Automated Detection of Tropes In Short Texts

Alessandra Flaccavento
Università Roma Tre
alessandra.flaccavento@uniroma3.it

Youri Peskine
EURECOM
youri.peskine@eurecom.fr

Paolo Papotti
EURECOM
papotti@eurecom.fr

Riccardo Torlone
Università Roma Tre
riccardo.torlone@uniroma3.it

Raphael Troncy
EURECOM
raphael.troncy@eurecom.fr

Abstract

Tropes — recurring narrative elements like the "smoking gun" or the "veil of secrecy" — are often used in movies to convey familiar patterns. However, they also play a significant role in online communication about societal issues, where they can oversimplify complex matters and deteriorate public discourse. Recognizing these tropes can offer insights into the emotional manipulation and potential bias present in online discussions. This paper addresses the challenge of automatically detecting tropes in social media posts. We define the task, distinguish it from previous work, and create a ground-truth dataset of social media posts related to vaccines and immigration, manually labeled with tropes. Using this dataset, we develop a supervised machine learning technique for multi-label classification, fine-tune a model, and demonstrate its effectiveness experimentally. Our results show that tropes are common across domains and that fine-tuned models can detect them with high accuracy.

1 Introduction

A trope is an easily recognizable device used in narratives to convey a specific theme or idea (Gala et al., 2020). This mechanism is widely used in the movie industry to generate effects and emotions in the audience as it can be traced back to the familiar feeling a person may sense when knowing what is coming up next in a given scenario (Su et al., 2021). Examples of tropes in this context are “the girl next door”, “the love triangle”, and “the damsel in distress”. In fact, tropes are used today in almost any form of communication, given their ability to convey attitudes and beliefs. In particular, just as storytellers in media use tropes to make stories more understandable and relatable, online content producers use them to communicate news and opinions, exploiting tropes’ familiarity and preconceived notions.

It has been observed, however, that this mechanism used in movies and literature to impact the audience’s perception is often used online to manipulate and deceive audiences (DiResta, 2021). Notably, the pervasive use of tropes in online anti-vaccine discourse holds significant potential for dangerously shaping public opinion, as it can lead individuals making uninformed vaccination decisions (Hughes et al., 2021). These tropes persist over time, recurring across various vaccines and contributing to ongoing anti-vaccine dialogues. For instance, in the 1800s, some people argued that natural methods were better than getting inoculated with cowpox-derived smallpox vaccines (Kata, 2012). Fast forward to today, we see similar claims that traditional cures are more effective than mRNA vaccines. Indeed, in spite the differences in details, most underlying tropes are consistently used across time and topics as well. For example, the narrative that “authorities cannot be trusted to make a decision that will benefit people” can be found both in the context of immigration (e.g., border control) and vaccine (e.g., vaccination policies) to invoke skepticism towards those authorities.

Therefore, the development of techniques for detecting and understanding these deceptive narrative elements is crucial to monitoring public discourse and promoting evidence-based communication. To this aim, we note that tropes are used not only in extended narratives but also in shorter forms of communication. For instance, a cinematic trope can be detected from a single scene, such as "love at first sight" in a fleeting glance exchanged by two characters. Similarly, the underlying message of a trope can be discerned from brief textual content – without any explicit mention of the trope itself, e.g., the “love triangle” in “Paul likes Anne, but his friend Harry met her first.”

Building on this observation, in this paper we strive to address the challenge of detecting *online* *tropes* in short text segments from social media. Automatically detecting online tropes from short text presents a challenging technical problem, as it requires not only the accurate identification of nuanced narrative elements but also the ability to extract and interpret context-dependent patterns within limited textual information. To address this problem, we define the general task of automatic trope detection. We start by providing the definition for nine tropes after an iterative qualitative coding process of online social posts discussing vaccines and immigration. These tropes are general, as they are common in discourses on any matter, but we found out that they are often used in these specific domains. Given the trope definitions, we create the first corpus of labeled short texts. This dataset highlights the prevalence of this problem and its distinct nature compared to other text classification tasks. Leveraging supervised machine learning techniques for multi-label classification, we present methods that can identify tropes even with limited textual information.

Numerous works focus on enhancing online information quality through text content analysis, including computational fact-checking (Guo et al., 2022; Nakov et al., 2021), identification of conspiracy theories (Shahsavari et al., 2020; Peskine et al., 2023), and detection of propaganda/persuasion techniques deployment (Da San Martino et al., 2021; Peskine et al., 2024). However, although trope identification is a powerful means to enhance our understanding of storytelling techniques, and effectively uncover implicit biases in many contexts, the task of trope detection has been ignored by the research community. In this work, we aim to bridge this gap.

Our contributions can be summarized as follows:

- We define the task of automatic trope detection and discuss its distinctions from prior research, focusing on the context of vaccine and immigration discussions on social media.
- We develop and provide a dataset of 3.3K vaccine and immigration related Twitter posts labeled with tropes.
- We demonstrate how supervised machine learning techniques for multi-label classification perform in this new task.
- We show that tropes are widely used online and analyze how these labels correlate with other popular tasks in text classification.

Code and datasets are available at `https://github.com/Tireswind/ADTIST24`

2 Task Definition

We start with a definition of online tropes, then list the tropes we identified, and finally present our problem formulation.

Definition. We use the term *trope* as defined in the media studies: “a storytelling device or convention, a shortcut for describing situations the storyteller can reasonably assume the audience will recognize”[1] (Gala et al., 2020).

By *online trope*, we mean a trope used in online discussions. These tropes are not used to refer to plots, but rather to human situations. Even if the general behavior, habit or issue is not stated explicitly, the reference is clear to the reader.

As examples of the trope “Natural is better”, which is often used in discussions about a variety of topics, consider the following texts :

t_1: `“Not sure I will get the vaccine, natural immunity is the best immunity”.`

t_2: `“GMO food is created by corporations to make profit, cannot be better than natural food”.`

The writers of these messages are both advocating for natural solutions as the most healthy. Online tropes appeal to popular concepts, common experiences, or part of a culture that is known by the target audience.

Online Tropes. We outline the definitions for nine online tropes used in short texts that we have identified through our analysis of tweets related to two major topics: vaccine and immigration. We point out that we focus on tropes that can be found in general discussions, not necessarily involving the two topics at hand. To pinpoint these tropes, we employed a systematic and iterative qualitative coding process consisting of four phases: familiarization (reviewing the literature on tropes and examining thousands of topic-related tweets), open-coding (labeling tweets with potential trope codes), framework development (organizing codes into themes and higher-level categories), and finally verification (re-validating the established categories by applying them to the tweets examined during the open-coding phase).

We list below the online tropes and we refer to Table 1 for the corresponding complete examples.

