Multimodal Online Manipulation: Empirical Analysis of Fact-Checking Reports

Olga Uryupina[1]

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

Abstract

This paper presents an in-depth exploratory quantitative study of the interaction between multimedia and textual components in online manipulative content. We discuss relations between content layers (such as proof or support) as well as unscrupulous techniques compromising visual content. The study is based on fakes reported and analyzed by PolitiFact and comprises documents from Facebook, Twitter and Instagram. We identify several pervasive phenomena currently, affecting the impact of manipulative content on the reader and the possible strategies for effective de-bunking actions, and discuss possible research directions.

Keywords

fact checking, multi modal, annotation,

1. Introduction

Manipulative online content (fake news, propaganda, among others) is growing at an alarming rate, hindering our access to truthful and unbiased information and thus threatening principles of the democratic society. The problem has been addressed by professional journalists, who – with the help of crowd-workers – fight a never-ending battle to prevent information contamination. To enable a large-scale response to the misinformation threat, the AI community has invested a considerable effort into building competitive models for identifying non-transparent content, such as false claims or altered videos (deep fakes). However, we still lack a thorough understanding of the manipulative content and multiple aspects affecting its perception and impact on the reader. This paper aims at an in-depth analysis of one of such aspects, namely, the interaction between different (multimedia) layers of the manipulative message. More specifically, we study the semantics underlying the relation between multimedia and textual parts of the fake news. Our study is based on around 800 fakes from January till September 2022, as identified and analysed by PolitiFact.[1]

Multimedia content, such as videos, reels, photos, screenshots or images is becoming increasingly popular in social media: it is an appealing and powerful way of expressing and/or enhancing one's message. Nevertheless, as a scientific community, we still have little understanding of the way the authors integrate multimedia into their content: most research so far has focused on a specific component and not on their interplay. Our study aims at identifying the role of multimedia part of manipulative messages.

Figure 1 shows some examples from potential fakes analyzed by PolitiFact. We observe different relations between the text and the image. In particular, in (1a), the video is supposed to *prove* the claim by providing direct evidence, whereas in (1b), the image provides a *support* (appeal to authority). In (1c), the image is a visual *paraphrase* of the claim, enhancing its appeal but not providing extra proof, support or informational material. Finally, in (1d), the photo is an *illustration* that, while depicting the discussed person, does not aim at being relevant to the claim's veracity or impact. While understanding the relation between the image and the text is interesting from the scientific perspective, it is also a crucial prerequisite for efficient and meaningful fact-checking response. For example, if a supposed *proof* is a compromised photo, the response should highlight this fact (e.g., the video in (1a) has been cropped misrepresenting the quote, which should be highlighted in the fact-checking report). On the contrary, if a compromised photo is used as a mere *illustration*, the effective fact-checking report should focus on the textual claim per se.

Another important angle is the issue with the multimedia part. In our example, the video in (1a) is *cropped*. On the contrary, (1b) represents an authentic screenshot, yet, it has been *miscaptioned* by the claim: an older content, irrelevant for the current events/topics, has been repurposed.

The current paper focuses on these two aspects to analyze empirically the interplay between multimedia and textual components in fake news, as identified by Politi-

CLiC-it 2024: Tenth Italian Conference on Computational Linguistics, Dec 04 — 06, 2024, Pisa, Italy

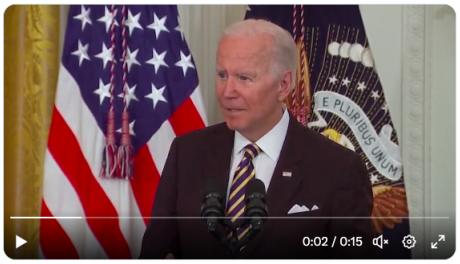
uryupina@gmail.com (O. Uryupina)

© 2024 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

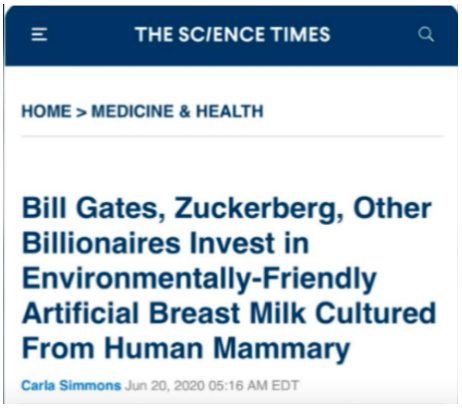
[1]PolitiFact (https://www.politifact.com/) is an independent journalistic agency and one of the most experienced fact-checking organizations, providing detailed analytics for non-transparent online content since 2007.

