Modelling Persuasion through Misuse of Rhetorical Appeals

Amalie Brogaard Pauli
Aarhus Universitet
Denmark
ampa@cs.au.dk

Leon Derczynski
IT University of Copenhagen
Denmark
ld@itu.dk

Ira Assent
Aarhus Universitet
DIGIT Aarhus University
Centre for Digitalisation,
Big Data and Data Analytics
Denmark
ira@cs.au.dk

Abstract

It is important to understand how people use words to persuade each other. This helps understand debate, and detect persuasive narratives in regard to e.g. misinformation. While computational modelling of some aspects of persuasion has received some attention, a way to unify and describe the overall phenomenon of when persuasion becomes undesired and problematic, is missing. In this paper, we attempt to address this by proposing a taxonomy of computational persuasion. Drawing upon existing research and resources, this paper shows how to re-frame and re-organise current work into a coherent framework targeting the misuse of rhetorical appeals. As a study to validate these re-framings, we then train and evaluate models of persuasion adapted to our taxonomy. Our results show an application of our taxonomy, and we are able to detecting misuse of rhetorical appeals, finding that these are more often used in misinformative contexts than in true ones.

1 Introduction

People are exposed to a large amount of online text that is quickly scrolled through, but which may have an inherent agenda to persuade or convince the reader. As a mitigation strategy, we hypothesise that automatic detection of persuasion in a text can help the reader navigate more critically online: like a skilled rhetorician spotting how something is trying to persuade and how an argument might be faulty (Rapp, 2022). With this social motivation, we study how to computational model persuasion in text. Computational modelling of persuasion techniques and strategies is a raising field in the area of computational argumentation. We establish the working term 'undesired persuasion' to be when the execution of persuasion in a text is unsound, e.g by using fallacies or tricks. Prior research in this directions has, among others, focused on propaganda techniques (Martino et al., 2020a), logical fallacies (Jin et al., 2022), and personal attacks (Zhang et al., 2018; Habernal et al., 2018). The field of persuasion detection consists of a variety of focuses and different classification schemes. However, prior work shares commonalities, and we argue that the problem can be tackled in a more unified way and thereby benefit the computational modelling and understanding of persuasion. In this work, we propose to model problematic and undesired persuasion by targeting rhetorical appeals –

Fallacy of Pathos
Appeal to Fear: "'Without this additional insurance, you could find yourself broke and homeless"
Appeal to Pity: "I know I missed assignments, but if you fail me, I will lose my financial aid and have to drop out."
Appeal to Popularity: "Nine out of ten shoppers have switched to Blindingly-Bright-Smile Toothpaste."
Fallacy of Ethos
False authority:"Dr. X is an engineer, and he doesn't believe in global warming."
Ad Hominem:"Why should we think a candidate who recently divorced will keep her campaign promises?"
Name-calling:"These rabble-rousers are nothing but feminazis."
Fallacy of Logos
False dilemma:"Either we pass this ordinance or there will be rioting in the streets"
Circular argument:"This legislation is sinful because it is the wrong thing to do."
Red Herring or Smoke Screen:"My opponent says I am weak on crime, but I have been one of the most reliable participants in city council meetings."

Table 1: Examples of fallacy types grouped into fallacies of ethos, pathos, and logos. These are from two sources of educational material [1]and Kashyap (2022)

[1] https://www.mvrhs.org/englishdept/shark/links/General%20Information/Rhetorical%20Fallacies%20U.%20Texas%20@%20Austin.pdf

or rather, misuse of rhetorical appeals. Rhetoric is the discipline of persuading or influencing others through speech or text. Rhetorical appeals are described by Aristotle as the three modes of persuasion through writing or speaking, where: Ethos is to persuade through the credibility of the speaker, pathos through the emotions of the listener, and logos through the soundness of the argument itself (Rapp, 2022). We will use the working-term *misuse* of rhetorical appeals to denote when an appeal becomes unsound or exaggerated in its reasoning, e.g using fallacies – with fallacies understood as making a reasoning seem better than it is (Hansen, 2022). Table 1 shows examples of different logical fallacy types grouped into the broader categories of fallacies of logos, ethos and pathos. Based on such a framework, we will discuss how to re-frame existing resources and on the basis of this, develop models to detect misuse of rhetorical appeals.

Following our social motivation, we hypothesise that misusing rhetorical appeals to argue or present some evidence is correlated with misinformation in broader terms. Therefore, this paper examines whether the misuse of rhetorical appeals are more often used in, e.g., mis/disinformation. This is carried out by applying the models for detecting misuse of rhetorical appeals on a variety of data sets targeting this. At the same time, misusing the appeals may be correlated with losing arguments. We, therefore, test our models on a dataset from a debate forum where users upvote and downvote comments (Chang and Danescu-Niculescu-Mizil, 2019). In this paper, we:

- Propose modelling persuasion through rhetorical appeals,
- Re-frame existing resources and reorganise resources to target misuse of rhetorical appeals: ethos, pathos and logos,
- Experiment with developing models for detecting misuse of rhetorical appeals and link it to misinformation,
- Find a tendency showing that misuse of rhetorical appeals appears more frequently in misinformation, but also that a notable amount of fallacies of ethos and pathos are used in reliable news as well.

In general, we hope with this work to increase the focus on using rhetorical appeals in computational modelling persuasion, both on desired and undesired persuasion.