- **Skepticism Towards Authority (STA)**. Text appeals to skepticism towards scientific experts

[1] `https://tvtropes.org/`

Posts	Tropes	Vaccine	Immig.	Total
The illegal migrant scandal is bigger than anybody realises.Our government should stop the boats from coming, not help them to shore.Mark my words this issue is not going away	Time Proves Me Right (TPMR)	43 / 2.1%	33 / 2.7%	76 / 2.3%
You are not, actually, interested in hearing the other side of the debate. And unless you can give birth, you can MYOB. Vaccines are a little different than abortions but you're a deacon, not a doctor.	Skepticism Towards Authority (STA)	194 / 9.4%	30 / 2.4%	224 / 6.8%
As Trudeau still goes on pushing this untested experimental vaccine using mRNA that has never been used successfully before on people!	Too Fast (TF)	142 / 6.8%	0 / 0%	142 / 4.3%
maybe you aren't looking in the right places. These vaccines are a negative cost/benefit for most people, particularly those with natural immunity.	Natural is Better (NIB)	63 / 3.0%	3 / 0.2%	65 / 2.87%
My body my choice no vaccine for me, but that woman over there? I decide her medical operations' - Everyone okay with the SCOTUS decisions.	Liberty, Freedom (LF)	325 / 15.7%	19 / 1.5%	344 / 10.4%
Lots of tweets saying I made up a story about migrants throwing their phones into the Channel yesterday. Well here is the exclusive footage.Why would legitimate refugees with nothing to hide throw their mobile phones into the sea?	Hidden Motives (HM)	244 / 11.8%	58 / 4.7%	302 / 9.1%
Well, it HAS TO BE either Climate Change or Putin! It can't possible be anything related to the mRNA vaccines, right?!	Scapegoat (SC)	58 / 2.8%	19 / 1.5%	77 / 2.3%
Publix Declines to Offer Coronavirus Vaccine to Children Under 5 PUBLIX IS PROTECTING OUR BABIES FROM THE POISON IN THE VACCINE	Defend the Weak (DTW)	99 / 4.8%	78 / 6.3%	177 / 5.4%
Migrants are being ferried around the country in taxis costing the taxpayer millions of pounds. How can the Government justify this expense with over 6,000 homeless veterans?	Wicked Fairness (WF)	0 / 0%	68 / 5.5%	68 / 2.1%
these vaccines becoming like those goddamn app updates.	None	1100 / 53	968 / 78.7%	2068 / 62.6%

Table 1: Examples of tropes occurring in tweets and frequency of their presence in our dataset.

or political authorities, with statements such as "They should know/do better" and "They don't know what they are doing". An example message is "authorities have failed now and before".

- **Defend The Weak (DTW)**. Text emphasizes the negative effects of something (e.g., vaccine, immigration policy) on vulnerable populations, with statements like "it is especially harmful to children". Example messages: "we must protect the weak", "they put the weak ones in danger".
- **Hidden Motives (HM)**. Text alludes to underlying agendas, suggesting that something (e.g., vaccines, illegal immigrants) is promoted by individuals with malicious intentions (such as hypocrites and tyrants) and concealed motives ("There is clearly an untold story behind it"). Examples of messages are "we must stop this scam" and "they are lying for their interest".
- **Liberty, freedom (LF)**. Text emphasizes personal autonomy and rights, using statements such as "my body, my choice", "not anti-something but pro-choice", and "people were stripped of their rights, jobs, freedom and forced against their will." Examples of messages are "I should be able to do what I want" and "They are forcing on me something I don't want".
- **Natural Is Better (NIB)**. Text promotes the idea that natural or traditional approaches are superior, with assertions like "natural immunity is the best immunity", "traditional solutions are more effective and secure", and "nature had a solution for this". Examples of messages are "I trust tradition more than innovation" and "They want to force non-natural solutions".
- **Time Proves Me Right (TPMR)**. Text appeals to the eventual validation of one's argument over time ("time will prove me right") and asserting foresight ("I told you this would happen"). Examples of messages are "I knew it / I know what is gonna happen" and "They don't see the problem coming"
- **Too Fast (TF)**. Text implies that something

(such as vaccines) is unsafe or unreliable because it is experimental, untested, developed too quickly ("haste makes waste"), or not yet fully approved by authorities. Example of messages is "They rushed the decision".

- **Scapegoat (SC)**. Text that attributes blame for a (possibly under-specified) problem to a person or entity not directly involved, such as "They claim it's A or B's fault, but it's really X's fault", or assigning responsibility for an issue to a popular entity, such as Bill Gates. Example of message is "It is all their fault!".
- **Wicked Fairness (WF)**. Text compares to how two entities are being treated, highlighting application of different principles for similar situations (i.e., double standard). Some examples use questions, "Why can't X have access to Z while Y can?", if/then statements "If X can be punished for that, then Y should be punished as well", or the claim "It's not fair".
- **None**. Texts that do not fit clearly into any trope category. A portion of these tweets contains misinformation and conspiracy theories related to vaccination or immigration without involving tropes. For instance, content suggesting that vaccines cause autism.

Problem Definition. Given a short text, our goal is to assign one or more labels corresponding to the online tropes used in expressing the message, if any. Identifying the trope category can be a complex task, posing challenges for automated methods.

Notice that tropes can be seen as a tool used in persuasion techniques to achieve their goals (Da San Martino et al., 2021). For example, tropes such as "Defend The Weak" can be used to implement the "Appeal to fear" persuasion technique. Similarly, the "Antivax" conspiracy theory could use a "Hidden Motives" trope.

3 Dataset

We opted to use supervised learning to detect tropes automatically. Thus, we created a ground truth for the model to learn from, focusing on topics that have been strongly debated in recent years and in which they can oversimplify complex matters and deteriorate public discourse. Specifically, we built a dataset comprised of short texts retrieved from Twitter (now X) by using the keywords "vaccine" for one topic, and "migration", "migrant" and "asylum" for the other. The retrieval did not take into account the specific user when scraping for texts, but it was keyword centered. We keep only posts written in English. We point out that we did not check the presence of misinformation in these posts, we simply collected tweets in which the keywords occurred at least once.

3.1 Annotation process

The annotation activity was guided by the following general criteria:

- A trope is a storytelling device, which exploits a shortcut for describing situations the storyteller can *reasonably assume* the audience will recognize. For this reason, if the presence of a trope in a text is likely but not evident, the text has to be annotated with the label "none".
- A short text can involve more than one trope. Hence, the labelling has to include all relevant tropes, not just the one that appears to be the "strongest".

To start, four co-authors[2] reviewed independently about 200 tweets and annotated them according to the nine tropes mentioned in the previous section. Next, they compared and discussed the tweets with disagreement in the labels to refine the labeling process. The Cohen's kappa coefficient agreement of annotation of the sample before the refinement was 0.62.

We realized that we encountered difficulties in labelling the texts with certain features, such as the use of sarcasm (which is very difficult to detect without context), references to different cultural aspects, and generally mixed-up topics brought into the argumentation. Moreover, we realized that some posts involved tropes we had not defined with precision: thus, we refined and redefined the labels each time this happened.

This initial activity was followed by another round of labeling, by four independent annotators, of 3.1K additional posts with the refined labels. A subsequent consolidation meeting with all authors on all the posts resulted therefore in a set of around 3300 annotated tweets with unanimous agreement. 2,074 tweets (63%) are about Vaccine and 1,230 (37%) are about Immigration.

During the annotation process, we made sure that the data did not contain any information that identifies individual people.

[2] The pool consisted of three males and one female. Their ages spanned between mid 20s and mid 50s. Annotators span two nationalities.

