The Times They Are A-Changin':
Characterizing Post-Publication Changes to Online News

Chris Tsoukaladelis
Stony Brook University

Brian Kondracki
Stony Brook University

Niranjan Balasubramanian
Stony Brook University

Nick Nikiforakis
Stony Brook University

***Abstract*—The current news landscape is in the middle of a major transition. Digital news are quickly overtaking legacy media (such as, newspapers and TV programs), offering a slew of benefits to consumers including ease and immediacy of access. They also, however, allow publishers to arbitrarily modify the articles they publish, at any time after the article has been released. Little is known about how often this happens and to what extent these post-publication edits change an article's original message.**

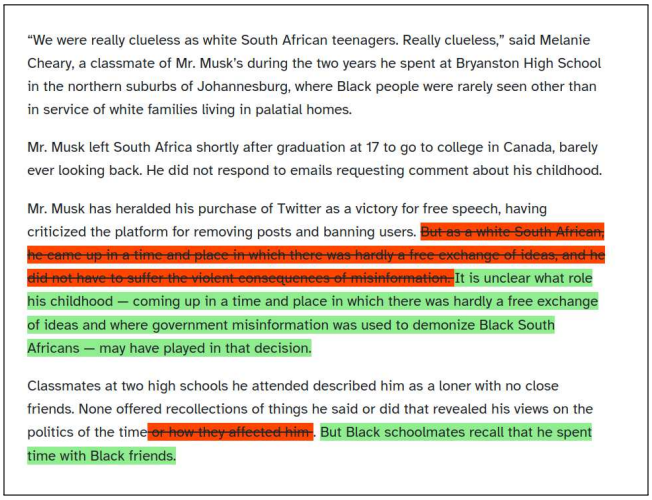
In this paper, we shine light to this previously ignored phenomenon by collecting and analyzing a corpus of more than 600k online news articles, published by tens of U.S. news publishers over a period of nine months. We discover that 165k articles exhibit post-publication changes and use natural language processing tools to identify the magnitude of these changes and their effect. Among others, we find that different publishers modify their articles at different rates, with a publisher's ranking and political bias affecting the frequency of changes and that over 15% of changed paragraphs do not "follow" their original versions. Finally, we discover that most of the evaluated publishers do not properly note these changes to their articles, using non-descriptive notices and updated timestamps that cannot be used by readers to assess what has changed.

1. INTRODUCTION

There is a constantly shifting landscape when it comes to news; more people than ever are reading their news online [13], eclipsing the collosi of old - Television, Radio, Newspapers. This shift enables readers to not only get the news whenever and wherever through their digital devices, but also follow stories as they develop. Such advantages are hard to turn down, but they also come with a new set of shortcomings.

News publishers have found themselves in control of a new kind of power; that of post-publication edits to their articles. In the past, once a version of a newspaper was in the press, there was little that could be done to change its content. Similarly, radio and television reporters couldn't alter words they had already spoken. However, in this new Internet landscape, a publisher can publish an article and effortlessly change parts of it at will.

Unless the original version was archived through a third party, there are few ways to identify that this happened, let alone compare the new version of the article to the original one. The security community has conducted a large number of studies on how attackers weaponize the lack of integrity in the context of expired domain names [55], [69], [74], blackhat search-engine optimization [42], [51], [77], cloaking [47], [57], [58], and maliciously-edited JavaScript [49], [68], [71] yet has largely ignored the effect of post-publication content changes in the context of misinformation and disinformation.


"We were really clueless as white South African teenagers. Really clueless," said Melanie Cheary, a classmate of Mr. Musk's during the two years he spent at Bryanston High School in the northern suburbs of Johannesburg, where Black people were rarely seen other than in service of white families living in palatial homes.

Mr. Musk left South Africa shortly after graduation at 17 to go to college in Canada, barely ever looking back. He did not respond to emails requesting comment about his childhood.

Mr. Musk has heralded his purchase of Twitter as a victory for free speech, having criticized the platform for removing posts and banning users. ~~But as a white South African, he came up in a time and place in which there was hardly a free exchange of ideas, and he did not have to suffer the violent consequences of misinformation.~~ It is unclear what role his childhood — coming up in a time and place in which there was hardly a free exchange of ideas and where government misinformation was used to demonize Black South Africans — may have played in that decision.

Classmates at two high schools he attended described him as a loner with no close friends. None offered recollections of things he said or did that revealed his views on the politics of the time ~~or how they affected him.~~ But Black schoolmates recall that he spent time with Black friends.

Figure 1: **A visualization of the changes on a NYT article regarding Mr. Musk.**

To understand how post-publication edits could affect a user's perception of the world, we draw attention to one popular example published by the New York Times [6] which is also captured in our dataset. Figure 1 shows the changes that were made to an article *days* after it was first published. These changes are meant to neutralize the originally negative sentiment of these specific paragraphs, adding nuance and additional viewpoints to the later version. While the authors of the article later included a correction, that correction was regarding other changes to the article (such as the correction of a misspelled name), leaving the highlighted issue unaddressed. We argue that these types of changes constitute an unwanted element of online news since they have the potential to create split worldviews in those that read a story *before* and *after* it has been changed. These changes are also unwanted in the context of social

media and information amplification where news stories get to keep their "likes" and "shares" regardless of how many times they have been edited. Overall, we argue that careless use of this editing power has the potential to further erode the already damaged trust that people place in the media where half of those surveyed [2] already feel that national news organizations *intend* to "mislead, misinform or persuade the public."

In this paper, we report the results of analyzing such post-publication changes, following articles from multiple popular publishers over time. We build a corpus of 608,723 articles published over a period of 9 months, of which 27.39% (i.e. 166,712 articles) exhibit one or more post-publication changes. We leverage our data-collection system to not just identify modified articles but to also categorize these changes based mostly on two large categories; one for semantic and one for syntactic classification[1]. This way, we track the changes in sentiment, the degrees of editing, as well as the positions of these changes in the article. In doing so, we surface multiple less-than-ideal practices regarding online news which have to be considered in a digital-first world.

Our paper's primary contributions are as follows:

1) **Article Data Set**: We collected 608,723 articles, of which 166,712 exhibited some sort of textual change since first published. We leverage this dataset to obtain a unique look into how major modern publishers operate when it comes to article publishing. This dataset will be made available upon publication, to further encourage and assist future work on web integrity.
2) **Syntactic Classification**: Using our pipeline, we assess both the type and extent of changes, trying to understand why and when they happen most. We collect various metrics on articles, spanning edit distance, topology of changes and paragraph additions/removals. In this way, we measure 51,722 articles that exhibit more paragraphs removed than added, as well as an average of 13.8% of edit distance on a changed paragraph.
3) **Semantic Classification**: We also assess more semantic aspects of each article, such as whether the changed article follows contextually from the original article, for which in 22.18% cases it doesnt, but also the sentiment swings of those changes, discovering that nearly 6.91% of all the paragraphs changed also change their sentiment. For the sentiment analysis, we make use of a roBERTa-base model [53] trained for sentiment analysis on Twitter, while for the entailment analysis we make use of the OpenAI GPT-3 DaVinci model [37].
4) **Stealth Edit Analysis**: Finally, we discuss stealth edits, as well as the attempts by various publishers to address them, in one way or another. We analyze how often they occur and ways to deal with them going forward. This way we notice that on average, less than 5% of the articles that exhibit some change get an "Editor's Note" or other such explicit correcion.

1. In the Natural Language Processing (NLP) field, these terms often refer to capturing parts-of-speech or dependency relations. In our case, we use the term syntactic to refer to surface aspects, such as, the amount and location of change, aspects which do not directly relate to the change in meaning. For changes that affect the sentiment or overall meaning, we use the term semantic.

2. BACKGROUND AND MOTIVATION

In this section, we define key terms and concepts, along with the necessary background on news-media edits, to assist with reader's understanding.

2.1. News Article Editing

Print media is inherently limited in the kinds of edits that can be made to published articles. Once a physical newspaper is printed and distributed, modifications are no longer possible, other than the issuance of corrections and updates in subsequent print editions. This limitation has made correcting a published article a purposeful act. These corrections are traditionally located in an explicitly labeled section of a printed newspaper, with a digital version of this also occasionally available on the publisher's website [27]. Some publishers have evolved this concept to all digital articles, with dedicated areas of sites listing all recently modified articles [15], [28].

Digital news articles, however, can be changed in a number of ways post-publication, including with and without a note acknowledging the change. These include the *addition* of new content, the *deletion* of previous content, and the *modification* of existing content. We note that these changes may not necessarily be malicious in nature. For instance, it is expected that an article following an on-going story will add new content throughout the duration of the event. Additionally, minor changes to correct grammar and spelling errors often occur. However, notorious cases of post-publication news-article modifications do exist [6], [30], revealing less-than-ideal behavior from publishers. Some publishers try to preempt backlash from cases such as these by including edit notes on articles that have been changed. It is unclear, however, what kinds of changes warrant an edit note from publishers, since guidelines are often vague, and open to interpretation [29].

When the text of an online news article changes in any way, while making no note of this change, we consider it to be a *stealth edit*, or *silent edit*. As before, while this does not imply malicious intent, it does reveal a lack of transparency. The aforementioned example shown in Figure 1 demonstrates that silent changes create ample room for readers to walk away with different interpretations of a story, depending on when they happened to read it.

At their core, post-publication edits (silent or not) are a problem of *information integrity*. Traditionally, the security community focused its attention on the integrity of content

with a threat model involving explicitly-malicious actors. These include the study and creation of tools and frameworks to prevent the loading of web resources with compromised integrity [49], [64], [68], as well as the complex web of trust associated with expired domain names on the Internet [41], [56], [66]. Only a single recent study [46] focused on the more subtle problems associated with seemingly benign online content changes. With this in mind, we set out to analyze the extent to which these practices are commonplace today - and if so, gather enough data to be able to surmise a reason behind their adoption.

