Life and Death of Fakes: on Data Persistence for Manipulative Social Media Content

Olga Uryupina[1]

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

Abstract

This work presents an in-depth investigation of the data decay for publicly fact-checked online content. We monitor compromised posts on major social media platforms (Facebook, Instagram, Twitter, TikTok) for one year, tracking the changes in their visibility and availability. We show that data persistence is an important issue for manipulative content, on a larger scale than previously reported for online content in general. Our findings also suggest a (much) higher data decay rate for the platforms suffering most from online disinformation, indicating an important area for data collection/preservation.

Keywords

fact checking, replicability,

1. Introduction

Manipulative online content is rapidly becoming a more and more pervasive issue for the modern society: by deliberately biasing our information flow, unscrupulous content writers can and do affect our emotional state, beliefs, reasoning and both online and offline behaviour. It is therefore not surprising that this has become a central issue for various stakeholders, from journalists and fact-checkers to NLP researchers both in academia and in the industry. Given the current rapid growth in data-driven studies of manipulative content, it is essential to have a reliable overview of data persistence issues in this specific domain: compromised content is often very dynamic and changes or becomes unavailable over time, raising reproducibility concerns,

From the readers' perspective, the visibility of compromised content over time affects directly its impact: a removed or strongly downgraded document is unlikely to be read/recovered and cannot be used to promote or support other fakes. From the research and development perspective, data persistence is crucial for benchmarking, ensuring fair comparison between models as well as even simply providing them with high-quality real-life training and testing examples.

Starting from already a decade ago, NLP benchmarking campaign studies [1] report data persistence issues for online content, as used in various shared tasks, reporting around 10% of entries missing compared to the original dataset (gold standard). These shared tasks, however, are based almost exclusively on Twitter and do not focus specifically on compromised content. We believe that a large proportion of manipulative content is created on purpose by professional copywriters who might have different goals and motivations to keep their texts online (e.g., for click-bait purposes) or remove them (e.g., to reduce the reputation loss from being exposed as unreliable).

Our work focuses specifically on the lifespan of fact-checked compromised content. We go beyond the naive binary *present* vs. *removed* view, studying more nuanced cases as well. In particular, we track compromised online posts over time for the appearance of explicit platform-specific reliability labels (e.g. "out of context"), obfuscation (the common situation when the online content is – fully or partially – rendered either very blurred or as a black/white box, with a message raising awareness of its limited reliability; this content, however, is still accessible to the user upon an extra click), and author-generated edits, as well as complete content removal.

More specifically, we address the following research questions:

RQ1: How persistent is the compromised content? How does its visibility and availability change over time?

RQ2: What is the typical timeline for interaction between the content generators and fact-checkers? How – if at all – do content writers alter their posts after being exposed as problematic by fact checkers?

RQ3: Are the trends different across platforms?

To this end, we analyze two datasets (in English) of social media documents, fact-checked by PolitiFact.[1]

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

uryupina@gmail.com (O. Uryupina)

[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.

2. Related Work

Multiple studies report on data persistence issues for online content. These works, however, mostly focus on Twitter datasets, as used for various challenges and shared tasks.

Zubiaga [2] provides an exhaustive report on data persistence for multiple Twitter datasets, showing an average data decay of around 20% over 4 years.

Küpfer [3] argues, always for Twitter, that data persistence is not random, becoming drastically more of an issue for emotionally charged or controversial content. Indeed, both Bastos [4] and Duan et al. [5] report much higher tweet decay rates for #Brexit and #BlackLivesMatter, content respectively.

To our knowledge, there have been no studies assessing explicitly data persistence issues for fakes. For some datasets, the creators provide estimations of content decay. For example, Bianchi et al. [6] estimate that around 25% of the tweets in their corpus on harmful speech online were no longer available at the paper publication time. It is, however, unspecified, how this estimation was obtained.

We hope to bring new insights to our understanding of the data persistence issues for compromised content by addressing the following novel angles: (i) we aim at a targeted analysis of manipulative content (fake news), (ii) we provide a more nuanced approach, tracking subtler changes in data availability for users and machines (e.g., obfuscation) and (iii) we go beyond Twitter, targeting all the major social media platforms.

