Abstract—Misinformation, propaganda, and outright lies proliferate on the web, with some narratives having dangerous real-world consequences on public health, elections, and individual safety. However, despite the impact of misinformation, the research community largely lacks automated and programmatic approaches for tracking news narratives across online platforms. In this work, utilizing daily scrapes of 1,404 unreliable news websites, the large-language model MPNet, and DP-Means clustering, we introduce a system to automatically identify and track the narratives spread within online ecosystems. Identifying 55,301 narratives on these 1,404 websites, we describe the most prevalent narratives spread in 2022 and identify the most influential websites that originate and amplify narratives. Finally, we show how our system can be utilized to detect new narratives originating from unreliable news websites and to aid fact-checkers like Politifact, Reuters, and AP News in more quickly addressing misinformation.

1. Introduction

Over the last decade, spurious, misleading, and outright false information has spread throughout online ecosystems [1]. Digital misinformation has influenced elections [2], promoted bogus health cures leading to unnecessary deaths [3], and incited mob violence throughout the world [4], [5]. Worsening the problem, misleading stories have been shown to spread at over ten times the rate of true information [6].

The security community, likening disinformation to an attack similar to spam, phishing, and censorship [7], [8], has applied a range of methodologies to ameliorate its spread [9], [10], [11], [12], [13], [14]. For example, by examining features similar to those used to identify spam accounts, researchers have identified networks of state-propagandists throughout Reddit and Twitter [11], [15]. However, despite these advances, most investigations into false narratives remain limited in scope and retroactive, primarily conducted through time-consuming, qualitative approaches [16], [17]. To make fundamental progress in combating the threat posed by disinformation, we argue that the security community needs to build technical approaches for tracking the online spread of false narratives globally and in real time.

In this work, we present an NLP-based approach for programmatically identifying and tracking the spread of narratives and stories across unreliable news websites and social media platforms. Between January 1 and November 1, 2022, we crawl a set of 1,404 known misinformation, state-propaganda, biased, and otherwise unreliable news websites as well as two fringe forums, 8kun and 4chan. We extract passages from these news articles, which we embed using an MPNet model [18] fine-tuned with contrastive learning on the semantic textual similarity task. Then, employing a modified version of nonparametric DP-Means, we cluster these embeddings to identify specific narratives/stories.

Our approach enables us to isolate and track 55,301 narrative threads that spread on unreliable news websites during 2022. We do not attempt to determine whether individual stories are factual, which is a qualitative task that ML-based approaches have failed to reliably achieve. Rather, we track all narratives from these specious sites across online ecosystems and quantify these websites’ influence. We find, unsurprisingly, that many of the most prominent narratives in 2022 concerned the Russo-Ukrainian War and inflation, with websites like southfront.org, rt.com, and zerohedge.com dominating these topics. Next, identifying which websites play outsized roles in originating and amplifying narratives across our set of unreliable news websites, we find that a website’s popularity has a small correlation with its ability to propagate narratives with seemingly minor websites like infostormer.com or barenakedislam.com playing massive roles in popularizing stories.

Next, we investigate how our method can be used to focus limited investigative resources on the most pernicious narratives. We show that, like an early detection alarm, our approach can identify when new narratives emerge. Comparing when popular false narratives appeared to when three major organizations (AP News, Reuters, and Politifact) fact-checked them, we show that our system can prioritize checking misleading narratives months before their peak when they first start to gain traction. We hope that this type of real-time visibility can enable fact-checkers and journalists to more efficiently track and respond to new, potentially problematic narratives as soon as they emerge.

Similar to past large-scale empirical analysis in the security community (e.g., [19], [20], [21], [22], [23], [24], [25], [26], [27], [28]), our work shows that a programmatic approach to tracking narratives at scale uncovers a set of
online propagation patterns that would have been difficult to uncover through manual, small-scale investigations. We also discuss how having a continuous tracking process can help analysts uncover and track the most worrisome influence operations. We stress that our approach does not make factual judgments on individual stories or on the reliability of websites. Indeed, our system takes as input websites that human experts previously labeled as unreliable. Rather, our approach provides the critical, real-time visibility into the spread of news narratives online that human experts need to effectively identify and respond to misinformation.

2. Background

Unreliable information often spreads through multiple avenues as individual users, state-supported actors, and even entire platforms participate in the dissemination of falsities. Unreliable information can take the form of misinformation, disinformation, fake-news, propaganda, among others [29]. Misinformation is any information that is false or inaccurate regardless of the author’s intent [29], [30], [31], [32]. The term “fake news” is often used interchangeably with misinformation. Disinformation, in contrast, is inaccurate information spread with the deliberate purpose to mislead [29], [33]. Similar to disinformation, propaganda refers to “deliberate, systematic information campaigns, usually conducted through mass media formats” regardless of whether the information is true or false [29]. A single narrative can be considered disinformation when spread by a state actor and as misinformation when spread by users. We refer readers to Jack et al. [29] for a more detailed taxonomy.

3. Methodology

The goal of our work is to programmatically track how narratives spread amongst “unreliable” news websites. In this section, we define a narrative and then describe how we collect news articles and extract narratives. We emphasize that while we focus on websites known to publish misleading, false, or state-controlled narratives, we do not assume that all narratives from these sites are “misinformation”. Indeed, many stories are not. We do not seek to label new narratives as “misinformation”, which is a qualitative, investigative task. As such, we refer to only individual stories that have been previously identified by experts as misinformation or disinformation, as false. We label the websites that spread these verified false narratives as unreliable.

3.1. Narrative Definition

Tracking misinformation narratives requires a high degree of specificity. Unlike traditional topic modeling, which seeks to identify themes and/or statistical word correlations [34], [35], [36], misinformation tracking requires distinguishing between specific narratives and stories. Within this work, we define a narrative/story using the same definition as the Event Registry [37], Hanley et al. [38], and Miranda et al. [39]: collections of documents that seek to address the same event or issue. For example, two example events in the Event Registry are “Felix Baumgartner’s jump from a helium balloon on October 14, 2012” and “bombings during the Boston Marathon on April 15, 2013.” Within our dataset, events constitute ideas like “election fraud in the 2020 US election” and “the COVID-19 vaccine leading to mass death.” An example of two ideas—while related—that we do not consider to be the same narrative are “US funds Ukrainian War” and “Russia attacks Ukraine.”

3.2. System Architecture

As shown in Figure 1, our system (1) collects news articles from unreliable news sites through web scraping on a daily basis and from Common Crawl data [40]; (2) parses out the news articles, extracting article text, which segmenting them into constituent passages; and (3) embeds text passages into a shared subspace utilizing the large language model MPNet [18]. To track the spread of narratives, our system then (4) clusters semantically similar content using DP-means, and (5) extracts keywords and auto-generated summaries to provide human-interpretable labels.
We opt for this LLM-based approach because our system needs to track specific narratives. Prior approaches that utilize simpler, more generic keyword-based topic modeling tools like LDA fall short in identifying specific narratives across different news websites [41]. Furthermore, keyword-based approaches often rely on pre-existing expert knowledge of disinformation campaigns and largely cannot adapt to the rapid pace of the news ecosystem [42]. By utilizing this new approach, our system can update and track news stories without a priori or domain-specific knowledge in an efficient and fine-grained manner.

