Reap the Wild Wind: Detecting Media Storms in Large-Scale News Corpora

Anonymous ACL submission

Abstract

Media storms, dramatic outbursts of attention to a story, are central components of media dynamics and the attention landscape. Despite their importance, there has been little systematic and empirical research on this concept due to issues of measurement and operationalization. We introduce an iterative human-in-the-loop method to identify media storms in a large-scale corpus of news articles. The text is first transformed into signals of dispersion based on several textual characteristics. In each iteration, we apply unsupervised anomaly detection to these signals; each anomaly is then validated by an expert to confirm the presence of a storm, and those results are then used to tune the anomaly detection in the next iteration.

We make available the resulting media storm dataset. Both the method and dataset provide a basis for comprehensive empirical study of media storms.

1 Introduction

Media storms - dramatic increases in media attention to a specific issue or story for a short period of time (Boydstun et al., 2014) - are central components of media dynamics. Such outbursts include, for example, news reports on acts of terrorism, public scandals, or major political decisions. They usually begin with a specific trigger event (e.g., Wien and Elmelund-Præstekær, 2009), and then surge to disproportionate levels of coverage - *hype* (e.g., van Atteveldt et al., 2018). Storms intensify nearly all media-related effects (e.g., Boydstun et al., 2014; Walgrave et al., 2017). In addition, being pivotal moments in the public agenda, storms can be critical junctures for political actors (Gruszczynski, 2020; Wolfsfeld and Sheafer, 2006).

However, we still lack a systematic and comprehensive understanding of such outbursts of media attention. One reason is that it is not clear how to operationalize this concept into a concrete measurable object (Boydstun et al., 2014, 518-519). Essentially, previous researchers are left devising "arbitrary" thresholds for their studies (Boydstun et al., 2014, 519). In addition to this amorphousness, an additional challenge is that media storms are relatively sporadic phenomena. Boydstun et al. (2014) approximate that they consist about 11% of all media coverage, a finding that was later corroborated by Nicholls and Bright (2019). These properties make it extremely difficult to create a gold-labeled data-set to train a model, or to even begin reading the raw articles to identify media storms directly, necessitating the development of a different strategy to solve this challenging task.

Traditionally, communication researchers employed manual content analysis to label and measure issue attention over short periods (e.g., Boydstun et al., 2014; Wolfsfeld and Sheafer, 2006). Recent computational work has utilized topic modeling (van Atteveldt et al., 2018; Nakshatri et al., 2023) and keyword analysis (Lukito et al., 2019) for the task. However, the drawback of these approaches is their sensitivity to research design - the keyword choice or delineation of topics. A researcher might choose a model with broad topics - hampering the ability to recognize deviances of specific outburst. Conversely, an overly complex model might cause a media storm to be dispersed across several topics, diluting attention peaks. This could make significant media events less discernible. Meanwhile, focusing on keywords may obfuscate the actual story behind the tokens.

Another approach adopted in recent computational communication research has been to focus on *news story chains*. Such methods utilize clustering to identify news events - articles describing the same event or story (e.g., Nicholls and Bright, 2019; Trilling and van Hoof, 2020). These techniques 'uncover' the stories occurring in the corpus - groups of documents discussing the same, specific event. Recently, Litterer et al. (2023) fine-tuned a model to generate document embeddings for the clustering.

While these methods identify media *stories*, they do not encompass the theoretical concept of media *storms*. Rather than capturing prominent stories, we are aiming for periods where the media coverage is not structured normally (Boydstun et al., 2014), but rather characterized by 'hype'—dramatic and anomalous levels of coverage of a story (van Atteveldt et al., 2018; Vasterman, 2005). However, it is impossible to determine hype when only taking into account the structure of a single story at a single time-step without noting long-term trends and cycles as baselines. Tellingly, such methods tend to identify many more instances than we do in our experiments here. For example, Litterer et al. (2023) identify 98 cases over nearly two years, while we detect 221 over a 20 year period.

