal complement → nominal subject* 2. *direct object → nominal subject* 3. *direct object → open clausal complement → nominal subject* 4. *object of preposition → prepositional modifier → clausal complement → nominal subject* 5. *nominal subject → clausal complement → clausal complement → nominal subject*

Faulty Generalization	$\mathcal{A}$ has attribute $\mathcal{B}$. $\mathcal{A}$ is a subset of $\mathcal{C}$. Therefore, all $\mathcal{C}$ has the attribute $\mathcal{B}$.	1. *nominal subject → adverbial clause modifier → nominal subject* 2. *nominal subject → clausal complement → nominal subject* 3. *nominal subject → conjunct → nominal subject* 4. *direct object → conjunct → nominal subject* 5. *nominal subject → conjunct → direct object*
Ad Hominem	$\mathcal{A}$ is claiming $\mathcal{B}$. $\mathcal{A}$ is a moron. Therefore, $\mathcal{B}$ is not true.	1. *adjectival modifier → attribute → nominal subject* 2. *attribute → nominal subject* 3. *adjectival complement → nominal subject* 4. *adjectival modifier → direct object →nominal subject* 5. *attribute → clausal complement → nominal subject*

Intentional Fallacy	$\mathcal{A}$ knows $\mathcal{B}$ is incorrect. $\mathcal{A}$ still claim that $\mathcal{B}$ is correct using an incorrect argument.	1. *adjectival complement → nominal subject* 2. *clausal complement → nominal subject* 3. *direct object → nominal subject* 4. *nominal subject → clausal complement → nominal subject* 5. *adjectival complement → clausal complement → nominal subject*
Appeal to Emotion	$\mathcal{A}$ is made without evidence. In place of evidence, emotion is used to convince the interlocutor that $\mathcal{A}$ is true.	1. *nominal subject* 2. *direct object* 3. *adverbial clause modifier → nominal subject* 4. *conjunct → adverbial clause modifier → nominal subject* 5. *conjunct → nominal subject*

Table 4.3: Most discriminative paths for logical fallacies.

Let's analyze the results presented in Table 4.3:

1. **Fallacy of Relevance**: Dependency paths having clausal complement with the nominal subject and the direct object are most unique to fallacy of relevance because this fallacy involves presenting irrelevant information in order to divert attention from the topic at hand. In other words, it is an attempt to change the subject or distract the listener from the main issue.

   The clausal complement with nominal subject and direct object in a dependency path represents a clause that provides additional information about the nominal subject and direct object, respectively. This additional information can be used to support or clarify the argument being made in the sentence. However, in the context of the fallacy of relevance, the additional information provided by the clause is not relevant to the main issue and is used to distract or mislead the listener.

For example, consider the sentence “The company is doing well financially, but the CEO is a great singer.” The clausal complement “the CEO is a great singer” is not relevant to the main issue of the company’s financial performance and is used to distract the listener. Therefore, dependency paths with this type of structure are most unique to the Fallacy of relevance, as they indicate the use of irrelevant information to distract from the main issue.

2. **Faulty (Hasty) Generalization**: Dependency paths with nominal subject and direct object connected by a conjunction are often associated with faulty generalization because they can represent a generalization made without proper consideration of all relevant factors. For example, consider the sentence “I ate a hamburger and it was delicious. Hence, all hamburgers are delicious.” This sentence includes a conjunction (*and*) connecting the subject (*I*) and direct object (*hamburger*) of the verb *ate*. However, the use of the conjunction to connect the subject and object suggests that the speaker is making a generalization based on a single experience, which may not be representative of all hamburgers.

   Similarly, in other sentences, the use of conjunctions with nominal subject and direct object can also signal a hasty generalization. For instance, “I met two doctors and they were both rude. Hence, all doctors are rude.” suggests that the speaker has met only two doctors, and is generalizing about all doctors based on a small sample size.

   Therefore, dependency paths with conjunctions connecting nominal subject and direct object are more common in sentences that are associated with hasty generalizations, and hence they can be used to detect this type of logical fallacy.

3. **Ad Hominem**: In ad hominem fallacy, the argument attacks the person instead of the issue or the argument itself. Hence, this type of fallacy is characterized by the use of language that is insulting or defamatory to the opponent.

Adjectival modifiers and adjectival complements are types of dependencies that provide descriptive information about a noun or a pronoun. They are often used to add adjectives or adjectival phrases to a noun, which help to characterize or qualify the noun.

In ad hominem fallacy, adjectival modifiers and complements are used to insult or defame the opponent, by attaching derogatory or negative adjectives to the noun that refers to the opponent. For example, consider the sentence "We should not listen to John's argument, since he is a dishonest person." Here, the adjective *dishonest* is an adjectival modifier of the noun *person*, and it is used to attack the character of John.

Therefore, the use of adjectival modifiers and complements in a dependency path is an important feature for identifying the ad hominem fallacy.

4. **Intentional Fallacy**: Intentional Fallacy is a type of logical fallacy where the author's intention is used to judge the meaning of a text rather than focusing on the text itself. This type of fallacy often involves making unsupported claims about the author's intent and using that as evidence to support a particular interpretation.

   Adjectival complements and clausal complements are more prominent in intentional fallacy because they provide additional information about the author's intent and can be used to support a particular interpretation of the text. Adjectival complements modify adjectives and provide additional descriptive information, while clausal complements provide additional context and can be used to clarify the meaning of a sentence.

