International Journal of Communication 13(2019), 4173–4197 1932–8036/20190005

Detecting Textual Reuse in News Stories, At Scale

TOM NICHOLLS[1]
University of Oxford, UK

Motivated by the debate around "churnalism" and online media, this article develops, evaluates, and validates a computational method for detecting shared text between different news articles, at scale, using *n*-gram shingling. It differentiates between newswire copy, public relations material, source-to-source copying, and common-source and incidental overlaps. I evaluate the method, quantitatively and qualitatively, and show that it can effectively handle newswire content, copying, and other forms of reuse. Substantively, I find lower levels of news agency and press release copy reuse than is suggested by previous studies, and conclude that the news agency finding is robust, but the lack of press release copy found might reflect limitations of the method and the changing practices of journalists.

Keywords: computational methods, news production, churnalism, news agency, automated content analysis, online news

Working journalists and scholars of news production are concerned about the extent to which news production is characterized by so-called churnalism (Davies, 2008)—the wide-scale practice of sourcing articles from lightly rewritten press releases and competitors' articles to maintain a volume and speed of news output that is not obtainable by conventional means.

More broadly, there is substantial interest in the notions of content isomorphism and professional imitation (Boczkowski, 2009; Graves, Nyhan, & Reifler, 2016); monitoring and imitation among journalists, coupled with other homogenizing pressures, can result in the outputs of news organizations becoming more homogeneous. This is an important assumption of the sociology of news, but in practice it is difficult to measure empirically at any wide scale.

Content reuse and imitation is difficult to directly measure. To do so comparatively or at scale requires either an extensive supply of labor or the development of effective computational methods.

This article develops, evaluates, and validates a method for detecting shared text among different news articles, at large scale. It collects and analyzes a new corpus of 163,297 news articles from the United

Tom Nicholls: tom.nicholls@politics.ox.ac.uk
Date submitted: 2018–06–20

[1] This work was supported by a grant from Google UK as part of the Digital News Initiative (CTR00220).

States and United Kingdom, together with 125,508 articles from four major news wires and 24,675 press releases from the leading press release agencies in both countries.

Several distinct sources of overlap are differentiated. I discuss the basing of stories on newswire material, press releases, competitors' stories (including by lightly rewriting their text), and common-source material such as quotations. I also differentiate various kinds of incidental overlap such as stock phrases and clichés, and artifactual noise such as newswire attribution notices and imperfectly cleaned Web-crawl text.

Results are presented showing the extent of textual overlap. Despite expectations from qualitative studies of journalists' working practices and previous quantitative analyses, there is little evidence of the large-scale recycling of press releases, or of the direct lifting of copy from competing news sources. Extensive use of wire copy is detected (though less than previous studies have found), but the majority of it is directly attributed.

This new method is scalable, sensitive, and reliable. It therefore represents a good foundation for future empirical work.

Literature Review

Within the literature on churnalism and journalistic imitation, there is a particular assumption that problems have become more acute with the crisis of journalism funding and the coincidental broader and more rapid availability of other producers' news output via the Internet (Johnston & Forde, 2011), with digital-born and other online-only media in particular seen as a conduit of derivative news (Jackson & Moloney, 2016; Johnston & Forde, 2017).

One concern is direct copying. We know from qualitative research on journalists' working practices that some proportion of stories are sourced from competitors, often rewritten in patchwork style, and then issued; rapidly scanning competitors' outputs in search of stories is routine (Boczkowski, 2009). There is anecdotal evidence that this legitimized process of identifying stories that other outlets are running and quickly reporting them can degenerate into the "JCB technique—lifting material out of [other] papers and dropping them into their own columns as if they were using a JCB mechanical digger" (Davies, 2008, p. 103).

A second issue is news agency wire copy. Opinions on wire copy are more mixed and nuanced. Although merely republishing wire content is not adding value to the news ecosystem in the same way that original reporting is, and some declare it as part of the churnalism problem (Beckett, 2008; Johnston & Forde, 2011), the (attributed) use of wire copy is directly in line with its purpose: to provide a quality source of news that allows news organizations to cover the news more widely than they can with their staff alone. As Lorenz (2017) notes, news agencies "are indispensable suppliers of news, allowing media to fill their newspaper pages and news bulletins every day with news from all over the world without their having to maintain bureaux in every relevant news centre" (p. 948).

Many studies have tried to quantify the degree of reliance on news agency copy, but it has seldom been done at a scale that allows generalization (Boumans, Trilling, Vliegenthart, & Boomgaarden, 2018).

Nevertheless, in conducting a manual content analysis of several UK midmarket and broadsheet newspapers, Lewis, Williams, and Franklin (2008) report a very high volume of newswire copy:

> 30 per cent of published items were wholly dependent on agency copy with a further 19 per cent strongly derivative from agency materials. In a further 13 per cent of stories agency copy was evident along with information from other sources . . . [and] in [only] 25 per cent of stories there was no evidence of dependence of agency copy. (p. 29)

Critically, much of the wire copy was found to be unattributed, making it difficult to identify without extensive cross-referencing with the wires themselves.

There is, similarly, a prevailing view that uncritically adopting press release material does not represent high journalistic standards, though levels may be high: "19% of newspaper stories . . . were verifiably derived mainly or wholly from PR material, while less than half the stories we looked at appeared to be entirely independent of traceable PR" (Lewis, Williams, Franklin, Thomas, & Mosdell, 2006, p. 3). Boumans (2017) identified that around 10% of Dutch newspaper articles mentioning a sample of corporations or NGOs were initiated by a press release from that organization.

Lewis et al. (2006) are not alone in finding large volumes of textual reuse. Cagé, Hervé, and Viaud (2015) found that a mean of 39% of the text in their sample of French news stories was nonoriginal, whereas Saridou, Spyridou, and Veglis (2017) found that 56% of articles in their sample of Greek stories contained at least 40% nonoriginal content.

Similar results have come from studies examining newswire use by news sources. Scholten and Ruigrok (2009) identified around 25% of their sample of Dutch stories to be at least partially based on the ANP newswire, whereas Boumans et al. (2018) highlighted 48% of their (large) sample of Dutch news articles to be predominantly based on ANP agency copy, with online newspapers more heavily based on the wire than their off-line versions.

There may also be overlap between the categories of wire and PR content; Kiernan (2003) quotes a PR source observing that one of the most effective PR strategies was to get material into Reuters or the Associated Press.

Press releases are harder than newswire content to detect reliably, because there is no central source of them as with the major newswires. We know that journalism based on PR activity is increasingly based on the direct supply of material to individual journalists, rather than more generally distributed press releases, which makes it even more opaque to us:

> As an MD of a small PR agency explains: "we never, ever, send press releases out any more, we do everything by email . . . news copy is in the email." (Jackson & Moloney, 2016, p. 769)

There are many legitimate reasons for news articles to contain the same text as others, and we must be careful not to attribute to improper activity that is reasonably considered journalistic. This

consideration was raised in the digital humanities context by Ganascia, Glaudes, and Del Lungo (2014), who noted that the boundaries between plagiarism, reference, and comment are fluid and contextual. They observed that parody, comment, reference, and copying were all distinct, and telling them apart using off-the-shelf plagiarism tools was problematic.

Lifting a whole article from a competitor and republishing it is a breach of professional ethics, but this is not the case if the source is a newswire and the republishing body is properly licensed. Quotation (even extensive quotation) for journalistic comment or to stand up a story is standard, meaning that the quotes in multiple papers may well be the same.

For research interested in the normative evaluation of textual reuse, being able to differentiate between the different forms of overlap is critical, whereas for other questions (perhaps tracking the overlap between versions of the same text over time) it is less important.