2.2. Sentiment Analysis

A potential negative outcome of post-publication news article modifications is a change in overall sentiment. This can result in two readers having completely diverging opinions on a particular subject simply because they consumed a particular article at different times. Prior work investigated this phenomenon in the context of online news article headlines, measuring the frequency of headline changes resulting in an increase or decrease in emotional language [46]. The authors demonstrated that the sentiment of popular news publishers change in the hours to days after first publication. For instance, the authors discovered that the publisher BuzzFeed removes emotionally-charged words from 18.9% of observed articles.

These results provide a coarse-grained view into the types of sentiment changes common in news headlines. However, they are limited in that they rely on the presence or absence of particular emotionally-charged words associated with such headlines [48]. Modern Natural Language Processing (NLP) techniques are capable of determining the overall sentiment of text using not only the individual words it is composed of, but also the context created by the relationship between those words. Prior work in this area has utilized NLP techniques to determine the sentiment of content reviews like products [40], [43], [73] and movies [60], [67], [70], as well as posts on social media platforms [33], [59]. Moreover, work has been done to identify the sentiment expressed between entities in news articles [38].

3. DATA COLLECTION AND PREPARATION

In this section, we discuss in detail the process for collecting and preparing our news article dataset. We describe the technical details of our data collection infrastructure, as well as the pre-processing steps we took to extract and normalize news article content from raw webpage data.

3.1. Online News Publishers

To ensure our news article dataset contains a comprehensive and unbiased representation of the overall online news ecosystem, we take great care to collect data on publishers

Figure 2: **News publishers' popularity and political bias.**

representing a wide range of both popularity and affiliation. Taking inspiration from the work of Guo et al. [46], we utilized the Allsides media bias chart [12] to determine the affiliation, and the Tranco ranking [22] to determine the popularity/platform size, of each publisher. We chose a subset of all popular news publishers, equally distributed along this two-dimensional popularity/bias representation. We illustrate these chosen publishers in Figure 2.

3.2. Article Collection Pipeline

In order to detect post-publication article changes, we design and develop a distributed crawling system that is capable of recording the state of a news article immediately after it is published, as well as at the end of our data-collection period. These article versions are then used to conduct a differential analysis of each article's text. Figure 3 shows a general overview of this system.

For each chosen publisher, we manually identify a series of publisher-provided RSS feeds that alert to the publishing of articles representing various categories (e.g., Politics, Entertainment, Business, etc.). For each article appearing on these RSS feeds, our data collection infrastructure records metadata provided in the RSS entry, and appends the URL to the underlying article onto a queue ①. Each entry in this queue is serviced by one of many distributed crawlers that fetch each article's HTML code and save it to a centralized database along with relevant metadata ②. To identify changes to crawled articles, the recrawl module of our data collection infrastructure re-queues article URLs on a predefined schedule, or on-demand.

Many popular websites utilize anti-bot services to fingerprint and block automated browsers. To account for this possibility, we ensure that our crawlers do not overload the servers of any particular publisher by only requesting articles as they are published. In practice this means that the average news publisher in our dataset received between 15 and 1700 requests each day from our crawlers, depending on their rate of publishing new articles. On average, we sent about 2300 requests for article crawls every day. Moreover, while we utilized the Python Requests library [17] to initiate each request, we changed the provided HTTP User-Agent

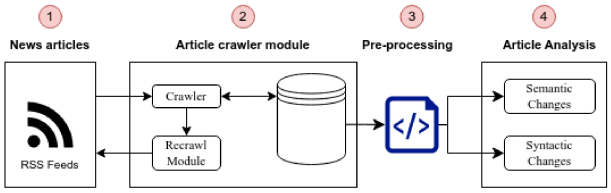


Figure 3: **Overview of news article processing pipeline.**

to a popular browser, and set the HTTP `Referrer` header to https://google.com, simulating a user who discovered the article in search results and clicked on it.

Raw article HTML webpages are consumed by the pre-processing module that extracts and normalizes article text (process detailed in Section 3.3) ③. Finally, processed article text is fed into our analysis module to identify potential semantic and syntactic changes ④.

3.3. Pre-processing

Next to article text and metadata, news-publisher websites contain a multitude of additional content, such as advertisements, promotions, and related articles. This data is not only redundant, but would actively harm our analysis were it to be kept, as it is highly dynamic in nature, capable of changing with each page load. Thus, our pre-processing module must preserve only the article text as well as some relevant metadata, such as the title, timestamp, and edit indicators (e.g., "Updated-at" tags). For this purpose, we use HTML parsers that allow us to isolate publisher-specific element identifiers and attributes from the raw HTML code.

Next to extracting article text and metadata from webpages, it is also necessary to apply a series of normalization steps to the extracted content to reduce false positives stemming from unrelated dynamic content embedded into the text of news articles. For instance, it is common for publishers to place ads inside of article text, with the content of these ads often changing between subsequent visits. Since these ads unrelated to the articles in which they appear, we remove them from all article text. We do this by recording common HTML identifiers associated with this dynamic content, and utilize regular expressions to remove them.

It is also common for publishers to make small imperceptible changes to article text post-publication. For instance, publishers periodically change particular characters in an article like replacing doubles quotes with single quotes, or swapping dash characters article-wide. In order to address this, we programatically normalize every crawled article's text by removing all whitespace, capitalization, and punctuation marks that we found to be of little significance when it comes to characterizing changes, such as "-" and "...". We replace these characters either a single space, or an empty string as appropriate. We also removed or replaced various crawler artifacts that were later discovered during analysis. For example, there was a case when the spaces between words in some articles used the unicode symbol "\xa0", which is the equivalent for "non-breaking space". Our crawler recorded the unicode text for that symbol (instead of the space) which needed replacing before further analysis. Finally, we split each article in individual paragraphs, which becomes our micro unit of operations, allowing us to look deeper into individual article changes.

4. ARTICLE CHANGES

We deployed our article data collection pipeline for 9 months from February 2022 to November 2022, recording 608,723 newly-published articles from the 17 selected publishers. To detect post-publication changes, we recrawled all articles in December 2022, capturing the most current version of each article. In total, we observed 166,712 (or 27.39%) of the articles exhibiting some sort of change (i.e., the text of the article at $t_0 \neq t_1$).

To better understand the discovered changes, we utilized multiple metrics to analyze and measure each article change. For ease of analysis, we broke each article down to its paragraphs, removing those that had the exact same text at each crawl. Then, in looking at the ones exhibiting differences, we used the Levenshtein distance [52] of the paragraph before and after its changes as a way to match those paragraphs to one another. Furthermore, we heuristically set a cut-off point for each paragraph at 50%. That is to say, if a paragraph of the original article had a Levenshtein distance of more than 50% compared with every paragraph exhibiting some sort of change of the later version of the article, then that paragraph is considered to be removed (and, vice versa, added).

4.1. Syntactic Measurements

Edit distance. To track the extent of a change in a given paragraph, we again rely on the paragraph's Levenshtein (edit) distance, breaking each paragraph down to its word elements and finding the distance between the two versions of the same paragraph. The greater the extent of changes made in a paragraph, the larger the distance between the two versions of the paragraph. For example, a New York Post article [5], changed the paragraph

---

Biden on Monday urged Americans to leave Ukraine because of the threat of invasion

---

to

~~*Biden on Monday*~~ ***On Monday, however, Biden*** *urged Americans to leave Ukraine because of the threat of invasion.*

In our analysis, we strip all commas and thus the only visible change is just the ordering of words, for which the Levenshtein distance is 0.2 (or 20%). Analyzing all articles across all publishers, we note that, as shown on Figure 4, publishers make changes on many of their articles,

but that does not necessarily translate to extensive changes on the articles' paragraphs. Inversely, some publishers may make extensive (i.e. larger edit distances when comparing paragraphs), yet more sparse changes to their articles (i.e. less frequent post-publication modifications).

Another interesting connection is in relating Figure 4 to each publisher's ranking (one dimension of Figure 2). On average, the publishers that perform the most changes to their articles tend to be the ones with a higher Tranco Rank. One possible explanation behind this observation is that higher-ranked websites have more writers/editors and therefore more opportunities to keep revising articles after they have been published. Furthermore, we can see a correlation between the political leanings of publishers and their likelihood of changing an article after it has been published. We calculate R^2 for the linear regression between the fraction of articles that were changed per publisher, and that publisher's Tranco ranking finding it to be 0.14. At the same time, the R^2 of the multiple linear regression predicting the same fraction of changed articles using the publisher's Tranco ranking *and* political bias, is 0.28 showing that a publisher's bias carries a significant signal beyond what could be explained merely by their ranking.

For a more detailed breakdown of Figure 4, Table 8 in the Appendix lists per-publisher statistics regarding the number of changed articles. Similarly, Table 9 (also in the Appendix) presents the median and mean edit distances of paragraphs, per publisher.

We emphasize that these changes are not necessarily representative of the effects present on the various articles. A publisher could change a single word in a paragraph (for instance, adding "not" to a sentence) and completely change its meaning, while the edit distance remains small. On the other hand, a publisher might rewrite/re-order a large fraction of an article (resulting in a large edit distance) with no perceptible change to what is being communicated. To account for this, we perform both sentiment as well as entailment analysis later in this section.