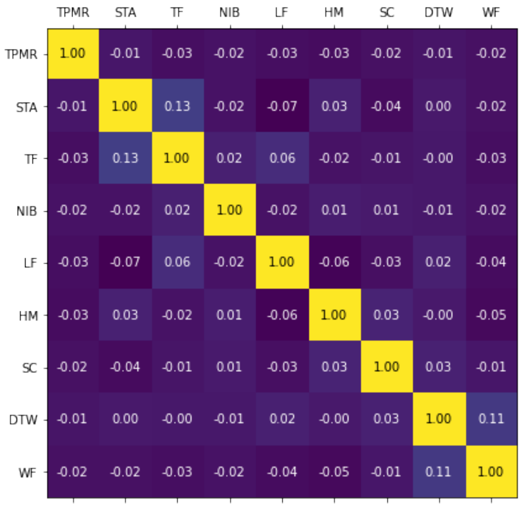

Figure 1: Correlations between tropes using the Pearson coefficient.

3.2 Data Analysis

Table 1 shows the distribution for each label. Despite the sampling of the scraped dataset being totally random, the tweets resulted to be fairly balanced after assigning the labels, in terms of texts with tropes and texts without them.

Interestingly, the result aligns with previous findings on key elements narratives, where studies (Rory Smith, 2020) found that the most frequent conversation about vaccines on social platforms involved a concept they labeled *liberty and freedom*. Conversely, the least numerous labels are Time Proves Me Right, Natural Is Better, Scapegoat, and Wicked Fairness. These tropes probably require a deeper dialectic, as the speaker tries to bring forth a kind of reasoning, making them more sporadic throughout the dataset compared to other, more direct arguments that characterize other tropes.

We investigated the correlation among the tropes we defined, to show that they would not overlap. As shown in Figure 1, no significant signal was detected. It is possible, however, given the nature of the problem and the way to express opinions, that some tropes are used together more often, like, for instance, the feeling of distrust towards science (Skepticism Towards Authority) that developed a solution too quickly (Too Fast), or pointing out a double standard (Wicked Fairness) while referring to a vulnerable target (Defend The Weak).

3.3 Vaccine vs Immigration

Topics such as Vaccine and Immigration inherently trigger different discourse on social media. Even though most Tropes are found on both topics, there are some significant differences between the two sets of tweets. We notice that tropes appear twice more often on the Vaccine topic than on the Immigration topic. We also found that the tropes STA, TF, NIB, LF and HM are shared more in Vaccination topics, the tropes DTW and WF in Immigration topics, and the tropes TPMR and SC are shared equally in both subsets. In fact, WF is only found in Immigration tweets and TF is only found in Vaccine tweets.

4 Models

We model the problem of trope detection as a multi-label classification task, which focuses on categorizing instances into several non-exclusive classes, with each associated class of an instance referred to as a label. We describe below the four models used in our study. Fine tuned models are trained on 80% of the examples in the annotated dataset. All models are tested on the remaining 20%. All reproducibility settings of our experiments (hyper-parameters, prompts, etc) are shared in Appendix 7.

Bert-FT. To predict one or more tropes for a given tweet, we fine-tune a BERT-large-uncased pre-trained language model using HuggingFace (345M parameters). We save the model with the best average F1-score on the validation set out of 20 epochs.

CovidBert-FT. Given that we analyze tweets and most of them discuss covid-related topics, such as vaccine hesitancy, we also fine-tune a second language model, COVID-TwitterBERT (CovidBert), which is a BERT-large model pre-trained on COVID-related tweets (Müller et al., 2020) (336M parameters). We follow the same fine-tuning process used for Bert-FT described above.

ChatGPT-ZeroShot. We model the trope detection task with the ChatGPT-3.5 turbo[3] engine (175B parameters). We use the OpenAI APIs to request ChatGPT to label all the texts from our dataset with the tropes we have identified. We use a Zero Shot approach by prompting, to obtain the labels, using only the tweet at hand and the trope definitions. The definitions of the labels prompted to ChatGPT are the ones reported in Section 2. The prompt itself is in the Appendix section.

Llama-3-ZeroShot. We also use an open weight LLM to perform the Trope classification task. We

[3] gpt-3.5-turbo-0125

chose the 'Meta-Llama-3-8B-Instruct' model from Huggingface[4] (8B parameters) using the same prompt used with ChatGPT-3.5, baring a few adjustments to fit the Llama prompting syntax.

5 Experiments

We first report results for the trope detection task over our annotated dataset (**Tropes**). We then show how tropes might be correlated to other textual features, namely conspiracy theories and persuasion techniques. Finally, we discuss the results.

5.1 Models for Trope Classification

In the first experiment, we evaluate and compare the four alternative trope detection models over our annotated validation dataset. We compute, for each trope, the F1-score, as well as for the 'None' category. We also report the weighted average of the F1-score across the dataset.

Table 2 reports the F1-scores and overall results of our models. It shows that CovidBert-FT has the best performance with a weighted average F1-score of **0.65**. Both supervised models perform better than LLMs with zero-shot prompting.

Results also show that some tropes are easier to detect than others. Indeed, both LF and TF obtain high F1-scores. However, models struggle to detect the TPMR trope. One explanation of this difference can be found in the most frequent bi-grams of each trope, where clear messages exist in LF ('body, choice', 'experimental, vaccine', or 'vaccine, mandates') and TF ('clinical, trials', 'trials, future' or 'emergency, use') while no clear insights appear for TPMR ('covid, vaccines', 'long, term' or 'wait, til'). Another reason as to why some classes are harder to detect than other, is because of the low number of samples in the training set. Some tropes, such as TPMR, have a low number of examples (around 2.3% in our dataset). This makes it more challenging for supervised models to properly learn how to detect them. Conversely, tropes with a high number of samples tend to be easier to detect, such as LF (10.4%).

Overall, models perform well in the binary detection of the 'None' class.

Additional results are available in the Appendix, giving further insights on the difference between training supervised models only on Vaccine or only on Immigration data. Finally, in the Appendix, we also report an in-depth error analysis, giving false positive examples for every class. The main takeaway is that false positives come from model over-fitting on certain keywords.

5.2 Persuasion Techniques and Conspiracy Theories

This section is devoted to studying the relationship between tropes and two detection tasks in misinformation analysis: (i) the use of persuasion techniques and (ii) the presence of text discussing and promoting conspiracy theories about COVID-19.

For this study, we select two datasets for which we have a ground truth, specifically:

- **Persuasion techniques**: a set of 7k texts extracted from online memes annotated with human-provided labels indicating the use of persuasive techniques[5] (Dimitrov et al., 2024).
- **Conspiracy**: a set of 2k tweets about Covid manually annotated with labels for nine conspiracy theories (Langguth et al., 2023).

Given that the CovidBert-FT shows the best results for trope detection, we use it in the rest of the experiments. To detect the use of persuasion techniques and conspiracy theories in our **Tropes** dataset, we rely on state-of-the-art models from the literature, specifically PERSUASION TECHNIQUE DETECTION (Peskine et al., 2024) and CONSPIRACY DETECTION (Peskine et al., 2021).

Tables 3 and 4 show the results from the execution of the models for the **Conspiracy** and **Persuasion Technique** datasets, respectively. In both tables, we report the human-labelled results (ground truth) in italic. The other results are obtained by running the detection models and can therefore be noisy. We remark that our model is trained only for trope detection: any text containing conspiracy theories or persuasion techniques, but without tropes, is labelled as "None". In this experiment, we compare all tasks at a binary level, i.e. the use of Tropes, Persuasion Technique, or Conspiracy Theories in text.