For this paper, we use data from January 1 to November 1, 2022, but we emphasize that our system runs continuously, enabling us to identify new narratives in near real time. In the rest of this section, we detail each step of our methodology and ensure that our system captures specific and coherent narratives.

### 3.3. Data Collection

Our study is based on scraping and parsing articles from websites known to spread unreliable information.

**Unreliable News Websites.** We collect articles from 2,514 candidate websites that have been labeled as “politically biased”, “misinformation”, “disinformation”, “conspiracy”, “fake news”, or “state-based propaganda” by past studies (Iffy Index [43], OpenSources [44], Politifact [45], Snopes [46], Melissa Zimdars [47], and Hanley et al. [48]). This list includes politically-biased websites like dailywire.com, conspiracy-oriented websites like x22report.com, and state-propaganda outlets like rt.com.

**Scraping Articles.** We crawl websites using Colly\(^1\) and Headless Chrome orchestrated with Python Selenium. For each website, we collect the homepage and linked pages daily from January 1 to November 1, 2022. To ensure full coverage of each site’s published articles, we additionally gather the HTML pages indexed by Common Crawl [49] for each site during this same period. We emphasize that under 1% of articles were only found in the Common Crawl dataset, indicating that our scrapes found the vast majority of published content on each site. We then parse each HTML page to collect the published articles using the Python libraries newspaper3k and htmldate. There were several instances (e.g., sputniknews.com) where this approach failed; in these cases, we built custom parsers based on site-specific HTML elements.

Of our 2,514 candidate websites, 1,404 were operational and published articles during our 2022 measurement period (many sites that spread unreliable information are short-lived [12], [50]). Altogether, we collect 2,126,738 articles. We provide this data to researchers upon request.

### 3.4. Preprocessing and Embedding

To prepare data for embedding, we first remove any non-English articles using the Python langdetect library and

1. https://github.com/gocolly/colly

---

**Table 1:** Evaluation—based on Pearson Correlation—of our MPNet-contrastive model and other models on the SemEval STS-Benchmark [53]. Data for GloVe [54], BERT [55], the Universal Sentence Encoder (USE) [56] is from Reimers et al. [57].

<table>
<thead>
<tr>
<th>Model</th>
<th>Precision</th>
<th>Recall</th>
<th>F1-Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>GloVe</td>
<td>0.580</td>
<td>0.464</td>
<td>0.749</td>
</tr>
<tr>
<td>BERT</td>
<td>0.749</td>
<td>0.840</td>
<td>0.856</td>
</tr>
<tr>
<td>USE</td>
<td>0.856</td>
<td>0.924</td>
<td>0.871</td>
</tr>
<tr>
<td>All-MPNet</td>
<td>0.856</td>
<td>0.924</td>
<td>0.871</td>
</tr>
<tr>
<td>Our Model</td>
<td>0.856</td>
<td>0.924</td>
<td>0.871</td>
</tr>
</tbody>
</table>

---

**Figure 2:** Evaluation of our model’s precision, recall, and F1 scores on the English portion of the SemEval22 dataset [60] (using 3.0 as the cut-off for the two articles being about the same event [52]).
We specifically adapt Dinari et al.’s efficient and parallelizable version of DP-Means [64] (Appendix C), making four main alterations. First, we cluster embeddings based on their cosine similarity rather than their Euclidean distance. We set $\lambda = 0.60$ (the minimum cosine distance an embedding can be from a cluster before a new cluster is created) to ensure that clusters have high semantic similarity, informed by our prior manual investigation (Table 2). Second, we perform partial fits over each day’s worth of news article embeddings. Specifically, on each day throughout 2022, we embed that day’s passages and update the previous day’s cluster centers (i.e., we update our given clusters on a daily basis with the DP-Means algorithm until convergence utilizing that day’s article embeddings). Third, we remove the random reinitialization of clusters added by Dinari et al. [64] from the algorithm; we find that this step often led to over-clustering given that many website passages are slight variations of each other. Lastly, we note that rather than relying on Dinari et al.’s released code, we re-implement their algorithm to take advantage of the matrix multiplication speedups that come from utilizing a GPU (3 times speedup with an Nvidia RTX A6000).

For this work, we utilize the clusters from November 1, 2022. From January 1 to November 1, clustering all embeddings required the equivalent of 1.5 days. We filter out clusters where 50% or more of the passages are from only one website (e.g., author bios) or there were fewer than 25 passages to remove spam, similar to the methodology specified by Leskovec et al. [17]. After this removal, we identify 55,301 narrative clusters. Each article’s passages are part of an average of 5.02 narrative clusters (4.0 median). On average, each embedding has an average similarity of 0.686 to its cluster center, which shows that our embeddings are assigned to clusters with high semantic similarity. Each narrative cluster has an average cosine similarity of 0.0169 with other identified narrative clusters, which indicates that our approach identified distinct narratives.

### 3.7. Interpretability and Narrative Specificity

We create human-interpretable identities for our narrative clusters using two approaches. First, we extract the most distinctive and representative keywords of the cluster using pointwise mutual information [65], [66] (detailed in Appendix D). Pointwise mutual information (PMI) is an information-theoretic measure for discovering the associations amongst different entities [66]. As in Kessler et al. [67], rather than finding the pointwise mutual information between different words, we utilize the measure to understand individual words’ association with narrative clusters. In this manner, we find the set of words most distinctive/associated with each cluster. Second, after identifying the top five passages closest (i.e., with the largest cosine similarity) to the center of the cluster, we subsequently use an off-the-shelf state-of-the-art BART [68] summarization tool from Huggingface fine-tuned on news data, to summarize the contents of the cluster. We utilize this approach because while keywords provide an identifiable “handle” for...
each cluster, keywords typically do not fully capture the full semantic meaning or the specificity of our narrative clusters. For example, the auto-generated summary for the cluster with keywords *Age, Pfizer, Booster, Children, Vaccine* is:

The U.S. Food and Drug Administration FDA in October 2021 authorized the PfizerBioNTech COVID vaccine for children 5 through 11. Children under 5 remain the only segment of the US population that isn’t eligible for one of the COVID vaccines.

where a random passage from the cluster states:

As of now, U.S. children aged five and older are eligible for the COVID19 vaccine, though only Pfizer’s shot has received authorization. The Pfizer jab is also available as a booster for children 12 and older.

However, the auto-generated summary for a similar cluster with keywords *Children, Risk, Adult, Covid, Immunity* is:

Children have a minuscule risk of COVID mortality. There is very limited safety data for vaccines from the trials on children. If the risk of adverse reactions is the same as for adults, the harms outweigh the risks.

where a random passage from the cluster states:

COVID poses no danger to children. They have a statistically zero chance of dying from that disease. The COVID shots, however, are already linked to innumerable adverse reactions, and their longer term side effects are unstudied.

This illustrates the need for further specificity using summarization to understand the narratives being spread. We provide several additional examples in Appendix F.

### 3.8. Validating Narrative Clusters

We evaluate our narrative clustering technique by validating whether a random sample of 500 passages (or maximum present) for a random set of 30 narrative clusters are about the same narrative using the methodology outlined in Section 3.5. Our methodology identifies coherent story/narrative clusters with an overall 98.9% precision and a minimum precision of 90.97% for Topic 22 (Table 3).