CEUR-WS.org/Vol-3878/114_main_short.pdf

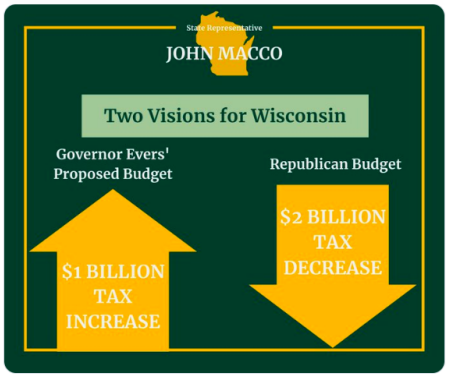
CEUR Workshop Proceedings ceur-ws.org ISSN 1613-0073



(a) Biden to teachers: “They’re not somebody else’s children. They’re yours when you’re in the classroom." (VIDEO)



(b) Now you know why there’s suddenly "a formula shortage". The new age robber barrons have conveniently invested in some unholy breast milk made from human organs.



(c) In honor of #TaxDay, I remind you that Governor Evers wanted to increase your taxes by $1 billion just for heating your homes. Instead, Republicans cut your taxes by more than $2 billion.

(d) Italian football agent Mino Raiola has died after suffering from an illness. RIP

Figure 1: Different uses of layered/multimedia content

Fact. To this end, we reannotate the PolyFake dataset [1] with fine-grained labels reflecting multimedia aspects.

2. Related Work

While fact checking has been receiving an increasing amount of attention recently both from NLP and Vision communities, only very few studies focus on the interaction between different modalities.

A breakthrough approach by Vempala and Preoţiuc-Pietro [2] focuses on two dimensions of the relationship between text and image on Twitter: whether the text is represented in the image and whether the image adds extra content to the textual message. Cheema et al. [3] propose a dataset of multimodal tweets, annotated for visual relevancy and checkworthiness. Finally, Biamby et al. [4] propose a larger-scale dataset of multimodal tweets, where "falsified" claims have been added synthetically to address the image repurposing problem.

These studies have paved the way for evaluation campaigns and benchmarking resources, for example, [5]. Yet, these studies rely on rather straightforward annotation guidelines to reduce the per-claim cost. Moreover, the annotators are not professional fact-checkers: while they can assess some aspects of the compromised content, they still can get deceived by more challenging cases – after all, the manipulative content has been created on

Layer	Facebook		Twitter		Instagram		TikTok		YouTube		Total	
none	64	12.7%	80	41.9%	4	3.9%	-	-	-	-	149	18.2%
video	195	38.6%	25	13.1%	40	38.9%	11	100%	6	100%	277	33.9%
photo	92	18.8%	31	16.2%	10	9.7%	-	-	-	-	133	16.3%
screenshot	114	22.5%	19	9.9%	45	43.7%	-	-	-	-	178	21.8%
link	29	5.7%	15	7.8%	-	-	-	-		-	44	5.4%
image	14	2.8%	6	3.14%	6	5.8%	-	-	-	-	26	3.2%
thread	-	-	17	8.9%	-	-	-	-	-	-	17	2.1%
total	506	100%	191	100%	103	100%	11	100%	6	100%	818	100%

Table 1
Types of layered content.

purpose to influence and bias the reader.

In a recent survey, Mubashara et al. [6] highlight the importance of an interdisciplinary approach to fact-checking, proposing a framework to model different axes of online manipulation, most importantly, fusing the textual and visual fact-checking and survey benchmarks and models developed by respective communities. Our study is built upon the same motivation – and our main goal is to study empirically the interplay between different modalities, based on real-world (i.e., not simulated or synthesized) fakes data.