3. Data

For our study, we use two data sets of real-life suspicious online posts, analyzed by PolitiFact. A 2-months dataset (PolitiFact reports from 15 May – 15 July 2023, around 200 entries) has been thoroughly monitored for data visibility and persistence up till now. A larger and older dataset (PolitiFact reports from January – September 2022, around 800 entries) has been analyzed twice to assess longer-term trends.

The two datasets include all the posts in English from the major social media platforms as reported by PolitiFact during the above mentioned periods (i.e., the original publications slightly predate May 15, 2023 and Jan 1, 2022, respectively).

The analysis involves the following dimensions:

- **visibility:** visible (possibly with a warning), obfuscated, removed;
- **persistence:** original, edited, removed;
- **extra labelling:** any platform-specific add-ons, e.g. "missing context".

source	total docs	min fc time	max fc time	median fc time
all	192	0	56	4
fb	86	1	56	4
twitter	16	1	30	4
tiktok	17	1	30	6
instagram	72	0	44	4

Table 1
Assessing the time required for professional fact-checking (fc): statistics for the 2-month dataset, days.

While some of these aspects are crucial for algorithmic NLP (e.g., data persistence is important for benchmarking and – in critical cases – even training ML models), others are more relevant for understanding the impact of manipulative content on human readers (e.g., obfuscation is an unambiguous warning the platform sends to the reader on a low reliability of the information).

The 2-months dataset has been analysed every two days for the first two months and then on a weekly basis for the following year. The 8-months dataset has been analyzed in May and October 2024, when the documents were 1.5-2 and 2-2.5 years old respectively.

4. Compromised content: timeline

4.1. From publication to fact-checking

For this project, we start monitoring the content the day it appears on PolitiFact. Obviously, this doesn't happen the very moment the content gets published by its creators: it takes some time for the content to reach PolitiFact and then an extra period to perform fact-checking. This lag may depend on numerous factors: for example, some fakes are simple and repetitive, thus requiring less investigative effort, whereas some others lead PolitiFact journalists to request third-party expert analytics, involving time-consuming communications with various public figures and organizations.

Table 1 shows time lag statistics (in days) between the content publication date (as reported by the platforms) and the appearance of the corresponding fact-checking report. It suggests that PolitiFact is doing an outstanding job at timely reacting to online misinformation: an average suspicious post is analyzed in 4 days, with a large bulk of reports appearing on the next day already. We observe no platform-based difference in PolitiFact reaction times, thus confirming their neutrality in this respect.

PolitiFact stays in active collaborations with major social media platforms.[2] As a result, in most cases the content is marked by the platform as somewhat spurious

[2] For example, https://www.facebook.com/help/1952307158131536?helpref=related and https://www.tiktok.com/safety/en/safety-partners/

	% d0	% d7	% d30	% d100	% d365	total
all	88.02%	80.72%	75.52%	69.27%	61.97%	192
fb	83.72%	80.23%	75.58%	70.93%	63.95%	86
twitter	93.75%	93.75%	87.5%	93.75%	93.75%	16
tiktok	94.11%	82.35%	76.47%	64.7%	58.82%	17
instagram	90.27%	77.77%	72.22%	63.88%	54.16%	72

Table 2

Statistics for the 2-moths dataset: data availability at fact-checking day and one week, 1, 3 and 12 months afterwards: % of available (visible or obfuscated) documents.

	% day0	% day7	% day30	% day100	% day365	total
all	48.43%	46.87%	43.22%	40.1%	36.97%	192
fb	41.86%	39.53%	36.04%	32.55%	27.9%	86
twitter	93.75%	93.75%	87.5%	93.75%	93.75%	16
tiktok	94.11%	82.35%	76.47%	64.7%	58.82%	17
instagram	34.72%	36.11%	33.33%	31.94%	30.55%	72

Table 3

Statistics for the 2-months dataset: data visibility at fact-checking day and one week, 1, 3 and 12 months afterwards: % of visible documents.