With these points in mind, we sought a different approach we believe better reflects the theoretical conception of storms. We return to the basic definitive property of media storms: a dramatic, temporary spike in attention to an issue (above the norm). In other words, storms are anomalies in news coverage, so we turn to anomaly detection to identify them. We create several signals representing the daily dispersion of texts across the time frame. These signals are the basis for a two-step procedure. First, an unsupervised anomaly detection model identifies *media storm candidates*—anomalous periods of news convergence. Then, a domain expert labels true media storms from these candidates. This human-in-the-loop process iterates until convergence, uncovering media storms over the period.

Our approach offers several advantages. First, methodologically speaking, it integrates the temporal features of topic- or keyword-based outlier detection described above, without relying on or being limited by idiosyncrasies of researcher design. Additionally, the utilization of unsupervised anomaly detection allows us to overcome the huge quantities of data, presenting experts with a small set of candidates to focus on in determining the existence of media storms. Furthermore, our approach attempts to bypass inherent amorphousness by offering a solution that is not based on predefined statistical thresholds designed for specific texts, but rather relies on the overall dynamics of news coverage for any given period. The use of unsupervised anomaly detection allows media dynamics to reveal themselves in the data. Our expert input comes into play in validating these patterns, confirming they correspond to the theoretical concept. This expert input in the choice of seeds and dispersion signals can also allow researchers to integrate their own research perspective within the process - an advantage when dealing with an inherently amorphous concept.Thus, we are able to uncover additional, more diverse media storms than in previous studies.

We utilize a large-scale corpus of news articles spanning 20 years of media coverage (1996-2016) to demonstrate our method. We employ two distinct experimental setups, addressing a broad spectrum of potential research applications. The first setup utilizes a seed list of media storms to uncover additional occurrences within the same time frame. The second setup utilizes an analyzed time frame to detect media storms in a new, unlabelled target period. We conclude with a preliminary analysis of our findings from both setups, underscoring the efficacy of our method and its potential for media storm research. We then test the capability of a generative Large Language Model (LLM) to perform the expert validation. The results justify the human-in-the-loop approach, while pointing to the possibility of further automation in the future.

Finally, we contribute these findings as a media storms dataset for the years 1996-2016. We believe that this dataset opens up a wide array of exciting research avenues. While the concept of media storms holds great significance to social actors, politicians and social scientists from various fields, empirical exploration has been limited. As the classification of storms within large-scale news coverage data improves, we can enhance our understanding of how these news hypes unfold from a single story or event to a cascade of public interest. In an era marked by heightened concern over the media's impact on the information landscape – highlighted by issues like polarization, the spread of misinformation, and the prominence of social media – such insights into these significant elements should offer important contributions.

2 Data

2.1 News Articles

To track the media coverage, we assembled a corpus of 1,187,607 news articles taken from three major news outlets – the *New York Times*, the *Los Angeles Times* and the *Washington Post* – between

Media Outlet	Articles	Tokens	Tokens/Article (Avg.)
New York Times	520,648	373,980,075	718.30
Washington Post	360,788	293,024,961	812.18
Los Angeles Times	306,171	240,119,545	784.27
Total	**1,187,607**	**907,124,581**	**763.83**

Table 1: Corpus Statistics

1996 and 2016. All full-length texts for this time period purchased and downloaded via a license agreement with LexisNexis. [1] These were filtered to include only articles from the News and Editorial sections. Corpus statistics are detailed in Table 1.

2.2 Seed list of Media Storms

To initialize our method, we build upon a seed list of media storms to begin calibrating the hyperparameters of the unsupervised anomaly detection. We begin with a list of storms from Boydstun et al. (2014) that has been widely used in media storm research. The researchers labeled the New York Times front page for a 10-year period to manually identify media storms. However, their effort contained several self-acknowledged constraints: they focused solely on domestic issues, measured only one national newspaper, and chose arbitrary statistical thresholds for operationalization. We wish to capture the essence of a media storm through a small set of mega-stories of national and global significance (expected to be present in the three outlets included in our corpus).

Consequently, we started with the items on their list as media storm *candidates*, which we could use for our first experimental setup of the method (within the 10-year period overlapping with our corpus collection: 1996-2006). However, we adjusted their list to better suit our use-case. First, since they analyzed only the New York Times, we included only national-level stories. For example, storms regarding local sports teams or municipal politics were removed. Second, we extended the list to include significant international stories, such as wars and foreign disasters, which also meet our conception 'media storms'. The end result is a modified list of 48 media storms between the year 1996 and 2006. We used this list to initialize the first calibration iteration of our unsupervised analysis of the full corpus in the first experimental setup (described in Section 3). We note that these are seed storm candidates used to begin the exploration of our data; we are aware that some of these events might not register as media storms after running our automated method, and that they do not represent all media storms occurring in the time period.

