

Qualitative Analysis of Persuasive Emotion Triggering in Online Content

Olga Uryupina¹

¹Department of Information Engineering and Computer Science, University of Trento

Abstract

This paper presents a qualitative analysis of the emotional component in manipulative online content (fakes). We show that emotion triggering is a crucial persuasion technique widely employed by unscrupulous content generators. Based on a dataset of real-life fakes analyzed by fact-checking professionals, we identify the most common types of triggered emotions to be used as a taxonomy for further annotation.

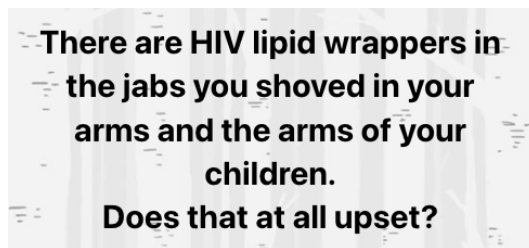
Keywords

persuasion, fact-checking, sentiment analysis

1. Introduction

The manipulative content, ranging from propaganda to hate campaigns, fake news, trolling and similar, is becoming more and more widespread, threatening our access to truthful and unbiased information and thus undermining our rights to make informed decisions as individuals and as members of the society. While there is a growing body of multidisciplinary research on identifying untruthful content, there is still very limited understanding of the manipulative techniques the unscrupulous content writers employ to convince the reader and ultimately change their point of view. We believe that this manipulation occurs through multiple channels: careful selection of fact-checkable and non-fact-checkable claims, biased yet seemingly solid argumentation/analytics, multimedia support and, most importantly, emotional component. Our current study focuses on *emotion triggering* – a technique widely used by content writers: when the reader is experiencing a strong feeling, they become less critical and thus easily overlook deficiencies in the argumentation and get more prone to manipulation.

Fig. 1 shows examples of manipulative textual content with strong emotional triggering. In (1a), the message makes a very strong appeal to fear, by mentioning HIV. Moreover, this triggering effect is intensified by mentioning "children". The fact-checking report¹ informs the reader that the COVID-19 vaccines do not contain any HIV material, but do contain other lipids to protect the mRNA. The distressed users, however, might not trust this information fully, due to such a strong emotion as a fear for their children's health. Ex-



(a) from Facebook



Rep. Marjorie Taylor Greene  
@RepMTG

Murkowski, Collins, and Romney are pro-pedophile.

They just voted for #KBJ.

3:13 AM · Apr 5, 2022


(b) from Twitter

Figure 1: Emotion triggering in manipulative content.

ample (1b) shows a typical manipulative message not addressed properly by the state of the art verification-oriented technology. The message combines a verifiable true claim ("Murkowski, Collins, and Romney voted for Ketanji Brown Jackson") with a statement that looks like a similarly factual claim ("Murkowski, Collins, and Romney are pro-pro-pedophile"), but in reality is an explanation/opinion offered by the writer. This triggers a rather strong anger at the powers/authorities under the spotlight, their presumed hypocrisy and their presumed (lack of) values. Here again, the triggering is intensified by bringing up a topic related to children. The fact-checking report² debunks this claim stating that "Sens. Murkowski,

CLiC-it 2023: 9th Italian Conference on Computational Linguistics,
Nov 30 – Dec 02, 2023, Venice, Italy

uryupina@gmail.com (O. Uryupina)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).
CEUR Workshop Proceedings (CEUR-WS.org)

¹<https://www.politifact.com/factchecks/2022/apr/12/facebook-posts/no-covid-19-vaccines-do-not-contain-hiv-lipid-wrap/>

²<https://www.politifact.com/factchecks/2022/apr/06/marjorie-taylor-greene/greene-twists-logic-and-facts-pedophilia-charge-a>

Collins and Romney have clear track records of acting against child exploitation, whether online or in person" and, moreover, the implied related accusations of Judge Johnson are "misleading". However, a reader driven by emotions, might still remain manipulated ("no smoke without fire"), even if only partially.

These examples show that fake news are way more complex than simply untrue messages. They might combine true facts with partially false or impossible to check statements, provide biased analytics on top and add very strong emotional messages to manipulate readers. We believe that while the NLP community is making an impressive progress on the fact verification task, our understanding of other phenomena related to manipulative content are still rather limited. The goal of our study is to get a deeper and more realistic insight on the emotional component of fakes. As a first step, we provide a qualitative data-driven analysis of emotion triggering.