2 Computational Persuasion

This section sets the background of computational modelling of persuasion. We first outline the broad spectrum of different ways of understanding and modelling persuasion, to both map the field and to clarify concepts. From here the scope is reduced to existing classification schemes, focusing on their connections to rhetorical appeals.

2.1 Mapping Persuasion Modelling

The literature takes different perspectives and distinctions to model persuasion in text. To create an overview, we group the approaches in three directions and discuss connections and overlappings. The first direction is on text units linguistic defined. The second on pre-defined categories driven by the intention behind persuasion. The third direction is based on an audience's response to a text.

In the first direction, we have rhetorical figures treated as linguistic style units. These are relevant as they aim at producing a rhetorical effect, or in other words, to persuade an audience by e.g. utilising cognitive bias in humans e.g. with rhythm and repetition. Studies include the detection of repetitive figures (Dubremetz and Nivre, 2018), exaggeration (Troiano et al., 2018; Kong et al., 2020) and of syntax figures (Al Khatib et al., 2020).

In the third direction, research is trying to capture what people perceive as persuasive, without resorting to predefined style units or other predefined concepts of persuasion. For example, one study attempts to answer what makes a text persuasive by extracting a lexicon based on people's responsive action to a text (Pryzant et al., 2018). Another example is the discipline of automatic argument quality assessment, which could, for example, include a dimension of rhetorical quality with a score of how persuasive an argument is (Wachsmuth et al., 2017).

In the second direction, work is dealing with what is denoted as persuasion techniques or persuasion strategies using predefined categories. This line of research focuses more on the intention behind persuasion than on linguistic style units. For example, some studies for propaganda detection did not treat repetition and exaggeration as style units as seen above but instead as propaganda techniques (Martino et al., 2019, 2020a).

This direction can be subdivided into two since studies often makes a distinction between whether the intention or execution of persuasion is 'desired' or 'undesired'.

The desired persuasion line covers topics such as rhetorical strategies (Yang et al., 2019; Shaikh et al., 2020), convincing and winning arguments (Tan et al., 2016; Habernal and Gurevych, 2016) and 'persuasion for social good' (Wang et al., 2019). Under undesired persuasion, papers talk about propaganda (Martino et al., 2020a; Vorakitphan et al., 2021; Da San Martino et al., 2021), logical fallacies (Habernal et al., 2017; Jin et al., 2022) and personal attacks (Habernal et al., 2018; Sheng et al., 2020). Propaganda can be seen as the intention to persuade in a political context with opposing groups (Guess and Lyons, 2020). Propaganda techniques can therefore overlap with e.g logical fallacies and emotional appeals as in (Martino et al., 2020a). Different classification schemes in this direction of pre-defined categories are further outlined in subsection 2.2.

In addition, research frequently distinguishes between whether the persuasion is mediated through monologue or dialogue.

2.2 Classification Schemes

Prior research on desired and undesired persuasion applies a variety of different annotation schemes and denotations for (respectively) persuasion techniques and strategies. The following attempts to summarise it by focusing on the relation to rhetorical appeals. We start with desired persuasion.

Various classification schemes have been applied to rhetorical strategies. Several papers have proposed to use schemes guided form social psychology on persuasion (Young et al., 2011; Yang et al., 2019; Chen and Yang, 2021). Chen and Yang (2021) argue that their taxonomy can be used to unify the modelling of persuasion strategies. Their scheme uses the following labels: Commitment, Emotion, Politeness, Reciprocity, Scarcity, Credibility, Evidence and Impact (Chen and Yang, 2021). The strategy labels "credibility" and "emotion" are linked to respectively ethos and pathos. Other labels correspondence to rhetorical appeals are seen in Iyer et al. (2017) where among their 14 labels is VIP Appeal to Authority (ethos), Empathy and popularity (pathos). The rhetorical appeals are specifically targeted in Wang et al. (2019) but on the same terms with a list of more domain-specific strategies for convincing others to donate to charity. Lastly, the Hidey et al. (2017) also annotated rhetorical appeals; here on the premise in arguments posted in the discussion forum, Change My View.

There is less research on problematic and undesired persuasion with persuasion techniques and fallacies. Habernal et al. (2017) was the first within NLP research to work with fallacies, using a crowdsourcing game to create different types of fallacious arguments. Martino et al. (2019) created a corpus for detecting propaganda in news with 18 different techniques. This evolved into a shared task at SemEval 2020 (Martino et al., 2020a) with 14 categories. Two datasets for Logical fallacy detection were created in Jin et al. (2022) with 14 categories. The first is crafted by collecting logical fallacy examples from online educational materials, and the second is crafted by annotating real discussions on climate change. In addition to these, attention has especially been paid to Ad Hominem Fallacies, which are to attack the person instead of the stand. For example, Habernal et al. (2018) studied Ad Hominem Fallacies in an online debate forum with data from Change My View, and Sheng et al. (2020) studied it in Twitter responses, and Zhang et al. (2018) in Wikipedia talk pages where editors discuss article content. The different resources mentioned above are outlined in Table 2. The next section discusses whether undesired persuasion can be addressed in a more unified way by re-framing existing resources to target rhetorical appeals.

3 Re-framing Persuasion

We discuss how to computationally model persuasion through the lens of a framework detecting rhetorical appeals. By this we examine whether problematic and undesired persuasion can be addressed in a more unified way by re-framing existing resources (Table 2). We propose that persuasion techniques should be grouped with respect to the rhetorical appeals they rely on, as it is outline in e.g. the educational material on rhetoric from Kashyap (2022).