#### 3.9. Ethical Considerations

Our analysis is based on analyzing publicly posted news articles. We limit the load that each news site experiences by checking for new articles daily at a maximum rate of one request every 10 seconds. We further follow the guidelines as outlined by others for scraping data [5]. The hosts that we scan from are identifiable through WHOIS, reverse DNS, and an HTTP landing page explaining how to reach us if they would like to be removed from the study. We received no requests from websites to opt-out.

#### 3.10. Positionality Statement

The misinformation websites we study often covered contentious political issues including election denial, the Russo-Ukrainian War, and US abortion rights. As US-based English-speaking researchers, we inevitably bring some biases to discussing these issues. We attempt to remain as neutral as possible. We do not take any stance on political issues and when labeling specific stories as being misinformation, we rely fully on cited prior work from other researchers and/or news groups.

### 4. Narratives on Unreliable News Sites

In the last section, we presented and validated our methodology for programmatically extracting the narratives promoted by unreliable news websites. Here, we describe the most prolific narratives, trace three misinformation/propaganda stories, and derive communities of topically-related websites.

#### 4.1. The Largest Narratives

We start by analyzing the narratives most prolifically covered by our set of unreliable news sites in 2022. As can be seen in Table 4, the most popular narratives concerned the Russo-Ukrainian War, inflation, and Elon Musk’s criticism and later acquisition of the social media platform Twitter.
As can be seen in Figure 4, we further observe peaks in coverage of specific stories, as well as narratives that maintained consistent coverage throughout our study. For example, stories about abortion peak both before the US Supreme Court decision ( Dobbs v. Jackson) about federal abortion rights was leaked and following the official decision [69]. In contrast, a narrative about the EU’s role in NATO has seen steady stream of articles throughout the year, with a slight uptick following the Russian invasion of Ukraine. Analyzing the specific news sites that post about each narrative, we find that many Russian-backed and controlled websites [70] such as rt.com and southfront.org, in addition to several UK-based tabloids express.co.uk and dailymail.co.uk were the most prolific in writing about the Russian invasion of Ukraine (Narratives 1, 2, 5 in Table 4). This largely matches previous studies of the Russian-controlled media in influencing discussions on the war [38].

4.2. Misinformation/Propaganda Case Studies

As seen in the last section, many of the most common narratives are mainstream news topics. However, one of our goals is to track the spread of misinformation narratives. In this section, we show that our technique is capable of tracking known unreliable narratives by investigating the
evolution of one confirmed propaganda and two confirmed misinformation stories.

Ukrainian Nazis (Keywords: Azov, Battalion, Regiment, Far-right, Ukraine): One of the most prominent propaganda narratives utilized by Russian media in justifying the Russian Federation’s invasion of Ukraine was that the Ukrainian government was controlled by “neo-nazis” [65]. This is despite Ukraine’s relatively low level of antisemitism [71]. Our method is able to find that even before the Russian invasion of Ukraine on February 24, 2022, there were heavy references to Nazism in Ukraine by Russian-controlled or influenced outlets. For example, on January 27, glasbiresearch.ca penned:3


If we are to draw parallels between the current crisis on the Ukraine border and WW2 we should compare the Neo-Nazi ideology which dominates Ukrainian nationalism with that of Nazi Germany.

However, as seen in Figure 5, the major increase in the number of articles promoting this narrative occurred in the weeks prior to the Russo-Ukrainian War (specifically jumping in volume on February 8, 2022). The most prominent websites that pushed this narrative were unsurprisingly known Russian propaganda outlets including globalresearch.ca (68 articles), sputniknews.com (55), and rt.com (46). Beyond these known pro-Russian websites, we find US-based websites like veteranstoday.com (64 articles), sott.net (62), and thegatewaypundit.com (21) repeating this narrative.

Killer Covid-19 Vaccines (Keywords: Vaccine, Safe, Adverse, MRNA, Effect): One prominent misinformation narrative about COVID-19 that we identify is that COVID-19 vaccines are “killer vaccines” and a major cause of death around the world. For example on lewrockwell.com, an author wrote:4


Whatever they may be, these vaccines are most definitely not safe. We can very clearly see this from the explosion of reports of death to the Vaccine Adverse Event Reporting System (VAERS), which coincided with the introduction of the Covid injections in late 2020.

As seen in Figure 5, stories about “killer vaccines” have remained prominent throughout 2022, increasing in popularity several times throughout the year. The sites that most prominently echoed this narrative were theepochtimes.com (53 articles), pandemic.news (36), and vaccines.news (31). This is consistent with prior studies [72], [73].

2020 Election Denialism (Keywords: Fraud, Election, 2020, Irregular, Voter): The narrative that the presidential election was stolen and that current President Biden is illegitimate [74] spread throughout social media and was a key aspect of the January 6, 2021 attack on the US Capitol [5]. We see in our dataset that this false narrative maintained a substantial presence amongst unreliable news sites (Figure 5). For example, the fringe website thetrumpet.com wrote on January 6, 2022:


The insurrection hoax is a cover-up for the stolen election

The websites that most prominently repeated this narrative were velovertump.com (143 articles), thegatewaypundit.com (55), and votefraud.news (14).

4.3. Communities of Unreliable News Websites

To begin to capture the semantic communities that exist within the unreliable news ecosystem, we utilize each website’s distinct distribution of articles among our discovered set of 55,301 stories/narratives. To compare each website’s reporting choices and semantic content, we represent each website’s narratives as a multinomial distribution. For example, if we had three narratives (rather than 55.3K) and a website that wrote five articles about Narrative 1, four articles about Narrative 2, and one article about Narrative 3, the website’s distribution would be [0.5, 0.4, 0.1]. We do this for all 55,301 narratives and 1,404 websites, thus representing each website as a 55,301-dimensional vector of probabilities. We then use Jensen-Shannon Divergence [75] (detailed in Appendix E) to compare websites’ probability vectors. For example, the JS-Divergence of rt.com and sputniknews.com.com, two Russian state-sponsored websites [38], [70] is 0.413 while the JS-Divergence of rt.com and nypost.com, a US-based website, is 0.600.

After calculating each website’s narrative similarities using JS-Divergence with every other website in our dataset, we build an undirected graph with edge weights based on these values (i.e., an edge between a website P and Q is given a weight of 1 − JS(P||Q)) where JS(P||Q) is the JS-Digerece between websites P and Q. We determine communities of websites using the Louvain clustering algorithm [76]. Louvain clustering identified 3 communities, and from these communities, we qualitatively identified the corresponding three semantic communities: US-focused,
International, and Conspiratorial. We label these clusters based on the top topics found within each cluster, with the US-focused cluster writing about Abortion and the Biden Administration, the International cluster about the Russo-Ukrainian War, and the Conspiratorial cluster heavily writing about COVID-19 vaccines.