Removed/added paragraphs. At times, an article exhibits a larger (or smaller) number of paragraphs compared to its original version. In that case, we mark these paragraphs as added (or removed). Moreover, when testing the edit distance between two paragraphs, there comes a point at which we mark those two paragraphs as no longer edited, but rather one as removed and another as added in its place. Heuristically, we set that point to be at 50% - as in, more than half the paragraph has been changed in some way. In the same vein, we are interested in shedding more light in situations where more paragraphs were removed than added - in other words, when a paragraph "disappears" from the article. One instance of this behavior is observed in a New York Times article [31] regarding Federal Reserve interest rates. The original article stated the following:

> *[...] how investors expect the Fed to react to inflation.*
> *Mortgage rates have already been ticking higher as a result of inflation, even though they remain historically low: Rates on 30-year fixed-rate mortgages averaged 3.85 percent with 0.8 points as of March 10, according to Freddie Mac, up from 3.76 last week and 3.05 a year ago. (A point is a one-time fee, equal to 1 percent of the mortgage amount, paid to the lender to buy down the mortgage rate.)*
> *The pain to the consumer[...]*

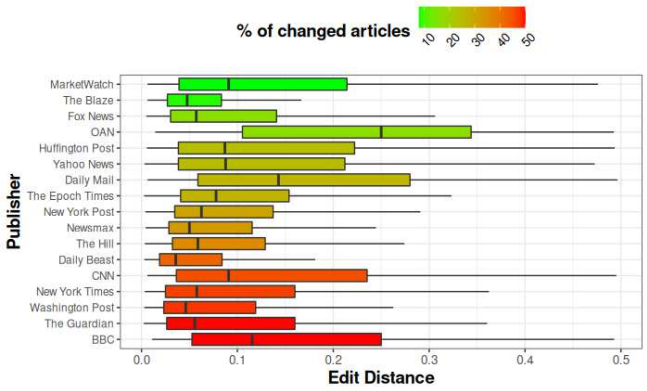


Figure 4: *A summary of the average amount paragraphs change on various publishers, as well as what proportion of articles change in the same publishers.*

The middle paragraph ("Mortgage rates have already...") disappeared from this article when it was updated after a period of about two months, as was the case for two more paragraphs that also were removed. We also note that there is no indication from the publisher that the article had been updated despite the extensive changes, other than the timestamp of the article having changed. Figure 5 shows a breakdown of the extent of removed paragraphs across publishers, in proportion to their changed articles. Figure 6 gives us a different look, this time in proportion to the total number of articles per publisher, regardless of whether they have been changed or not. We can notice by contrasting the two figures that certain publishers, such as *"Yahoo News"*, overall lean towards this practice in general, even if they generally edit less articles than i.e. *"The Guardian"*, which changes about 50% of their total articles. For this reason this seems to be a tendency more related to individual publishers than one that is tied to the frequency of changes a publisher makes to their articles.

Live updates. Across publishers, there are cases in which the article is expected to change drastically since its first appearance, commonly encountered when dealing with articles that follow a news story as it develops over time (i.e. "Breaking news"), categorized here as "live updates". We mark these as separate cases and handle them differently than other articles, although we do find added benefit in exploring the changes in sentiment in both the paragraphs that have potentially been changed, but also in the paragraphs being added. It is entirely possible for a story to become more

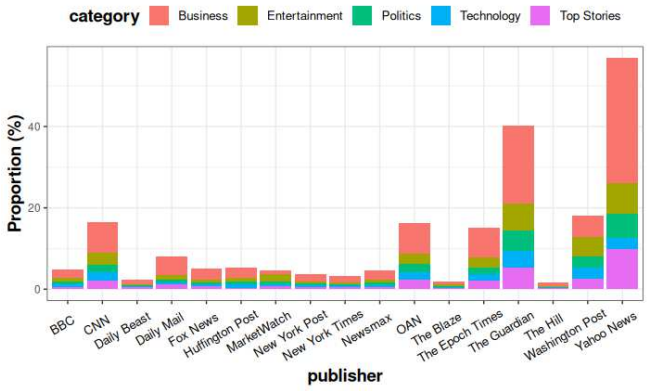


Figure 5: *A breakdown of the percentage of articles exhibiting more paragraphs removed than added, by category and publisher, in proportion to the changed articles of each publisher.*

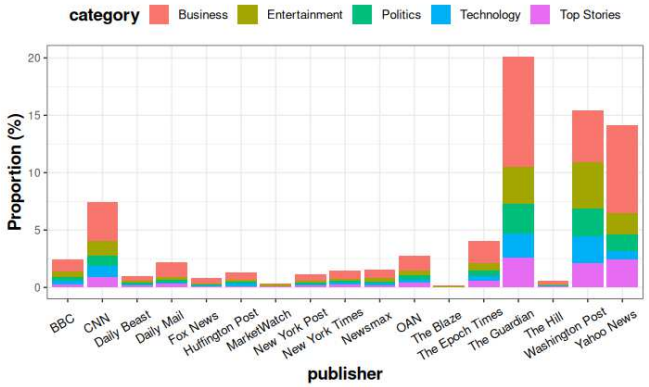


Figure 6: *A breakdown of the percentage of articles exhibiting more paragraphs removed than added, by category and publisher, in proportion to the total articles of each publisher.*

neutral as additional details are added, trying to counter the publisher's initial bias, as new evidence (possibly counter to some of their initial claims) emerge. Cases in which paragraphs are removed from these live-updates articles are also of interest. In our analysis, 1,282 (or 0.2%) of our articles are live-update articles, with 137 of them end up having more paragraphs removed than added.

Topology of changes. Adding to our understanding of the various ways in which publishers modify articles, is the location within the article where the change occurs. Intuitively, we can tell for instance that changes near the beginning of an article should be more common, as that is the area most readers focus on. Changes in the middle or end of an article are not automatically suspicious in nature, but they do represent unusual behavior that warrants further analysis. Figure 7 shows the distribution of the position of changes within articles in our dataset. We find that, generally, most changes do indeed occur at the beginning of articles. However, publishers such as OAN and The Washington Post make changes in the bottom third of articles at a higher rate than others.

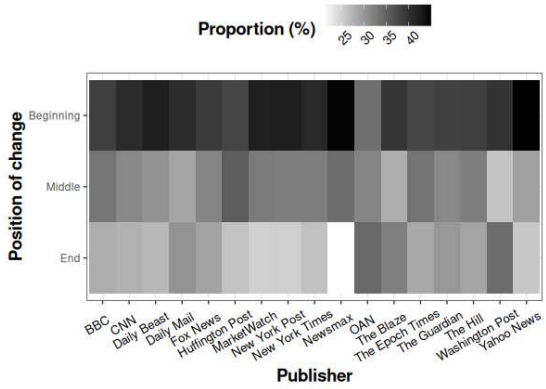


Figure 7: *Position of changes in articles, per publisher.*

Interestingly, we also observe publishers such as The Huffington Post and The Epoch Times modify the middle of articles nearly 30% of the time. Even though modifications to the beginning/end of an article could be explained by live updates to an ongoing story, changes in the middle of an article are likely related to the correction of errors or rewording of statements. Such changes may be more harmful to readers as they are more likely to go undetected if not explicitly called out by the publisher. In Section 4.3, we analyze how often publishers include update notes to changed articles, as well as how much detail is included in such notes to explain the nature of the change.

Length. We also measured the length of the articles (in terms of number of words) before and after any changes had been made to them. This helps us understand how publishers change an article overall, by adding or subtracting text to it. As demonstrated in Figure 8, the most common theme is erring on the side of adding rather than removing text when making changes to an article. This is consistent with the notion that more detail will be added to an article as new information is discovered regarding a particular event. It also bears mentioning that, for the sake of clarity, we removed 14 outliers from this plot, specifically live update articles that sprawled to tens of thousands of words while starting from a few hundred. We also need to note that specific publishers, such as *"Daily Mail"*, sometimes do not label their articles as "live update" articles even though they fit the exact criteria of what would be considered a "live update" article. Next, we will further analyze three specific articles that demonstrate interesting update behavior.

Point "A" in Figure 8 represents a live updating article -although not labelled as such- from the *The Daily Mail* discussing a heat wave in the United Kingdom in July 2022 [10]. Between the time we first crawled this article and when we re-crawled it, 30 paragraphs had been added and almost 400 paragraphs had been removed. Analyzing the removed paragraphs, many of them contain information that became out-of-date throughout the event such as changing weather temperatures, a number of quotes, as well as a number of articles that were embedded in their entirety. These embedded articles contained information involving

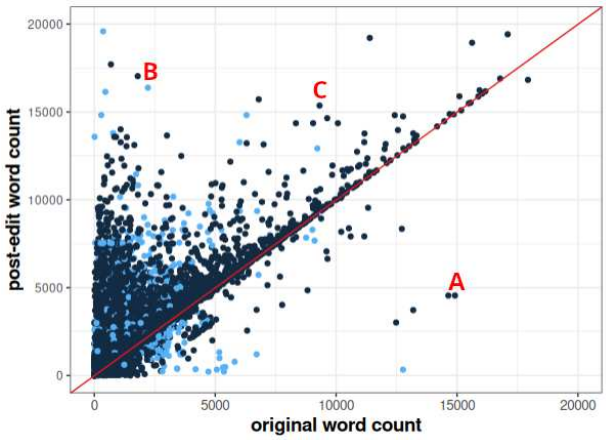


Figure 8: *A summary of the article length in words, before and after the article has been changed. The red line signifies a 1:1 match. The blue points indicate live update articles.*

casualties and damages incurred due to the heat wave, as well as health advice.

Point "B" represents another live update article from *Yahoo News* involving Russia's invasion of Ukraine from March 2022 [4]. Between our two crawls of this article, 600 paragraphs were added. Changes such as these are expected from explicitly-labeled live update articles, especially those that cover events spanning over large periods of time.