(e.g. "false" or "out of context") shortly after or even before the publication on the PolitiFact website. This marking, as we will see below, often leads to immediate content modification or withdrawal.

4.2. Content availability after fact-checking

Tables 2 and 3 illustrate data availability over time for the 2-months set. We distinguish between two categories: visible and available. Available content can be accessed by either a human or a machine, possibly with some effort (e.g., an extra click). Visible content can be accessed as-is. In other words, non-visible accessible content includes fully or partially obfuscated posts.

We see several important trends here. First of all, already at the fact-checking date, around 12% of documents are no longer available. This number grows rapidly: after one year, the unavailable content comprises 38% of datapoints for our 2-month set.. This number is much more pessimistic than common estimations of online data persistence [2]. This raises an important and a very urgent issue: as a community, we should invest a more focused and consistent effort in timely saving samples of compromised documents for ongoing and future research/benchmarking. From the human reader perspective, only one third of posts are clearly visible after one year (and even in such cases, they might contain explicit markings, such as "partially false").

We also observe a striking difference across platforms: while most tweets remain online, almost a half of compromised Instagram posts are no longer available after 12 months. This is truly problematic: while the NLP community focuses mainly on Twitter data, fakes on other platforms are more prevalent—and keep appearing and disappearing at an alarming rate, leaving us virtually no opportunity to model the underlying trends.

4.3. Content adjustment

As we have seen above, once a document has been fact-checked and deemed false, the most typical reaction is its – rather fast – removal. This would be a rather natural reaction: most creators do not enjoy having their content (and their name) marked as unreliable. In some cases, however, the users[3] prefer keeping the compromised content online. Such content – proven do be problematic by a publicly available fact-checking report – would trigger a reaction from (a) the hosting social media platform, (b) the community and (c) the authors themselves. The observed reactions for *visible* documents are summarized in Table 4.

Facebook and Instagram adopt their own labels to mark questionable content, distinguishing between "false", "out-of-context" and "partly false" documents.[4] Although PolitiFact stays in an active collaboration with the both platforms, there is no direct correspondence between the labels. The labels get assigned rather quickly and stay unchanged (almost all of the observed label change is due to the complete removal of the document).

Twitter relies on its own community to highlight problematic content. This measure was introduced after the start of our project and therefore we cannot assess di-

[3] We do not have any reliable estimations on the content removal by the major online platforms themselves. In this study, we assume, albeit unrealistically, that the content gets removed by the users.

[4] The exact labels vary across platforms (e.g. "out of context" vs. "missing context").

	% day0	% day7	% day30	% day100	% day365	at some point
Platform labels						
missing context	11.5%	10.9%	12.0%	10.4%	8.9%	13.5%
partly false	8.9%	8.9%	9.4%	9.4%	8.9%	11.5%
Community labels						
reader's context	0.5%	1.0%	2.1%	3.1%	3.1%	3.1%
Authors' intervention						
editing	1.6%	2.6%	2.1%	1.6%	1.6%	2.6%

Table 4
Reactions to fact-checking by social media platforms, community and users.

all	visible				obfuscated				removed				total
	May 2024		Oct 2024		May 2024		Oct 2024		May 2024		Oct 2024		
all	363	44.21%	346	42.14%	128	15.59%	107	13.03%	330	40.19%	368	44.82%	821
fb	170	33.53%	164	32.35%	106	20.9%	90	17.75%	231	45.56%	253	49.90%	507
twitter	156	81.25%	157	81.77%	3	1.56%	2	1.04%	33	17.18%	33	17.8%	192
tiktok	3	25%	1	8.33%	0	0	0	0	9	75%	11	91.67%	12
instagram	29	28.15%	23	22.33%	19	18.44%	15	14.56%	55	53.39%	65	63.11%	103
youtube	5	83.33%	5	83.33%	0	0	0	0	1	16.66%	1	16.66	6

Table 5
Statistics for the 8-months dataset: data persistence across platforms, assessed in May 2024 (1.5-2 years after the publication).

rectly how quickly the posts become marked as potentially problematic.