Our study aims at an in-depth exploratory analysis of the multimodal online content. To this end, we focus on more specific labels to describe the relationship between different layers/modalities. We extend the scope of our study to cover all the three major platforms (Facebook, Instagram and Twitter). Moreover, our input is not only the claim per se, but the professionally created fact-checking report from PolitiFact. In our experience, PolitiFact reports contain a wealth of information about online manipulation: as opposed to 2-3 binary labels of common NLP fact-checking benchmarks, PolitiFact characterizes each claim with 1-3 pages of analytics. This analytics, however, comes in a free textual form. While it might be still impossible for the NLP community to encode these reports for building high-quality fact-checking systems, we believe that we should at least learn from them to get better insights, stop trivializing the task and highlight understudied, yet impactful, subtasks.

3. Analyzing Multimedia Content

3.1. PolyFake

Our study is based on the PolyFake dataset [1] covering fake news from 2022, as analyzed by professional fact-checkers from the PolitiFact agency.[2] The current study is based on the first nine months of PolyFake (818 entries). Each entry has been re-assessed by two annotators, with further adjudication by the supervisor. The original PolyFake labels are binary and encode more generic properties of fake news (e.g. whether the reasoning is fallacious or whether the document triggers emotions). For the present study, we have designed and iteratively refined annotation guidelines for labelling multimedia aspects of manipulative content.

The annotation process is based on consulting jointly not only the original content, but the PolitiFact report as well. This way we make use of the wealth of analytics provided by experienced professional fact-checkers by encoding it in more structured annotation labels.

PolyFake covers fakes from different social media (Twitter, Facebook, Instagram, TikTok, Threads and YouTube). Note that manipulative content often gets propagated across platforms through re-posts, sharing, linking or just copying. For example, a large proportion of Facebook videos originates from TikTok (in this case, PolitiFact typically analyzes the Facebook message, hence a low number of TikTok entities in the table). In the following study, we omit TikTok, YouTube and Telegram as largely underrepresented categories with rather straightforward patterns.

3.2. Multimedia and Layered Content

Layer Types. Table 1 shows the distribution of different media types for each platform. We have identified several types of layered content: parts of the message rendered together with the initial post. The most common ones are *videos* (including reels), *photos* and *screenshots* (typically, complex visual objects combining textual content with photos/images and referring the reader to a different source). We have also observed *images* (infographics, maps or drawings), *links* (this content typically is rendered with a photo/stillshot, yet it explicitly points to a different online location, for example, promotion website) or *threads* (characteristic for Twitter, this type of layering helps to contextualize the message). On rare occasions, social media posts might contain more than

[2]PolyFake annotation guidelines cover a wide range of phenomena related to online manipulation: from fallacious/propaganda reasoning to emotive appeals, factual veracity etc. Current study aims at an in-depth analysis of a specific angle. The Appendix discusses the distribution of veracity labels across PolyFake documents.

role	video		photo+		screensh.+	
	total	%	total	%	total	%
content	66	23.8	19	12.0	114	48.1
anchor	62	22.4	46	29.1	16	6.8
proof	86	31.0	36	22.8	39	16.5
support	14	5.1	4	2.5	16	6.8
paraphr.	30	10.8	6	3.8	23	9.7
context	8	2.9	3	1.9	21	8.9
illustr.	1	0.4	55	34.8	9	3.8
action	3	1.1	1	0.6	14	5.9
other	28	10.1	-	-	2	0.84
total	277		158		237	

Table 2
Role of mulimedia layers, per content type (photo+ includes photos and images, screenshot+ includes screenshots, links and threads/retweets), purely textual documents discarded.

Issue	video		photo+		screenshot+	
falsehood	93	33.6%	16	10.12%	130	54.9%
crop	12	4.3%	-	-	1	0.4%
miscaption	60	21.7%	47	29.7%	15	6.3%
altered/fake	17	6.1%	15	9.5%	29	12.2%
misperception	7	2.5%	5	3.2%	-	-
noproof	27	9.7%	3	1.9%	5	2.1%
explain	26	9.4%	6	3.8%	12	5.1%
none	13	4.7%	58	36.7%	43	18.1%
	277		158		237	

Table 3
Types of manipulative content for different multimedia layers.

one extra layer (e.g., videos and photos).