The contributions of this study are as follows: (i) we provide data-driven analysis, focusing on real data, combining original (source) fakes and high-quality reports by professional fact-checkers thus improving our insight, (2) we aim at a taxonomy of triggered emotions covering a majority of real-life fakes, departing from more theory-oriented labels and (3) we analyze perceived (i.e., triggered) emotions, as opposed to the common focus on expressed emotions, as we believe that induced sentiment plays a more important role in manipulation/persuasion.

2. Related Work

There is a rapidly growing body of studies on online misinformation detection. These works, however, mainly focus on the verification part (*Is the information truthful – i.e., supported by the evidence?*), and not on the persuasion (*How is the information presented to manipulate the reader?*). Thus, most computational models are built upon the FEVER corpus [1]: a large collection of true/false claims generated by human annotators, annotated as supported/refuted/unknown by the evidence. FEVER claims are originally extracted from Wikipedia (true) and then mutated (false). An example FEVER claim is "Shakira is Canadian". Note a strong difference between this example and (1a-b) above: the Shakira claim was generated with no manipulative purpose in mind and does not involve any specific persuasion/triggering techniques. The claims in (1), on the contrary, have a strong manipulative component and have been generated with a genuine unscrupulous intent. For example, (1a) cannot be fully accounted for by a simple mutation: the choice of "HIV" is crucial to induce fear and thus the same manipulative effect would not be achieved if "HIV lipids" were replaced with any other kind of lipids. In

our study, we focus on real-world data, analyzing fakes generated with a real purpose, albeit not always clear (and not necessarily malicious).

Giachanou et al. [2] address the impact of emotional signals on the credibility for fake news. This study shows that emotional signals are extremely important as the emotion-aware system outperforms their baseline by a large margin. This work, however, focuses on already existing generic resources for defining emotions: either lexicons of terms expressing specific sentiments or a corpus of triggered sentiments with labels corresponding to five different Facebook reactions (love, joy etc). We believe that findings of Giachanou et al. [2] are extremely important and show that emotions triggered by manipulative content should be studied in a more principled way. We hope that our study could help define a more triggering-oriented approach to emotions.

Several recent papers analyze emotion triggering as a part of propaganda persuasion techniques. For example, Da San Martino et al. [3] develop a taxonomy of propaganda techniques, whereas Piskorski et al. [4] propose a shared task build upon this taxonomy. These studies do not, however, focus on emotions specifically. For example, Piskorski et al. [4] group most emotions under the "manipulative" category, while some others (e.g., "appeal to patriotism/pride" also known as "flag-waving") are classified based on reasoning fallacies associated with them. Moreover, these studies focus on unscrupulous persuasion techniques introduced in the theoretical studies, e.g., on (in)formal argumentation fallacies. We advocate a more data-driven approach: the phenomenon of manipulative online content is rather new and evolving, thus, it is not clear how well more traditional labels describe it. We aim at decoupling emotions from (fallacious) argumentation and improving our insight into the variety of sentiments the content writers appeal to.

Finally, some of the discussed triggers, especially "fear", have been a focus of multidisciplinary studies, ranging from psychology (see an overview in [5]) to ethics [6]. At the same time, there exist much less research on more complex triggers.

3. Data

Our study aims at a qualitative analysis with the end goal of developing reliable annotation guidelines that provide good coverage for triggered sentiments. We have therefore opted for in-depth analysis of a small number of documents. Our analysis relies on both the documents themselves and their corresponding fact-checking reports by PolitiFact. This way, we make sure that we ourselves do not fall victim to the manipulation techniques and can assess them impartially.

We rely on PolitiFact reports from mid-March to mid-

INDIVIDUALS WHO WERE IN STANDING ROCK OCTOBER THROUGH NOVEMBER 2016

If you were in Standing Rock the months of Oct to Nov 2016, you were intentionally poisoned by the Governor of North Dakota Jack Dalrymple, Kyle Kirchmeier of Morton County Sheriffs Department and the pilot who knowingly sprayed poisonous chemicals over the Standing Rock Oceti Sakowin and Sacred Stone Camps.

Figure 2: Appeal to fear, from Facebook.

May 2022. We filter out fakes that originate on TV, interviews and other sources outside of social media. This leaves us with 160 "claims", each associated with their corresponding social media post and high-quality PolitiFact report, written by professional fact-checkers. We then annotate them with metadata, overall professional fact-checking judgement, atomic fact veracity, reasoning flaws (e.g., "simplification") and, most importantly, triggered emotions. The latter is done in data-driven bottom-up fashion, with the set of considered emotions under constant refinement.