Finally, point "C" represents yet another live update article from *The Daily Mail* which did not receive any such labelling regarding a service in remembrance of Prince Phillip from March 2022 [7]. This article saw over 200 paragraphs added, including quotes from various individuals at the ceremony, as well as some background information on Prince Andrew. As the article was initially published before the ceremony, it is also to be expected that details would be added during and after the event.

4.2. Semantic Measurements

Sentiment analysis. An important factor in our understanding of article changes is the difference in overall sentiment between the two versions of the same paragraph, in which an edit has taken place. In noting the extent to which sentiment swings happen in articles, either from positive/negative to neutral or vice versa, we can detect patterns among news publishers, and build a contextual hypothesis as to why these patterns may emerge. To gauge a paragraph's sentiment, we used a public Twitter dataset [63] to train a roBERTa-based model for classification. This model classifies text as either positive, neutral, or negative. Working under the assumption that this classification is not perfect, we also sample random paragraphs and cross-check with the classification given by roBERTa to verify these results to the best of our ability. More specifically, we found that in our sampling, roBERTa correctly assesses the sentiment 80% of the time.

In Table 1, we can see the sentiment fluctuations of articles on a paragraph level. While many paragraphs retain

TABLE 1: *The sentiments present in individual paragraphs, as a proportion of the total amount of changed paragraphs. The row sentiments are associated with the original sentiment, while the column sentiments with the post-change sentiment.*

Original		*Post-change*		
		Negative	**Neutral**	**Positive**
	Negative	23.71%	2.26%	0.03%
	Neutral	2.21%	60.54%	1.42%
	Positive	0.03%	0.96%	8.84%

their original neutral sentiment, there are a number that move between adjacent sentiment states. For example, we find 2.26% of paragraphs began in a "negative" sentiment state, before changing to a "neutral" state. In our paper's leading example about Mr. Musk in Figure 1, for instance, the last paragraph's change moves its sentiment from "negative" to "neutral". This sentiment swing has us as well as the roBERTa-trained model in agreement. Overall, 6.91% of all paragraphs that exhibited some update across the articles in our corpus changed their sentiment after a post-publication edit.

In our analysis, the top 3 publishers that tend to change their sentiment were *The Daily Mail*, with an average sentiment swing of 10.7% followed by *MarketWatch* with 8.3% and *Huffington Post* with 8.1%. In Table 2 we present examples of the different types of sentiment swings.

In the first example, we find a paragraph in which the sentiment became more positive after it was changed [20]. Here, the addition of the text "the greatet lyricist of his generation" has a more positive impact to the sentiment of this paragraph, leading the tool we used to classify it as "Positive", while the original paragraph was "Neutral".

Another article contained a paragraph that became more negative post-edit [18]. The original paragraph in this article, while it could be reasonably interpreted as mostly negative in the context of the wider article, is classified as neutral when analyzed on its own. With the addition of the extra quotation to the paragraph, it becomes more critical of the subject, changing its label to "Negative" sentiment.

Finally, the last example of Table 2 is from an article in which a paragraph became more neutral post-edit [26]. While the original paragraph mostly conveys a positive sentiment, with terms such as "warm welcome", the updated paragraph takes a more critical tone, adding references to events such as 9/11 and Pearl Harbor. This shifts the overall sentiment from "Positive" to "Neutral".

Entailment. An important dimension regarding modified news articles is whether post-publication changes to the text alter its original meaning. Being able to reason about meaning helps us understand some of the more extensive changes to articles, since a modification that would make the changed paragraph no longer "follow" from the original paragraph, would be one that fundamentally alters the meaning of the original paragraph. To tackle this question effectively at scale, we used both manual as well as automated techniques. The first heuristic approach involved manually analyzing

TABLE 2: *Examples of the different sentiment swings according to our roBERTa-based classification model.*

Sentiment swing	Original Paragraph	Post-Edit Paragraph
More Positive	By the time Eminem appeared, anticipation was at fever pitch. A quick blast of Forgot About Dre's chorus paid homage to the producer who first brought Marshall Mathers to worldwide attention, and then he was off.	By the time Eminem appeared, anticipation was at fever pitch. A quick blast of Forgot About Dre's chorus paid homage to the producer who first brought Marshall Mathers**, the greatest lyricist of his generation,** to worldwide attention~~,and then he was off~~.
More Negative	"That's really the M.O. of this administration," said Rep. Fred Keller, R-Pa., while serving as a Thursday guest on the "American Agenda" show.	"That's really the M.O. of this administration," said Rep. Fred Keller, R-Pa., while serving as a Thursday guest on the "American Agenda" show. **He added that the Biden administration is certainly "behind the curve on many things. Not putting Americans first."**
More Neutral	Ukrainian President Volodymyr Zelenskyy addressed Congress Wednesday morning, receiving a warm welcome from both sides of the aisle as he called on the United States to do more.	Ukrainian President Volodymyr Zelenskyy addressed Congress Wednesday morning, ~~receiving a warm welcome from both sides of the aisle~~ **invoking the Sept. 11 attacks and Pearl Harbor** as he called on the United States to **"**do more**"**.

random pairs of paragraphs, before and after modification, from randomly chosen articles. For our automated analysis, we used an easy-to-use and effective large language model GPT-3 (named Davinci) to which we provided 500 randomly chosen paragraphs before and after they had been changed, and asked it whether the latter followed the former. In this way, we aim to automatically establish whether the modified paragraph is restating the facts present in its original version or whether it is introducing additional details that do not "follow" from the earlier version of the same text. Much like roBERTa's sentiment analysis, we sampled 50 randomly chosen paragraphs per publisher manually as well, in order to have a frame of reference with regards to GPT-3's accuracy, which we measured to be 90% on average.

We present our findings regarding GPT-3, along with the results of our manual analysis of a sample of pairs of paragraphs (N=50) in Table 3. In this case, the ratio of positive entailment describes the percentage of all paragraphs where the changed paragraph follows from the original one. We can see that most of the publishers are in the 75%-85% range. Outliers do exist, such as *OAN*, which also has the highest average edit distance (as shown earlier in Figure 4), and *MarketWatch*, whose changes mostly have to do with market moves. The takeaway, however, is that in all the surveyed publishers, 15-25% of changed paragraphs *do not* follow from their original paragraph. This implies that the changes they perform on their paragraphs change their meaning enough from the original one so as to imply something new, not before considered, which more so than anything implies these changes would at the very least warrant a dedicated section in the article explaining why these changes took place.

An example of what we consider to be positive entailment is as follows [19]:

> *That force includes reconnaissance and artillery troops and medics, as well as about 100 howitzer cannons and other military vehicles.*

was changed to

> *That force includes reconnaissance and artillery troops and medics, as well as about 100 ~~howitzer cannons~~* ***howitzers*** *and other military vehicles.*

Here, the only change is from the original "howitzer cannons" to "howitzers", which have the same meaning. Therefore, the entailment test is successful.
An example of negative entailment is as follows [25]:

> *[...] This week, Russian occupation officials began efforts to force some 60,000 people from Kherson to the western side of the Dnipro, ahead of the Ukrainian push.*

was changed to

> *[...] This week, Russian occupation officials began efforts to force some 60,000 people from Kherson to the ~~western~~* ***eastern*** *side of the Dnipro, ahead of the Ukrainian push.*

While a single word was changed in this example, it is one that inverses a key point the paragraph was originally making, and so it fails both our and the AI model's entailment test.

Overall, we can see that entailment is an important metric when considering the semantic changes to a piece of text, and it gives us insights regarding the more important changes present in an article. Considering the capabilities of the latest NLP models, entailment analysis can be integrated into a larger toolset leveraged in order to provide more integrity to the average person reading the news.

Time-related analysis

One may wonder to what extent the articles that *do* change after publication, are changed immediately after they

TABLE 3: *The ratio of sampled articles that GPT-3 flagged positively regarding entailment, as well as the ratio of manually sampled articles that we flagged positively, sorted by GPT-3 positive entailment ratio.*

Publisher	GPT-3 (N=500)	Manual (N=50)
The Blaze	87.2%	90%
BBC	84.4%	88%
New York Post	84.2%	84%
New York Times	83.2%	80%
Newsmax	83%	80%
The Guardian	82.8%	86%
Epoch Times	82.6%	74%
CNN	81.8%	72%
Washington Post	81%	84%
Daily Beast	80.4%	80%
Huffington Post	77.2%	78%
Yahoo News	76.8%	82%
Fox News	76.4%	74%
Daily Mail	76.2%	68%
MarketWatch	67.4%	78%
The Hill	60.4%	72%
OAN	58%	54%

TABLE 4: *Comparison of the sentiment swing and lack of entailment between changed paragraphs, with regards to the specified time delta*

	Delta < 2 days	Delta ≥ 2 days	Statistical Significance
Sentiment	9.05%	7.51%	✓(< 0.01)
Entailment	18%	22%	✓(0.02)

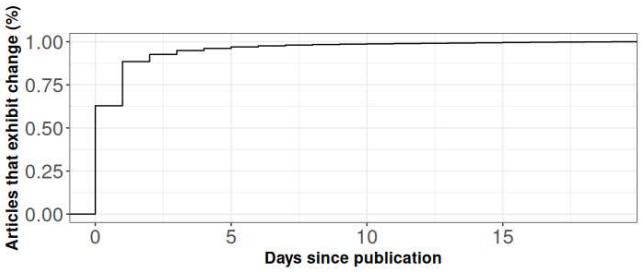


Figure 9: **The time after which an article has stopped being edited**

are first published (e.g. to issue a correction that went unnoticed through the editorial process) or at a later time. While our pipeline did not capture multiple snapshots per indexed article, we can take advantage of the publishers that report both a publication timestamp as well as an updated timestamp. Figure 9 shows the distribution of deltas of these two timestamps across articles published by these publishers. We observe that for most edited articles, changes occur in the first couple of days since publication.