**US-focused Community.** 364 websites fall into our US-focused community including sites like dailymail.com, breitbart.com, and weloveotrump.com. The most common narrative in the community concerned the US Supreme Court Dobbs v. Jackson decision to overturn the 1973 Roe v. Wade decision that provided the federal right to abortion (Keywords: Roe, Abortion, Wade, Overturn, 1973). To further examine the role of this website community, particularly in regard to its most prominent narrative, we collect a larger set of narratives that more broadly relate to the topic of abortion by aggregating all 284 narrative clusters whose centers have a 0.50 similarity to the Abortion/Roe cluster.

We consider a website to *originate* a narrative if they published an article about the narrative on the first day that the narrative appeared in our dataset (more than one website can originate a narrative). Altogether, we find that 61.27% of Roe/Abortion narratives originated from this community, with the website rawstory.com originating the most Roe/Abortion narratives (36). In addition to originating most of the narratives about abortion, these websites contribute 76.93% of the articles on the 121 narratives about abortion; theepochtimes.com (13,30 articles across 146 Roe/Abortion narratives) and breitbart.com (1,264 articles across 136 Roe/Abortion narratives) have the most. Largely expected, many International websites such as dailymail.co.uk (1,104 articles across 116 narratives) and Conspiracy websites like evil.news (65 articles across 25 narratives) also picked up on these US-centered political narratives, evidencing the spread of stories from this community.

**International Community.** 408 websites fall into our International community including rt.com and dailymail.co.uk. The top story was one of our top overall narratives and Russian invasion of Ukraine (Keywords: Ukraine, Kyiv, Troop, Russian, Donbas). We gather a larger set of 504 narrative clusters that discuss the Russo-Ukrainian War using the same methodology outlined in the prior section.

We find that 38.2% of Russo-Ukrainian War narratives started from the International community of websites with 5.4% of these narratives specifically starting on nine pro-Russian propaganda websites [70]. Southfront.org.com (7 Ukraine narratives) and tass.com (10 Ukraine narratives) originate the most narratives among these Russian websites. We again find that this cluster of websites is responsible for a large portion (56.12%) of articles about the war. Again, largely expected, other websites such as nypost.com (2,934 articles across 160 Ukraine narratives) or treason.news (32 articles across 23 Ukraine narratives) write extensively about the conflict as well.

**Conspiratorial Community.** 632 websites belong to our Conspiratorial community, including popular sites known for spreading conspiracy theories about QAnon and COVID-19 [5], [48] like unz.com, qresearch.ch, and radiopatriot.net. Unsurprisingly, the top narrative within this community concerns COVID-19 (Keywords: Children, Risk, Adult, Covid, Immunity). We gather a more extensive set of narrative clusters that discuss the COVID-19 and/or COVID-19 vaccines using the same methodology outlined before; altogether gathering 162 narratives. Most prominently, the website childrenshealthdefense.org, the nonprofit run by Robert F. Kennedy Jr., wrote about nearly every COVID-19 story in our dataset (1250 articles across 94 COVID narratives).

We find that our set of Conspiratorial websites originate 37.3% of narratives about COVID-19, the most prominent of these being nvc.org (18 COVID narratives) and healthimpactnews.com (7 COVID narratives). We note that COVID-19 narratives originated not only from these websites but from our International (23.8%) and US-focused cluster (38.9%) as well, with the Conspiratorial cluster being responsible for only 20.3% of COVID-19 articles. This evidences that COVID-19 narratives spread and originated among a wide variety of different websites in 2022.

**5. Originating and Amplifying Narratives**

As seen throughout the last section, several websites play dominant roles in perpetuating and promoting certain types of stories. In this section, we quantify and identify which websites have the most pivotal roles in originating and amplifying narratives throughout the ecosystem of unreliable websites. As before, we consider a website to *originate* a narrative if they published an article about the narrative on the first day that the narrative appeared in our dataset (more than one website can originate a narrative). We consider a website to have *amplified* (i.e., increased the popularity of) a narrative if it (1) posted an article about the narrative before the narrative peaked in popularity, (2) did not originate the narrative, and (3) if the posted article appeared in the first 15% of the total volume of that given narrative. We utilize the 15% cutoff as it ensures that the vast majority of a narrative’s articles have not been published yet (i.e., the story has not dramatically increased in popularity already), allowing us to observe how amplification affects the narrative’s popularity. This threshold is consistent with prior work [17].

Using this approach, we further investigate how the popularity of a website influences its effectiveness in originating and amplifying narratives. To do so, we utilize the website rank data provided by the Google Chrome User Report (CrUX) from October 2022, which Ruth et al. showed to be the most reliable website popularity metric [77].

**5.1. Originating Narratives**

To measure the efficacy of websites in originating narratives, we perform a correlational comparison of the number of external non-origin articles that are written about a given narrative in the week after a given website originated a narrative vs. the number of non-origin external articles that are written in the week after origination if the website did not originate the narrative but still eventually wrote about
that narrative. We note that for this analysis, we weight the number of articles by the log inverse of its CrUX popularity ranking [78], [79] to ensure that we do not consider an article from a highly popular website such as breitbart.com the same as from a relatively obscure website such as welovetrump.com.

To ensure each website has a marked effect on the full unreliable news ecosystem and improve the robustness of our approach, we utilize a bootstrapping procedure $(B = 250)$ to measure the influence of each website by taking a random subset of 100 websites in each bootstrap and then measuring the increased weight in the number of articles across this set of a random set of 100 websites [80]. We provide the average effect size (Cohen’s D) statistical measure of the increase in articles and the p-value to check for the significance (using Mann-Whitey U-tests) for the change in the number of external articles in Table 5. For this section, we limit our analysis to websites that consistently change in the number of external articles in Table 5. For this measure of the increase in articles and the p-value to check the effect of the increase in articles (Cohen’s D) we limit our analysis to websites that consistently change in the number of external articles in Table 5.

We observe a small correlation (Pearson correlation $\rho = 0.213$) between a website’s popularity and its ability to originate and perpetuate narratives amongst other unreliable news websites. For example, considering express.co.uk and dailymail.co.uk, two tabloids known to engage in sensationalism and biased reporting with the highest CrUX popularities, while dailymail.co.uk is fairly effective at originating narratives (Cohen’s D = 0.334), express.co.uk is one of the worst at originating new narratives (Cohen’s D = 0.007). Further, a seemingly unpopular website, infostormer.com, is one of the best at propagating narratives it originates to other sites. This illustrates many different types of websites can originate and propagate narratives in the misinformation news ecosystem. Infostormer.com, with a header labeled the “Jewish Problem”, writes heavily sensationalist and antisemitic perspectives on the news that is taken up by other websites. For example, after writing an article on how the CNBC news host Jim Cramer was promoting Meta stock, this news story was later covered by more popular websites like activistpost.com7 and hannity.com.8

Here, a negative Cohen’s D indicates that the “time for a narrative to peak” occurs faster. This metric, combined with the previous metric, describes how effective a website is at reorienting online conversations to its own narratives. We see only a slight correlation ($\rho = 0.125$) with a website’s CrUX-defined popularity. Rather, we again see that many smaller websites are highly effective in originating narratives that peak quickly (i.e., writing about narratives that become of immediate interest). For example, we see that when the small website infostormer.com originates narratives, those narratives peak in popularity 155 days earlier than when infostormer.com does not originate narratives (Table 5). We similarly observe that the right-wing and conspiratorial website ussanews.com, which had often written about QAnon [5], is also highly effective at quickly landing its narratives on other websites, with one of the lowest “time to peak” in our dataset.