In light of the above, I identify three research goals:

RG1: To quantify the extent of textual reuse in news articles, to triangulate against previous estimates.

RG2: To develop a more detailed picture of textual overlap by differentiating between news agency use, press release use, quotations, and coincidental or artifactual overlaps.

RG3: To develop a method to answer RG1 and RG2 that is suitable for future comparative use at large scale.

Existing Approaches to Automated Analysis

In general, effective computational analysis of content overlap requires an efficient way of comparing large volumes of news content.

If text in news articles were tagged with metadata indicating the source of each part, then detecting overlaps would be trivial. In practice, even the much easier task of detecting significant news agency copy by examining the credit lines "is a notoriously inaccurate measurement given news organizations' inconsistent attributing practices" (Boumans et al., 2018, p. 1770). Without reliable tagging, a method is needed to reverse engineer this information.

One set of approaches measures textual similarity between article pairs by counting the frequency of each word in each article irrespective of word order[2] and modeling the article in vector space using these term-frequency vectors. This permits the use of standard vector space to document measures such as cosine similarity (Manning, Raghavan, & Schütze, 2008, p. 111). Boumans et al. (2018) use a cosine similarity approach to detect articles "initiated by" news agency coverage; a similar technique (using tf–idf-weighted

[2] This is known as a "bag of words" assumption; it works well in many text processing and analysis tasks, despite the obvious loss of information from word ordering.

vectors after parts-of-speech-based word selection) was used by Welbers, van Atteveldt, Kleinnijenhuis, and Ruigrok (2018) to measure the influence of news agencies on subsequent coverage choices.

Cosine similarity measures need fairly high thresholds to ensure that the similarity is a result of reuse. Boumans et al. (2018) propose a 0.65 cosine similarity threshold and note that in a manual analysis of content overlaps at their threshold of 0.65 they found that the "majority of the content is identical" (p. 1775). This limits the sensitivity of the cosine similarity measure to detect small pieces of reused text. Although this more than suffices for the given purpose of detecting articles heavily based on agency copy, it is not likely to be effective at detecting the overlapping use of quotations, for example, or the straightforward copying of smaller parts of a source article as the core of a partially rewritten new article. Nevertheless, textual similarity methods have wide applicability to various communications questions. A similar approach was built on by Nicholls and Bright (2019) to group related stories in an article corpus together.

Another approach, originating in computer science and not common in communications scholarship, involves modeling articles as groups of comparable overlapping textual units (*n*-grams[3]) and recording these sets of *n*-grams directly in an index. Each article can then be checked for overlaps against the whole corpus of existing articles by looking up each of its *n*-grams. If only one index entry exists for a given *n*-gram, then the text is unique. If not, the other articles that share the same piece of text can be identified.

This is a useful method because it systematically and exhaustively checks articles for exact text overlaps. Because the detection unit length is several words (rather than a whole article) long and the word order within each *n*-gram is preserved, specific multiword sequences are precisely detected rather than relying on a less direct measurement of the similarity of word frequency vectors.

A naïve implementation of *n*-gram comparison would require each part of each text to be repeatedly compared against each other text in turn—which is computationally infeasible at scale because the time taken grows geometrically as the number of documents increases. This can be avoided with an efficient implementation: By using a hash map index, for example, as in Citron and Ginsparg (2015), overlap lookup time for each entry can be made essentially constant. As the total number of entries to be looked up is proportional to the total volume of content, the overall time needed for overlap detection becomes linear and easily manageable at scale.

An *n*-gram overlap approach was employed by Cagé et al. (2015) and is the basis for an ongoing analysis of the reuse of scientific text in the arXiv (Citron & Ginsparg, 2015). An early computational study of news agency copy use by Scholten and Ruigrok (2009) used a similar technique.

Some research uses a proprietary plagiarism detector to detect overlap (e.g., Saridou et al., 2017), but tools designed for a nonresearch purpose do not generally have the controls needed for social scientific work. Saridou et al. (2017) report that their chosen detector did not provide reliable evidence of which source published first, requiring further manual work.

---

[3] An *n*-gram is simply a number of consecutive words in a document taken as a unit—so 3-grams are groups of three consecutive words, 5-grams are groups of five consecutive words, and so on.

Method

I implement an *n*-gram-based approach to detecting and measuring textual reuse, with the aim of it being both scalable and sensitive to small textual overlaps. Specific methods are employed to detect newswire content, press releases, and quotations. These methods aim to allow nuanced differentiation between different kinds of overlap and allow researchers to distinguish between ethically problematic forms of overlap (such as straightforward copying) and those that are morally unobjectionable (such as the coincidental use of stock phrases and the common use of quotes from public figures).

The Python code used to extract and process the *n*-grams and to conduct the analyses below is available.[4]

Data Collection and Preprocessing

I built a new corpus of news articles, comprising 162,297 articles. This is the full population of articles published online from April 13 to May 12, 2017, by 10 leading news sources in both the United States[5] and United Kingdom,[6] with articles fetched using the RISJbot Web crawler (Nicholls, 2017). To detect wire copy and PR content directly, I collected a secondary corpus of 125,508 articles from the four major news wires[7] (crawled from their websites and also collected from the Factiva database) and 24,645 press releases from the leading press release agency (PR Newswire, also collected by Web crawling) in both countries.

Each news article in the corpus was fetched and preprocessed by converting text to lowercase, stripping punctuation, and storing in a database. Accompanying metadata, including news source, date, and time were also stored.

Each article was then modeled as *w*-shingles (Manning et al., 2008, pp. 400–403), sets of distinct overlapping *n*-grams. Each *n*-gram was recorded in an indexed database table.

To model the texts, a decision was needed about *n*-gram length. There is not, at present, a standard approach for selecting the *n*-gram length; Cagé et al. (2015) modeled text as 5-grams, Citron and Ginsparg (2015) as 7-grams, and Saridou et al. (2017) as 4-grams. The shorter the *n*-gram length, the less likely that a given piece of text is genuinely from the same source,[8] whereas with longer *n*-grams, text with minor editing may not be detected if long runs of words are not identical between the sources.

---

[4] See Nicholls (2019).

[5] ABC, CBS, CNN, Fox News, NBC, *The New York Times, USA Today, The Washington Post*, and *Buzzfeed News US.*

[6] BBC News, Mail Online, *The Guardian, The Independent, Metro, The Mirror, The Sun, The Telegraph, Buzzfeed News UK,* and *Huffington Post UK.*

[7] AP, AFP, PA, and Reuters. I excluded articles less than 250 words long or consisting of market pricing data.

[8] The degenerate case is the 1-gram, where every article featuring the word *the* will be detected as having overlap.

For this analysis, I calculated both 5-grams and 7-grams, expecting the 5-gram analysis to have greater recall (picking up more examples of reuse, but also picking up more noise) and the 7-gram analysis to have greater precision (more reliably featuring only articles with substantial reuse, but being less sensitive to smaller and more edited overlap). The two approaches are evaluated in the "Validation" section below.

Inferring Publication Times

Although the method measures textual overlap between any article pairs, our substantive interest is normally more focused: We are interested in which was the original source of the text, and which articles are derivative. Although we cannot see the underlying mechanisms of copying, we can assume that only the second and subsequent articles with a given piece of text are potential copiers.[9]

Unfortunately, publication and modification times published by most news sites are deeply unreliable; pages' announced publication times can be later than the actual time of fetch, and many servers' time zones are misconfigured. Both these problems lead to false absolute times being given, in ways that cannot be systematically corrected.

Publication times are therefore detected using the Web-crawler's page fetch time. As it crawled essentially continuously, each article was fetched within 15 minutes of going live, and article fetch times more than 15 minutes apart therefore reflect the actual order of Web publication. A separate analysis showed very little nonwire reuse faster than 15 minutes; as wire content is detected separately, this constraint is not a significant limitation.