As we focus on problematic persuasion, we group fallacies based on whether they are making a faulty appeal to logos, ethos or pathos (Kashyap, 2022). Examples of fallacies are presenting something as a false dilemma, making an appeal to fear or attacking the person instead of the argument. Table 1 shows examples of fallacies related to rhetorical appeals. However, this grouping is

Corpus	Labels	Grouped to
Martino et al. (2020a)	Black-and-white fallacy, causal oversimplification	Misuse of logos
	Doubt, Appeal to authority, Name calling or labelling, Flag-waving, Bandwagon & reduction ad hitlerum	Misuse of ethos
	Loaded language, Appeal to fear/prejudice, Thought-terminating cliché	Misuse of pathos
	Repetition, Exaggeration or minimization, (mixed category: Whataboutism, straw man, red herring), slogans	Others
Jin et al. (2022)	Intentional fallacy, faulty generalization, fallacy of relevance, deductive fallacy, false causality, fallacy of extension, false dilemma, circular claim	Misuse of logos
	Fallacy of credibility, Ad Hominem	Misuse of ethos
	Appeal to emotion, Ad populum	Misuse of pathos
	Equivocation	Others
Habernal et al. (2017)	Red herring, hasty generalisation	Misuse of logos
	Irrelevant authority	Misuse of ethos
	Appeal to emotion	Misuse of pathos
Zhang et al. (2018)	Personal Attack	Misuse of ethos
Sheng et al. (2020)	Ad Hominem	Misuse of ethos
Habernal et al. (2018)	Ad Hominem	Misuse of ethos

Table 2: Re-framings of different labels from varies sources into the taxonomy of misuse of rhetorical appeals.

not straightforward, for multiple reasons. There is no absolute or final list of fallacies types. This is reflected in the variety of labels used in different prior works (Table 2). Some types might be a subcategory of others or contain a mix. At the same time, a type of fallacy can be argued to be a mix or use a different appeal depending on the utterance. Our over-arching principle is to group fallacies based on their fallacy type, along with a discussion of the noise it creates in the data. To create an overview of the grouping proposed by this paper, a colour scheme is applied to the categories from the different studies in Table 2. The Other category contains different linguistic or rhetorical devices that, based on their labels, cannot directly be grouped into appeals of logos, ethos and pathos. In the following, the grouping is discussed, starting with ethos.

Misuse of Ethos Ethos is an instrument of persuasion by appealing to credibility or authority. The fallacy of ethos is to unjustly strengthen one's own or associate's character or credibility, or to unfairly undermine or attack the opponent's character or credibility (Kashyap, 2022). From the previous resources listed in Table 2, we map the following fallacies to ethos: Appealing to irrelevant authority. Name-calling or labelling, which is to use negative connotations in relation to the opponent in an attempt to undermine her. Doubt, which is to question somebody's credibility (Martino et al., 2020a). The fallacy of flag-waving is a corner case, as it can both be an attempt to call upon authority in the form of a country, or disparages another country, while, on the other hand, it could also relate to pathos e.g. with an appeal to the emotion of national feeling. Lastly, we consider Ad Hominem, which is to make a personal attack. The annotation of Ad Hominem fallacy or personal attack category might be a source of noise, since it might target rude behaviour in general and not specific attacks on credibility.

Examples of positive cases tagged with the Ad Hominem fallacy that contain a faulty appeal to ethos: *Fine be that way, just to let you know you are very rude* and *So only Falun Gong practitioners are allowed to edit on this board is that right?*, and one example where it is rude but where it does not attack credibility directly: *The article clearly sucks* (Zhang et al., 2018).

Misuse of Pathos Pathos is an instrument of persuasion by appealing to emotion in the audience. To misuse it is to use it excessively or unfairly, e.g. creating strong positive emotions for one's stand or negative emotions associated with the opponent's argument (Kashyap, 2022). In the resources listed in Table 2, we argue that the follow-

ing fallacies types belong to the broader category of fallacies of pathos: Appeal to emotion. Appeal to fear/prejudice. Loaded language which is to use strong emotional words or phrases (Martino et al., 2020a) to create an emotional effect. Ad Populum is the fallacy of making something appear more real or better because more people think so (Jin et al., 2022), and can therefore be thought of as waking emotions, for example belonging. The thought-terminating cliché is perhaps a mixed category that could contain different appeals; however, in Wikipedia, it is described as a form of loaded language [2], and we map this to pathos.

Some positive examples from existing resources: *Because if this crisis continues, many people will go to hell*, Appeal to fear / prejudice (Martino et al., 2020a) and *How could someone oppress our women? They are our mothers, our lovers, our everything.. nobody would be so cruel*, Appeal to emotion (Habernal et al., 2017), and *"Everyone is wearing the new skinny jeans from American Eagle. Are you?"* Ad populum (Jin et al., 2022).