Most importanly, only 18% of PolyFake documents are purely textual: adhering to the popular adage that a picture is worth a thousand words, manipulative content creators use visuals for a variety of purposes, from increasing the outreach to improving the credibility. Moreover, the prevalence of multimedia content is way more critical for Facebook and Instagram – the two platforms not typically addressed by NLP practitioners. This alone suggests that we need to pay much more attention to joint models and start with deeper understanding of relevant phenomena.

A large percentage of documents are re-using or spreading already existing information. This is true for screenshots (21% in total) and links (5%), but also for many videos – only very few videos represent original content. While there exist some studies on identifying previously fact-checked claims, they are restricted to the textual content. We believe that a more complex multimodal approach would be beneficial here.

For presentation issues, in what follows we merge our underrepresented categories *link*, *image* and *thread* with roughly functionally similar major categories *screenshot*, *photo* and *screenshot* respectively.

Layer Roles. Table 2 shows different roles multimedia levels play in PolyFake documents. We distinguish between the following roles: *content* (the essential part of the content is presented on the multimedia layer, whereas the textual layer just adds minor details or suggests opinions), *proof* (the multimedia layer offers a physical proof – cf. Example (1a)), *support* (the multimedia layer provides some material to support the claim, from a reputable source – cf. Example (1b)), *paraphrase* (the multimedia layer paraphrases the claim without adding any extra angle – cf. Example (1c)), *context* (while the textual claim is generally self-contained, it cannot be interpreted without the context given by the multimedia part (e.g., the claim contains pronouns and the image presents their referents)), *illustration* (the multimedia layer shows some objects/persons mentioned in the claim without any connection to its semantics – cf. Example (1d)) and *action* (the multimedia layer suggests an appropriate reaction to the claim, for example, a scam website). Finally, a rather common role for videos and photos is *anchor*: in such cases, the textual claim is about the multimedia itself (for example, "the sharpest image of the sun ever recorded."; here, the multimedia is not compromised per se and the textual claim contains no falsehoods about the world, yet the combination might be very misleading.

In more than half of the documents, multimodal layers

provide essential content. This is true for all the media types (videos, photos and screenshots). We have observed several possible factors contributing to this effect: in general, social media users tend to repost existing "fancy" content and not create their own texts. Even in authentic self-created posts, the message is often put in a visual, whereas only some emotions are added in a text. We believe that there is a wide variety of potential reasons for this behaviour (e.g., videos and photos get more *likes*, whereas texts are mostly ignored by peers), requiring a more specialized study.

Almost one third of multimedia layers, especially videos, supposedly present proofs. Such compromised proofs are out of reach for the modern evidence-based automatic fact-checking: while a fact-checking model can provide extensive evidence to refute a claim, the user would still trust the video/photo and not the model. Human fact-checkers address such proofs from a different, more promising, perspective: they try to explicitly attack and debunk the proof. We believe that this is a very important and largely unaddressed research direction.

Issues with multimedia layers. Finally, we have identified the most common unscrupulous techniques relevant for multimedia layers. Those include: *crop* (essential part(s) of the original message are omitted to render it out of context – cf. Example (1a)); *miscaption* (while the image/video is authentic, the textual claim misleads w.r.t. some crucial details, e.g. events or timeline – cf. Example (1b)); *altered/fake* (the image/video has been altered – beyond cropping – with the specialized software, including deep fakes); *misperception* (the image/video is – deliberately or not – deceiving because of its low quality, unclear angle, optical effects etc); *noproof* (the – typically long – video does not contain any components relevant for the claim); *falsehood* (the video/image is authentic, yet its content is untrue – i.e., the textual claim spreads the original fake generated by the video/image); and *explain* (the textual part explains – misleadingly – what we are supposed to see in the video, often of a rather low quality).