### 5.2. Amplifying Narratives

To understand how effective websites are at amplifying narratives, we correlationally compare the number of external non-origin articles that are written about a given narrative when a given website amplifies the narrative versus when the website does not amplify the narrative but still eventually wrote about that narrative. We utilize the same weighting and bootstrapping procedure as in the previous section, again limiting our analysis to websites that amplify at least 25 narratives across our period of study. We again observe only a slight correlation between a website’s popularity and its ability to amplify narratives (Pearson correlation $\rho = 0.279$). As before, we see in Table 6 that websites across different CrUX popularities excel at amplifying narratives. One of the most effective websites is barenakedislam.com, an anti-Islam website with the slogan “It isn’t Islamaphobia when they really ARE trying to kill you.” For example, after echoing a narrative about how Muslim men were targeting Ukrainian refugees,9 this news story traveled to an additional 12 other unreliable news websites including more popular websites like americanthinker.com10 and breitbart.com.11 This illustrates many different types of websites can amplify and and help propagate narratives in the misinformation news ecosystem.

Finally, we determine the effect of narrative amplification by each website on how quickly the narrative peaks. There is again a small correlation between website popularity and amplification (Pearson correlation $\rho = 0.169$). As seen in Table 6, the most effective website at quickly amplifying narratives to their peak popularity is populistpress.com, a drudgy-style news website that hosts hyperlinks to different news articles. Despite not hosting many articles itself, we see that when it does mention a narrative, this news story is more likely to peak in popularity sooner.

#### Trend Setting

Despite not seeing clear discernible patterns in how website popularity correlates with a website’s ability to originate and amplify narratives, we do observe differences in when these websites write about given narratives. As seen in Figure 6, across all narratives, more popular websites tend to write fewer articles on the day that a given narrative peaks. A slightly higher percentage (33.0%) of articles from websites with a CrUX rank $<100K$ come before the peak versus the 28.76% of articles with a rank above 1M. This indicates, as also found by Leskovec et al. [17], that across all narratives, popular websites have some ability to set the agenda for what smaller websites write.

### 5.3. The Role of Fringe Forums/Social Media

We now analyze the relationship between our set of unreliable news websites and the fringe social media sites 8kun and 4chan. As with our set of unreliable news websites, we scrape 8kun and 4chan/pol posts published between January 1 and November 1, 2022 (Section 3.4). 8kun data is readily available on their website 8kun.top; 4chan/pol posts are archived through the website archive.4plebs.org. Altogether, we gather 632,091 posts from 8kun.top and 4.69 million posts from 4chan (Table 7).

To find the correspondence of 8kun and 4chan comments between news narratives, we preprocess, embed, and assign each 8kun post to its most similar narrative cluster. As before, we utilize a threshold of 0.60 for matching a comment to its corresponding news article narrative. Altogether, we find that 35,227 (5.57%) of comments on 8kun.top and 455,327 (9.71%) of comments from 4chan.org correspond to a narrative on our set of unreliable news websites (Table 7).

#### Table 7: Our dataset of social media posts and their relationship to the narratives published by unreliable news websites.

<table>
<thead>
<tr>
<th>Social Media Platform</th>
<th>Posts</th>
<th>Posts With Corresponding News Article Narrative</th>
</tr>
</thead>
<tbody>
<tr>
<td>8kun.top</td>
<td>632,091</td>
<td>35,227 (5.57%)</td>
</tr>
<tr>
<td>4chan.org</td>
<td>4,690,669</td>
<td>455,327 (9.71%)</td>
</tr>
</tbody>
</table>


TABLE 8: The top topics from our set of unreliable news websites present on 8kun and 4chan /pol.

<table>
<thead>
<tr>
<th>Platform</th>
<th>Narratives Originated</th>
<th>Wtd. Ext. Art.Δ</th>
<th>Cohen’s D</th>
<th>To Peak Cohen’s D (Days)</th>
<th>To Peak Cohen’s D (Days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8kun</td>
<td>392</td>
<td>0.416</td>
<td>1.781</td>
<td>-72.94</td>
<td>-1.149</td>
</tr>
<tr>
<td>4chan</td>
<td>2,432</td>
<td>0.438</td>
<td>0.961</td>
<td>-28.60</td>
<td>-0.445</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Platform</th>
<th>Narratives Amplified</th>
<th>Wtd. Ext. Art.Δ</th>
<th>Cohen’s D</th>
<th>To Peak Cohen’s D (Days)</th>
<th>To Peak Cohen’s D (Days)</th>
</tr>
</thead>
<tbody>
<tr>
<td>8kun</td>
<td>2,695</td>
<td>0.903</td>
<td>0.267</td>
<td>-20.28</td>
<td>-0.372</td>
</tr>
<tr>
<td>4chan</td>
<td>12,184</td>
<td>0.370</td>
<td>0.938</td>
<td>0.646</td>
<td>0.006</td>
</tr>
</tbody>
</table>

TABLE 9: The influence of 8kun and 4chan on the ecosystem of unreliable news websites and their similarity (by JS-Divergence) to the unreliable news dataset. We star values that are not significant, according to the Mann-Whitney U-test.

5.3.1. 8kun. Examining the top narratives posted on 8kun that corresponded with a news narrative (Table 8), we see that the most commonly shared narratives on 8kun concern the Russo-Ukrainian war, the investigation of Donald Trump by Special Counsel Robert Mueller [81], antisemitic beliefs, and the 2022 Truck Convoy in Ottawa Canada [82]. This largely corresponds with 8kun being known as the home of hard-right, conspiratorial, and antisemitic posts [83]. As in Section 4.3, we determine the distribution of narratives from our set of unreliable websites that are present on 8kun to understand the similarity between 8kun and the collective narratives on our set of unreliable news websites. Altogether, 8kun has a JS-Divergence of 0.242 with the collective narrative distribution of unreliable news websites (Table 9). Performing this on an individual site level, we observe several of the websites with the most similar narrative distributions prominently discuss conspiratorial ideas (e.g., lucianne.com and radiopatriot.net) [5]. Having examined similarities between the narratives discussed on 8kun and those on particular websites in our dataset, we next determine the influence of 8kun on our unreliable news website ecosystem. We utilize the same definitions of originate and amplify as well as the same methodology as in Sections 5.1 and 5.2. As seen in Table 9, 8kun originating or amplifying a particular narrative has a modest effect on the number of articles written about that narrative. In the week after 8kun originates a given narrative, we see an average Cohen’s D of 1.781. In contrast, while not as effective as some of the websites in our dataset (Tables 5 and 6), when 8kun users comment on narratives, this correlates with a slight increase in the narrative’s popularity. We see this further mirrored in the effect that 8kun has in expediting narratives to peak earlier. On average, if 8kun originates a narrative, it peaks in popularity 72.4 days earlier than if 8kun did not originate the narrative. Similarly, if 8kun amplifies a narrative it peaks in popularity 20.8 days earlier on average.