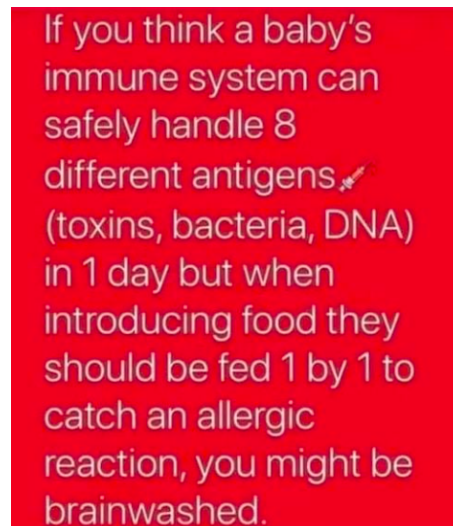
4. Appealing to Emotions

In this section, we discuss the emotions triggered in manipulative online messages. We start with commonly acknowledged and studied triggers, such as "fear" and expand the label set to accommodate data-driven categories not sufficiently covered in the literature.

Appeal to Fear is the most studied and widely used manipulative technique: by making the readers believe that they are in imminent personal danger, the author can influence their attitude toward the message, suppress critical thinking, instill doubt and ultimately manipulate their behavior. There are multiple studies showing the efficacy of this persuasion technique, see [5] for an overview. From the data-driven perspective, however, it is not always easy to define the boundaries of "personal danger".

Thus, our example (1a) shows a clear case of appeal to fear, since the governments' policies strongly suggest all the population to be vaccinated. Consider our example in Figure 2. This post informs a rather limited group of people of the alleged imminent danger, thus inducing fear. However, when going viral, it might have a fear-triggering effect on the whole population, stating that the authorities are able to and, in practice, do employ carcinogenic chemicals against humans.

Bandwagon and Anti-bandwagon. Another relatively widely studied technique is an appeal to common practice/belief ("safe choice"), also known as "bandwagon fallacy". This technique urges the reader to adopt specific



If you think a baby's immune system can safely handle 8 different antigens (toxins, bacteria, DNA) in 1 day but when introducing food they should be fed 1 by 1 to catch an allergic reaction, you might be brainwashed.

Figure 3: Anti-bandwagon (appeal to uniqueness), from Instagram.

choices, because everybody is doing so. For example, bandwagon is commonly used in advertisement, where a lot of products are marketed as a must since everybody buys them. Surprisingly, we haven't found a single example of an appeal to common practice in manipulative online content in our data. However, we have observed the opposite appeal: the authors urge the reader *not* to follow the common practice, appealing to their uniqueness and superiority.

Figure 3 shows a very common example of appeal to uniqueness/superiority: the authors state that while most people are brainwashed by mainstream information channels and left to believe in some fake reality, the readers should – and is definitely capable of – avoid falling for the same trap. This boosts the readers' ego, improves their trust in fake news while, at the same time, undermines mainstream media and paves the path for various conspiracy theories. We have observed this opinion framing strategy on a variety of polarized topics, ranging from vaccination to government spending or climate.

While this direct appeal to readers' uniqueness/ego is very widespread and seemingly rather effective, we are not aware of any in-depth studies of this phenomenon, especially from the NLP perspective.

Appeal to Populism is an emotionally-loaded technique triggering strong antagonising feelings between "us" ("the good people") and "them" ("the corrupt powers: government, rich, media etc"). Populism plays an ever rising role in the modern political discourse, affecting and polarizing people's views. While it is widely studied in

Now you know why there's suddenly a "formula shortage."

The new age robber barons have conveniently invested in some unholy breast milk made from human organs.



HOME > MEDICINE & HEALTH

Bill Gates, Zuckerberg, Other Billionaires Invest in Environmentally-Friendly Artificial Breast Milk Cultured From Human Mammary

Figure 4: Appeal to populism, from Facebook.

political science, the related psychological mechanisms are still underresearched [7]. We have observed multiple cases of appeal to populism throughout the data.

Thus, in a Facebook post on Figure 4, the author makes it pretty clear that the rich are responsible for and benefiting from the suffering of "us" – in this specific case, the formula milk crisis. The same strategy is used throughout our data to implicate different kinds of powers: the administration, the rich or the media and sometimes a mixture or just a generic/underspecified "power". The appeal to populism is often combined with other emotions: for example, triggering the fear or unfairness/injustice for the outcome of "their" actions as well as uniqueness/ego for uncovering the plot.

Appealing to (Un-)Fairness is a very strong technique, often used in combination with appealing to populism (see an example on Figure 5).