Detecting Newswire Copy

Detection of newswire content is important because the use of newswire copy is such a large proportion of reported textual reuse. I use three separate techniques that combine to give good coverage.

The first is to crawl the websites of the wires directly. However, only a small proportion of wire content is directly published by the wire providers themselves, and some wires (e.g., the Press Association) do not run public sites.

The second is to collect newswire content from a database, such as Factiva or LexisNexis.

The third is to rely on markers from news sites themselves. *Mail Online*, for example, republishes large volumes of wire content at URLs such as https://www.dailymail.co.uk/wires/ap/index.html; others will byline articles to "Agence France-Presse" or "John Smith at Reuters"; many will use a "(PA)" dateline; and yet others run a newswire copyright and attribution notice at the foot of the article. Where these are detectable algorithmically, the content of these articles can also be credited as wire content.

---

[9] This gives a number of false negatives: If the original source is outside the data set, we will treat the first-in-data-set publisher as presumptively original when it is a copy. This can be mitigated by expanding the data set with more sources, but cannot be entirely overcome (see the discussion in the Validation section).

Consequently, I code an article (and its component *n*-grams) as being attributed to a newswire if the article was from the newswire's website or Factiva content, if it possesses an Associated Press copyright notice, if it was published in a special section for wire content such as the /wires/ URLs on *Mail Online*, if its byline contains the name of any of the wires (or their abbreviations), if it starts with a newswire dateline, or if it contains a statement of the form "[newswire] contributed to this [report/story/article/post]."

Detecting Press Releases and Public Relations Copy

If PR material is available to the researcher, it can simply be added to the database in the same way as newswire copy and its presence detected in news articles in the same way. In my data set, I have done this for the PR Newswire feeds in the U.S. and UK. As credit is rarely given to PR agencies, the heuristics for detecting newswire attribution do not work here.

Where press releases have not been collected and added to the database, the first article using the text will be treated as the "originator," which represents a false negative. If more than one article has been based on the same underlying block of source text, then subsequent publishers will be correctly picked up, but with the text credited to the first publishing source in the database rather than to the PR author.

Identifying Quotations

Quotations are the raw material of many forms of reporting, and are routinely used when doing further reporting on a broken story. As such, it is valuable to be able to separate out their reuse from the copying of other forms of text. On the other hand, they also provide a useful marker of content being reused, and therefore should not be discarded.

I detect quotations and flag the *n*-grams that have occurred within them. Quotations are identified as text contained within matching pairs of quote markers ("", ‘’, or “”).[10]

Other Forms of Reuse

The most potentially problematic textual reuse is not from newswires, PR, or quotations, but from other chunks of text appearing in multiple articles. Straightforward copying within the corpus, with or without light editing, will be detected directly. If at least one phrase at least *n* words long (depending on the *n*-gram length chosen) is carried between articles unaltered, it will be detected and the second and subsequent appearances flagged as potentially derivative. After newswire-attributed, PR, and quotation *n*-grams have been identified, those remaining can be treated as unexplained and potentially problematic.

[10] This aims to detect (most) English and American-style quotations; the test will need adjustment for multinational or multilingual corpora. Reliably detecting quoted text is a surprisingly difficult problem, particularly multilingually (de La Clergerie et al., 2009; Pouliquen, Steinberger, & Best, 2007).

A small proportion of these overlaps are the result of independent uses of clichés and stock phrases. Two otherwise unrelated articles, for example, may both contain the phrase "in the first quarter of the year 2017," creating a few matching *n*-grams. Many of these can be detected by frequency analysis: If a phrase is in dozens of different articles, it is likely to have an independent source rather than being copied from one to the other. An approach to this, informed by the analysis of this data set, is given in the "Establishing Thresholds" subsection of the "Validation" section below. Similarly, artifactual overlaps such as newswire copyright notices appear in many articles. These are conceptually different from clichés and stock phrases in that they are not generally part of the original article text, but a result of the news sites' publication methods or researchers' data collection methods. Nevertheless, they can be treated similarly.

A further way to immunize analysis against these small incidental overlaps is to impose a threshold before treating articles as containing substantially overlapping material. This is also discussed in the "Establishing Thresholds" subsection below.

Analysis

In this section, I conduct a high-level analysis of this corpus and quantify different kinds of reuse. This contributes to the debates on churnalism and wire reuse and also serves as a demonstration of the method.

I conduct neither the most far-reaching nor the most sophisticated analysis possible using these techniques as a foundation. Comparative approaches could be used in future to explore differences within and between news outlets and/or countries, analyses of the spread of repurposed texts over time within organizations would give insight into the dynamics of updating online stories throughout the day, and analyzing the temporal dimension of cross-outlet textual reuse would potentially offer insights into the dynamics of repurposing and the extent to which publishing patterns are consistent with the perceived need to quickly follow breaking news.

For analysis, summary tables were constructed comprising all the distinct *n*-gram hashes found in the corpus and summarizing the extent of overlaps between each article and the rest of the corpus.

Descriptive Analysis of Data

Table 1 breaks down the total volume of content, counting each distinct 5-gram and 7-gram in the news corpus. This is broken down by quotations, attributed newswire content, and PR content picked up in the PR Newswire crawls. Three quarters of text is neither wire, PR, nor quote.

Table 1.* N-*Grams by Origin.

Origin	5-grams	7-grams
Not wire, PR or quote	44,242,266	48,123,103
In wire	9,114,382	9,769,114
In quote	4,927,902	4,766,652
In wire, in quote	1,348,848	1,152,195
In PR	55,999	9,586
In wire, in PR, in quote	42,602	9,212
In wire, in PR	42,182	7,596
In PR, in quote	36,553	6,368
Total	59,810,734	63,843,826

The source of the articles in the data set is shown in Figure 1; the high output for *Mail Online* and *The Washington Post* in particular is largely a consequence of their decision to publish large volumes of articles straight from the wire.

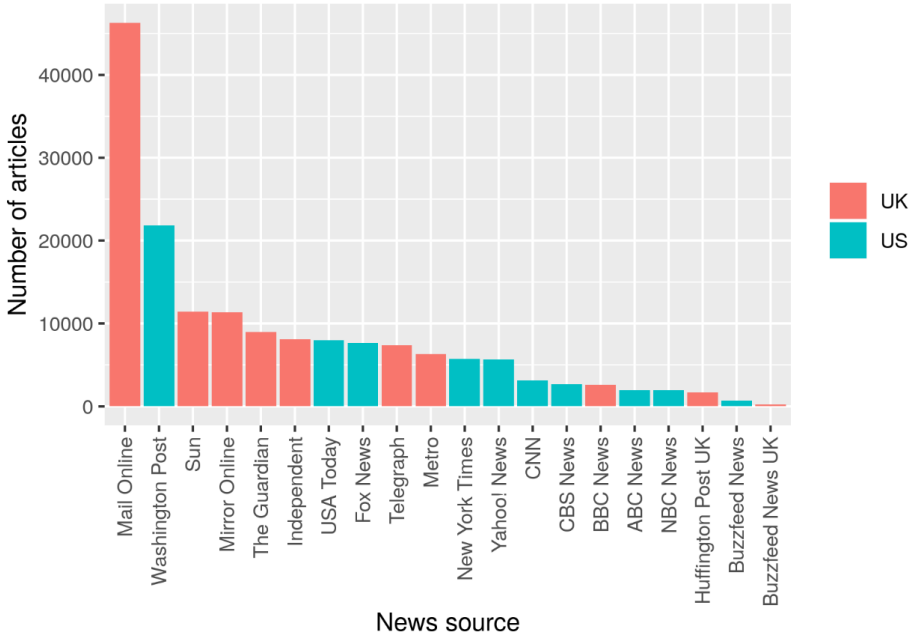

Figure 1. Articles by source.