5.3.2. 4chan /pol. Looking at the top corresponding shared narratives on 4chan /pol, we see several that target Judaism and the Jewish people (Table 8). As with 8kun, 4chan has a reputation for antisemitism and racist language. Besides the narratives that center on the Russo-Ukrainian war, we see this racism and antisemitism reflected in the top shared narratives on the website [84]. Determining the distribution of narratives from our set of unreliable news websites that are present on 4chan, altogether, 4chan has a JS-Divergence of 0.250 with the collective narrative distribution of unreliable news websites (Table 9). Examining the most similar websites to 4chan, we observe several with known conspiratorial reputations. As documented by Medias-Bias/FactCheck, the most similar website to 4chan, unz.com is a conspiratorial and hate-oriented website that often cites white nationalist groups in its articles [85]. Similarly, beforeitsnews.com [86] and thetruthseeker.co.uk [87] are known to “promote conspiracy theories and pseudoscience.”

Finally, we determine the role 4chan has in promoting and amplifying narratives within our ecosystem of unreliable news websites. We observe a similar effect to 8kun, in terms of the weighted increase of articles when 4chan originates a narrative compared to when it does not (Cohen’s D of 0.438). However, unlike 8kun, we do observe that 4chan is better at amplifying narratives, with an effect size of Cohen’s D of 0.938 (Table 9). However, in contrast to 8kun, 4chan is relatively less effective at getting the narrative to peak earlier. If 4chan users originate a narrative, it peaks in popularity 28.6 days earlier compared to 72.9 days earlier when 8kun originates a narrative. When 4chan amplifies a narrative, it has little effect on when that narrative peaks.

6. Detecting Narratives and Fact-Checking

In the last two sections, we analyzed the narratives and behavior of unreliable news websites during 2022. In this section, we present two case studies that highlight how our programmatic approach can also identify new narratives and assist in focusing fact-checking efforts.

6.1. Identifying New Trending Narratives

By examining the week-over-week percentage increases in story volumes, we programmatically determine which narratives are receiving new or renewed focus on unreliable
news websites, which is imperative for ameliorating the spread of specious information [9], [10], [11]. The narratives that increased most in volume during the last week of our experiment (October 26 to November 1, 2022) were:

The Attack of Paul Pelosi, Keywords: Pelosi, Depap, hammer, Nancy, Paul. On October 28, 2022, the husband of Congresswoman Nancy Pelosi was attacked in his home [88]. Largely due to the proximity of the time of the attack to the 2022 US midterm elections, the attack became a source of conspiracy theories and wild speculation. For example, one user wrote on thegatewaypundit.com: 12

Whenever bad things happen to Paudie P, his wife always manages to have an alibi.

364 articles (compared to zero the week before) were written about the event within our dataset across 175 websites. The thegatewaypundit.com had 44 articles, dailymail.co.uk.com had 40, and nypost.com had 37.

The Seoul Halloween Stampede, Keywords: Halloween, Seoul, Itaewon, Festivity, Stampede. On October 29, 2022, a crowd rush in the Seoul neighborhood of Itaewon resulted in the death of 158 people. Across our dataset, we see 132 articles across 46 websites written about this event, with 18 articles from dailymail.co.uk, 11 from republicworld.com, and 8 from mirror.co.uk.

Elon Musk’s First Visit to Twitter Headquarters, Keywords: Sink, Headquarters, Musk, Carry, Twitter. After officially purchasing the social media company Twitter, on his first visit to the company on October 26, Elon Musk carried a sink into the headquarters with him. This prop humor by Musk was supposed to be a play on “let that sink in” but with a real sink. We see 176 articles from 84 domains about the story, with 10 articles from the dailymail.co.uk, 9 from westernjournal.com, and 8 from conservativeangle.com.

6.2. Fact-Checking

One approach to combating the spread of new misinformation stories that many organizations have adopted is fact-checking. Fact-checking a story requires hours to deeply understand its context and nuance [89]. Unfortunately, this means that propaganda and misinformation often spread significantly in a rapidly evolving media landscape before journalists can respond. Our approach can serve as a way to programmatically identify new misinformation narratives as they appear and begin to gain traction, ideally curbing the behaviors of particular narratives before being fact-checked by fact-checking organizations, we utilize our approach to analyze the behaviors of particular narratives before being fact-checked by three organizations: Politifact, Reuters, and AP-News [90], [91], [92]. For the three agencies, we gathered the set of fact-checking articles that each published in 2022.

Table 10: Efficacy of fact-checking websites. All three websites most commonly fact-check—by the number of articles with the same narrative as the fact-check—the articles of resear.ch (433 articles), gateofvienna.net (414), dailymail.co.uk (406).

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>Politifact</td>
<td>6,231</td>
<td>6</td>
<td>55.0</td>
<td>4.0</td>
<td>110</td>
</tr>
<tr>
<td>Reuters</td>
<td>9,604</td>
<td>3</td>
<td>49.0</td>
<td>0.0</td>
<td>647</td>
</tr>
<tr>
<td>AP News</td>
<td>230</td>
<td>15</td>
<td>83.0</td>
<td>3.0</td>
<td>8</td>
</tr>
</tbody>
</table>

Altogether we scraped 1,524, 3,090, and 140 articles from Politifact [93], Reuters [94], and APNews [95], respectively. To augment our system to perform fact-checking (i.e., determine whether a fact-check article refutes a given narrative), we additionally train a DeBERTa-based [96] classifier on the FEVER [97] dataset that takes a claim (i.e., an article) and a query (i.e., a fact-check) and labels the query as either supporting the claim, refuting the claim, or not having enough information to say anything about the claim. Using 10% of the FEVER dataset as a held-out test set, our DeBERTa-based model achieves an overall 90.7% accuracy on this test set (90.5% precision in labeling refutations).

For each fact-checking article, as with articles from unreliable websites, we divide the article into its constituent passages and embed them utilizing our MPNet model. We consider a narrative to have been addressed by a fact-checker if the fact-checker writes about the narrative. We note that articles frequently “fact-check” or “add context” to multiple narratives. To provide fact checkers with the greatest number of “opportunities” to fact-check a narrative, we map each fact-check passage to all articles above our cosine similarity threshold of 0.60 rather than map the fact-check passage to only the single closest narrative. After mapping these fact-checking passages to our set of articles, we utilize our DeBERTa-based fact-checking classifier to ensure that the corresponding “fact-check” refutes the information of the corresponding unreliable news article passage. We provide an example of an identified fact-check below. To ensure that our model is able to properly identify “fact-checks”, we manually validate 100 random fact-check-article refutation pairs, finding that 94% of them are indeed refutations.

| Article Passage: This is a shining example and a small part of why it is so vitally important to find the underlying cause of the fraud that took place both in November 2020 and the lead-up to that election. |
| Fact-Check: THE FACTS: To be clear, no widespread corruption was found and no election was stolen from Trump. |

As seen in Table 10, on average, narratives can spread one to two months before being fact-checked by these reputable websites. Furthermore, on average both Politifact and AP News write about stories after they have peaked in popularity; AP News, with the fewest fact-checks, writes about narratives right as they peak in popularity. We further see a heavy overlap between the narratives that each website fact-checks. Reuters and Politifact have an overlap of 2,646 stories/narratives; AP News and Politifact, have an overlap of 137 narratives; and Reuters and AP News have an overlap of 141 narratives. Furthermore, the unreliable websites that have articles most commonly fact-checked
by the three fact-checking organizations are the same: g-check (433 articles), gatesofvienna.net (414), and daily-mail.co.uk (406). This underscores that these fact-checking websites are duplicating effort, often fact-checking the same narratives [98]. We note, however, that while narratives often spread for long periods before being fact-checked, the number of articles, on average, is often low (3–15 articles). We thus see that many fact-checkers are effective at fact-checking narratives when they peak in popularity, but often understandably do not fact-check narratives that have just begun to spread among different unreliable news websites.