Building on top of our semantic analysis, both in terms of sentiment and entailment, we assess the extent to which small-delta edits behave any differently than long-delta edits. To decide on a threshold, we look to the analysis of Guo et al. who report that propagation of news over Twitter happens mostly in the first few hours since an article's publication [46]. Given that not all outlets show timestamps at the granularity of hours, we conservatively choose a threshold of two days to include corner cases where an article was published near midnight and edited in the early hours of the next day.

Overall, we find that 9.05% of articles change their sentiment within the first two days of publication, while 7.51% change their sentiment after this threshold. We note that these averages are slightly elevated as compared to those of all publishers in Table 1.

Regarding entailment, we sampled 1,000 changed paragraphs from articles that were changed within the first two days of publication, finding that about 18% those changes do not follow from their original paragraphs. Similarly, when sampling 1,000 changed paragraphs from articles changed over a larger time horizon, the proportion changes to 22%.

Finally, looking at Table 4, we conduct a two-sided hypothesis test for the comparison of two independent proportions. For each row, our null hypothesis (H_0) was that the true proportion of sentiment swings or lack of entailment was the same between the two sets, while the alternative hypothesis (H_A) was that the true proportions of these observations over the two sets were different from each other. In the third column we can see the reported p-value for each hypothesis test. Following standard conventions, the cut-off point for a computed p-value is 0.05, over which the null hypothesis is maintained. We further note that these proportions are for different sampling sizes. In the case of entailment, the sample for each delta category was 1,000, while in the case of sentiment we sampled 30K changed paragraphs per delta category.

As we can see, both p-values are smaller than the cut-off value, thereby rejecting the null hypothesis. We can therefore conclude that the articles are changed in different ways, when these changes happen early in their lifetime vs. at a later time.

4.3. Stealth Edits

One of the most important features we keep track of, is whether or not the publisher indicates the article has been changed in any way since its original publication. When a publisher updates an article to some extent, but does not notify the reader of doing so, we consider that to be a silent change or *stealth edit*. In our analysis, we found that 67,449 articles (i.e. 40.5%) noted an update out of the 166,712 changed articles.

Through our analysis, we identified a number of ways in which publishers attempt to communicate updates to their readers. For instance, when changes to the article are extensive (albeit not always), a publisher would add a correction section to that article, either at the very beginning or end of it. We generally consider this to be a *Proper* update, even though it is not possible for us to distinguish for all such articles if the publisher is making note of every change that occurred in the article within the correction. Below is an

TABLE 5: *The ratio of articles exhibiting an update notice, compared to the total number of changed articles per publisher. The breakdown here is for the proportion to the changed articles of each publisher for what we defined as "Proper" and "Technical" updates. Sorting is done based on the ratio of "Technical" updates.*

Publisher	Technical Updates	Proper Updates
The Guardian	100%	0.7%
Daily Beast	100%	1.2%
Huffington Post	100%	4.2%
MarketWatch	100%	0.8%
OAN	100%	-
Daily Mail	99.9%	-
New York Post	98.5%	0.5%
CNN	97.7%	25%
Epoch Times	92.3%	1.5%
New York Times	38.8%	0.7%
Washington Post	35.6%	2.4%
Fox News	-	10.2%
Newsmax	-	0.2%
The Blaze	-	7.0%
The Hill	-	3.5%
Yahoo News	-	2.5%
BBC	-	-

example of one such "proper" update, from a *Huffington Post* article [21]:

> *This story has been updated to reflect the Senate's vote to discharge Jackson's nomination from the Judiciary Committee.*

Another, and much more prevalent way for publishers to communicate updates, is through adding a generic *Updated* tag somewhere in the article page, usually next to the timestamp. As such, in the vast majority of these cases, there is no further elaboration provided to the reader, and for some publishers it is possible that this tag is auto-generated every time the body of an article is changed. We also note that on some publishers, most articles displayed this tag, regardless of us detecting a change or not. This may be due to the change involving something we filtered out in pre-processing, such as, whitespace manipulation.

Interestingly, *The Guardian* will notify the reader that the article was "last modified", only if the reader clicks on the timestamp of the article (shown in Figure 10). For our purposes, we conservatively treat all of these to be non-silent updates even though we argue that informing readers about what was changed is far more meaningful and actionable. For this reason, we differentiate between these *"Technical"* updates and the aforementioned "Proper" updates.

Finally, some publishers only change the original timestamp of the article to that of the time of the update. We intuitively consider it nearly impossible for a reader to keep track of what timestamp an article had on it when they first read it, especially if that change happened days or weeks later. For that reason, we do not consider this to be a valid update note in our analysis.

With these points in mind, Table 5 lists the rate with which publishers announce to their readers that an article

Figure 10: **The Guardian: In order for the reader to notice an article they are reading has been modified, they need to click on the timestamp.**

was updated, differentiating between technical and proper updates. Despite extensive efforts, we were unable to locate any visible way of communicating a change for some publishers, either because they rely solely on updating the article timestamp, or because an update-communicating mechanism does not exist. These publishers are marked with "-" in Table 5.

Overall, we find that "proper" updates are much less prevalent in digital news media than in print. Some publishers, such as the *New York Times*, have a dedicated "Corrections" section on their website [15], but these corrections appear to be present mostly -if not solely- because of user-provided comments. Furthermore, a concerning trend is that most publishers are relying more and more on small update tags to their article, instead of specifically addressing the changes made within the article. We can see that 8 of the 17 publishers that we studied include an update tag in all or nearly-all articles. However, these are typically "technical" updates, as their ratio of "proper" updates is significantly smaller. While we do note that a portion of these involve minor grammatical corrections, "proper" updates are highly uncommon. In the original example in Figure 1, there is only a "technical" update with no proper issued correction. As a result, at the time of that change, the authors of that article were publicly criticized for "stealth editing" it [16], [23] even though no correction has been issued at the time of this writing.

4.4. Sampling

As previously mentioned, we also manually sampled N=50 randomly selected paragraphs from every publisher. This sampling consisted of analyzing a pair of changed paragraphs and gauging the magnitude of the change. For this, we took inspiration from Faruqui et al. [44], who performed a sampling of edits on wikipedia articles. We also tested each pair of paragraphs for entailment as demonstrated in Table 3, to be able to bind an error range on the results of the GPT-3 model.

The sampling results concerning the magnitude of changes in paragraphs, which we break down to "Equivalent", "Minor Change/Ambiguity", "Significant Change" and "Dynamic Update" are present in Table 7. Some examples regarding what we consider to be equivalent, minor change/ambiguity and major change are listed in Table 6.

TABLE 6: *Examples of what we consider to be Equivalent [32], Minor Change/Ambiguity [8], Significant change [1] or Dynamic Update [14].*

Publisher	Original Paragraph	Post-change Paragraph
Equivalent	Dorries, who is backing Liz Truss, accused Sunak's team of "dark arts" for allegedly trying to engineer getting a candidate into the final two with him who was easily beatable and suggested Dominic Cummings was supporting him.	Dorries, who is backing Liz Truss, accused Sunak's team of "dark arts" for allegedly trying to engineer getting ~~a candidate into the final two with him who was easily beatable~~ **an easily beatable candidate into the final two with him**, and suggested Dominic Cummings was supporting him.
Minor Change/Ambiguity	The National Institute of Forensic Medicine continues to investigate the death of the immensely popular musician who died at the age of 50.	The National Institute of Forensic Medicine ~~continues~~ **and attorney general's office will continue** to investigate the death of the immensely popular musician who died at the age of 50.
Significant Change	Pappas, who is openly gay, served as a member of the New Hampshire Executive Council and was first elected in 2017.	Pappas~~, who is openly gay,~~ served as a member of the New Hampshire Executive Council and was first elected in 2017.
Dynamic Update	Loren & Alexei Brovarnik, "Loren & Alexei: After the 90 Days"	Loren & Alexei Brovarnik, "Loren & Alexei: After the 90 Days" ***WINNER**

We observe a strong correlation between entailment and how each change is classified. That is, equivalency of two paragraphs implies entailment, while a significant difference implies lack of entailment. In the cases where we had minor differences or ambiguity, we used our best judgement incorporating contextual information to arrive at a conclusion regarding entailment. This also mostly explains some of the discrepancies between our sampling and GPT-3's, such as when a person's name is replaced with a more generic pronoun or their title. Finally, it bears noting that our manual sampling regarding entailment mostly matches that of GPT-3, and potentially opens a path towards more massive entailment analysis using a model such as that in the future, or potentially even a more powerful one. That possibility is important when taking into consideration individual readers that might appreciate a tool that does this automatically for articles they read.

Finally, we sampled $N = 10$ articles per publisher and reviewed the parsed HTML code after our pre-processing step has taken place, comparing that text with an article's text when accessed via a web browser. This was done to ensure that no article content was removed during HTML parsing in a way that biases the rest of our analysis. During this process, we did not observe any significant or systematic problems with our parsing or pre-processing, outside of the occasional element that escaped our filters (e.g. a reporter's affiliation and prompts to contact the publisher). These issues do not change our findings and further motivate the need for more inter/intra-publisher standardization (Section 5).

5. DISCUSSION

Standardization of updates

In our dataset of over 600k articles, we made every effort possible both during pre-processing but also during sampling, to bring all these articles from distinct publishers to a reasonable level of uniformity. When we first started working on this project, we did expect that there would be differences across publishers, such as, their use of specific HTML tags and their specific ways of issuing corrections. What we did not expect, however, was the lack of uniformity we encountered even across the articles of the same publisher.