In some cases, the authors trigger this sentiment in a positive way, inviting the reader to celebrate the victory of fairness.

In both cases, however, the content writers trigger a very strong and deep desire for (social) justice, that deflecting the readers' attention from inconsistencies and misrepresentation in the presented facts and arguments.

To our knowledge, appealing to fairness is acknowledged as a powerful technique by a variety of practicing professionals, e.g., negotiators or copywriters. However, there is still virtually no research on this specific emotion. We believe that since this is one of the most frequent and



Figure 5: Appeal to unfairness, from Facebook.

Big talk...he's hoping NC Republicans will forget that when he was Governor, @PatMcCorryNC appointed the "Republican" judge who sided with Democrats in the partisan Democrat lawsuit/power-grab over redistricting. #ncsen #ncpol

(a) attacking a specific person, Twitter

This DID NOT happen yesterday. The south lawn of the WH was completely empty, no chairs or anything set up. ABSOLUTELY nothing. Also, it was FREEZING cold yesterday. 30° and extremely windy. I was bundled up and my cheeks were numb after about 15 min. More lies. FAKE NEWS.

President Biden and Vice Pr... See More



(b) undermining trust ("everybody lies"), Facebook

Figure 6: Appeal to honesty.

efficient triggers in manipulative content, an urgent attention from the research community, including NLP, might have a considerable impact and help fight online misinformation.

Appeals to honesty are very popular in manipulative content. This category includes allegations of hypocrisy, inconsistency or accusations of lying, aimed at casting a doubt on specific persons (Figure 6a).

However, a far more widespread appeal to honesty is the technique where some information coming from mainstream media or official sources is presented as a lie,



Figure 7: Appeal to values, from Facebook.

with no clear and specific purpose (Figure 6b). This type of fakes promote the idea of everything being unreliable and slowly but steadily push the readers to become less critical of various conspiracy theories.

Values. Certain online posts make appeal to values, promoting responsible choices or condemning someone else’s behavior as unethical. This type of triggering is often used in polarized contexts to attack the opposite side and thus misrepresent their position (Figure 7).

Appeals to values are often used as a part of the reduction/simplification fallacy: the fact-checkable facts in the message are true (e.g., the statement above is focused on "A National Terrorism Advisory System bulletin", addressing the threats of online misinformation), yet their interpretation is fallacious and manipulative, introducing loaded lexica ("attack", "criminalize") to misrepresent these facts, substituting objective reporting with moralistic judgement. This type of fakes are therefore particularly problematic for state-of-the-art NLP models, based on fact verification.

Disasters. We have observed a large number of fakes focusing on natural and man-made disasters. Media coverage of disasters has been shown to attract a large number of readers/viewers, triggering a wide variety of inter-related negative emotions, in particular fear and anxiety [8]. Unscrupulous content generators abuse the users’ interest in catastrophic events for their own purposes (e.g. click-bait). We label this specific type of fear/anger as "disaster" for the lack of better term, since a more precise analysis is still an open research issue in psychology.

5. Emotions in Fakes

In this section, we discuss the distribution of triggered emotions in the manipulative content collected and analyzed by PolitiFact. Most importantly, we have observed that a vast majority of fakes trigger emotions: 128 documents (80%) in our collection unambiguously aim at affecting the readers’ emotional state. For comparison, only 88 documents (55%) contain clearly untrue atomic facts and 95 documents (59%) employ fallacious argumentation. We believe, once again, that these numbers suggest that the efficient approach to manipulative content analysis should expand from mere fact verification to modeling fallacious argumentation and emotion triggering.

Trigger	#documents	%
populism	62	38.7
fear (personal)	18	11.3
fear (empathy)	16	10
fairness	27	16.9
honesty	22	13.8
values	15	9.4
uniqueness	18	11.25
disaster	8	5
other	6	3.8

Table 1
Triggers in the PolitiFact data.

Table 1 shows the document statistics for each of the triggers discussed in this section. The most common category is *populism*, which might be due to the political orientation of our domain. Note that populism is also relatively easy to identify: our preliminary experiments show very little disagreement on this label. Appeal to *fear* is the second most popular category: unscrupulous content writers are well aware of its efficiency. Annotating it reliably, however, requires extra work on guidelines, since the boundaries between personal fear and empathy for others are very subjective. Depending on the definition of fear, we observe 11-21% of such documents. Fairness, honesty and values are also rather common. Finally, only 6 documents (4%) appeal to other emotions that are not covered by our taxonomy.