Finally, the users themselves might react verbally to fact-checking reports or consequent actions by social media platforms, editing their original posts. The modifications might range from acknowledging the fact-checking findings and putting clear and unambiguous updates all the way to claiming being ironic or actively attacking fact checkers and arguing against their findings. We have also observed a higher percentage of edits from non-anonymous accounts.

4.4. Longer-term trends

Table 5 shows similar statistics for our 8-months dataset, covering PolitiFact reports published from January to September 2022. We have computed them in May and October 2024 when most posts were almost 2 and 2.5 years old respectively.

These numbers support our initial findings: almost half (44.8%) of compromised documents are no longer available after 2 years. The decay is more pronounced for TikTok and Instagram.

A considerably larger percent of Facebook posts remains visible (non-obfuscated) in our 8-months dataset: this might be attributed to a rendering policy change.

Finally, the 2022 dataset (8-months) contains a larger share of tweets. The decay rate for Twitter is at 17% after 2 years (compared to just 6% after 1 year for the 2-months 2023 dataset). We believe that the considerable change in the platform guidance in the past two years has affected the way content writers use Twitter (both publishing and removing). A larger-scale study is needed to provide more reliable Twitter-specific estimates under the new policies.

5. Conclusion

This paper aims at an in-depth analysis of data persistence for publicly fact-checked online content. After one year of monitoring thoroughly online posts fact-checked by PolitiFact, we have observed the following findings. First, the data persistence is a crucial and underrated issue for compromised content, with considerable decay rates. Second, the decay trends differ across platforms, with Facebook, TikTok and Instagram showing much less data persistance. Third, the decay starts immediately, with 12% of the compromised posts getting deleted at (or before) the publication of the PolitiFact report and 20% becoming unavailable within a week. This suggests an urgent need for a concentrated effort on timely collecting real-life fakes if we want to go beyond synthetic or simplistic datasets and train impactful fact-checking models.

In the future, we want to analyze further aspects of the decay issues for the compromised content. Thus, we plan to add more fact-checking outlets beyond PolitiFact to see if there are any effects due to the report itself. Second, we plan to study in more detail the difference in online behaviour (content removal) between anonymous users, non-anonymous users and public figures. Finally, we plan to expand our research on interaction between content writers and fact-checkers ("editing").

Acknowledgments

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

References

[1] I. Alegria, N. Aranberri, P. Comas, V. Fernández, P. Gamallo, L. Padró, I. San Vicente, J. Turmo, A. Zubiaga, Tweetnorm: a benchmark for lexical normalization of spanish tweets, Language Resources and Evaluation 49 (2015) 1–23. doi:10.1007/s10579-015-9315-6.

[2] A. Zubiaga, A longitudinal assessment of the persistence of twitter datasets, Journal of the Association for Information Science and Technology 69 (2018). doi:10.1002/asi.24026.

[3] A. Küpfer, Nonrandom tweet mortality and data access restrictions: Compromising the replication of sensitive twitter studies, Political Analysis (2024) 1–14. doi:10.1017/pan.2024.7.

[4] M. Bastos, This account doesn't exist: Tweet decay and the politics of deletion in the brexit debate, American Behavioral Scientist 65 (2021) 000276422198977. doi:10.1177/0002764221989772.

[5] Y. Duan, J. Hemsley, A. O. Smith, "this tweet is unavailable": #blacklivesmatter tweets decay, AoIR Selected Papers of Internet Research (2023). URL: https://spir.aoir.org/ojs/index.php/spir/article/view/13414. doi:10.5210/spir.v2023i0.13414.

[6] F. Bianchi, S. HIlls, P. Rossini, D. Hovy, R. Tromble, N. Tintarev, "it's not just hate": A multi-dimensional perspective on detecting harmful speech online, in: Y. Goldberg, Z. Kozareva, Y. Zhang (Eds.), Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates, 2022, pp. 8093–8099. URL: https://aclanthology.org/2022.emnlp-main.553. doi:10.18653/v1/2022.emnlp-main.553.