Figures 2 and 3 show 5-gram and 7-gram frequencies in our data set; *n*-grams attributed to newswires are shown separately from those which are not. Both are log-log plots and show similar distributions: Tens of millions of *n*-grams appear once, whereas only a handful appear more than 100 times each.

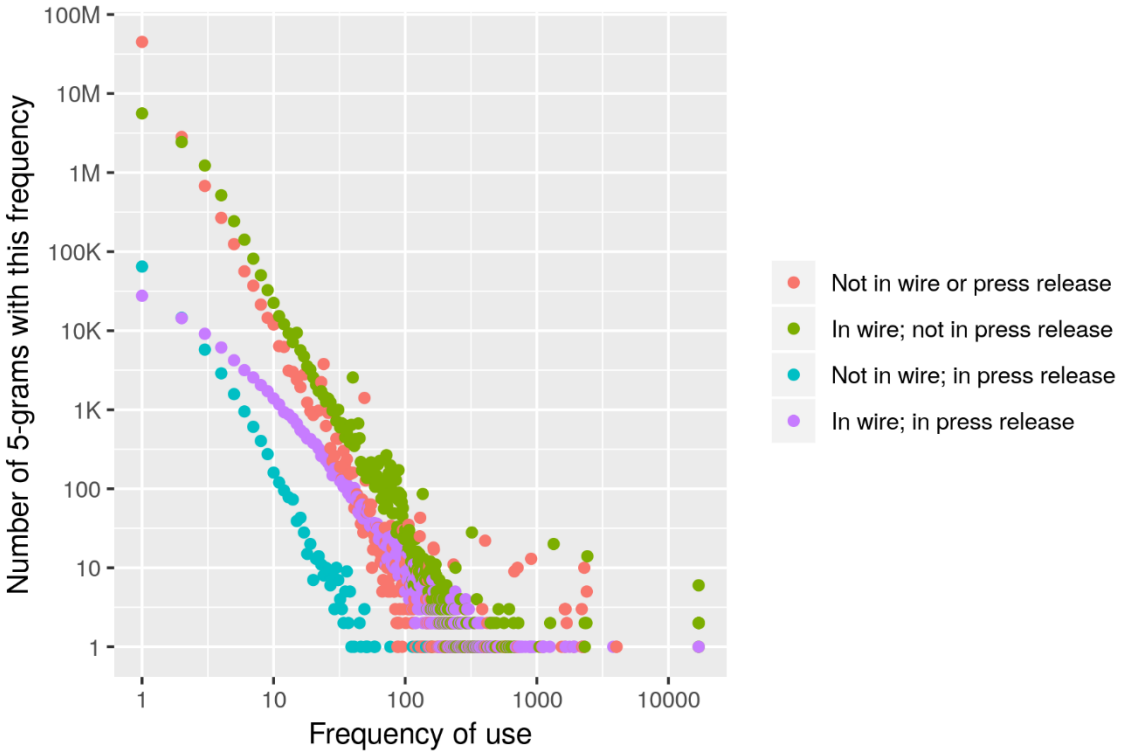

Figure 2. 5-gram frequency of use by news outlets.

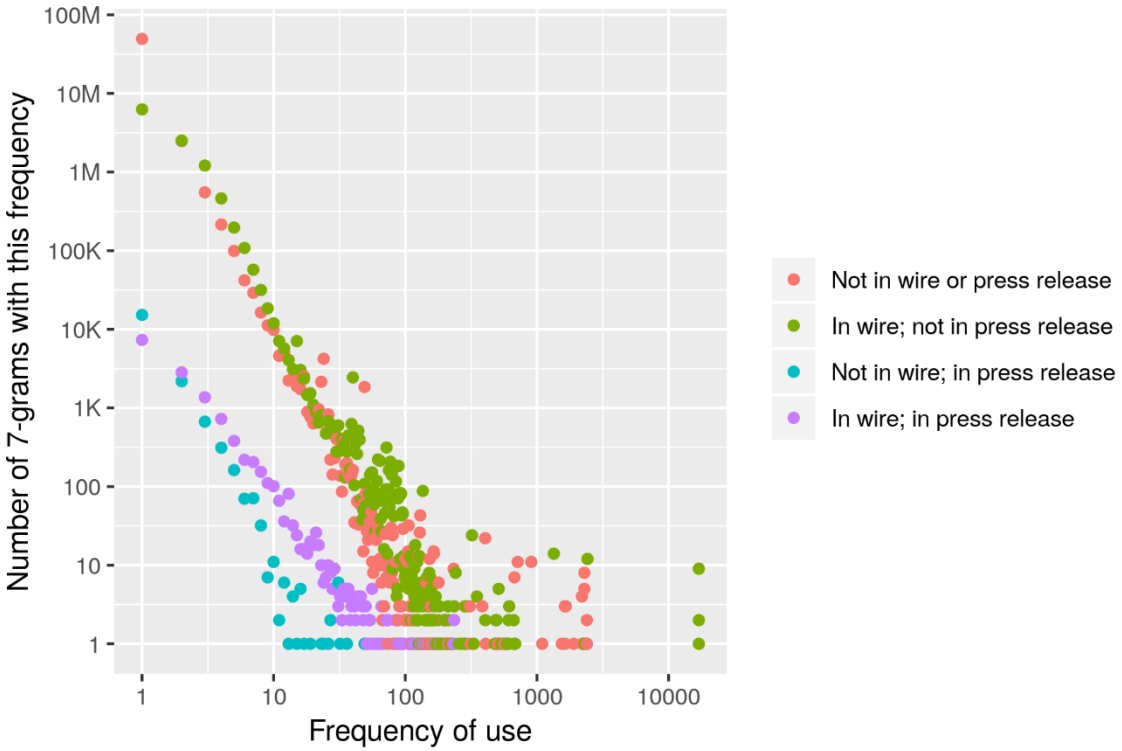

Figure 3. 7-gram frequency of use by news outlets.

The distribution at the right-hand tail is also instructive. The trend falls straightforwardly until *n*-grams have between 50 and 100 uses, beyond which there are a small number of cases with huge number of uses—the largest being twelve 7-grams and fourteen 5-grams, which appear in more than 17,000 news articles each. These most extreme cases represent newswire copyright notices[11] and not substantive reuse.

There are much lower levels of press release use, with the PR-and-newswire *n*-grams following a similar distribution to the newswire-only *n*-grams and the PR-only material following a similar distribution to non-PR news outlets.

[11] Specifically, the text "Copyright 2017 The Associated Press. All rights reserved. This material may not be published, broadcast, rewritten or redistributed." that appears at the end of some, but not all, republishing of AP material.

Reuse of Newswire and Press Release Text

Where newswire content has been heavily used, it has generally been credited: 75% of articles crediting the wires have at least 75% of wire content in them. For nonattributed content, results are much lower: 75% of articles have 1% or less of wire content (and 90% have 13% or less).

I also examine the supply side: the proportion of PR and newswire content disseminated that was actually reused by news sources. Figure 4 shows the 37.8 million distinct 7-grams identified as newswire and/or PR Newswire content. In particular, it distinguishes those that only appear in newswire or PR material and not in the main news corpus ("unused") and those that have also appeared in news outlets at least once ("used").