The same post can trigger multiple emotions. In particular, appeals to populism ("they are bad") are often combined with any other trigger ("they are bad: they are threatening our existence, imposing unfair policies and lying"). A rather common combination throughout all the fakes we have analyzed is "*they* (the media/administration) are lying, but *you* are smart and *you* don’t believe them, *we* will tell you the truth" (anti-bandwagon + honesty + populism). Note that this trigger makes it very difficult to respond to and counter the effect

of manipulative content: if the readers are convinced that "they" are lying, they can simply discard a fact-checking report since they perceive fact-checkers as liars (paid by "them") or, at the very least, brainwashed (by "them").

6. Conclusion

This study focuses on emotional component of manipulative online content. Analyzing real-life fake content from PolitiFact, we have observed a variety of emotional triggers used to promote unscrupulous content by agitating the users and making them less critical of the deficiencies in the fact selection and argumentation of the manipulative discourse.

We have seen that emotions play a crucial role in pushing through different kinds of manipulative agenda and it is therefore extremely important for the scientific community to extend state-of-the-art verification-based approaches to fact-checking and incorporate models for emotion triggering and fallacious argumentation.

Our study identifies the most common types of emotions triggered by manipulative content. However, defining them accurately is not a trivial task, as we have already observed with *fear*. Our current work focuses on refining the definitions of the most common triggers to provide reliable annotation guidelines and create a dataset of appeals.

Triggered emotions (reactions) have so far mostly been out of the scope of the NLP community, where the vast body of research is focused on emotions *expressed* in the document. We believe that our research can contribute to a better understanding of perceived emotions, crucial for modelling a text's impact on the reader. In particular, we plan to study the relation between expressed and triggered emotions and investigate possibilities of transferring high-performing state-of-the-art (expressed) emotion recognition models to account for triggered emotions.

Finally, we believe that multi-factor understanding of manipulative content is essential to generate adequate response and thwart the misinformation. Emotionally-loaded fakes are particularly hard to debunk since they render the user less receptive to the rational argumentation of fact-checkers. As a part of our future work, we want to investigate strategies for automatic response generation that take into account the emotional component and try to produce an adequate reaction, regaining the users' trust.

Acknowledgments

We thank the Autonomous Province of Trento for the financial support of our project via the AI@TN initiative.

References

- [1] J. Thorne, A. Vlachos, C. Christodoulopoulos, A. Mittal, FEVER: a large-scale dataset for fact extraction and VERification, in: Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers), Association for Computational Linguistics, New Orleans, Louisiana, 2018, pp. 809–819. URL: <https://aclanthology.org/N18-1074>. doi:10.18653/v1/N18-1074.
- [2] A. Giachanou, P. Rosso, F. Crestani, The impact of emotional signals on credibility assessment, Journal of the Association for Information Science and Technology 72 (2021) 1117–1132.
- [3] G. Da San Martino, S. Yu, A. Barrón-Cedeño, R. Petrov, P. Nakov, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 5636–5646. URL: <https://aclanthology.org/D19-1565>. doi:10.18653/v1/D19-1565.
- [4] J. Piskorski, N. Stefanovitch, G. Da San Martino, P. Nakov, SemEval-2023 task 3: Detecting the category, the framing, and the persuasion techniques in online news in a multi-lingual setup, in: Proceedings of the The 17th International Workshop on Semantic Evaluation (SemEval-2023), Association for Computational Linguistics, Toronto, Canada, 2023, pp. 2343–2361. URL: <https://aclanthology.org/2023.semeval-1.317>.
- [5] M. Tannenbaum, J. Hepler, R. Zimmerman, L. Saul, S. Jacobs, K. Wilson, D. Albarracín, Appealing to fear: A meta-analysis of fear appeal effectiveness and theories, Psychological Bulletin 141 (2015) 1178–1204.
- [6] D. Arthur, P. Quester, The ethicality of using fear for social advertising, Australasian Marketing Journal (AMJ) 11 (2003) 12–27. URL: <https://www.sciencedirect.com/science/article/pii/S1441358203701153>. doi:[https://doi.org/10.1016/S1441-3582\(03\)70115-3](https://doi.org/10.1016/S1441-3582(03)70115-3), social Marketing.
- [7] S. Obradović, S. Power, J. Sheehy-Skeffington, Understanding the psychological appeal of populism, Current Opinion in Psychology 35 (2020) 125–131.
- [8] B. Pfefferbaum, E. Newman, S. Nelson, P. Nitiéma, R. Pfefferbaum, A. Rahman, Disaster media coverage and psychological outcomes: descriptive findings in the extant research., Current Psychiatry Reports 16 (2014) 1178–1204.