Sometimes when an article is corrected or updated and a correction is issued, that correction is at the top of the article, while other times it is at the bottom of the article. Occasionally, it is located entirely outside the body of the article. Finally, it bears noting that while some publishers note an article has been generally updated (by adding an *"Update"* tag next to the timestamp), others do not include this tag, but update only the timestamp showing the last edit time. It is clearly unreasonable to expect readers to remember what timestamp was on every article they read, particularly given the fact that some articles are modified *months* after they are first published. As shown earlier in Figure 10, publishers like The Guardian do have the ability to show that article was modified but do not make that available by default.

This lack of uniformity and sometimes even lack of attempt to inform that an article has been updated, poses definite challenges to readers, who are likely to be reading tens of articles each day, without the ability to later recall what each article said. Therefore, it is not enough, in our opinion, to simply perform what we refer to as a "technical" update to an article, and there is a need for a more objective and ideally standard, cross-publisher way of communicating the presence of updates. We describe some possible avenues for future work at the end of this section.

Publisher Update Policies

In our study, we discovered that online news publishers commonly update articles post-publication, with many offering little (if any) notice to the reader regarding the nature of the changes. Our findings highlight the need for online news publishers to refine and publicly clarify their article-update

TABLE 7: Proportion of each magnitude of change acquired during our sampling with N=50, per publisher.

Publisher	Equivalent	Minor Change/Ambiguity	Significant Change	Dynamic Update
BBC	76%	16%	4%	4%
The Blaze	74%	20%	6%	0%
New York Post	70%	20%	8%	2%
Yahoo News	70%	16%	6%	8%
Washington Post	70%	18%	12%	0%
Huffington Post	68%	16%	8%	8%
New York Times	68%	20%	8%	4%
Newsmax	64%	24%	12%	0%
The Guardian	64%	26%	10%	0%
MarketWatch	62%	22%	8%	8%
Epoch Times	60%	26%	14%	0%
Daily Beast	56%	42%	2%	0%
Fox News	52%	28%	14%	6%
The Hill	52%	36%	12%	0%
CNN	50%	36%	8%	6%
Daily Mail	26%	42%	20%	12%
OAN	22%	54%	24%	0%

policies to ensure that readers understand the conditions under which the articles they consume may change *after* they read them.

Many publishers have publicly-listed update policies for their content [9], [29], but typically these only serve to inform readers that they can report errors which will be updated in the corresponding articles. Moreover, in the few cases in which publishers specify update conditions, they are often vague and open to interpretation. For instance, the Washington Post mentions that they will include an update note whenever they correct a "significant mistake". However, a clear definition on what is considered significant is never mentioned.

Inconsistencies across publishers further exasperate this problem. We encourage the development and wide-scale adoption of general update guidelines to be used by all online news publishers, such as those of the *Independent Press Standards Organisation* (IPSO) [11]. The IPSO *Editors Code of Practice* lists correction standards to be followed by its members, but these are still limited in their clarity, as well as the number of member publishers. The large-scale standardization of such policies would make it easy for a reader to identify when and how an article has changed, and prevent multiple readers from walking away with different takeaways even when they all read the same sets of articles.

Limitations

As with any large-scale web data collection study, the effect of anti-bot services must be taken into account. As mentioned in Section 3.2, we design our data collection infrastructure in such as a way as to maximize crawling efficiency, while also taking measures to prevent bot detection. These measures, along with our infrequent crawling of any particular site (at the time of publication of each article), allow us to be confident that our crawlers were able to collect most if not all article data. To ensure that this is the case, we sampled 850 articles from our dataset (50 for every publisher) and gauged whether the entire article text was collected. Through this process, we did not identify any obvious cases where text was missing (e.g. an article suddenly stopping half-way a developing story).

Also, as mentioned earlier in this section, there is an obvious lack of standardization among the publishers, which is reflected in our attempts to extract data regarding *silent changes*. With no standardization in place, there is little we can do short of manually going over every changed article to assess whether or not a correction has been issued, so our current attempts rely on manually curated regular expressions and searching for specific keywords in the first and last paragraphs of each changed article. Similarly, when it comes to "live updates"-type articles, we have to rely on contextual clues (i.e. the article's URL and title) to label them as such. This heuristic-based detection is imperfect, as some publishers do not treat their "live updates" articles any different than their regular articles, and thus a better scheme for keeping track of them could yield more accurate results in the future.

We also note that due to language barriers (both ours as well as those of the utilized NLP models) we focused solely on English news outlets. As Martins et al. [54] have shown, this type of work can applied to other languages just as much as it can to English. The main difficulty of such an undertaking is keeping individual parsers for each news publisher updated, which can be potentially achieved through the bazaar model [62] of open-source software.

Lastly, the precision of our analysis is tied to the specific tools that we used. It is impossible to manually analyze more than 600k articles, which means that we must rely on automated tools thereby inheriting their limitations. Even though, at the time of this writing, GPT-3 is considered to be the state-of-the-art in Natural Language Processing, there can be no guarantees about its responses regarding entailment. Similarly, the roBERTa-based model we used for sentiment analysis has its own limits. To deal with these limitations, we manually sampled articles to gauge how often human analysts would agree with the output of these tools.

Going forward

We consider this paper to be a problem-showing one, rather than problem-solving one. Despite the importance of integrity in computer security, integrity of content in the context of online news has largely escaped the attention of our community. Our findings clearly show, not just that post-publication edits are an issue, but also that they occur in many different ways across publishers, some blatant, others more subtle. We therefore hope that, other than expecting publishers to engage in better self-governance, our community can start building tools and systems for measuring post-publication changes, differentiating between appropriate and inappropriate changes, and devise disincentives for behavior that is deemed unwanted. These could range from browser-level systems tracking consumed articles and warning users when previously-read articles are substantially changed, to standard formats for news publishing that can be automatically consumed by software, and third-party watchdogs alerting when articles are silently modified. The further we move from printed media and the physical world, the more we have to explicitly guard the integrity of digital content.

6. RELATED WORK

In this era of diminishing trust in news publishers, there have been extensive attempts to catalogue what is commonly referred to as "fake news" [45], [50], [72]. Vosoughi et al. [75] went on to measure the spread of rumor cascades on Twitter from 2006 to 2017, and found that the top 1% of it diffused between 1,000 and 100,000 people, while more accurate and factual news stories rarely ever diffuse to more than 1,000 people, effectively demonstrating that lies spread faster than the truth. There have also been other studies exploring the various ways that fake news can be harmful not only to individual readers, but to society at large [34], [36]. As a result, there have been multiple attempts to detect, study, and stem the propagation of fake news [39], [61], [65], [80]. One major difference between our work and past work is that we do not automatically assume that large established news outlets do not engage in unwanted behavior regarding online news. We thereby find subtle (as well as less subtle) issues even with the largest and most respected publishers.

Related to this work are also various ways of classifying text and recognizing sentiment. Yang et al. [76] focus on text edits in Wikipedia and create a taxonomy based on the perceived intentions behind these edits. Yin et al. [78] leverage natural language processing to capture both the structure and semantics, of text edits. Barbieri et al. [35] create an evaluation platform [24], which concentrates various previously fragmented classification tasks trained on Twitter corpora. One of those tasks belongs to Rosenthal et al. [63], who trained a roBERTa-base model [53] to perform sentiment analysis on tweets, categorizing them as positive, neutral or negative.

Guo et al. [46] focused on post-publication headline changes on a wide range of articles across various popular online news publishers, and calculate both the frequency and time since publication it takes for these changes to appear. Furthermore, they labeled these changes using BERTscore [79] and tracked their propagation over twitter. Because our work focuses on the entirety of articles (as opposed to just their headlines), we were able to extract significantly more insights from our dataset, including the characterization of silent edits and to what extent the modified paragraphs in each article "follow" from their earlier versions.

Martins and Mouro [54] identify post-publication article changes focusing on Portugese online news sources (specifically a Portugese web archive [3]), and also track whether these changes are disclosed by the publishers (i.e. whether the publishers engage in silent changes). Their methodology for identifying silent changes is unfortunately missing from their paper and they do not attempt to categorize the changes in the articles that they track. Leveraging our unique dataset of more than 165k changed articles, as well as Natural Language Processing tools that became only recently available, we attempt to understand the reasons and extent of article changes on a multitude of news publishers.

7. CONCLUSION

As we keep transitioning from a physical world to a digital world, we cannot keep taking integrity for granted. Users are steered towards custodial platforms where they can consume content but that content always stays within the control of the publisher. In this paper, we investigated the phenomenon of post-publication, news article edits. By collecting and analyzing more than 600k articles from 17 publishers over a period of 9 months, we detected that 27.39% of these articles experienced some degree of post-publication edits. We used multiple syntactic and semantic techniques to identify the magnitutude of changes, their location within an article, and whether the publisher notified readers that the article was changed since it was first published. We found that only 67,449 (or 40.5% of all changed articles) note they have been changed in any way, and that is typically done through vague "technical" update notices that do not explain what exactly was changed. Finally, we employed both a GPT-3 model as well as a roBERTa-based model to gauge entailment and sentiment of changes. Using these models we discovered that, on average, 22.18% of the changed paragraphs do not "follow" from the original paragraph, and 6.91% of all paragraphs exhibiting some post-publication change, affecting the text's original sentiment.

We view this work as a problem-showing work, aiming to start a discussion between users, researchers, and news stakeholders related to acceptable behavior around the concept of data integrity. There are multiple opportunities for future work to assess how users react to these changes when told, to build technologies for standardizing the delivery of online news, and methodologies for incentivising transparency. We hope that our work can serve as motivation for all these future endeavors.

Availability

To assist the community in understanding and tackling misinformation on the web, we will be open-sourcing our dataset of post-publication article modifications and our analysis at the following URL: https://changing-times.github.io/

Acknowledgments

We thank our anonymous shepherd and the reviewers for their helpful feedback. This work was supported by the National Science Foundation (NSF) under grants CNS-1941617 and CNS-2126654.