Misuse of Logos Logos is concerned with the nature of the argument itself. It appeals to logic by following valid reasoning and presenting of evidence. In this regard, a misuse of logos is to use faulty logic by e.g drawing a conclusion that is not supported by the premise. In that sense, this category is distinct from pathos and ethos which are in its definitions drawing attention away from the argument itself. An example is the fallacy of Red Herring which is to present irrelevant or misleading information to avoid the real issue (Kashyap, 2022) - this could often be the case by using an emotional appeal and it could therefore be grouped as a fallacy of pathos[3] and not logos. Nevertheless, we map it as logos along with the following fallacies from the previous resources in Table 2: Black-and-white fallacy, Casual oversimplification, Intentional fallacy, faulty or hasty generalisation, deductive fallacy, false causality, fallacy of extension, false dilemma, or circular claim.

One example of Red Herring that uses faulty logos: *You might be correct. The best era for European economy was 60s and 70s when there were practically no immigrants* (Habernal et al., 2017).

[2] https://en.wikipedia.org/wiki/Thought-terminating_clich%C3%A9

[3] https://www.mvrhs.org/englishdept/shark/links/General%20Information/Rhetorical%20Fallacies%20U.%20Texas%20@%20Austin.pdf

4 Detecting Rhetorical Appeals

This section relates experiments on detecting misuse of rhetorical appeals. We develop models for detecting misuse of ethos, pathos and logos in English, based on the re-framing of existing resources discussed in Section 3. We then examine how misuse of rhetorical appeals links to misinformation. We understand misinformation as the working-definition from Guess and Lyons (2020): *as constituting a claim that contradicts or distorts common understandings of verifiable facts.* We posit the following hypotheses:

First, we hypothesise that misuse of rhetorical appeals appears more often in losing arguments – since a faulty argument only has a persuasive effect if it is not spotted, cf. Section 1. The second hypothesis is that in misinformation, not only incorrect information but also persuasive language are used, and so misuse of rhetorical appeals may appear more frequently in misinformative contexts.

In the following, we present training details on the machine learning models we develop, describe the datasets we experiment on along with results, and discuss limitations and uncertainties.

4.1 Training Details: Appeal models

This subsection describes how models are developed to detect misuse of ethos, pathos, and logos. Three binary transformer models are fine-tuned independently on the RoBERTa architecture (Liu et al., 2019) based on the implementation and pretrained RoBERTa-base model provided by HuggingFace. (Wolf et al., 2020) Each model is fine-tuned based on a re-constructed dataset built on some of the resources discussed in Section 3 and reformulated into a binary task - based on the labels re-grouped in Table 2. The labels not responding to the current task at hand are used as negative examples. The datasets for the re-framing are chosen based on accessibility and length of utterances. With these limitations, the data used to develop the models, comes from: Habernal et al. (2017), Martino et al. (2020a), Jin et al. (2022) (only the part of educational examples), and in addition, for detecting ethos the data from Zhang et al. (2018). Each of the three constructed binary datasets are split into train, validation and a hold-out test set. The hold-out test sets consist of 1.6K data points for the ethos dataset, 1.2K for pathos and 1.2K for logos. The training dataset for ethos contains of 4.7K positive and 8.8K negative examples, for pathos; 3.5K

	Accuracy	Micro-F1
Ethos_model	85.14 (0.47)	85.12 (0.49)
Pathos_model	80.51 (0.35)	80.48 (0.41)
Logos_model	88.32 (0.36)	88.25 (0.39)

Table 3: The hold-out test set accuracy and Micro-F1 score avaraged over five runs. Standard deviation in brackets.

positive and 6.9K negative examples, and for logos; 2K positive and 8.4K negative examples. Oversampling is used to balance the datasets. All training parameters are kept equal to the standard used in the implementation by HuggingFace.[4] The models are fine-tuned with five different seeds and the averaged results on the hold-out test set are shown in Table 3. Note the hold-out test set is also on the re-framings. The best model in terms of F1 on the positive class for respectively ethos, pathos and logos is chosen for the misinformation experiments. For short, in the following, the models will be just denoted as ethos-, pathos- and logos-model though they are detecting what we with the re-framing have denoted misuse of rhetorical appeals.

4.2 Losing Arguments

We experiment on one dataset containing indications of good versus bad argumentations from the user's perspective:

- **Change My View** (CMV) is a forum in Reddit featuring good-faith debates on various topics with the aim of changing the opinion of the original poster. In the forum, users have the option of upvoting or downvoting utterances. An extraction of these data is provided in Chang and Danescu-Niculescu-Mizil (2019) and distributed by ConvoKit [5] with the voting on each utterance turned into a score (upvoting minus downvoting). We remove outliers in the score if the score exceeds 3 times the standard deviation. The data contains around 40K utterances.

As the data is from a forum with the purpose of changing other users' views through good argumentation, we expect that the argument is well evaluated and that this is reflected in the score. Hence, we expect users to dislike utterances using a faulty

[4] https://github.com/huggingface/transformers/tree/main/examples/pytorch/text-classification

[5] https://convokit.cornell.edu/documentation/awry_cmv.html

	Predicted		Not predicted	
Misuse of	score	support	score	support
Ethos	8.51	33558	5.98	8923
Pathos	8.15	37550	6.63	4931
Logos	8.6	33183	5.73	9298

Table 4: Change My View dataset: The average score on comments i.e. up-vote minus down-vote from users. The comments are grouped by whether the models have found a fallacy or not.

appeal. The hypothesis is, therefore, that utterances which contain a misuse of appeal should be less liked by the users resulting in a lower score.