Table 3 summarizes the distribution of problematic issues across the three main multimedia types, showing several trends. First, video layers provide more possibilities for unscrupulous content generators: cropped, otherwise altered or low quality videos are pervasive in manipulative content. While most of the research focuses on images, they do not exhibit such a variety of manipulative strategies. Screenshots – authentic or fake – are largely used to disseminate falsehoods. At the same time, an increasing amount of authentic videos, mostly originating from TikTok, is created to spread falsehoods and promote "critical thinking" (i.e., conspiracy theories as opposed to rational argumentation). These remain largely understudied, despite their large impact on the audience. Another rather unstudied area are explanatory claims: authentic videos/photos accompanied by misleading explanations of what we see and what it means; in such cases, the factual component might be non-compromised, yet the biased explanation makes the whole message an impactful and hard to debunk propaganda tool. Finally, unlike videos and screenshots, most photos represent true authentic information – the textual claims either rely on them as illustrations or use them as building blocks to support fallacious argumentation.

4. Conclusion

We have presented an in-depth analysis of the interaction between textual and multimedia components of compromised social media documents. We have identified several high-impact issues, insufficiently studied by the community at the moment. These include the interaction between different modalities, the role of the multimedia part and its impact on selecting the successful fact-checking strategy, the difference between platforms and media types (current NLP studies predominantly focus on Twitter and images) and the importance of a more principled approach to content re-use. We hope that this study, motivated by human fact-checking expertise, can sparkle a meaningful discussion and improve automatic modeling.

Acknowledgments

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

References

[1] Anonymous, PolyFake: Fine-grained multi-perspective annotation of fact-checking reports, in: Accepted for publication, 2024.

[2] A. Vempala, D. Preoţiuc-Pietro, Categorizing and inferring the relationship between the text and image of Twitter posts, in: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Florence, Italy, 2019, pp. 2830–2840. URL: https://aclanthology.org/P19-1272. doi:10.18653/v1/P19-1272.

[3] G. S. Cheema, S. Hakimov, A. Sittar, E. Müller-Budack, C. Otto, R. Ewerth, MM-claims: A dataset for multimodal claim detection in social media, in: Findings of the Association for Computational Linguistics: NAACL 2022, Association for Computational Linguistics, Seattle, United States, 2022, pp. 962–979. URL: https://aclanthology.org/2022.findings-naacl.72. doi:10.18653/v1/2022.findings-naacl.72.

FC label	Facebook		Twitter		Instagram		TikTok		YouTube		Total
pants-on-fire	95	18.6%	18	9.4%	29	28.2%	2	18.2%	2	33.3%	146
false	353	69.8%	97	50.8%	64	62.1%	9	81.8%	2	33.3%	526
mostly false	34	6.7%	36	18.8%	6	5.8%	-	-	1	16.7%	77
half true	17	3.3%	18	9.4%	-	-	-	-	1	16.7%	36
mostly true	6	1.2%	11	5.7%	3	2.9%	-	-	-	-	20
true	1	0.2%	10	5.2%	1	1.0 %	-	-	-	-	12
total	506	100%	191	100%	103	100%	11	100%	6	100%	818

Table 4
Manipulative content on social media fact-checked (FC) and reported by PolitiFact (Jan-Sept 2022).

[4] G. Biamby, G. Luo, T. Darrell, A. Rohrbach, Twitter-COMMs: Detecting climate, COVID, and military multimodal misinformation, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States, 2022, pp. 1530–1549. URL: https://aclanthology.org/2022.naacl-main.110. doi:10.18653/v1/2022.naacl-main.110.

[5] A. Bondielli, P. Dell'Oglio, A. Lenci, F. Marcelloni, L. Passaro, Dataset for multimodal fake news detection and verification tasks, Data in Brief 54 (2024) 110440. URL: https://www.sciencedirect.com/science/article/pii/S2352340924004098. doi:https://doi.org/10.1016/j.dib.2024.110440.

[6] A. Mubashara, S. Michael, G. Zhijiang, C. Oana, S. Elena, V. Andreas, Multimodal automated fact-checking: A survey, 2023. arXiv:2305.13507.

A. True vs. Fake content and multimedia layers

Our dataset by construction contains mostly untrue claims: even though PolitiFact occasionally fact-checks statements that turn out to be true, most of their materials are "false", "mostly false" or even "pants on fire". Moreover, even true claims often exhibit signs of user manipulation. In this appendix, we show statistics for fake vs. true content in PolitiFact reports (Table 4).