Comparison with Conspiracy Theories. First, we can see in Table 3 that the proportion of tweets that contain conspiracy theories is constant across both datasets (around 50%). This holds for the proportion of tweets containing Tropes. This shows that datasets are not biased towards the textual features they annotate. We also analyze how Tropes

[4] https://huggingface.co/meta-llama/Meta-Llama-3-8B-Instruct

[5] Even though the data contains memes, only the textual content was used for the annotation.

Model	STA	DTW	HM	LF	NIB	TPMR	TF	SC	WF	None	*Avg*
Bert-FT	0.54	0.57	0.42	0.78	0.50	0.33	0.75	0.48	0.55	0.83	0.58
CovidBert-FT	0.60	0.68	0.59	0.80	0.55	0.27	0.77	0.64	0.57	0.87	**0.65**
ChatGPT-3.5-ZeroShot	0.19	0.36	0.27	0.66	0.27	0.00	0.31	0.20	0.44	0.55	0.32
LLAMA3-8B-ZeroShot	0.15	0.29	0.20	0.38	0.27	0.12	0.16	0.10	0.10	0.24	0.23

Table 2: F1-score results for our models for each trope, the 'None' class, and weighted average across the dataset.

	Consp.	Trope	Both	Trope	Consp.	None
Dataset				**Only**	**Only**	
Tropes	49.9%	*37.4%*	24.1%	13.3%	25.8%	36.8%
Consp.	*51.9%*	30.4%	19.9%	10.5%	32.0%	37.6%

Table 3: Proportions of conspiracy and tropes in respective datasets. Ground truth in italics.

	Pers.	Trope	Both	Trope	Pers.	None
Dataset				**Only**	**Only**	
Tropes	91.7%	*37.4%*	35.9%	1.5%	55.8%	6.8%
Persuas.	*81.9%*	11.4%	10.6%	0.7%	71.3%	17.3%

Table 4: Proportions of persuasion techniques and tropes in respective datasets. Ground truth in italics.

and Conspiracies appear together on those datasets. In both datasets, more than 60% of tweets contain at least a Conspiracy or a Trope, showing the prevalence of such features in social media posts. We also analyze the correlations between tropes and conspiracy theories using Matthews correlation coefficient. The only positive correlation found is with the 'Hidden Motives", even though the coefficient is low (0.19). This confirms that Tropes and Conspiracy Theories are orthogonal concepts.

In order to evaluate that our Tropes model can be applied to Conspiracy data, we perform manual validation by checking 40 tweets positively labeled by our model with high confidence. We obtain a binary F1-score of 0.943, highlighting that our model can be used outside its training distribution.

Comparison with persuasion techniques. Table 4 shows the proportions of both Tropes and Persuasion techniques in both datasets. We notice that Persuasion Techniques are used a lot more than any other textual features, however, tropes seem to be used less in online memes. We also note that their appearance together is not consistent across both datasets. This may be due to the fact that both data are coming from different sources (social media textual posts for Tropes and memes for persuasion techniques) and about different topics. Indeed, the persuasion dataset contains a significant amount of memes heavily biased towards US politics, most of them being offensive to certain groups of people. We have found no positive correlations between tropes and persuasion techniques (Matthews correlation coefficients are less than 0.07).

We also evaluate the performance of our Tropes model on persuasion technique data by manually labeling the textual content of 60 memes positively labeled with high-confidence by our Tropes detection model. We found a binary F1-score of 0.843, showing that our model can safely be used on this other kind of content.

5.3 Discussion

Results from these experiments highlight some interesting insights. First, we see that Tropes exist independently from Persuasion Techniques and Conspiracy Theories in the online discourse about Vaccines and Immigration. They therefore provide new information that can be used to understand written language better, in addition to existing textual features. Indeed, we show that tropes are orthogonal to conspiracy theories and persuasion techniques. As much as mentioning a conspiracy or using a certain persuasion technique does not necessarily implies spreading misinformation, we believe that tropes are yet another dimension of analysis that should be studied.

One more important aspect separates tropes from conspiracy theories and persuasion techniques. Tropes can be used to polarize opinions either way, in a more neutral manner: in this context, most of the time, they are used to belittle the efforts of experts but it is not the only way. Consider for instance the following sentence: t_3: "Great point - collectively, we failed to get the vaccine to hundreds of millions of people who needed it because Canada, the USA, the UK, and others supported windfall profits of drug companies over people's health." Here, the "Hidden Motive" that led to negative consequences was used to support the argument for the failure of vaccine availability.

Results also show that LLMs struggle to detect Tropes, but supervised models reach convincing performance. However, not all Tropes are detected with the same precision, giving us insights about the difficulty of the task. For example, the trope TPMR shows poor performance from the models.

Lastly, we manually annotated conspiracy tweets and persuasive memes on a high-confidence threshold as we believe that precision is a more important metric than recall in an out-of-distribution setting. This way, we can reliably detect documents with tropes, which provide useful information for our study. The classes detected the most out-of-distribution are **Defend The Weak**, **Hidden Motives** and **Liberty, Freedom**.

6 Related Work

Several works study the impact of false information and misleading narratives on online social media platforms. For identifying and addressing misleading information in online text, current techniques focus on detecting veracity (fact-checking) (Guo et al., 2022; Nakov et al., 2021), the use of propaganda or persuasion techniques (Da San Martino et al., 2021), data voids (Mannino et al., 2024), and support for conspiracy theories (Shahsavari et al., 2020). Some of these works specifically focus on vaccine-related content (Jamison et al., 2020; Du et al., 2021), but to our knowledge, there is no work yet on trope detection. Persuasion techniques are methods employed to manipulate public opinion and promote a specific agenda, while tropes are communication devices that are not inherently tied to misinformation. Examples of persuasion techniques are "reductio ad hitlerum", to discredit an idea that is popular in groups hated by the audience, and "bandwagon", to appeal to the popularity of an argument (Da San Martino et al., 2019). A trope is also different from a text that supports a conspiracy theory. The latter is focused on content, i.e., on the entities and arguments for the topic at hand, while the former is rather a tool for achieving a communication goal. Indeed, a given text that refers to a conspiracy theory (the "what") can use different techniques to convince the audience, including tropes and propaganda techniques (the "how"). Similarly, tropes are different from themes (Islam and Goldwasser, 2024; Pacheco et al., 2023): while themes are the central messages conveyed by a narrative, tropes provide familiar and concise elements that can be used to implement multiple themes, e.g., the same tropes are found in both the Vaccine and Immigration posts. Tropes can be seen as a new factor to characterize online content (Burel et al., 2024).

Tv tropes have been widely studied, given their persuasive use to simplify narratives and improve communication (Su et al., 2021; Gala et al., 2020) and the problem of trope detection has been studied on a TVTropes dataset of 5.6k movie synopses and 95 tropes (Chang et al., 2021).

There are several studies that focus on analyzing the public discourse surrounding vaccines and vaccine hesitancy, as well as the use of tropes and misinformation in this discussion (DiResta, 2021; Hughes et al., 2021). It has been observed that a multitude of narratives, including tropes, converge to create an environment of extreme uncertainty in the vaccine information ecosystem (Rory Smith, 2020; Kata, 2012). Studies have been done also for the problems with the immigration discourse on social media (Mendelsohn et al., 2021; Ekman, 2019). None of them, however, propose methods for the automatic detection of tropes in this context.