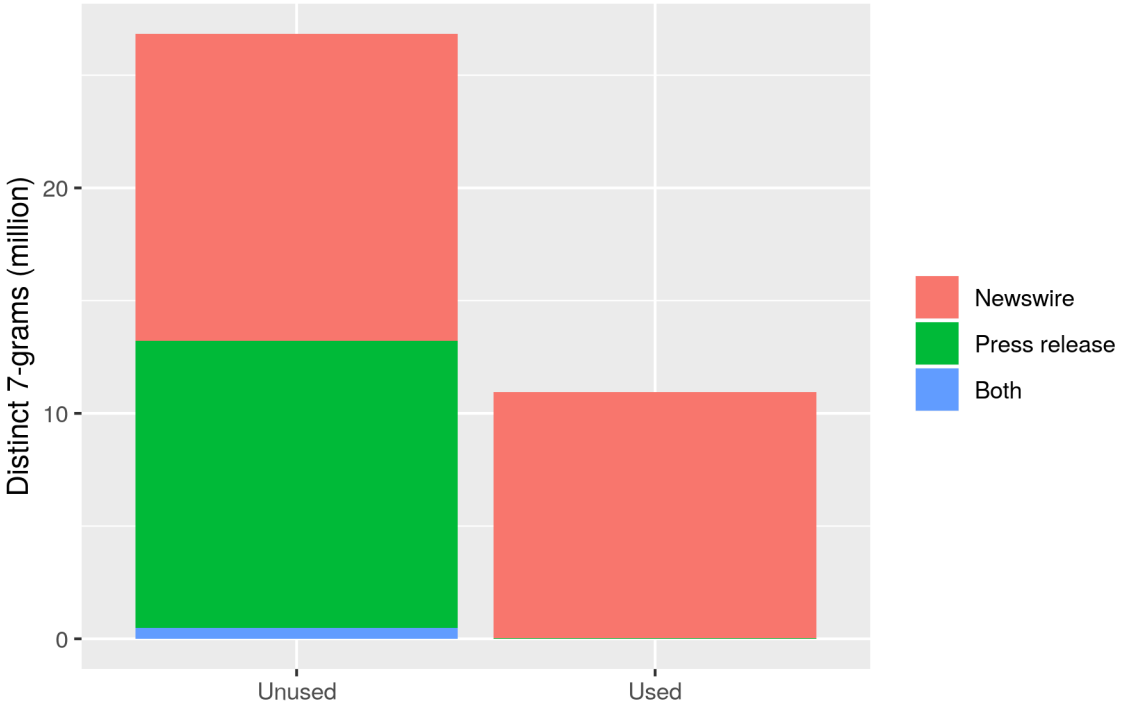

Figure 4. Use of PR and newswire copy.

About half of the content that went out on the four wires was used at least once by one of the main news outlets. A noticeable, but small, proportion of newswire content has been sourced from the press releases that was also collected (2.6% of 5-grams, 2.0% of 7-grams).

Almost none of the PR Newswire material made it into the main news sources: only 33 thousand of 13.2 million PR Newswire 7-grams also appeared in the main corpus (0.25%); 14 thousand of these were also in (and therefore potentially reached the press via) one of the newswires.

Total Text Reuse by Article

The analysis now moves from the individual *n*-gram to the article, reflecting the more usual unit of analysis in communication scholarship. The proportion of each article that consists of reused content is examined, broken down by category, by aggregating each article's component *n*-grams.

Several kinds of content are specifically excluded from each article for this analysis. First, *n*-grams that first appeared in other articles published by the same news source; these may represent earlier iterations of the same story, but are often noise and structural elements from the source's website that were imperfectly cleaned by the Web content extraction process.

Second, the very few *n*-grams that were first published less than 15 minutes before the article being examined were disregarded, as it is uncertain who published first in these cases.

Finally, any article with less than 5% of its content overlapping is treated as wholly original; low levels of reuse are also in practice coincidental or artifactual (see the "Validation" section below).

Overall, 70.1% of content is identified as original, 27.3% as newswire content, 1.4% as quotations, 1.1% as unexplained source-to-source overlap, and only 0.05% of content as being from PR Newswire press releases. This is graphically presented in Figure 5.

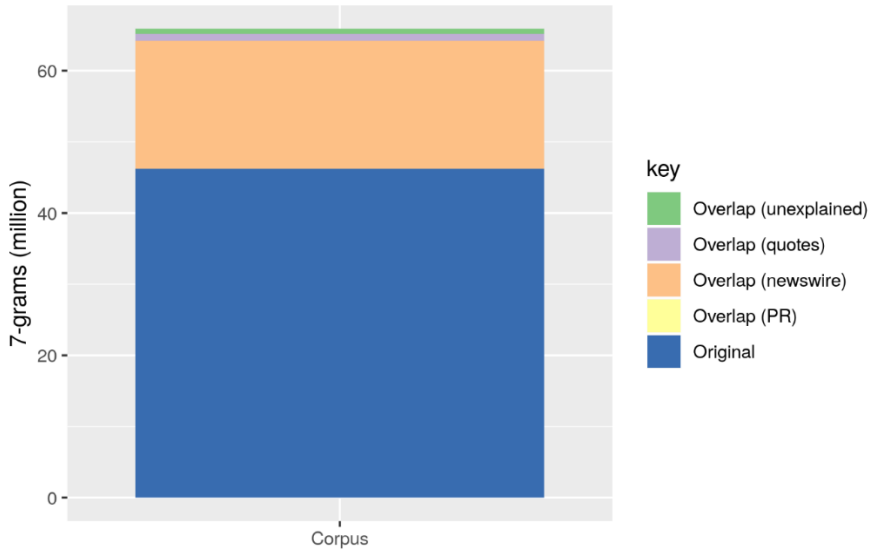

Figure 5. Overall levels of content reuse.

Finally, we look by article at the proportion of each article's 7-grams that are unexplained nonquote source-to-source variation.[12] A total of 74% of articles have zero unexplained source-to-source reused 7-grams, with only 4.2% having 5% or greater. The distribution is plotted in Figure 6.

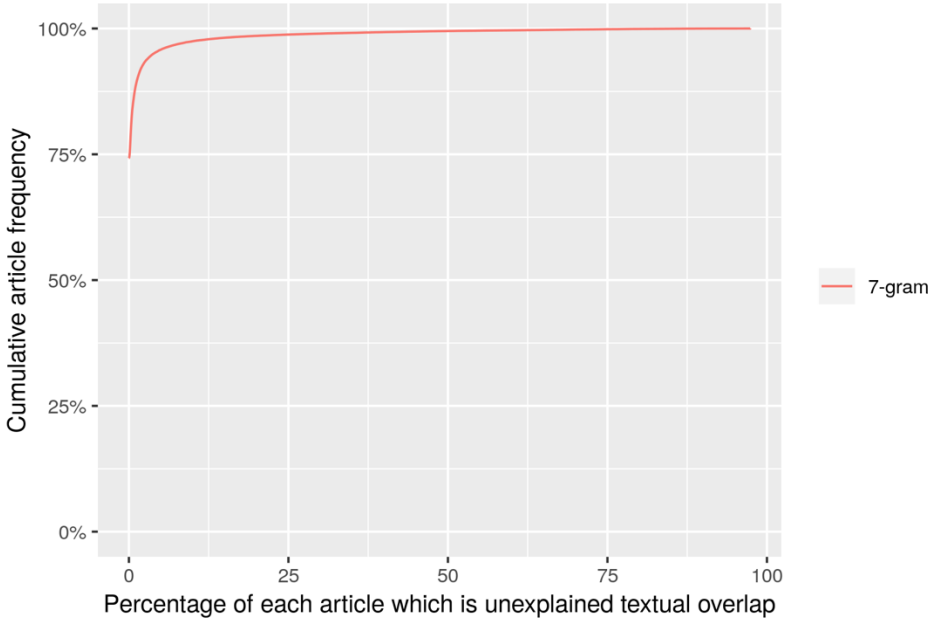

Figure 6. Percentage of nonoriginal, non-PR, nonquote material per article.