We apply the three models for detecting misuse of ethos, pathos and logos described in Subsection 4.1 on each utterance from the dataset Change My View. Based on each model's predictions, the utterances are divided into groups of whether they contain a misuse of appeal or not, separately for the three models. The mean score is calculated for each group and is reported in Table 4. It shows for all three models, that the utterances where a misuse of appeal is detected on average have a lower score. To validate these differences, we conduct a statistical test. The data fails the normality test of Shapiro-Wilks (Shapiro and Wilk, 1965), and, therefore, we use the nonparametric Mann Whitney U test (Mann and Whitney, 1947). In all three cases, we can reject the null with a p-value < 0.01. We can conclude that the distribution of the scores regarding whether an appeal is predicated on the utterance or not is different. Hence, we can say that the utterance is less liked by users when it contains a misuse of appeal.

4.3 Misinformation

Misinformation and manipulation are rife on the web (Derczynski et al., 2015). We apply our models on to misinformation datasets to examine whether the misuse of appeals appears more frequently in the category of false claims than genuine ones. We use the following datasets, which contain both text and false/true annotations for veracity.

- **ISOT Fake News Dataset** (Ahmed et al., 2018) is a collection of news articles distant labelled with fake or true based on the sources. The unreliable news sources were flagged by Politifact.com or Wikipedia and the reliable news was crawled from Reuters.com. It counts 21K articles labelled real and 23K arti-

Fallacies of Ethos		
	True	False
ISOT Fake News Dataset	35.45	**49.96**
Liar	15.34	**19.75**
FakeNewsNet	13.62	**16.49**
COVID19-FAKE	2.14	**19.15**
PUBHEALTH	**17.72**	16.69

Table 5: The percentage of examples predicted by the ethos model to contain a misuse of ethos in different datasets grouped by the dataset's labels of false or true. The highest percentage is marked in bold. The sizes of the datasets are specified in the list describing each dataset.

Fallacies of Pathos		
	True	False
ISOT Fake News Dataset	22.27	**57.81**
Liar	15.75	**16.39**
FakeNewsNet	**42.32**	40.80
COVID19-FAKE	21.93	**24.90**
PUBHEALTH	**22.59**	18.66

Table 6: The percentage of examples predicted by the pathos model to contain a misuse of pathos in different datasets grouped by the dataset's labels of false or true. The highest percentage is marked in bold. The sizes of the datasets are specified in the list describing each dataset.

cles labelled false. In the experiments of this paper, the title is used to predict on.

- **Liar** (Wang, 2017) is a dataset for claim verification consisting of short utterances taken from Politifact.com and manually annotated into six fine-grained labels of truthfulness: pants-fire, false, barely-true, half-true, mostly-true, and true. We follow Upadhayay and Behzadan (2020) at convert it to binary labels with mostly-true and true in true (3.6K training data) and the rest in false (6.6K training data).

- **FakeNewsNet** (Shu et al., 2018) is a resource for claim verification with a set of metadata from Social Media. The news is fact-checked mainly by gossipcob.com. We use the title of the news article and labels for fake or real, and work with 23K examples imbalanced in labels, with around 6K labelled fake and 17K labelled real.

- **PUBHEALTH** (Kotonya and Toni, 2020) is a corpus on fact-checking of public health-related short claims enriched with explanations. It originally uses four labels, but we use only the annotations for False (3K) and True (5K).

- **COVID19-FAKE** (Patwa et al., 2021) is a manually annotated corpus of news tweets related to the Covid19 pandemic with fake or real. We use the fairly balanced train part of about 6K posts.

The three models for detecting misuse of ethos, pathos and logos described in Subsection 4.1 are applied to the misinformation datasets. Results are reported on how many percentages in each predefined group of either 'false' or 'true' in each misinformation dataset contain a predicted misuse of the appeal in question. For ethos, the results are reported in Table 5, for pathos in Table 6 and for logos in Table 7. In general, we see a tendency for more cases of misuse in the false columns than in the true - regarding all three appeals. However, there are some variations.

Regarding the ethos-model, we see large differences in COVID-19-FAKE and in the ISOT Fake News Dataset, but in the rest of the datasets, the differences between false and true are less distinct. In the PUBHEALTH dataset, we even see more cases of misuse of ethos in the true-labelled group than in the false-labelled group, although the numbers are quite close.

The pathos-model also spots a notable distinction in the ISOT Fake News Dataset with more cases of misuse of pathos among misinformation than in true news. In fact, the pathos model distinguishes the data to a degree that it obtains an accuracy on the true/fake labels on 0.6731. This can be compared to the dummy baseline of a majority vote on 0.5230. However, in contrast, the difference in pathos appeals between false and true in the remaining datasets is quite smaller. The PUBHEALTH and FakeNewsNet datasets have a few more cases of misuse of pathos among the true statements.

Both the pathos- and ethos-model find a notable amount of misuse in true news as well: e.g. over 40% of the titles in the FakeNewsNet are predicted to contain a fallacy of pathos and around 35% in the ISOT Fake News Dataset to contain a fallacy of ethos.