In this paper, we focus on supervised ML algorithms for detecting tropes in short texts. We model the problem as a multi-label classification task and report results for state-of-the-art methods using pre-trained language models, such as BERT (Devlin et al., 2019) and GPT (Brown et al., 2020), using fine-tuning and zero-shot learning.

7 Conclusions

This paper addresses the challenge of automatically detecting general tropes in short texts. We have defined the task of trope detection and demonstrated its distinct nature compared to other text classification tasks. We have created and shared a unique ground-truth dataset of 3,300 vaccine (63%) and immigration (37%) related Twitter posts labeled with common tropes, which can be used to further advance this area of research. Results show that supervised approaches for multi-label classification achieve significant success in detecting tropes. Our work contributes to a better understanding of public opinions and biases through the lens of tropes.

We plan to extend the scope of our dataset by incorporating additional topics, to better understand tropes' usage in different domains. Furthermore, we aim to integrate our approach with interdisciplinary efforts in fact-checking (Nakov et al., 2021), persuasion technique detection, and conspiracy theory prevention to devise a comprehensive framework for better online communication.

Limitations

Firstly, our analysis focuses on a subset of tropes relevant to vaccine and immigration-related discussions on one social media platform (Twitter/X). Despite our report on some results on short texts present in memes, our focus raises questions about the generalizability of our findings. Tropes present in extended narratives or other social media platforms might not correspond directly to those identified in our work. To achieve a broader understanding, future research should involve annotated datasets encompassing a wider array of topics and platforms, potentially revealing new tropes beyond those identified here. We also do not think that the proposed list of tropes is exhaustive. We observed that some tropes only appear on Immigration (WF) or Vaccine (TF) topics, and it is possible that other tropes may appear on other topics. We focus on the 9 tropes defined in this work as they are the ones that appeared from the examination of the tweets. Our process was iterative as we refined the list of tropes as well as their definitions over time to provide a reliable resource. Those 9 tropes are the ones we can reliably define from our dataset and are representative therefore for this dataset. Also, we do not study the full online discourse exhaustively and it could be studied more extensively in a follow-up work. We focus on different domains such as conspiracy theories in tweets and persuasion techniques in memes as they represent a large portion of the research topics on online misinformation.

Secondly, our model is developed primarily for English texts, a language with relatively simple morphological structures. Consequently, its applicability to languages with more complex morphology is limited. This restriction underscores the necessity for further evaluations and adaptations for multi-lingual settings.

Third, our models are optimized for short textual content, and their performance on longer texts remains unverified. The extension to longer formats like articles or essays could introduce different narrative structures that might not be aptly captured by the model tuned for brevity.

Additionally, the LLMs used in our experiments have inherent sensitivity to prompt variations and hyperparameters such as temperature settings. This sensitivity may lead to inconsistent labeling of the same tweet under slightly different prompts. Furthermore, these models have a tendency to produce false positives, often erring on the side of identifying at least one trope even when it is not present. We present these LLMs experiments as baselines, with two very popular options. GPT-3.5 as a proprietary and top-performing LLM accessible through an API and Llama3-8B as an open-weight, lighter LLM that could be run on consumer GPUs. We are aware of the many techniques to improve LLMs performance (few-shot, definitions, chain of thoughts, etc), but decided not to explore them as this is not the focus of this paper.

Potential Risks

The first risk is the potential for misuse. The ability to automatically detect tropes in social media posts can be a double-edged sword. While this capability can be instrumental in identifying problematic content, it can also be misused. Malicious actors could leverage this technology to craft more sophisticated campaigns by avoiding recognizable tropes or adapting their messaging in ways that are harder to detect. To safeguard against such misuse, we consider to continuously update the model to recognize emerging patterns of malicious usage.

The second risk is related to bias and fairness. The models we have developed might inadvertently reinforce existing biases present in the training data. For instance, if the dataset used for training predominantly includes tropes associated with certain socio-political contexts, the model might underperform or produce biased results when applied to different cultural settings or new topics. This can lead to the exclusion or misrepresentation of certain groups, amplifying disadvantaged voices, or even perpetuate stereotypes and result in unequal treatment of certain demographics. To mitigate this risk, it is crucial to audit the model across diverse datasets and contexts. Future research should focus on developing more inclusive training datasets that represent a wide array of cultural and linguistic backgrounds, as well as implementing fairness-aware algorithms that detect biases.

Acknowledgements

This work has been partially supported by CHIST-ERA within the CIMPLE project (CHIST-ERA-19-XAI-003), by ANR within the ATTENTION (ANR-21-CE23-0037) and ECLADATTA (ANR-22-CE23-0020) projects, and by MASE within the CSSE project (PTR22-24).

References

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *FAccT '21: 2021 ACM Conference on Fairness, Accountability, and Transparency, Virtual Event / Toronto, Canada, March 3-10, 2021*, pages 610–623. ACM.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *NeurIPS*.

Gregoire Burel, Martino Mensio, Youri Peskine, Raphaël Troncy, Paolo Papotti, and Harith Alani. 2024. Cimplekg: A continuously updated knowledge graph on misinformation, factors and fact-checks. In *ISWC 2024, 23rd International Semantic Web Conference, 11-15 November 2024, Baltimore, USA*, Baltimore.

Chen-Hsi Chang, Hung-Ting Su, Jui-Heng Hsu, Yu-Siang Wang, Yu-Cheng Chang, Zhe Yu Liu, Ya-Liang Chang, Wen-Feng Cheng, Ke-Jyun Wang, and Winston H. Hsu. 2021. Situation and behavior understanding by trope detection on films. In *Proceedings of the Web Conference 2021*, WWW '21, page 3188–3198, New York, NY, USA. Association for Computing Machinery.

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 Joint Conference on Artificial Intelligence*, IJCAI'20.

Giovanni Da San Martino, Seunghak Yu, Alberto Barrón-Cedeño, 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, Hong Kong, China. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*, pages 4171–4186. Association for Computational Linguistics.

Dimitar Dimitrov, Firoj Alam, Maram Hasanain, Abul Hasnat, Fabrizio Silvestri, Preslav Nakov, and Giovanni Da San Martino. 2024. SemEval-2024 Task 4: Multilingual Detection of Persuasion Techniques in Memes. In *18th International Workshop on Semantic Evaluation (SemEval)*, SemEval 2024, Mexico City, Mexico.

Renee DiResta. 2021. 'Prebunking' Health Misinformation Tropes Can Stop Their Spread. *Wired*.

Jingcheng Du, Sharice Preston, Hanxiao Sun, Ross Shegog, Rachel Cunningham, Julie Boom, Lara Savas, Muhammad Amith, and Cui Tao. 2021. Using machine learning–based approaches for the detection and classification of human papillomavirus vaccine misinformation: Infodemiology study of reddit discussions. *J Med Internet Res*, 23(8):e26478.

Mattias Ekman. 2019. Anti-immigration and racist discourse in social media. *European Journal of Communication*, 34(6):606–618.

Dhruvil Gala, Mohammad Omar Khursheed, Hannah Lerner, Brendan O'Connor, and Mohit Iyyer. 2020. Analyzing gender bias within narrative tropes. In *Proceedings of the Fourth Workshop on Natural Language Processing and Computational Social Science*, pages 212–217, Online. Association for Computational Linguistics.

Zhijiang Guo, Michael Schlichtkrull, and Andreas Vlachos. 2022. A survey on automated fact-checking. *Transactions of the Association for Computational Linguistics*, 10:178–206.