References

[1] 8 key house races that could flip this november. https://www.theepochtimes.com/these-8-house-seats-could-flip-in-november_4749200.html. Online.

[2] American views 2022: Part 2 - trust,media and democracy. https://knightfoundation.org/wp-content/uploads/2023/02/American-Views-2022-Pt-2-Trust-Media-and-Democracy.pdf. Online.

[3] Arquivo.pt. https://arquivo.pt/. Online.

[4] As it happened: Cities pulverized and thousands dead in first 3 weeks of russia's invasion of ukraine. https://news.yahoo.com/live-updates-germany-halts-pivotal-143310043.html. Online.

[5] Biden oks pentagon plan to help americans flee ukraine if russia invades: report. https://nypost.com/2022/02/09/biden-oks-plan-to-evacuate-americans-if-russia-invades-ukraine/. Online.

[6] Elon musk left a south africa that was rife with misinformation and white privilege. https://www.nytimes.com/2022/05/05/world/africa/elon-musk-south-africa.html. Online.

[7] Emotional queen returns to windsor with prince andrew by her side after royals rallied round her at moving westminster abbey memorial for prince philip - as european royals and other guests are hosted at london receptions in honour of the late duke. https://www.dailymail.co.uk/news/article-10663365/Prince-Charles-Camilla-arrive-Prince-Philips-memorial-Westminster-Abbey.html. Online.

[8] Foo fighters drummer taylor hawkins had 10 different substances in system when he died, authorities say. https://www.theblaze.com/news/taylor-hawkins-cause-of-death-drugs. Online.

[9] The guardian - editorial code. https://www.theguardian.com/info/2015/aug/05/the-guardians-editorial-code. Online.

[10] Heatwave to peak on tropic-hell tuesday with 43c predicted – forcing hospitals to cancel operations, threatening power cuts and bringing more travel chaos after uk sweated through 'warmest night on record'. https://www.dailymail.co.uk/news/article-11025787/UK-weather-hot-highest-temperature-today-hot-tonight.html. Online.

[11] Ipso - editors code of practice. https://www.ipso.co.uk/editors-code-of-practice/. Online.

[12] Media bias chart. https://www.allsides.com/media-bias/media-bias-chart. Online.

[13] More than eight-in-ten americans get news from digital devices. https://www.pewresearch.org/fact-tank/2021/01/12/more-than-eight-in-ten-americans-get-news-from-digital-devices/. Online.

[14] Mtv movie & tv awards: See the full list of winners. https://www.cnn.com/2022/06/05/entertainment/mtv-movie-tv-awards-winners/index.html. Online.

[15] New york times - corrections. https://www.nytimes.com/section/corrections. Online.

[16] Nyt makes stealth edits to elon musk piece tainting him with apartheid smears; writer lashes out at critics. https://www.westernjournal.com/nyt-makes-stealth-edits-elon-musk-piece-tainting-apartheid-smears-writer-lashes-critics/. Online.

[17] Python requests library. https://pypi.org/project/requests/. Online.

[18] Rep. keller to newsmax: Biden admin 'behind the curve' on american crisis matters. https://www.newsmax.com/politics/pennsylvania-baby-formula-shortage-house/2022/05/19/id/1070628/. Online.

[19] Russia fm says talks will continue as west's ukraine attack fears grow. https://nypost.com/2022/02/14/russia-says-talks-to-continue-as-ukraine-attack-fears-grow/. Online.

[20] Super bowl: Dr dre and eminem pack in the hits at half-time show. https://www.bbc.com/news/entertainment-arts-60354066. Online.

[21] Supreme court nominee ketanji brown jackson clears senate committee. https://www.huffpost.com/entry/supreme-court-ketanji-brown-jackson-senate-committee_n_6248919ae4b098174501a8c6. Online.

[22] Tranco. https://tranco-list.eu. Online.

[23] Tweet regarding the nyt elon musk story. https://twitter.com/tomgara/status/1522284258496233472. Online.

[24] Tweeteval. https://github.com/cardiffnlp/tweeteval. Online.

[25] U.s. sees opportunity for ukraine to capitalize on russian weakness. https://www.nytimes.com/2022/10/20/us/politics/us-ukraine-war.html. Online.

[26] Volodymyr zelenskyy invokes 9/11, pearl harbor in powerful address to congress. https://www.huffpost.com/entry/volodymyr-zelenskyy-congress-ukraine_n_6231e11fe4b0fe0944de6d55. Online.

[27] Wall street journal - corrections. https://www.wsj.com/news/types/corrections. Online.

[28] Washington post - corrections. https://www.washingtonpost.com/national/corrections/2016/04/25/5dac200a-0679-11e6-a12f-ea5aed7958dc_story.html. Online.

[29] Washington post - corrections policy. https://www.washingtonpost.com/policies-and-standards/#corrections. Online.

[30] Washington post issues two corrections after stealth-edit scrubbed false claim from taylor lorenz report. https://www.foxnews.com/media/washington-post-issues-correction-stealth-edit-scrubbed-false-claim-taylor-lorenz-report. Online.

[31] What the fed's interest rate increase means for you. https://www.nytimes.com/article/federal-reserve-rate-increase.html. Online.

[32] 'dark arts' and betrayal: one week in the tory leadership race. https://www.theguardian.com/politics/2022/jul/15/tory-leadership-race-contenders-dark-arts. Online.

[33] Ahmed Abbasi, Hsinchun Chen, and Arab Salem. Sentiment analysis in multiple languages: Feature selection for opinion classification in web forums. *ACM transactions on information systems (TOIS)*, 26(3):1–34, 2008.

[34] Hunt Allcott and Matthew Gentzkow. Social media and fake news in the 2016 election. *Journal of Economic Perspectives*, 31:211–236, 5 2017.

[35] Francesco Barbieri, Jose Camacho-Collados, Leonardo Neves, and Luis Espinosa-Anke. Tweeteval: Unified benchmark and comparative evaluation for tweet classification. 10 2020.

[36] Alexandre Bovet and Hernán A. Makse. Influence of fake news in twitter during the 2016 us presidential election. *Nature Communications*, 10:7, 1 2019.

[37] 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. Language models are few-shot learners. 5 2020.

[38] Eunsol Choi, Hannah Rashkin, Luke Zettlemoyer, and Yejin Choi. Document-level sentiment inference with social, faction, and discourse context. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 333–343, 2016.

[39] Myojung Chung and Nuri Kim. When i learn the news is false: How fact-checking information stems the spread of fake news via third-person perception. *Human Communication Research*, 47:1–24, 2 2021.

[40] Kushal Dave, Steve Lawrence, and David M Pennock. Mining the peanut gallery: Opinion extraction and semantic classification of product reviews. In *Proceedings of the 12th international conference on World Wide Web*, pages 519–528, 2003.

[41] Matteo Dell'Amico, Leyla Bilge, Ashwin Kayyoor, Petros Efstathopoulos, and Pierre-Antoine Vervier. Lean on me: Mining internet service dependencies from large-scale dns data. In *Proceedings of the 33rd Annual Computer Security Applications Conference*, pages 449–460, 2017.

[42] Kun Du, Hao Yang, Zhou Li, Hai-Xin Duan, and Kehuan Zhang. The ever-changing labyrinth: A large-scale analysis of wildcard dns powered blackhat seo. In *USENIX Security Symposium*, pages 245–262, 2016.

[43] Xing Fang and Justin Zhan. Sentiment analysis using product review data. *Journal of Big Data*, 2(1):1–14, 2015.

[44] Manaal Faruqui, Ellie Pavlick, Ian Tenney, and Dipanjan Das. Wikiatomicedits: A multilingual corpus of wikipedia edits for modeling language and discourse. 8 2018.

[45] Axel Gelfert. Fake news: A definition. *Informal Logic*, 38:84–117, 3 2018.

[46] Xingzhi Guo, Brian Kondracki, Nick Nikiforakis, and Steven Skiena. Verba Volant, Scripta Volant: Understanding Post-publication Title Changes in News Outlets. In *Proceedings of the ACM Web Conference*, pages 588–598, 2022.

[47] Luca Invernizzi, Kurt Thomas, Alexandros Kapravelos, Oxana Comanescu, Jean-Michel Picod, and Elie Bursztein. Cloak of visibility: Detecting when machines browse a different web. In *IEEE Symposium on Security and Privacy (SP)*, pages 743–758, 2016.

[48] Danielle K Kilgo and Vinicio Sinta. Six things you didn't know about headline writing: Sensationalistic form in viral news content from traditional and digitally native news organizations (2023). 2016.

[49] Deepak Kumar, Zane Ma, Zakir Durumeric, Ariana Mirian, Joshua Mason, J Alex Halderman, and Michael Bailey. Security challenges in an increasingly tangled web. In *Proceedings of the 26th International Conference on World Wide Web*, pages 677–684, 2017.

[50] David M. J. Lazer, Matthew A. Baum, Yochai Benkler, Adam J. Berinsky, Kelly M. Greenhill, Filippo Menczer, Miriam J. Metzger, Brendan Nyhan, Gordon Pennycook, David Rothschild, Michael Schudson, Steven A. Sloman, Cass R. Sunstein, Emily A. Thorson, Duncan J. Watts, and Jonathan L. Zittrain. The science of fake news. *Science*, 359:1094–1096, 3 2018.

[51] Nektarios Leontiadis, Tyler Moore, and Nicolas Christin. A nearly four-year longitudinal study of search-engine poisoning. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security (CCS)*, pages 930–941, 2014.

[52] Vladimir Iosifovich Levenshtein. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

[53] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach. 7 2019.

[54] Flávio Martins and André Mourão. Revisionista.pt: Uncovering the news cycle using web archives, 2020.