The misuse of logos-model detects much fewer

Fallacies of logos		
	True	False
ISOT Fake News Dataset	0.11	0.62
Liar	**16.52**	14.65
FakeNewsNet	2.00	2.36
COVID19-FAKE	9.94	**11.41**
PUBHEALTH	4.69	**11.83**

Table 7: The percentage of examples predicted by the logos model to contain a misuse of logos in different datasets grouped by the dataset's labels of false or true. The highest percentage is marked in bold. The sizes of the datasets are specified in the list describing each dataset.

cases of misuse in the datasets, in general, than the two other models. It predicts less than 1% of the cases in the ISOT Fake News Dataset and around 2% in FakeNewsNet. However, a few more cases are found in the remaining datasets. In addition, this stands in contrast to the experiments on the Change My View dataset, where the logos model found more cases than the two other models.

4.4 Discussion of Results

The models themselves are expected to be noisy as there are fine-tuned on re-framed resources with expected noise in the labels and without gold-annotations. However, we can see from the hold-out-test on reorganised datasets (Table 3) that the models learn to some degree to separate the re-grouped examples. Applying the models, we see the expected tendency: Misuse of appeal appears more often in misinformation than in reliable news, but with variations.

We notice a notable amount of fallacy of ethos and pathos in the true news. An explanation for this could be that even reliable news e.g. with their titles also aims at capturing the readers' attention: and doing so might appeal to the emotions of the reader or the credibility of the sources. At some point, this might be overdone and become faulty, related to the discussion: that it might at times be a thin line of when an appeal to emotions or credibility becomes faulty.

Another explanation for the different distributions of ethos and pathos across the datasets could be rooted in different topics the news is covering. One speculation is that some topics relate more easily to the use of e.g. pathos than others.

We expect some uncertainties in the results: There is a domain shift between the different misinformation datasets and the training data - despite the training data also containing data from news articles, it also contains data from dialogues and educational examples of fallacies. Concretely, a mismatch in the data distributions could be caused by the representation of negative examples in the training data. To examine the robustness of the prediction on the misinformation datasets, a pathos model on a different seed than previously reported is used for predictions. This causes some relatively large variation in the results in some of the misinformation datasets. For example, a pathos-model on a different seed captures more cases in the ISOT Fake News data set. But it does so both for the fake-labelled and the true-labelled group, respectively 63.15% versus 57.81% and 24.78% versus 22.27% . The models differ a bit in recall and precision. But this variation might also be explained by a large uncertainty in some of the model predictions in some of the examples, i.e. for some examples, the probability scores lie close to the decision border on 0.5, which might explain why the prediction is subject to shift with a similar model just fine-tuned on a different seed. This uncertainty might be caused by the domain shift and the sources of error in the distribution of negative data examples, but these are speculations.

5 Societal Impact

It is said we live in an information age; even WHO Director-General Tedros Adhanom Ghebreyesus has called the Covid19 epidemic an infodemic (García-Saisó et al., 2021). In general, people are exposed to a lot of text that has an inherent agenda of convincing, persuading, or misleading readers, seen in websites (Mathur et al., 2019), political debates (Addawood et al., 2019) and news (Barrón-Cedeno et al., 2019; Martino et al., 2020b). In this paper, we follow the assumption that language use plays a role in how information and arguments are perceived. We already know that, for example, the stances people adopt in conversation can relief their belief in underlying claims (Dungs et al., 2018; Lillie et al., 2019). Our vision is that automatic detection of undesired persuasion can help an online reader navigate more critically in the vast amount of information online, e.g by surfacing or flagging such text. This relates to the discussion that a person skilled in rhetoric posits the competences to spot how and when a text is persuasive (Rapp, 2022).

At the same time, automatic analysis of misuse of rhetorical appeals could help a writer present stronger more convincing arguments. As an example, one qualitative study manually analysed the rhetorical tactics and appeals used in vaccine discussion in the New York Times comments (Gallagher et al., 2020). In the study, they categorised the arguments in pro-vaccines and anti-vaccines and analyzed the rhetorical tactics and appeals in the comments. They found that pro-vaccine comments more often contained ad hominem arguments, and that this was an ineffective strategy.

While this comes with a dual-use risk – technology for highlighting manipulation can e.g. help manipulative authors better hide their intent – we posit that putting computational power behind rhetorical analysis can have a positive impact on the information society.

6 Conclusion

In this paper, we unify the modelling on problematic persuasion by using rhetorical appeals - or rather misuse of these. We focus on the problematic use of rhetorical appeals and re-frame and re-organise existing resources into this taxonomy. However, it is relevant to spot rhetorical appeals in all kinds of persuasion, and we speculate that for future work it might be useful to model pathos and ethos with less the distinction of misuse.

We link misuse of rhetorical appeals to misinformation. We showed that misuse of appeals appeared more often around misinformation than true claims. However, we also saw that in some contexts, reliable news was frequently tagged with misuse of ethos and pathos. This indicates the relevance of assessing the use of persuasion in a broad spectrum of text.

Limitations

This paper discusses limitations regarding both framing and experiments. We summarise the main points. First, regarding the framing: The idea is to propose a unifying taxonomy that can utilise existing work and resources, and hence gives rise to the idea of detecting misuse of rhetorical appeals. Still, the re-grouping based on a variety of labels is noisy, and the definitions themselves have limitations. For example, the misuse of logos is not a fully disjoint category with pathos and ethos, which both in their essence draw attention away from the argumentation. Regarding ethos and pathos, it is not easy to determine when they are unwarranted and hence can be classified as misuse. The distinction in rhetoric between use and misuse, desired and undesired persuasion is fluid and hence open for discussion in further work. Regarding limitations of the experiments: Gold-annotations specific on misuse of rhetorical appeals is lacking to better verify the trained models. In general, the results are preliminary, in the sense that e.g. a manual study could better demonstrate the models' detection of misuse of appeals in misinformation.