Brian Hughes, Cynthia Miller-Idriss, Rachael Piltch-Loeb, Beth Goldberg, Kesa White, Meili Criezis, and Elena Savoia. 2021. Development of a codebook of online anti-vaccination rhetoric to manage covid-19 vaccine misinformation. *International Journal of Environmental Research and Public Health*, 18(14):7556.

Tunazzina Islam and Dan Goldwasser. 2024. Discovering latent themes in social media messaging: A machine-in-the-loop approach integrating llms.

Amelia Jamison, David A. Broniatowski, Michael C. Smith, Kajal S. Parikh, Adeena Malik, Mark Dredze, and Sandra C. Quinn. 2020. Adapting and extending a typology to identify vaccine misinformation on twitter. *American Journal of Public Health*, 110(S3):S331–S339. PMID: 33001737.

Anna Kata. 2012. Anti-vaccine activists, web 2.0, and the postmodern paradigm – an overview of tactics and tropes used online by the anti-vaccination movement. *Vaccine*, 30(25):3778–3789. Special Issue: The Role of Internet Use in Vaccination Decisions.

Johannes Langguth, Daniel Thilo Schroeder, Petra Filkuková, Stefan Brenner, Jesper Phillips, and Konstantin Pogorelov. 2023. Coco: an annotated twitter dataset of covid-19 conspiracy theories. *Journal of Computational Social Science*.

Miro Mannino, Junior Garcia, Reem Hazim, Azza Abouzied, and Paolo Papotti. 2024. Data void exploits: Tracking & mitigation strategies. In *Proceedings of the 33rd ACM International Conference on Information and Knowledge Management, CIKM 2024, Boise, ID, USA, October 21-25, 2024*, pages 1627–1637. ACM.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Martin Müller, Marcel Salathé, and Per E Kummervold. 2020. Covid-twitter-bert: A natural language processing model to analyse covid-19 content on twitter. *arXiv preprint arXiv:2005.07503*.

Preslav Nakov, David P. A. Corney, Maram Hasanain, Firoj Alam, Tamer Elsayed, Alberto Barrón-Cedeño, Paolo Papotti, Shaden Shaar, and Giovanni Da San Martino. 2021. Automated fact-checking for assisting human fact-checkers. In *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 4551–4558. ijcai.org.

Maria Leonor Pacheco, Tunazzina Islam, Lyle H. Ungar, Ming Yin, and Dan Goldwasser. 2023. Interactive concept learning for uncovering latent themes in large text collections. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 5059–5080. Association for Computational Linguistics.

Youri Peskine, Giulio Alfarano, Ismail Harrando, Paolo Papotti, and Raphael Troncy. 2021. Detecting covid-19-related conspiracy theories in tweets. In *MediaEval 2021, MediaEval Benchmarking Initiative for Multimedia Evaluation Workshop, 13-15 December 2021 (Online Event)*. CEUR.

Youri Peskine, Damir Korencic, Ivan Grubisic, Paolo Papotti, Raphaël Troncy, and Paolo Rosso. 2023. Definitions matter: Guiding GPT for multi-label classification. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 4054–4063. Association for Computational Linguistics.

Youri Peskine, Raphaël Troncy, and Paolo Papotti. 2024. Eurecom at semeval-2024 task 4: Hierarchical loss and model ensembling in detecting persuasion techniques. In *Proceedings of the 18th International Workshop on Semantic Evaluation*, SemEval 2024, Mexico City, Mexico.

Claire Wardle Rory Smith, Seb Cubbon. 2020. Under the surface: Covid-19 vaccine narratives, misinformation and data deficits on social media. *First Draft*.

Shadi Shahsavari, Pavan Holur, Tianyi Wang, Timothy R Tangherlini, and Vwani Roychowdhury. 2020. Conspiracy in the time of corona: automatic detection of emerging covid-19 conspiracy theories in social media and the news. *Journal of computational social science*, 3(2):279–317.

Hung-Ting Su, Po-Wei Shen, Bing-Chen Tsai, Wen-Feng Cheng, Ke-Jyun Wang, and Winston H. Hsu. 2021. Truman: Trope understanding in movies and animations. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, CIKM '21, page 4594–4603, New York, NY, USA. Association for Computing Machinery.

Appendix

Reproducibility

To encourage reproducibility of our experiments, we share our code at: `https://anonymous.4open.science/r/ADTIST24-768D`.

For the training of our Bert-FT and CovidBert-FT models, we used one Tesla K80 GPU. Training time is around two hours. We used the following hyper-parameters: batch size of 12, learning rate of $2 \cdot e^{-5}$, 20 epochs, AdamW optimizer, weight decay of 0.01. We use a Cross-Entropy Loss weighted with the inverse frequency of the class sample. We split the dataset into 80% training and 20% validation using a stratified split (according to the nine tropes).

We used the API provided by OpenAi to prompt the gpt-3.5-turbo-0125 model used in our experiments. We used the Llama-3-8B-Instruct model locally on a NVIDIA GeForce RTX 3090 GPU, with an inference time of around 4 seconds per annotation. For both GPT and Llama experiments, we use the following prompt:

```
The task is to label some texts
    according to these definitions:

Skepticism Towards Authority (STA): The
    text appeals to skepticism towards
    science and scientific experts or
    towards political authorities,
    featuring narratives such as '
    authorities have failed now and
    before', 'this political party does
    not know what they are doing' (I
    know better than experts; They
    should know better; They don't know
    what they are doing).
Defend The Weak (DTW): The text
    emphasizes the negative effects of
    something on vulnerable populations,
     e.g. children (it is especially
    harmful to the weak; I must protect
    the weak; They are putting the weak
    in danger).
Hidden Motives (HM): The text alludes to
     underlying agendas, suggesting that
     something is secretly promoted by
    individuals with malicious
    intentions (such as hypocrites and
    tyrants) and concealed motives (
    There is clearly an untold story
    behind it; I am being lied to; They
    are trying to hide their real
    motives).
Liberty, freedom (LF): The text
    emphasizes personal autonomy and
    rights (my body, my choice; I should
     be able to do what I want; They are
     forcing on me something I don't
    want; people were stripped of their
    rights, jobs, freedom and forced
    against their will).
Natural Is Better (NIB): The text
    promotes the idea that natural or
    traditional approaches are superior,
     with assertions like 'natural
    immunity is the best immunity' and '
    natural/traditional solutions are
    more effective and secure' (I think
    natural solutions are more effective
    ; The other solutions put us in
    danger).
Time Will Tell (TWT): The text appeals
    to the eventual validation of one's
    argument over time and asserting
    foresight (I know what is gonna
    happen; I knew it was gonna happen;
    They don't see the problem coming).
Too Fast (TF): The text implies that
    something is unsafe or unreliable
    because it is experimental, untested
    , developed too quickly ('haste
    makes waste'), or not yet fully
    approved by authorities (I currently
     don't feel safe without more
    evidence; They rushed the decision,
    it's dangerous).
Scapegoat (SC): Text that attributes
    blame or responsibility for a
    problem to a person or entity not
    directly involved, such as 'They
    claim it's A, B, or C's fault, but
    it's really X's fault' or assigning
    responsibility for an issue to a
    famous or popular entity, such as
    Bill Gates (I think this group of
    people/entity is to be held
    responsible; They are the biggest/
    only problem).
Wicked Fairness (WF): Text that hints to
     the fact that someone is receiving
    something they do not deserve,
    pointing to the unfairness of the
    situation (something feels unfair
    about one group of people/entity;
    They should receive the same
    treatment as someone else).
None: Texts that do not fit clearly into
     any trope category.

No other labels are allowed if you think
     the text should be labelled as a
    None. Labels are not mutually
    exclusive, there can be up to three
    but not necessarily.
```

Ethics Statement

Our research aims to address the issue of misinformation in the context of vaccination and immigration discussions, which has the potential to influence public opinion and health outcomes on a global scale. Detecting and understanding tropes can help combat the spread of misinformation and promote evidence-based communication.