[1] https://www.lexisnexis.com

3 Method

In this section, we present our method to detect media storms in a large corpus. First, we describe the representation of our texts into dispersion signals. Second, we detail the unsupervised anomaly detection model employed to analyze the signals. Finally, we outline the integration of the dispersion signals, anomaly detection and human-in-the-loop validation in a media storm detection method.

3.1 Representation

Our basic assumption is that during media storms, the news coverage converges surrounding a single story or event, decreasing its variance. Thus, we utilize the following method to refine the raw text into a one-dimensional signal representing the daily media dispersion. For each day in the duration of our research period, the corresponding news articles are converted into a multi-dimensional embedding. We calculate a covariance matrix based on this embedding, to capture the variance between all the day's articles over all of the embeddings' dimensions. However, since we are interested in capturing the dynamic of the dispersion over time, we calculate the commonly-used *trace* value (normalized by the number of articles published that day). This provides us with a single value for the daily dispersion of the news articles. These are then aggregated to compile one-dimensional dispersion signals for the full duration of the research corpus.

In identifying media storms, we seek to include multiple representations of the texts, capturing diverse discursive attributes. We do this due to the complexity of media storms. In some cases, they might correspond to a single event; in others, they might evolve to encompass multiple stories and news "angles". In some cases, such as in crises or scandals, we might expect to find specific textual styles expressing drama or surprise. However, in cases such as groundbreaking court cases or anticipated political events, the storm is signaled by the sheer volume of coverage rather than any specific reporting approach. With this complexity in mind, we incorporated four types of document embeddings to create four separate dispersion signals. This offers a level of robustness, ensuring that we rely on various types of discursive attributes.

3.1.1 Actors & Settings

Actors are integral components of news stories. Previous research on automated identification of news events does so by focusing on entities, assuming that texts referring to the same people, places and times in the same period, refer to the same news event (Nicholls and Bright, 2019; Trilling and van Hoof, 2020). Therefore, we include these same features in our own approach in order to identify convergence in coverage around specific events. We used the *spaCy* open-source natural language processing (NLP) named-entity recognition (NER) package (Honnibal and Montani, 2017) to extract the actors and settings of each article. For each document, we generated an embedding based on the frequency of each entity within an entity vocabulary computed over the full corpus.[2]

3.1.2 Topics

In many cases, news coverage focuses more on a general issue than a specific story. For instance, strings of unrelated violent incidents could trigger a general spike in attention to crime without any of the individual events being newsworthy on their own. Thus, we sought to include storms being expressed in categories as opposed to only distinct stories, aligning with previous studies identifying storms as dramatic increases in coverage to an issue (Boydstun et al., 2014; van Atteveldt et al., 2018). To generate embeddings for this feature, we utilized an unsupervised topic model – *top2vec* – which leverages joint document and word semantic embedding to find topic vectors in a corpus (Angelov, 2020). Such topics focus on the issues expressed in the news articles. We trained a model containing 100 topics, so each document was represented by a 100-dimensional vector. Each dimension's value was the cosine distance of the document from the corresponding topic's centroid.

[2]Documents were truncated to the first 200 tokens, is in accordance with previous work in media studies showing that the first section of the article contains the important and relevant information (Welbers et al., 2021)

3.1.3 Narrative plot elements

Plot refers to "the ways in which the events and characters' actions in a story are arranged" (Kukkonen, 2014), and thus provide more information on the structure and "tellability" (Shenhav, 2015) of stories at the heart of media storms. In order to include plot elements, we used *NEAT* – a multi-label classifier that was trained on a specially compiled dataset (Levi et al., 2022) to identify three plot-driven, narrative elements – *complication*, *resolution*, and *success*. Each document was represented by three dichotomous variables to include each of the three narrative elements.