Ethics Statement

Our work complies with the ACL Ethics Policy. As discussed in the section on the potential for scientific impact, we believe that battling misinformation could benefit from taking fallacies of pathos, ethos and logos into account. By making transparent the use of such argumentative structures, we contribute to a fair and transparent discourse.

Acknowledgements

This work was supported by the Danish Data Science Academy – which is funded by the Novo Nordisk Foundation (NNF21SA0069429) and VILLUM FONDEN (40516) – and by the Independent Danish Research Fund project Verif-AI.

References

Aseel Addawood, Adam Badawy, Kristina Lerman, and Emilio Ferrara. 2019. Linguistic cues to deception: Identifying political trolls on social media. In *Proceedings of the international AAAI conference on web and social media*, volume 13, pages 15–25.

Hadeer Ahmed, Issa Traore, and Sherif Saad. 2018. Detecting opinion spams and fake news using text classification. *Security and Privacy*, 1(1):e9.

Khalid Al Khatib, Viorel Morari, and Benno Stein. 2020. Style analysis of argumentative texts by mining rhetorical devices. In *Proceedings of the 7th Workshop on Argument Mining*, pages 106–116.

Alberto Barrón-Cedeno, Giovanni Da San Martino, Israa Jaradat, and Preslav Nakov. 2019. Proppy: A system to unmask propaganda in online news. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 9847–9848.

Jonathan P Chang and Cristian Danescu-Niculescu-Mizil. 2019. Trouble on the horizon: Forecasting the derailment of online conversations as they develop. *arXiv preprint arXiv:1909.01362*.

Jiaao Chen and Diyi Yang. 2021. Weakly-supervised hierarchical models for predicting persuasive strategies in good-faith textual requests. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 12648–12656.

Giovanni Da San Martino, Stefano Cresci, Alberto Barrón-Cedeño, Seunghak Yu, Roberto Di Pietro, and Preslav Nakov. 2021. A survey on computational propaganda detection. In *Proceedings of the Twenty-Ninth International Conference on International Joint Conferences on Artificial Intelligence*, pages 4826–4832.

Leon Derczynski, Kalina Bontcheva, Michal Lukasik, Thierry Declerck, Arno Scharl, Georgi Georgiev, Petya Osenova, Toms Pariente Lobo, Anna Kolliakou, Robert Stewart, et al. 2015. Pheme: Computing veracity—the fourth challenge of big social data. In *Proceedings of the Extended Semantic Web Conference EU Project Networking session (ESCW-PN)*.

Marie Dubremetz and Joakim Nivre. 2018. Rhetorical figure detection: Chiasmus, epanaphora, epiphora. *Frontiers in Digital Humanities*, 5:10.

Sebastian Dungs, Ahmet Aker, Norbert Fuhr, and Kalina Bontcheva. 2018. Can rumour stance alone predict veracity? In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3360–3370.

John Gallagher, Heidi Y Lawrence, et al. 2020. Rhetorical appeals and tactics in new york times comments about vaccines: Qualitative analysis. *Journal of medical internet research*, 22(12):e19504.

Sebastián García-Saisó, Myrna Marti, Ian Brooks, Walter H Curioso, Diego González, Victoria Malek, Felipe Mejía Medina, Carlene Radix, Daniel Otzoy, Soraya Zacarías, et al. 2021. The covid-19 infodemic.

Andrew M. Guess and Benjamin A. Lyons. 2020. *Misinformation, Disinformation, and Online Propaganda*, SSRC Anxieties of Democracy, page 10–33. Cambridge University Press.

Ivan Habernal and Iryna Gurevych. 2016. What makes a convincing argument? empirical analysis and detecting attributes of convincingness in web argumentation. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 1214–1223.

Ivan Habernal, Raffael Hannemann, Christian Pollak, Christopher Klamm, Patrick Pauli, and Iryna Gurevych. 2017. Argotario: Computational argumentation meets serious games. *arXiv preprint arXiv:1707.06002*.

Ivan Habernal, Henning Wachsmuth, Iryna Gurevych, and Benno Stein. 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.

Hans Hansen. 2022. "Fallacies", The Stanford Encyclopedia of Philosophy.

Christopher Hidey, Elena Musi, Alyssa Hwang, Smaranda Muresan, and Kathleen McKeown. 2017. Analyzing the semantic types of claims and premises in an online persuasive forum. In *Proceedings of the 4th Workshop on Argument Mining*, pages 11–21.

Rahul R Iyer, Katia P Sycara, and Yuezhang Li. 2017. Detecting type of persuasion: Is there structure in persuasion tactics? In *CMNA@ ICAIL*, pages 54–64.

Zhijing Jin, Abhinav Lalwani, Tejas Vaidhya, Xiaoyu Shen, Yiwen Ding, Zhiheng Lyu, Mrinmaya Sachan, Rada Mihalcea, and Bernhard Schölkopf. 2022. Logical fallacy detection. *arXiv preprint arXiv:2202.13758*.

Erika Kashyap, Athena ans Dyquisto. 2022. 2.5: Logical fallacies - how to spot them and avoid making them. In chapter *2: Writing and the Art of Rhetoric* from *Writing, Reading, and College Success: A First-Year Composition Course for All Learners*, https://human.libretexts.org.