Validation and Limitations

To understand in more depth the nature of the overlaps that are being detected, and to check whether the quantitative results have reasonable face validity, a manual qualitative analysis of a sample of article pairs with overlapping textual content was carried out, as well as some validation of the assumptions of the method against the data set collected.

This section discusses the results of the qualitative analysis and establishes empirical thresholds for maximum *n*-gram frequency and minimum article *n*-gram overlaps to more easily apply the method in practice with new data. I also identify limitations of the method that might affect its suitability for future work.

[12] Note that this is after filtering the excluded content mentioned above; in particular, articles with less than 5% nonoriginal content are treated as coincidental or artifactual and therefore not unexplained.

A Qualitative Examination of Overlapping News Articles

I randomly sampled 152 article pairs with at least one 7-gram overlap, excluding from comparison press releases, newswire articles, and pairs of articles from the same outlet. I also excluded those news articles that either directly attributed themselves as newswire output or had at least 50% of their *n*-grams identified as newswire content. As this produced a sample with a high number of low-overlap articles (reflecting the data set as a whole), I supplemented with an additional sample of 65 articles with between 3% and 25% of overlapping 7-grams. I did not differentiate between content detected as quotes and content that was not. Intuitively, sharing quotes is at least an indication of sharing a subject, and I wanted to examine in depth those articles that did so.

For each article pair, the 7-grams identified as overlapping were examined and each article also read as a whole. The pair was then coded according to the primary cause of overlap as "coincidental," common-source "quote," "reuse—attributed," "reuse—nonattributed," or "other." The attribution distinction depended on whether the reused text was attributed by the news source to another source (often a wire or a competing paper).

All the high-overlap articles (with at least 20% of *n*-grams overlapping) in the analysis represented substantial textual reuse, in various ways and from various sources. Some were independently built on the same underlying source material—as when two separate stories are written on the same political event, but featuring the same extensive quotes—and some were straightforwardly derived from each other. But all were "true positives" in some sense.

More specifically, many of the high-overlap article pairs were both derived from social media content, with different outlets picking up the story either from each other or from social media trackers.[13] In these cases, the quotes from social media were extensive and largely in common, but the relatively sparser surrounding material and comment tended to be entirely distinct. This is in keeping with our understanding of high-throughput news production practices.

It is important to differentiate this observation, that there was substantial overlap, from an assumption that such overlap constituted plagiarism. Because of the various difficulties in characterizing the source of common text discussed above, this represents an extreme upper bound for improper content rather than an estimate.

Likewise, each of the very low-overlap articles (less than 5% of *n*-grams) either featured purely artifactual or coincidental overlaps or were written on a similar topic and featured only a quote or two in common.

In some cases, the extent of reuse is underestimated as a result of light editing of quotes. In the following example, different editing of a quote by two sources (differences underlined) results in only the

[13] A typical example is a pair of celebrity pregnancy stories. ABC News published their story (Messer, 2017) in the evening, U.S. time, and the UK's *Metro* published a similar story (Baillie, 2017) a few hours later, at 8:30 UK time, based on a very similar selection of quotes from the couple's Instagram page.

italicized portion being considered to be overlapping. Were 5-grams the measure, "get out and about and meet" would also register as two extra overlapping 5-grams.

> "We won't be doing the *television debates. I believe in campaigns where politicians* get out and about and meet the voters." (*The Mirror*)

> "We won't be doing *television debates. I believe in campaigns where politicians* actually get out and about and meet with voters. …" (*The Telegraph*)

Critically, though, this editing of the quotes has not prevented the overlap from being detected, but simply reduced our estimate of its magnitude.

In relation to those stories I examined that are clearly derived from other publishers, extensive rewriting is evident, to the extent that it is difficult to tell whether particular articles were patchwritten from another news source's story. Sources and their quotes are generally identical, so they are not being meaningfully rereported in any strong sense, but nonquote overlaps in the actual text are difficult to differentiate from innocent coincidence—no doubt intentionally.

Looking more broadly at the rest, much of what remains is common source material, even after allowing for quotes: Social media comments republished by multiple outlets are one major contributor.

Finally, attribution for prior reporting proved to be hugely variable, in contrast to the higher levels of attribution found for wire text above. One pair of similar articles on the aftermath of a traffic accident were published by British tabloid websites: first by *The Sun* (Hodge, 2017) and slightly afterward by *The Mirror* (Thompson & Mills, 2017). The former looks like an original story, whereas the latter made clear that the video and eyewitness quotes came via the *Leicester Mercury* (a local paper that is part of the same ownership group as *The Mirror*). It is likely that *The Sun* also picked up the story from the *Mercury*, but was simply less willing to publicize it.

Detection of PR and Newswire Copy

In the "Method" section, heuristics were offered for identifying content attributed to newswire sources by the publishing news outlet. I have also downloaded the full content of four key newswires from Factiva, thereby identifying the actual wire text available for use by those outlets.

I validated this in both directions. Is the Factiva feed incomplete? If so, we would expect to see stories to the wires by the story's publisher with no matching content in the feed. Or are the heuristics developed ineffective? If so, we would expect to see articles that heavily overlap with newswire content, which I have not identified as attributed, but which are.

Wire copy attribution methods vary by news outlet, and I have not attempted to set up detection heuristics for all of them. Consequently, some of the (few) high wire content articles highlighted as "wire—nonattributed" are likely to be misclassified. Simultaneously, not every *n*-gram in a wire attributed article

will be from the wire, as local edits are made. There is, therefore, a small amount of outlier variation in both directions.

Nevertheless, I can successfully identify, on current attribution detection alone, the large majority of reused wire copy and detect no large reservoir of attributed content that is not in the wire corpus.

This gives mutual confidence both that the Factiva wire feed used as a ground truth reflects the actual text used by news organizations, and that it is reasonable to identify wire copy heuristically when a full feed is not available.

There is no such parallel mechanism for identifying PR content, and the PR Newswire archive does not content all press releases, so we must be far more cautious about this.

Establishing Thresholds

Empirical thresholds would be useful when deciding when high-frequency *n*-grams should be treated as artifactual or coincidental (and thus should be discarded) and establishing the minimum level of *n*-gram overlap in an article that suggests substantial (rather than incidental) reuse. These thresholds would both allow some incidental overlap to be removed from the data set, reducing noise.

Establishing Thresholds: High-Frequency N-*Grams*

To test the assumption that high-frequency *n*-grams can be discarded, a sample of three hundred 7-grams with more than 20 uses was manually examined by a single coder and their sources ascertained. A few of the most frequent are listed in Table 2.

Every one of the 7-grams with at least 100 uses examined was incidental, not reflecting source copying between uses. Most were noise from the data collection process, such as a set of links to a *USA Today* slideshow that appeared on multiple articles and had not been excised during the crawling process—"Cavaliers Forward LeBron James (23) dunks in." Others were entirely coincidental phrases reflecting either named entities that reoccur in multiple stories ("White House Chief of Staff Reince Priebus") or stock phrases ("in the early hours of the morning").

Further down the frequency scale, there were a handful of 7-grams of interest. The high watermark were quotes of President Trump tweets, which appeared up to 70 times. For the 7-grams with less than 50 items, those featured in multiple outlets were generally more relevant; those featured in only one outlet were all noise.

Table 2. A Sample of Heavily Used 7-Grams, With Manual Classification.