[55] Najmeh Miramirkhani, Timothy Barron, Michael Ferdman, and Nick Nikiforakis. Panning for gold. com: Understanding the dynamics of domain dropcatching. In *Proceedings of the 2018 World Wide Web Conference*, pages 257–266, 2018.

[56] Nick Nikiforakis, Luca Invernizzi, Alexandros Kapravelos, Steven Van Acker, Wouter Joosen, Christopher Kruegel, Frank Piessens, and Giovanni Vigna. You are what you include: large-scale evaluation of remote javascript inclusions. In *Proceedings of the ACM conference on Computer and Communications Security (CCS)*, pages 736–747, 2012.

[57] Adam Oest, Yeganeh Safaei, Adam Doupé, Gail-Joon Ahn, Brad Wardman, and Kevin Tyers. Phishfarm: A scalable framework for measuring the effectiveness of evasion techniques against browser phishing blacklists. In *2019 IEEE Symposium on Security and Privacy (SP)*, pages 1344–1361. IEEE, 2019.

[58] Adam Oest, Yeganeh Safei, Adam Doupé, Gail-Joon Ahn, Brad Wardman, and Gary Warner. Inside a phisher's mind: Understanding the anti-phishing ecosystem through phishing kit analysis. In *APWG Symposium on Electronic Crime Research (eCrime)*, pages 1–12, 2018.

[59] Alexander Pak and Patrick Paroubek. Twitter as a corpus for sentiment analysis and opinion mining. In *LREc*, volume 10, pages 1320–1326, 2010.

[60] Bo Pang, Lillian Lee, et al. Opinion mining and sentiment analysis. *Foundations and Trends® in information retrieval*, 2(1–2):1–135, 2008.

[61] Verónica Pérez-Rosas, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. Automatic detection of fake news. 8 2017.

[62] Eric Raymond. The cathedral and the bazaar. *Knowledge, Technology & Policy*, 12(3):23–49, 1999.

[63] Sara Rosenthal, Noura Farra, and Preslav Nakov. Semeval-2017 task 4: Sentiment analysis in twitter. pages 502–518. Association for Computational Linguistics, 2017.

[64] Ronak Shah and Kailas Patil. A measurement study of the subresource integrity mechanism on real-world applications. *International Journal of Security and Networks*, 13(2):129–138, 2018.

[65] Kai Shu, Amy Sliva, Suhang Wang, Jiliang Tang, and Huan Liu. Fake news detection on social media. *ACM SIGKDD Explorations Newsletter*, 19:22–36, 9 2017.

[66] Milivoj Simeonovski, Giancarlo Pellegrino, Christian Rossow, and Michael Backes. Who controls the internet? analyzing global threats using property graph traversals. In *Proceedings of the 26th International Conference on World Wide Web*, pages 647–656, 2017.

[67] Vivek Kumar Singh, Rajesh Piryani, Ashraf Uddin, and Pranav Waila. Sentiment analysis of movie reviews: A new feature-based heuristic for aspect-level sentiment classification. In *International mutliconference on automation, computing, communication, control and compressed sensing (imac4s)*, pages 712–717, 2013.

[68] Johnny So, Michael Ferdman, and Nick Nikiforakis. The more things change, the more they stay the same: Integrity of modern javascript. In *Proceedings of the Web Conference*, 2023.

[69] Johnny So, Najmeh Miramirkhani, Michael Ferdman, and Nick Nikiforakis. Domains do change their spots: Quantifying potential abuse of residual trust. In *IEEE Symposium on Security and Privacy (SP)*, pages 2130–2144, 2022.

[70] Richard Socher, Jeffrey Pennington, Eric H Huang, Andrew Y Ng, and Christopher D Manning. Semi-supervised recursive autoencoders for predicting sentiment distributions. In *Proceedings of the 2011 conference on empirical methods in natural language processing*, pages 151–161.

[71] Marius Steffens, Marius Musch, Martin Johns, and Ben Stock. Who's hosting the block party? studying third-party blockage of csp and sri. In *Network and Distributed Systems Security (NDSS) Symposium*, 2021.

[72] Edson C. Tandoc, Zheng Wei Lim, and Richard Ling. Defining "fake news". *Digital Journalism*, 6:137–153, 2 2018.

[73] Peter D Turney. Thumbs up or thumbs down? semantic orientation applied to unsupervised classification of reviews. *arXiv preprint cs/0212032*, 2002.

[74] Thomas Vissers, Timothy Barron, Tom Van Goethem, Wouter Joosen, and Nick Nikiforakis. The wolf of name street: Hijacking domains through their nameservers. In *Proceedings of the ACM SIGSAC Conference on Computer and Communications Security*, pages 957–970, 2017.

[75] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359:1146–1151, 3 2018.

[76] Diyi Yang, Aaron Halfaker, Robert E. Kraut, and Eduard H. Hovy. Who did what: Editor role identification in wikipedia. In *International Conference on Web and Social Media*, 2021.

[77] Ronghai Yang, Xianbo Wang, Cheng Chi, Dawei Wang, Jiawei He, Siming Pang, and Wing Cheong Lau. Scalable detection of promotional website defacements in black hat seo campaigns. In *USENIX Security Symposium*, pages 3703–3720, 2021.

[78] Pengcheng Yin, Graham Neubig, Miltiadis Allamanis, Marc Brockschmidt, and Alexander L. Gaunt. Learning to represent edits. 10 2018.

[79] Xichen Zhang and Ali A. Ghorbani. An overview of online fake news: Characterization, detection, and discussion. *Information Processing & Management*, 57:102025, 3 2020.

[80] Xinyi Zhou and Reza Zafarani. A survey of fake news. *ACM Computing Surveys*, 53:1–40, 10 2020.

Appendix A.

TABLE 8: **The total number of articles collected per publisher, as well as the number of articles exhibiting changes. Ratio refers to the ratio of changed articles to total articles.**

Publisher	Changed	Total	Ratio
BBC	3,721	7,417	50%
CNN	5,096	11,225	45%
Daily Beast	3,713	8,722	43%
Daily Mail	30,031	112,458	27%
Epoch Times	4,107	15,264	27%
Fox News	3,664	22,248	16%
Huffington Post	1,913	7,857	24%
MarketWatch	888	11,379	8%
New York Post	6,995	22,926	31%
New York Times	6,405	13,675	47%
Newsmax	3,696	11,394	32%
OAN	1,510	8,990	17%
The Blaze	624	7417	8%
The Guardian	6,024	12,032	50%
The Hill	6,352	17,626	36%
Washington Post	6,215	12,995	48%
Yahoo News	75,758	305,020	25%

TABLE 9: **The mean and median edit distance for paragraphs that have been changed, per publisher.**

Publisher	Mean	Median
BBC	16.5%	11.5%
CNN	14.7%	9.1%
Daily Beast	8%	3.6%
Daily Mail	18.1%	14.3%
Epoch Times	11.9%	7.8%
Fox News	11%	5.7%
Huffington Post	13.9%	8.7%
MarketWatch	14.3%	9.1%
New York Post	11.1%	6.3%
New York Times	11.5%	5.8%
Newsmax	10.3%	5%
OAN	24%	25%
The Blaze	7.6%	4.8%
The Guardian	11.7%	5.6%
The Hill	11.2%	5.9%
Washington Post	9.7%	4.6%
Yahoo News	14.1%	8.8%

Appendix B.
Meta-Review

B.1. Summary

This paper examines the post-publication updates of major US news outlets for a period of 9 months in 2022. Based on a corpus of 608K articles, the authors find that 27% of them change over time, with no direct indicators of change for over 60% of edits. Using GPT-3 to detect logical "entailment" and a custom sentiment analysis model, the authors find that "on average, 22.18% of the changed paragraphs do not "follow" from the original paragraph, and 6.91% of all paragraphs exhibiting [sic] some post-publication change, affecting the text's original sentiment." More broadly, the work points to evidence of silent changes to news story logic and sentiment and posits that they can create unwanted or even dangerous "split views" for different populations of news consumers.

B.2. Scientific Contributions

- Independent Confirmation of Important Results with Limited Prior Research.
- Provides a New Data Set For Public Use.
- Provides a Valuable Step Forward in an Established Field.
- Establishes a New Research Direction.

B.3. Reasons for acceptance

1) Provides a valuable step forward in the emerging field of information integrity.
2) Provides a useful dataset for public use.
3) Reiterates the fact that media accountability relies in part on detecting or preventing changes/removal of published digital news content.
4) Motivates future research into potential technical solutions.
5) The logical "entailment" technique could be useful for future studies.

B.4. Noteworthy Concerns

1) Low confidence in the evaluation and interpretation of the roBERTa sentiment analysis results. If we accept the assertion that the trained classifier is only 80% accurate, it is unclear to what degree deviations in Table 1 (which are all less than 2.3%) are due to model inaccuracy, since statistical confidence/likelihood values are not provided.
2) The paper does not sufficiently analyze/categorize the types of changes that are made: which edits can confuse readers and in what ways? As a result, the paper does not support and overstates claims about the danger of split views arising from changes that "go well beyond what should be appropriate of edits performed by news organizations" - as stated during the interactive discussion phase.
3) As a minor concern, given that prior work has already identified changes to news article titles, this work does not indicate the additional impact of changes to news article content.

Appendix C.
Response to the Meta-Review

We thank the reviewers for their insightful comments. We view this work as the first paper that identifies the issue of post-publication content changes in news articles. There are many directions for future work, including i) better NLP models to automatically characterize changes, ii) user studies to understand how real users react when shown the existence of post-publication content changes (and whether their reaction agrees with our automated sentiment-analysis/entailment models), and iii) what are true boundaries between acceptable vs. not-acceptable content changes. We are eager to keep working in this space and we hope that the community will join us.