Li Kong, Chuanyi Li, Jidong Ge, Bin Luo, and Vincent Ng. 2020. Identifying exaggerated language. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7024–7034.

Neema Kotonya and Francesca Toni. 2020. Explainable automated fact-checking for public health claims. *arXiv preprint arXiv:2010.09926*.

Anders Edelbo Lillie, Emil Refsgaard Middelboe, and Leon Derczynski. 2019. Joint rumour stance and veracity prediction. In *Proceedings of the 22nd Nordic Conference on Computational Linguistics*, pages 208–221.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Henry B Mann and Donald R Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *The annals of mathematical statistics*, pages 50–60.

Giovanni Da San Martino, Alberto Barrón-Cedeno, Henning Wachsmuth, Rostislav Petrov, and Preslav Nakov. 2020a. Semeval-2020 task 11: Detection of propaganda techniques in news articles. In *Proceedings of the fourteenth workshop on semantic evaluation*, pages 1377–1414.

Giovanni Da San Martino, Shaden Shaar, Yifan Zhang, Seunghak Yu, Alberto Barrón-Cedeno, and Preslav Nakov. 2020b. Prta: A system to support the analysis

of propaganda techniques in the news. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 287–293.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeno, Rostislav Petrov, and Preslav Nakov. 2019. Fine-grained analysis of propaganda in news article. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 5636–5646.

Arunesh Mathur, Gunes Acar, Michael J Friedman, Elena Lucherini, Jonathan Mayer, Marshini Chetty, and Arvind Narayanan. 2019. Dark patterns at scale: Findings from a crawl of 11k shopping websites. *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW):1–32.

Parth Patwa, Shivam Sharma, Srinivas Pykl, Vineeth Guptha, Gitanjali Kumari, Md Shad Akhtar, Asif Ekbal, Amitava Das, and Tanmoy Chakraborty. 2021. Fighting an infodemic: Covid-19 fake news dataset. In *International Workshop on Combating On line Hostile Posts in Regional Languages during Emergency Situations*, pages 21–29. Springer.

Reid Pryzant, Kelly Shen, Dan Jurafsky, and Stefan Wagner. 2018. Deconfounded lexicon induction for interpretable social science. 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 1615–1625.

Christof Rapp. 2022. "Aristotle's Rhetoric", The Stanford Encyclopedia of Philosophy.

Omar Shaikh, Jiaao Chen, Jon Saad-Falcon, Polo Chau, and Diyi Yang. 2020. Examining the ordering of rhetorical strategies in persuasive requests. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1299–1306.

Samuel Sanford Shapiro and Martin B Wilk. 1965. An analysis of variance test for normality (complete samples). *Biometrika*, 52(3/4):591–611.

Emily Sheng, Kai-Wei Chang, Premkumar Natarajan, and Nanyun Peng. 2020. " nice try, kiddo": Investigating ad hominems in dialogue responses. *arXiv preprint arXiv:2010.12820*.

Kai Shu, Deepak Mahudeswaran, Suhang Wang, Dongwon Lee, and Huan Liu. 2018. Fakenewsnet: A data repository with news content, social context and dynamic information for studying fake news on social media. *arXiv preprint arXiv:1809.01286*.

Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning arguments: Interaction dynamics and persuasion strategies in good-faith online discussions. In *Proceedings of the 25th international conference on world wide web*, pages 613–624.

Enrica Troiano, Carlo Strapparava, Gözde Özbal, and Serra Sinem Tekiroğlu. 2018. A computational exploration of exaggeration. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3296–3304, Brussels, Belgium. Association for Computational Linguistics.

Bibek Upadhayay and Vahid Behzadan. 2020. Sentimental liar: Extended corpus and deep learning models for fake claim classification. In *2020 IEEE International Conference on Intelligence and Security Informatics (ISI)*, pages 1–6. IEEE.

Vorakit Vorakitphan, Elena Cabrio, and Serena Villata. 2021. " don't discuss": Investigating semantic and argumentative features for supervised propagandist message detection and classification. In *Recent Advances in Natural Language Processing (RANLP 2021)*.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

William Yang Wang. 2017. " liar, liar pants on fire": A new benchmark dataset for fake news detection. *arXiv preprint arXiv:1705.00648*.

Xuewei Wang, Weiyan Shi, Richard Kim, Yoojung Oh, Sijia Yang, Jingwen Zhang, and Zhou Yu. 2019. Persuasion for good: Towards a personalized persuasive dialogue system for social good. *arXiv preprint arXiv:1906.06725*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

Diyi Yang, Jiaao Chen, Zichao Yang, Dan Jurafsky, and Eduard Hovy. 2019. Let's make your request more persuasive: Modeling persuasive strategies via semi-supervised neural nets on crowdfunding platforms. 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 3620–3630.

Joel Young, Craig Martell, Pranav Anand, Pedro Ortiz, Henry Tucker Gilbert IV, et al. 2011. A microtext corpus for persuasion detection in dialog. In *Workshops at the Twenty-Fifth AAAI Conference on Artificial Intelligence*.

Justine Zhang, Jonathan Chang, Cristian Danescu-Niculescu-Mizil, Lucas Dixon, Yiqing Hua, Dario Taraborelli, and Nithum Thain. 2018. Conversations gone awry: Detecting early signs of conversational failure. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1350–1361.