N-Gram	Classification
the east room of the white house	Coincidental
white house chief of staff reince priebus	Coincidental
out for the rest of the season	Coincidental
did not respond to request for comment	Coincidental
in the early hours of the morning	Coincidental
on facebook at www facebook com thesun	Noise
cavaliers forward lebron james 23 dunks in	Noise
you live updates as soon as they	Noise
this report copyright 2017 the associated press	Noise
org copyright 2017 the associated press all	Noise
usa today sports fullscreen golden state warriors	Noise
dan hamilton usa today sports fullscreen april	Noise
the latest news gossip rumours and done	Noise
32 11 of 32 12 of 32	Noise
on facebook and stay updated on foreign	Noise

Although only 18,100 of 90.7 million 7-grams in our database were used 50 times or more (0.02%), reuse on the scale of hundreds or thousands of times per *n*-gram will clearly have a disproportionate effect on measurements of reuse and should therefore be discounted. I therefore recommend as a rule of thumb for future work that 7-grams that appear 50 or more times should be disregarded as likely artifactual.[14]

N-grams that occur only once could also be discarded, to reduce database size, subject to two important caveats. First, if further articles might later be added, this will prevent accurate overlap analysis for the later texts. Second, the total *n*-gram count must be recorded before discarding, so the denominator on later-calculated statistics is still correct.

[14] As the frequency of *n*-gram reuse is roughly power-law distributed, the total number of *n*-grams in the corpus will only slightly affect the appropriate cutoff. For analyses with radically different volumes of content than the 90.7 million 7-grams in our study, a graph similar to Figures 2 and 3 can be plotted; the point at which the straight-line log-log relationship between frequency and volume flattens out should be a suitable cutoff.

Establishing Thresholds: Proportion of Reused N-*Grams in Each Article*

Intuitively, a thousand word article with only one five-word phrase in common with another thousand word article is unlikely to reflect copying or the use of substantially similar source material, but rather reflect coincidental language use or artifactual noise. By establishing the lowest level at which substantive reuse appears to be occurring, we can establish a minimum threshold to consider reuse potentially problematic, and thereby separate out some noise automatically.

It is necessary to strike a balance between a too-low threshold and a too-high one that discards potentially marginal cases such as those based on very similar source material, but a different write up. The distribution of reused *n*-grams in this data set is assessed in the "Analysis" section above.

This is not a straightforward question, as the best approach will depend on the research questions being answered, and consequently whether false positives or false negatives are preferred. On the basis of the manual examination above, I recommend that, for general purposes, articles with up to 5% of 7-grams overlapping should be disregarded. This is conservative and would have discarded very little of the potentially problematic material in the qualitative study. Given the higher noise floor for 5-grams, a higher threshold of 10% might be appropriate.

N-*Gram Length*

As predicted, the 5-gram measure is more sensitive but noisier, whereas the 7-gram measure by contrast features higher precision, but lower recall. The best example of the specificity trade-offs with shorter *n*-grams is the existence of stock phrases, appearing as textual overlap despite being independent. In the 7-gram corpus, phrases like "at the University of Texas at Austin" appear repeatedly, but there are relatively few of them, as most institutions have shorter names. With 4-grams, practically every appearance by academics at any university would generate overlaps. The level of noise is, therefore, visibly lower with the larger size. The challenge with large *n*-grams is that lightly edited material is less likely to trigger a detection. A 10-gram measure, for instance, only requires one word in every 10 to be changed for the text not to be noted as copied at all.

On balance, specificity is preferred for this kind of analysis, particularly given the opprobrium that can be attracted to charges of improper copying. I recommend the use of 7-grams, noting that this aligns with the existing practice of the arXiv (Citron & Ginsparg, 2015).

Limitations

Although the method is robust to lightly edited copied text, it is not necessarily robust against a copier aware of the method who wishes to evade detection. If a writer is prepared to systematically alter at least every seventh word, then the text will not be detected. I have seen no evidence of this in the manual validation, but it remains a potential limitation, as it does for uses of similar techniques, such as academic plagiarism detection.

When comparing news outlets, it is important to carefully avoid multiple comparison problems. Consider a group of several articles from various sources: the first being a news agency S_a and four other news outlets that run edited versions of the wire article, in time order S_1 . . . S_4. If comparisons are done pairwise, S_a will correctly be credited as the source of text for each of the other sources. But when S_1 is compared against S_2 . . . S_4, it will be erroneously also credited as the source of text for the later articles; ditto for S_2 (twice) and S_3 (against S_4 only).

This can be avoided by coding each *n*-gram appearance in each text as "first" (the first appearance of the *n*-gram, therefore assumed to be original) or "subsequent" (has appeared in an earlier time-stamped article, so assumed to be derivative). This avoids multiple comparisons and treats the first user of each *n*-gram as its presumptive source.

Nevertheless, it is not straightforward to determine if the first source in the database to publish a piece of text wrote it or obtained it from a source outside the corpus. If a story was first published by an out-of-sample local paper or a press release, we will not detect the reuse for the first source in the corpus (a false negative). We will underestimate copying in this case, by crediting the first-publishing source in our data set (though not the subsequent ones) with original content.

More generally, I treat the processes of reusing text as unobservable and focus only on the articles produced. By examining the differences induced in the texts by different authors as they are repurposed, it may be possible to model the routes by which each outlet found and edited the copied material (analogously to the work done by Leskovec, Backstrom, & Kleinberg, 2009, on meme tracking).

Conclusions

Overall, the data find less textual reuse than that of the previous studies discussed in the literature review above (Boumans, 2017; Boumans et al., 2018; Cagé et al., 2015; Saridou et al., 2017; Welbers et al., 2018). This is despite extensive data collection and conservative thresholds, meaning that little is excluded as incidental or artifactual.

Most of the identified textual reuse in our corpus is the attributed use of newswire copy; we found much less text overlapping with the PR Newswire press release archive than expected and also less wholesale overlap between news sources. Although many articles contain small percentages of overlapping text, these are largely explainable as common-source quotations, data collection noise, and stock phrases.[15]

Some of the differences in newswire use between this study and those reporting higher numbers may be the result of different study populations; Dutch, French, and Greek news sources could have different news sourcing practices from the wires than those in the UK and the U.S. For the remaining differences, some are likely to reflect differences in the study methodologies.

[15] Two articles in the data set, for example, shared the phrase "In the first quarter of the year 2017": four 5-grams and two 7-grams in common, but clearly unrelated.

For the newswire content, I validate the method and conclude it is robust. It offers a complementary measurement to the cosine-similarity-based analyses: By measuring reuse of each *n*-gram separately and precisely, rather than an overall similarity to newswire sources, it accurately and sensitively tracks low- and medium-scale textual reuse. That there are such different headline results from different measurement approaches and different sets of newspapers suggests that it is right to measure in different ways to triangulate answers.

The very small quantity of press release copy detected is more surprising, given previous findings that indicate, for example, "that 1 in every 10 newspaper articles is initiated by a press release; for the [news] agency this is slightly higher" (Boumans, 2017, p. 1). We can conclude that the low volume of PR Newswire text reuse does not necessarily reflect a low level of PR material use—just a low level of use of that publicly dispatched to the PR Newswire. This supports the analysis of the changing nature of PR work highlighted by Jackson and Moloney (2016) and reinforces the need for care when assembling data sets intended to be representative. Given the different results found by Boumans (2017) using press release data collected directly from organizations' websites and staff, differences in source data rather than method could be responsible.

Conversely, the overlap between newswire and press release content analyzed above and shown in Figure 4 reinforces the importance of analyzing both together, rather than assuming that press release and newswire content are disjoint sets, supporting the observation that PR professionals work through newswire journalists as well as those working for regular news organizations (Kiernan, 2003).

The methods developed and evaluated in this article have strong potential for future application in the analysis of news text at scale. They would support, for example, a temporal analysis of the extent to which news providers update breaking news reports after their initial release, a comparative analysis of the news writing processes of different news organizations, or (with suitable developments) the spread of particular soundbites or news framings among news organizations.