We see that many narratives spread for long periods on unreliable news websites before they are fact-checked near their narrative peak. However, our system can surface these narratives to fact-checkers long before they peak in popularity, aiding in the fact-checking organizations’ typical workflow in identifying potential misinformation. This can enable fact-checkers to identify and address misleading narratives concurrent to when they first rise in popularity.

7. Related Work

Our study builds on considerable prior work on both the spread of misinformation online and language models. There have been several past quantitative studies of the spread of information online. Leskovec et al. identify the trends in the propagation of “memes” [17]. They find that while the majority of memes originate from mainstream websites, key phrases that start on smaller blogs are often adopted by larger platforms. Gomez-Rodriguez et al. adopt a cascade transmission model and identify how best to estimate the relative influence of different news outlets in spreading stories [99]. Similar to our use of DP-Means, prior works have utilized CluStream among other clustering techniques to track information or news over time [100], [101], [102]. For example, Czuriski et al. [103] utilize document clustering based on dictionaries to track topics.

Analyzing the Spread of Misinformation. Several works have tracked the spread and impact of misinformation. Shu et al. [104] present the largest overall overview of misinformation detection issues, presenting various paradigms for tracking and labeling misinformation. These include tracking news content features and social content features. For example, Cao et al. [105] and Meel et al. [106] explore utilizing image and text-based features to label misinformation. Abdali et al., in contrast, use screen shots of websites to identify the trustworthiness of websites and label misinformation [107]. Extensive work has studied individual campaigns that spread unreliable information, on topics like QAnon [5], [48], Syrian White Helmets [108], the Russo-Ukrainian War [38], [65], and COVID-19 [109].

Recent work from the security community has focused on identifying and curbing misinformation. Kaiser et al. [13] studied how borrowing techniques from the security warning landscape might help to inform mis/disinformation warnings. Paudel et al. [14] recently demonstrated how techniques like Learning To Rank (LTR) can be used to soft-moderate misinformation on Twitter. On the human-level, Sharevski et al. identified folk models of misinformation on social media that could inform potential defenses [110].

Language Models, Semantic Search, and Topic Analysis. Many previous topic analysis methods have been built on Latent Dirichlet Allocation (LDA). Albalawi et al. show that LDA is one of the most effective methodologies for extracting topics from short text data compared to other computationally light alternatives proposed within the last decade (e.g., LSA, LDA, NMF, PCA, RP) [111]. Meng et al. [112], Angelov [113], and Grootendorst [63] have enabled users to perform topic modeling utilizing large language models. Utilizing these techniques and online document clustering [114], [115], others have performed robust, but smaller scoped semantic analysis (e.g., on Russian dis-information campaigns [38], [52]).

8. Discussion and Conclusion

In this work, we introduced and validated a new, scalable methodology for tracking news narratives online. Applying the methodology to study the stories published on 1,404 unreliable news websites during 2022, our work shows how a large-scale, quantitative analysis can identify propagation patterns and significant players that may otherwise have been difficult to uncover through qualitative investigations of individual disinformation campaigns. Specifically, we showed that less frequented websites and fringe social media platforms can have marked effects on amplifying the narratives discussed on unreliable news websites.

Our study also highlights the need to programmatically detect the rise of false narratives in real time. Prior work has shown that misinformation can spread ten times faster than legitimate news [6] and our analysis finds that false narratives can often start on small, seemingly unpopular websites. In many cases, these false narratives spread for months online before being fact-checked. As such, we are exploring how best to publicly and continuously release real-time updates of our narrative analyses on an online dashboard while protecting against misuse (e.g., use in AI training models [116] or targeted misinformation chatbots).

While our study illustrates the potential for programmatically tracking news narratives, it also simultaneously surfaces areas for further research. For example, as found in past works [38], [117], though our approach can identify precise stories/narratives within our dataset, medical information poses challenges for large language models like MP-Net. For example, one of the misclassifications (Narrative 6 in Table 3) concerned COVID-19 rather than Monkeypox. Given the high knowledge level needed in understanding medical misinformation, past works have recommended utilizing models specifically trained for medical misinformation for topic analysis of these stories [117]. We hope that the potential of and demonstrated need for programmatic approaches for tracking news narratives and misinformation online motivates further work on these topics.
References


Appendix A.
Training with Unsupervised Contrastive Loss

To train our MPNet model, we utilize unsupervised contrastive learning to better the quality of our embeddings [59]. For training, this is such that we embed each example \( x_i = \langle text_i, t_i \rangle \in D_{News} \) (where \( text_i \) is the text and \( t_i \) is whether the text is toxic or not) using a contextual word model by inputting \([CLS]text_i[SEP]\) and averaging the contextual word vectors of the resulting output as a hidden vector \( h_i \) for \( text_i \) as its representation. Then, given a set of hidden vectors \( \{h_i\}_{i=0}^N \), where \( N_b \) is the size of the batch, we perform a contrastive learning step on that batch. This is such that for each Batch \( B \), for an anchor hidden embedding \( h_i \) within the batch, the set of hidden vectors \( h_j, h_j \in B \), vectors where \( i = j \) are positive pairs. Other pairs where \( i \neq j \) are considered negative pairs. Within each batch \( B \), the contrastive loss is computed across all positive pairs in the batch such that:

\[
L_{\text{contrastive}} = \frac{1}{N_b} \sum_{h_i \in B} l^c(h_i)
\]

\[
l^c(h_i) = \log \frac{\sum_{j \in B} I_{|i|=j} e^{h_i^T h_j / \tau}}{\sum_{j \in B} e^{h_i^T h_j / \tau}}
\]

where, as in prior work [118], we utilize a temperature \( \tau = 0.07 \).

Appendix B.
Passage Pairs

**Appendix C.
DP-Means Algorithm**

DP-Means [119] is a non-parametric extension of the K-means algorithm that does not require the specification of the number of clusters \( a \ priori \). Within DP-Means, when a given datapoint is a chosen parameter \( \lambda \) away from the closest cluster, a new cluster is formed. Dinari et al. [64] parallelize this algorithm by delaying cluster creation until the end of the assignment step. Namely, instead of creating a new cluster each time a new datapoint is discovered, the algorithm instead determines which datapoint is furthest from the current set of clusters and then creates a new cluster with that datapoint. By delaying cluster creation, the DP-means algorithm can be trivially parallelized. Furthermore, by delaying cluster creation, this version of DP-Means avoids over-clustering the data (i.e., only the most disparate datapoints create new clusters) [64].