By focusing on vaccine and immigration discussions, our results should be treated cautiously by communication experts. It is vital not to undermine freedom of speech, ensuring that individuals have the right to express dissenting opinions, while mitigating the spread of misinformation.

While the developed model exhibits promising results in detecting tropes in vaccine discussions, it should not be extrapolated to assume the capability to identify tropes across all domains and languages without proper fine-tuning and validation for different contexts.

We are also aware of the biases coming from LLM and the risks in using them in classification tasks (Bender et al., 2021). To minimize the impact of these biases, it is necessary to ensure the application of appropriate evaluation metrics and

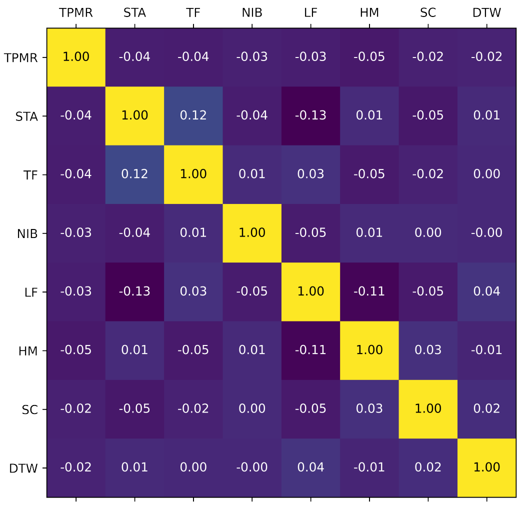

Figure 2: Correlations between tropes using the Pearson coefficient on the Vaccine subset.

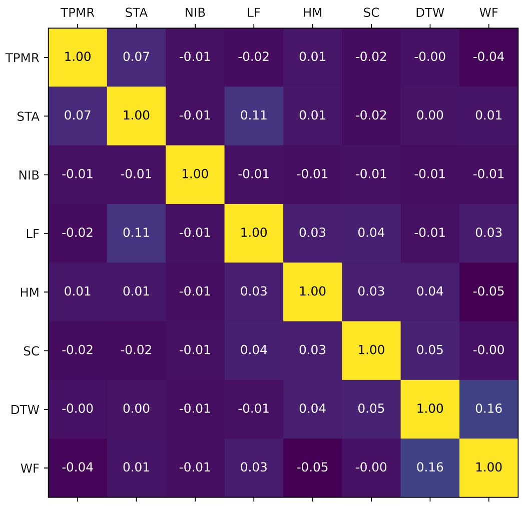

Figure 3: Correlations between tropes using the Pearson coefficient on the Immigration subset.

involve practitioners with domain expertise in a target application.

Data Collection

In constructing our dataset, we have focused on Twitter posts, which are publicly available data. We have removed personally identifiable information from the dataset, and the content of the post has been stripped of links; however, no profanity filter has been applied. The Twint Python library has been used for data collection. The library scraped, just by querying the keyword "vaccine", "migrant", "migration" and "asylum" filtering for results in the English language, resulting in about 50k tweets. We collected tweets posted throughout the 26th and 27th of June 2022 for the "vaccine" keyword and late November 2023 for the immigration domain. However, for the immigration topic, we realized soon enough that posts were too similar one another: thus, to have a less biased dataset as possible and to avoid a consequent bias in the training process, we went back in time to retrieve data since late 2019 up to 2022, and then chose about 300 tweets from each year. The vaccine domain did not encounter the same issue.

Additional Experimental Results

We share additional results, about the difference between Vaccine and Immigration topics on the training of supervised models in Table 5 and 6, as well as correlations between Tropes on Vaccine and Immigration subsets in Figures 2 and 3. We train Bert and CovidBert models on subset of the training set: full dataset (V+I), Vaccine only data (V) and Immigration only data (I). We report results on subsets of the test set: full dataset (V+I), Vaccine only data (V) and Immigration only data (I). Lines Bert-V+I and CovidBert-V+I correspond to the Bert-FT and CovidBert-FT in Table 2.

We can see that models trained on full data tend to obtain best results, even outperforming models trained and tested on specific subsets. For example, a CovidBert model trained only on Vaccine data performs worse on Vaccine tweets than a model trained on both Vaccine and Immigration data. This shows that Tropes can be generalized and can be transferred from one topic to another, as the information of Immigration tweets help the classification of Vaccine tweets. However, we see that models tend to over-fit: models trained on Immigration data perform poorly on Vaccine data, and inversely.

Another takeaway is that Tropes on the Immigration subset are more difficult to detect. Indeed, the average F1-score for Vaccine data is consistently higher than the average on Immigration data. This can be due to the number of Vaccine tweets in the training set being twice the number of Immigration tweets.

Error Analysis

In this section, we analyse some false positive examples for every class and try to identify the cause. Results are obtained using our best detection model (CovidBert-FT). We focus on false positive rather than false negative because we think precision is a more important metric than recall in this use case.

Model	TPMR			STA			TF			NIB			LF			HM		
	V+I	V	I	V+I	V	I	V+I	V	I	V+I	V	I	V+I	V	I	V+I	V	I
Bert-V+I	0.33	0.20	**0.43**	0.54	0.58	0.0	0.75	0.75	0.0	0.50	0.52	0.0	0.78	0.80	**0.29**	0.42	0.46	0.24
Bert-V	**0.35**	0.31	0.40	0.58	0.60	0.0	0.76	0.76	0.0	0.52	0.55	0.0	0.61	0.78	0.093	0.43	0.52	0.22
Bert-I	0.17	0.14	0.2	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.03	0.0	**0.29**	0.13	0.1	0.25
CovidBert-V+I	0.27	0.18	0.36	**0.60**	**0.63**	0.0	0.77	0.77	0.0	**0.55**	**0.57**	0.0	**0.80**	**0.83**	0.0	**0.59**	**0.64**	0.40
CovidBert-V	0.30	**0.36**	0.22	**0.60**	0.62	0.0	**0.78**	**0.78**	0.0	0.46	0.48	0.0	0.74	0.81	0.14	0.49	0.61	0.16
CovidBert-I	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.24	0.25	0.0	0.15	0.03	**0.48**
# support	16	8	8	45	42	3	28	28	0	13	12	1	68	64	4	61	51	10

Table 5: F1-score per class on subsets of the test data. Models are trained on different subset of the train set. V stands for Vaccine and I for Immigration.