Scope remains for the further development of these techniques. To give one example, authorship is not considered, although many news sources provide author credits. Analyzing these alongside the textual content would allow for the more straightforward differentiation of reuse of text by the same author in different publications (such as might be the case for op-eds or republishing in multiple papers that are part of the same ownership group) from repurposing by different authors. Likewise, Citron and Ginsparg (2015) treat the use of common *n*-grams in academic texts by many nonconnected authors as evidence of coincidental use; this may be an effective supplement to the frequency analysis demonstrated in this article if sufficiently reliable authorship data are available.

References

Baillie, K. (2017, May 5) Nikki Reed and Ian Somerhalder announce they are having a baby with cute Instagram snap. *Metro*. Retrieved from https://metro.co.uk/2017/05/05/nikki-reed-and-ian-somerhalder-announce-they-are-having-a-baby-with-cute-instagram-snap-6616984/

Beckett, C. (2008, March 31). How to end churnalism [Web log message]. Retrieved from http://blogs.lse.ac.uk/polis/2008/03/31/how-to-end-churnalism/

Boczkowski, P. J. (2009). Technology, monitoring, and imitation in contemporary news work. *Communication, Culture and Critique*, *2*(1), 39–59. doi:10.1111/j.1753-9137.2008.01028.x

Boumans, J. (2017). Subsidizing the news? *Journalism Studies*, *19*(15), 2264–2282. doi:10.1080/1461670X.2017.1338154

Boumans, J., Trilling, D., Vliegenthart, R., & Boomgaarden, H. (2018). The agency makes the (online) news world go round: The impact of news agency content on print and online news. *International Journal of Communication*, *12*, 1768–1789. Retrieved from http://ijoc.org/index.php/ijoc/article/view/7109

Cagé, J., Hervé, N., & Viaud, M.-L. (2015). *The production of information in an online world: Is copy right?* (Working Paper No. 15-05). Paris, France: NET Institute. Retrieved from https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2672050

Citron, D. T., & Ginsparg, P. (2015). Patterns of text reuse in a scientific corpus. *Proceedings of the National Academy of Sciences*, *112*(1), 25–30. doi:10.1073/pnas.1415135111

Davies, N. (2008). *Flat earth news: An award-winning reporter exposes falsehood, distortion and propaganda in the global media*. London, UK: Chatto & Windus.

de La Clergerie, É., Sagot, B., Stern, R., Denis, P., Recourcé, G., & Mignot, V. (2009). Extracting and visualizing quotations from news wires. In Z. Vetulani (Ed.), *Human language technology: Challenges for computer science and linguistics* (LTC 2009: Lecture Notes in Computer Science, *6562*, pp. 522–532). Berlin, Germany: Springer. doi:10.1007/978-3-642-20095-3_48

Ganascia, J.-G., Glaudes, P., & Del Lungo, A. (2014). Automatic detection of reuses and citations in literary texts. *Literary and Linguistic Computing*, *29*(3), 412–421. doi:10.1093/llc/fqu020

Graves, L., Nyhan, B., & Reifler, J. (2016). Understanding innovations in journalistic practice: A field experiment examining motivations for fact-checking. *Journal of Communication*, *66*(1), 102–138. doi:10.1111/jcom.12198

Hodge, M. (2017, May 8). THAT'S TEAM-MERC! Amazing moment 15 passers-by LIFT Mercedes A Class off a seven-year-old boy after road crash. *The Sun*. Retrieved from https://www.thesun.co.uk/news/3511699/passers-by-lift-mercedes-boy-leicester/

Jackson, D., & Moloney, K. (2016). Inside churnalism: PR, journalism and power relationships in flux. *Journalism Studies*, *17*(6), 763–780. doi:10.1080/1461670X.2015.1017597

Johnston, J., & Forde, S. (2011). The silent partner: News agencies and 21st century news. *International Journal of Communication*, *5*, 195–214. Retrieved from https://ijoc.org/index.php/ijoc/article/view/928

Johnston, J., & Forde, S. (2017). Churnalism: Revised and revisited. *Digital Journalism*, *5*(8), 943–946. doi:10.1080/21670811.2017.1355026

Kiernan, V. (2003). Embargoes and science news. *Journalism & Mass Communication Quarterly*, *80*(4), 903–920. doi:10.1177/107769900308000410

Leskovec, J., Backstrom, L., & Kleinberg, J. (2009). Meme-tracking and the dynamics of the news cycle. *Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (pp. 497–506). New York, NY: ACM. doi:10.1145/1557019.1557077

Lewis, J., Williams, A., & Franklin, B. (2008). Four rumours and an explanation. *Journalism Practice*, *2*(1), 27–45. doi:10.1080/17512780701768493

Lewis, J., Williams, A., Franklin, B., Thomas, J., & Mosdell, N. A. (2006). *The quality and independence of British journalism*. Cardiff, UK: Cardiff University.

Lorenz, H. (2017). News wholesalers as churnalists? The relationship between Brussels-based news agency journalists and their sources. *Digital Journalism*, *5*(8), 947–964. doi:10.1080/21670811.2017.1343649

Manning, C. D., Raghavan, P., & Schütze, H. (2008). *Introduction to information retrieval*. Cambridge, UK: Cambridge University Press. Retrieved from https://nlp.stanford.edu/IR-book/information-retrieval-book.html

Messer, L. (2017, May 4). Ian Somerhalder and Nikki Reed are expecting their 1st baby. *ABC News*. Retrieved from https://abcnews.go.com/Entertainment/ian-somerhalder-nikki-reed-expecting-1st-baby/story?id=47220578

Nicholls, T. (2017). RISJbot: A scrapy project to extract the text and metadata of articles from news websites (Version 1.0.0) [Computer software]. doi:10.5281/zenodo.1168542

Nicholls, T. (2019). Data processing and analysis code for 'Detecting Textual Reuse in News Stories, At Scale' in the International Journal of Communication [Computer software]. doi:10.5281/zenodo.3338003

Nicholls, T., & Bright, J. (2019). Understanding news story chains using information retrieval and network clustering techniques. *Communication Methods and Measures*, *13*(1), 43–59. doi:10.1080/19312458.2018.1536972

Pouliquen, B., Steinberger, R., & Best, C. (2007). Automatic detection of quotations in multilingual news. *Proceedings of Recent Advances in Natural Language Processing* (pp. 487–492). Retrieved from http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.63.1807&rep=rep1&type=pdf

Saridou, T., Spyridou, L.-P., & Veglis, A. (2017). Churnalism on the rise? Assessing convergence effects on editorial practices. *Digital Journalism*, *5*(8), 1006–1024. doi:10.1080/21670811.2017.1342209

Scholten, O., & Ruigrok, N. (2009). Bronnen in het nieuws: Een onderzoek naar ANP-berichten in nieuws en achtergrondinformatie in Nederlandse dagbladen 2006–2008 [Sources in the news: An investigation into ANP messages in news and background information in Dutch newspapers 2006–2008]. *Mediamonitor*. Retrieved from http://www.mediamonitor.nl/gastauteurs/otto-scholten-en-nel-ruigrok-2009/

Thompson, A., & Mills, K.-A. (2017, May 8). Strangers save little boy's life when they LIFT the two tonne car he was trapped under. *The Mirror*. Retrieved from https://www.mirror.co.uk/news/uk-news/strangers-save-little-boys-life-10379930

Welbers, K., van Atteveldt, W., Kleinnijenhuis, J., & Ruigrok, N. (2018). A gatekeeper among gatekeepers. *Journalism Studies*, *19*(3), 315–333. doi:10.1080/1461670X.2016.1190663