**Appendix D.
Pointwise Mutual Information**

The PMI of a particular word \( word_i \) in a cluster \( C_j \) is calculated as:

\[
PMI(word_i, C_j) = \log_2 \frac{P(word_i, C_j)}{P(word_i)P(c_j)}
\]

where \( P \) is the probability of occurrence and a scaling parameter \( \alpha \) is added to the counts of each word. This scaling parameter \( \alpha \) prevents single-count or one-off words in each cluster from having the highest PMI values. Given the scale of our dataset and the number of clusters within our dataset, we determine that a baseline count of 1 (\( \alpha = 1 \)) for each word in the full dictionary in each cluster led to the best results [120]. We utilize this approach rather than cluster-normalized TF-IDF as in other works [38], [63] because class TF-IDF is dependent on document classes being similar in length [63], [121] and the number of articles within each of our clusters varies widely. PMI finds the distinct characteristics of individual clusters and is not dependent on how often words appear in other individual clusters, avoiding this issue.

**Appendix E.
JS-Divergence**

Formally JS-Divergence between two distributions \( P \) and \( Q \) is calculated as follows:

\[
JS(P||Q) = \frac{1}{2} KL(P||\frac{P + Q}{2}) + \frac{1}{2} KL(Q||\frac{P + Q}{2})
\]

\[
KL(P||Q) = \sum_x P(x) \log \left( \frac{P(x)}{Q(x)} \right)
\]

For our purposes, given that every website does not address every topic, as recommended in other works, we add a small value \( \epsilon = 0.1 \) to the counts of every website’s topics before calculating each website’s probability distribution.

**Figure 7:** Example of passage pairs at different levels of cosine similarities.
### Appendix F.  
**Auto-Generated Summaries and Cluster Specificity**

<table>
<thead>
<tr>
<th>Narr.</th>
<th>Keywords</th>
<th>Auto-Generated Summary</th>
<th>Random Sample Passage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>trudeau, motion, 151, emergency</td>
<td>The Canadian Parliament voted Monday night to approve Prime Minister Justin Trudeau's motion to invoke the Emergencies Act by a vote of 185 for and 151 against.</td>
<td>On Monday night, Canada’s parliament voted to confirm Prime Minister Trudeau’s declaration of the Emergencies Act in response to the freedom protests that have swept across the nation for three weeks.</td>
</tr>
<tr>
<td>2</td>
<td>manchin, filibuster, schumer, sinema, senate</td>
<td>Republicans and other critics immediately started to wonder: If Democrats extract what they want out of Manchin, couldn’t even a small number of them promptly refuse to go along with the secondary assurances he’s been promised?</td>
<td>[The pipeline Manchin was promised] would require passage of legislation that would overhaul the permitting process for energy infrastructure, according to The American Prospect, a liberal website. Apparently, progressives in the House are not keen on supporting a measure that could undercut the IRA’s down payment on clean energy by accelerating approval for energy projects that could ramp up U.S. fossil fuel production and exports of natural gas. The American Prospect reported.</td>
</tr>
<tr>
<td>3</td>
<td>agrawal, musk, parag, ceo, twitter</td>
<td>Twitter CEO Parag Agrawal tweeted he was “excited” that Musk would join Twitter’s board after it was revealed that Musk bought a 9.2 percent stake in the company, and in doing so became its largest shareholder.</td>
<td>When Musk’s takeover of Twitter became official, Agrawal and Bret gave comments alongside the Tesla CEO.</td>
</tr>
<tr>
<td>4</td>
<td>capitol, committe, hearing, select, january</td>
<td>The House of Representatives committee investigating the Jan. 6, 2021, attack on the U.S. Capitol is planning to hold its next hearing on Sept. 28.</td>
<td>The U.S. House of Representatives select committee investigating the deadly Jan. 6, 2021, attack on the Capitol will conduct its next hearing on Oct. 13, the panel said in a statement on Thursday.</td>
</tr>
<tr>
<td>5</td>
<td>smith, slap, black, oscar, rock</td>
<td>Apparently, whites can’t be outraged by Will Smith’s slap without being racist. And never mind that plenty of blacks including Kareem Abdul-Jabbar were also outraged. He makes blacks look bad his slap reinforces the widely held stereotype that blacks are violent. He shamed AMPAS before the world.</td>
<td>[The pipeline Manchin was promised] would require passage of legislation that would overhaul the permitting process for energy infrastructure, according to The American Prospect, a liberal website. Apparently, progressives in the House are not keen on supporting a measure that could undercut the IRA’s down payment on clean energy by accelerating approval for energy projects that could ramp up U.S. fossil fuel production and exports of natural gas. The American Prospect reported.</td>
</tr>
<tr>
<td>6</td>
<td>extremist, maryland, walkby, virginia, protest</td>
<td>The group that calls itself Ruth Sent Us announced its plans on a website to harass the justices. It said: Announcing Walkby Wednesday, May 11, 2022! At the homes of the six extremist justices, three in Virginia and three in Maryland.</td>
<td>RSU subsequently announced a WalkBy Wednesday protest on May 11, to be held in front of the homes of the six extremist justices.</td>
</tr>
<tr>
<td>7</td>
<td>cornyn, boo, convert, texas, gop</td>
<td>U.S. Senator John Cornyn (R-Texas) was loudly booed at the Republican Party of Texas Convention in Houston, where the state GOP adopted a resolution condemning the bipartisan gun control framework he has negotiated in the Senate.</td>
<td>Very loud boos for John Cornyn as he takes the stage at the TexasGOP convention. Cornyn has faced opposition within the party for working with Democrats on a gun package after the shooting in Uvalde.</td>
</tr>
<tr>
<td>8</td>
<td>threat, truss, china, uk, britain</td>
<td>Prime Minister Liz Truss is for the first time due to officially declare China a threat to the UK within days. The designation would be a formal update to former PM Boris Johnson’s Integrated Review of Defense and Foreign Policy published in March 2021.</td>
<td>On October 11, The Guardian reported that the Liz Truss government is going to formally designate China a national “threat” to Britain in its upcoming strategic defense review. Under former Prime Minister Boris Johnson, China was named just a “systemic competitor.”</td>
</tr>
<tr>
<td>9</td>
<td>lithuania, vilniu, baltic, export</td>
<td>China has called for a corporate boycott of the small Baltic nation. The move is in retaliation for Lithuania’s decision to open a Taiwanese representative office in its capital of Vilnius in November 2021.</td>
<td>In a letter last month, the German-Baltic Chamber of Commerce demanded that Lithuania come to a constructive solution with the communist nation, saying per Reuters: The basic business model of the companies is in question and some will have no other choice than to shut down production in Lithuania.</td>
</tr>
<tr>
<td>10</td>
<td>curfew, quebec, province, legal, montreal</td>
<td>Quebec first imposed a COVID curfew on January 9, 2021, which was lifted on May 28. Quebec is the only province in Canada to have imposed a curfew during the pandemic.</td>
<td>Quebec is the first province to impose such a system on its citizens. It is also the only province in Canada that has a curfew in place.</td>
</tr>
</tbody>
</table>

Example of the auto-generated summaries and passages from a set of 10 random narrative clusters to illustrate the specificity and the precision of our approach.