Model	SC			DTW			WF			Weighted AVG			None		
	V+I	V	I	V+I	V	I	V+I	V	I	V+I	V	I	V+I	V	I
Bert-V+I	0.48	0.56	0.29	0.57	0.56	0.59	0.55	0.0	0.55	0.58	0.62	0.42	0.83	0.78	0.88
Bert-V	0.35	0.50	0.0	0.47	0.61	0.19	0.0	0.0	0.0	0.50	0.63	0.15	0.79	0.79	0.80
Bert-I	0.0	0.0	0.0	0.49	0.17	0.73	0.41	0.0	**0.59**	0.12	0.04	0.42	0.78	0.71	0.90
CovidBert-V+I	**0.64**	**0.70**	**0.50**	**0.68**	**0.58**	**0.82**	**0.57**	0.0	0.57	**0.65**	**0.68**	**0.50**	**0.87**	**0.84**	**0.91**
CovidBert-V	0.41	0.53	0.20	0.41	0.55	0.11	0.0	0.0	0.0	0.54	0.66	0.11	0.83	0.81	0.85
CovidBert-I	0.07	0.0	0.29	0.47	0.08	0.77	0.20	0.0	0.58	0.16	0.08	0.44	0.80	0.72	**0.91**
# support	16	12	4	36	20	16	14	0	14	100%	63%	37%	411	218	193

Table 6: (continued) F1-score per class on subsets of the test data. Models are trained on different subset of the train set. V stands for Vaccine and I for Immigration.

Obviously, a similar study could be done for false negatives.

Time Proves Me Right The model tends to classify time-related predictions ('it will soon be', '100 years ago was very similar to today') as Time Proves Me Right, even though it's not a sufficient condition, as it lacks the actual prediction of what "is to come".

$text_0$: It's not too late but we must act quickly to reduce immigration by a lot, or it soon will be.

$text_1$: A global pandemic 100 years ago was very similar to today I thought the alleged Spanish Flu started as an experimental vaccine gone wrong in a US military hospital where all the vaccinated soldiers went down with bronchial pneumonia.

Skepticism Towards Authorities In this case, even though authorities are mentioned in the text, it's not clearly implied that the user wants to promote skepticism or suspicion.

$text_0$: THE FDA IS ATTEMPTING TO 'FLU SHOT' FUTURE COVID VACCINES. NO TRIAL FOR THE VACCINES OF THE FUTURE

$text_1$: Last year, Home Secretary @sajidjavid set out plans for a new skills-based immigration system that would mark the end of free movement. Find out more: #Brexit.

Too Fast Here, the model understood that something declared as "emergency use" is a fast and temporary solution, thus developed too quickly. Also, the term "experimental" could have misled the model, but calling a product arbitrarily experimental does not represent a trope.

$text_0$: Where is the long term safety data for monkeypox vaccines? And why mass produce a vaccine for such a rare illness if not created out of a laboratory?

$text_1$: They sure pushed the fear....some fell for the scam... EMERGENCY USE ONLY VACCINE DID NOT STOP TRANSMISSION OR INFECTION....AND TRIPLE JABBED ARE GETTING INFECTED REPEATEDLY WITH COVID ????????

$text_2$: The hypocrisy is stunning. How can you not say vaccine mandates of an experimental product with NO liability and poor safety and efficacy is not 100% about bodily autonomy and free choice.

Natural is Better The model seems to trigger positively on the word 'immunity', which is surely strongly correlated with the trope Natural is Better.

$text_0$: Nonsense, many of those did not need the jab and would have recovered, fact. What we now know is that the vaccine has killed more than it has saved, fact. It has also undermined the natural immune system because it NEVER was a vaccine! FACT! So naff off Mr Village idiot!

$text_1$: Maybe the survival rate of 99.98% has something to do with people not being obsessed with it. And the fact that most have immunity now, through infection and vaccines

Liberty, Freedom The model activation seems to correlate with the word 'forced', which may not always be a cause of Liberty, Freedom trope.

$text_0$: How is depopulation possible through Forced Vaccines. I read the article, and yes it was said, and it is an agenda discussed as well as followed by everyone participating in Davos.

$text_1$: so the Pentagon can just ignore federal laws, but individuals in the armed forces can't ignore vaccine mandates?

$text_2$: And your Forced VACCINE prevents nothing! Only in your head! It has not stopped the SPREAD anywhere! And people like you never talk about the side effects nor natural immunity! BUT ABORTION is not Reproductive Rights...it is MURDER plain and simple!

Hidden Motives The word "expose" is quite often used to talk about something that is revealed through investigative reports: the model has learned this, and used it to wrongly label as Hidden Motive texts that had it in them (as shown for $text_3$). We also see a trend of mentioning organizations ('the Tories', 'Big Pharma', 'Bill gates') in false positive tweets, hinting that the model may have over-fit on the training data since these may be behind a Hidden Motive narratives, but not always.

$text_0$: Up to date? The old polio vaccine worked fine for 40 years until Bill Gates Corp created a new one for Africa which is a failed vaccine

$text_1$: Just a reminder that the Tories' betrayal over post-#Brexit #immigration is only part of the Establishment's treachery. 'All the same, all to blame'.

$text_2$: Here is why Big Pharma wants their vaccine in your kids. They know. Please retweet.

$text_3$: EXCLUSIVE: Citizen journalist 'who exposed Migrant Crisis' in bid to become MP @UKIP @Steve_Laws_ via @PoliticaliteUK

Scapegoat It is not sufficient to mention a famous person for it to be a Scapegoat, especially if those are just mentioned on the fly or to attributed conspiracies.

$text_0$: And if you were as smart as you think you are, you would know that birth rates are plunging throughout the vaccinated World. So, guess what? Babies are going to be rare and precious. Gates is achieving his dream of depopulation through vaccine.

$text_1$: Look at the spoilt immigrant rich brat advocating not just drag queen indoctrination for our children, but also acid attacks on our history (not Nelson Mandela's statue though).Looking at her, there's a song in Cabaret comes to mind.....

Defend the Weak These examples are incorrectly classified as "Defend the Weak". It seems like the model puts too much emphasis on the mention of "kids" and "children" when classifying this label.

$text_0$: Look at the spoilt immigrant rich brat advocating not just drag queen indoctrination for our children, but also acid attacks on our history (not Nelson Mandela's statue though).Looking at her, there's a song in Cabaret comes to mind.....

$text_1$: NY Post: Children, who ordinarily love shots, recoil in pain and horror from vaccine mandate forced on them by parents.

$text_2$: The truth is, this 'demographic' would have been better off not injecting their kids with 3-5 'vaccines' every single month which then ended up directly causing autism. That's a fact.

$text_3$: CDC Caught Using False Data To Recommend Kids' COVID Vaccine CDC showcased highly misleading data about the risk of COVID-19 to kids when its expert vaccine advisers voted to recommend vaccines for children under five years old.

Wicked Fairness This set of examples highlights tweets that mention comparison between two entities. However, they do not stress the unfair treatment they may have received, thus not qualifying as positive Wicked Fairness examples.

$text_0$: At last count, two thirds of 'child refugees' entering the UK were adults lying about their age in order to cheat their way into #BenefitsBritain. A reliable dental check to confirm their ages was suggested but was deemed 'racist' by the Woke mob.

$text_1$: RIDDLE ME THIS!HOW DO IMMIGRANTS STRENGTHEN OUR COUNTRY BUT NOT THEIR OWN ?!?

$text_2$: France is a wealthy country perfectly capable of affording refuge to those on its territory who are in need. Migrants there should either be considered for asylum in France or be returned to their own countries as economic migrants.