   For example, consider the following sentence: "The author's use of vivid language reveals a desire to manipulate the reader." Here, the adjectival complement *vivid* provides additional information about the type of language used by the author, while the clausal complement "reveals a desire to manipulate the reader" provides insight into the author's intent.

   Another example is the sentence: "The author's decision to include only positive examples in their argument shows a deliberate attempt to mislead the reader." In this sentence, the clausal complement "shows a deliberate attempt to mislead the reader" provides evidence for the claim that the author intentionally omitted negative examples to mislead the reader.

Overall, adjectival complements and clausal complements are prominent in intentional fallacy because they can be used to support claims about the author's intent, which is a key aspect of this type of logical fallacy.

5. **Appeal to Emotion**: The appeal to emotion fallacy involves manipulating emotions to influence beliefs or actions rather than using valid arguments or evidence. Therefore, it makes sense that dependency paths related to emotions are prominent in this fallacy.

   One common way to appeal to emotions is by using vivid language or describing emotional situations. Adverbial clause modifiers, which provide information about time, place, or manner, can be used to add detail to emotionally charged descriptions. For example, in the sentence "The innocent puppy was brutally beaten to death in broad daylight," the adverbial clause modifier "in broad daylight" adds a sense of shock and horror to the statement, appealing to the reader's emotions.

   Additionally, nominal subjects and direct objects can be used to evoke emotions in the reader or listener. For instance, in the sentence "Our brave soldiers risk their lives to protect our freedom," the nominal subject "brave soldiers" appeals to the reader's sense of patriotism and admiration. Similarly, the direct object "freedom" appeals to the reader's emotions of love for their country and appreciation for the sacrifices made by soldiers.

   Overall, these dependency paths can be used to create a strong emotional impact on the reader or listener, making them more susceptible to accepting the argument without critical evaluation.

To provide a summary, the experiments performed on dependency parsing and logical fallacy detection demonstrated that the dependency paths generated by the proposed framework were closely related to the logical structure of the respective fallacy. The framework was able to identify highly discriminating dependency paths which were specific to each type of fallacy, thus providing an alternative method for identifying and detecting fallacies. By extracting these unique dependency paths, it is possible to generate a set of rules that could aid in the identification and detection of logical fallacies.

5

Concluding Discussions

Logical fallacies refer to flawed patterns of reasoning that can compromise the soundness of an argument. They can occur intentionally or unintentionally, and can take various forms such as oversimplifications, irrelevant appeals, and generalizations. Detecting logical fallacies is a critical task in the field of argumentation and debate as it helps in constructing more persuasive and credible arguments.

In this study, we noticed that the individuals who used logical fallacies in CreateDebate [Perspective] discussions had more adversaries, which implies that their comments were more likely to be contested by other users. However, these same users also had a greater number of allies, suggesting the presence of a tightly-knit and supportive community. Our findings offer an intriguing insight into the structure of the CreateDebate network, which could be valuable in comprehending the conduct of users and the mechanics of diverse types of debates in online forums and social media platforms.

We introduced a novel approach for detecting logical fallacies through linguistic analysis, which proved successful in identifying distinct dependency paths specific to each type of fallacy. Our experiments in dependency parsing and fallacy detection revealed that the dependency paths generated by our proposed framework accurately reflected the logical structure of the corresponding fallacy. Despite the substantial overlap of popular dependencies, our method identified highly discriminative paths for each type of fallacy. This framework offers an alternative and effective approach to detecting fallacies by extracting unique dependency paths, which could be used to develop sophistical models for identifying and detecting logical fallacies.

Bibliography

Blair, J. (2006). *Logical Self-Defense.*

Damer, T. E. (2008). *Attacking Faulty Reasoning: A Practical Guide to Fallacy-Free Arguments.* Wadsworth/Cengage Laerning.

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

Fast, E., Chen, B., and Bernstein, M. S. (2016). Empath. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.* ACM.

Habernal, I., Hannemann, R., Pollak, C., Klamm, C., Pauli, P., and Gurevych, I. (2017). Argotario: Computational argumentation meets serious games. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 7–12, Copenhagen, Denmark. Association for Computational Linguistics.

Habernal, I., Pauli, P., and Gurevych, I. (2018a). Adapting serious game for fallacious argumentation to German: Pitfalls, insights, and best practices. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, Miyazaki, Japan. European Language Resources Association (ELRA).

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

Hansen, H. (2015). Fallacies. In Zalta, E., editor, *The Stanford Encyclopedia of Philosophy.*

Jin, Z., Lalwani, A., Vaidhya, T., Shen, X., Ding, Y., Lyu, Z., Sachan, M., Mihalcea, R., and Schoelkopf, B. (2022). Logical fallacy detection. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7180–7198, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

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

Mondal, M., Silva, L. A., and Benevenuto, F. (2017). A measurement study of hate speech in social media. In *Proceedings of the 28th ACM Conference on Hypertext and Social Media*, HT '17, page 85–94, New York, NY, USA. Association for Computing Machinery.

Patel, U., Mukherjee, A., and Mondal, M. (2023). “Dummy Grandpa, do you know anything?”: Identifying and Characterizing Ad hominem Fallacy Usage in the Wild. In *Proceedings of The 17th International AAAI Conference on Weblogs and Social Media (ICWSM'23).*

Sahai, S., Balalau, O., and Horincar, R. (2021). Breaking down the invisible wall of informal fallacies in online discussions. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 644–657, Online. Association for Computational Linguistics.

Sanders, J. A., Wiseman, R. L., and Gass, R. H. (1994). Does teaching argumentation facilitate critical thinking? *Communication Reports*, 7(1):27–35.