3.1.4 Large Language Model (LLM)

Finally, we chose to include document embeddings based on pre-trained, transformer-based LLMs. Such models uncover latent features and patterns found within texts, and have proven to be a standard for diverse NLP tasks. We used the *all-mpnet-base-v2* sentence-embedding model trained with a modified pre-trained BERT network that uses siamese and triplet network structures to derive semantically meaningful sentence embeddings that can be compared using cosine-similarity (Reimers and Gurevych, 2019).

We note significant correlations between the four signals (Table 2). However, the correlations indicate that there is not a complete 'overlap'. This attests to each signal's exclusive information.

	LLM	Entities	Plot
Topics	0.89	0.92	0.69
LLM		0.86	0.88
Entities			0.70

Table 2: Pearson correlations between signals

3.2 Unsupervised Anomaly Detection

With these media dispersion signals, we can begin the detection of anomalous convergence periods. To this end, we chose to utilize Facebook Prophet (Taylor and Letham, 2018). Prophet is an open-source library that is conceived to be a reliable "off-the-shelf" time-series forecasting model that could be easily applicable in a variety of use cases. Prophet fits an additive regression model to a time series while including components for a linear or logistic growth curve, yearly and weekly seasonality cycles, and user-designated holidays:

$y(t) = g(t) + s(t) + h(t) + \varepsilon_t$, where $g(t)$ represents the trend component, $s(t)$ denotes the seasonal component, $h(t)$ stands for the holiday effect at time t, and ε_t is the error term.

The model is fitted to the time series in question, flagging data points that significantly deviate from predicted values as anomalies. The deviation is determined by the *interval width* hyperparameter – the width of the uncertainty levels ascribed to the model. For example, a wider interval means only extreme values will be labeled anomalies. Two other hyperparameters - the *changepoint prior scale* and the *changepoint range* - are important for our application. The first sets the number of time-series changepoints to include in the model. The second specifies the proportion of the time series used to fit these changepoints. When working with decades worth of data, such values can significantly influence the model's predictions. For example, a lower changepoint range means that the model takes into consideration only the early portions of the time series, while a low changepoint prior leads to decreased sensitivity to fluctuations. We chose to focus on these three hyper parameters, fine-tuning them throughout our procedure to calibrate the unsupervised anomaly detection. For example, in Figure 1 we see the dispersion signals for the outbreak of Hurricane Katrina.

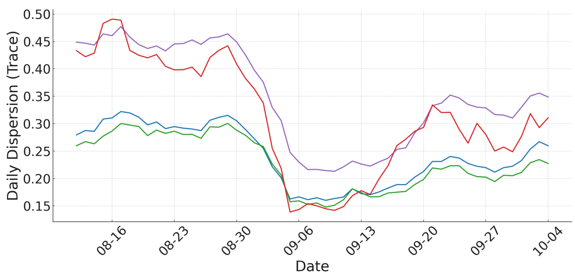

Figure 1: Hurricane Katrina – dispersion signals: entities (green), LLM (purple), narrative plot elements (red) and topics (blue).

3.3 Media Storm Detection

We define a two-step procedure for identifying media storms in our corpus.

Step 1: Take as input an initial list of media storms and a target corpus of media coverage represented as described in 3.1,[3] to run the anomaly detection. Treating the initial input list as the "ground truth" for the current iteration, we evaluate the model's precision and recall as follows:

Precision $= \frac{D}{A}$ and *Recall* $= \frac{D}{S}$, where D is the number of media storms from the initial list labeled as anomalies by the model, A is the total number of anomalies detected by the model, and S is the number of media storms in the initial list.

We conduct a random search (Bergstra and Bengio, 2012) of the hyperparameter space, running multiple instances of the anomaly detection with varying the three aforementioned hyperparameter values. We evaluate each instance by its precision and recall, seeking iterations with the highest scores in both metrics. In cases of ties, we prioritize recall. [4] For the optimal instance, we examine the results of the anomaly detection, noting the dates of all periods of consecutive anomalies of at least two consecutive days. We filter these to include only the time frames where a majority out of the four dispersion signals were flagged as anomalies. This criterion was added due to the inherently ambiguous nature of media storms; we want to focus on genuine media storms and not merely statistical noise originating in the anomaly detection model or borderline instances that might be contentious among researchers. This final, filtered list is our output: a collection of anomalies – media storm candidates.

Step 2: Take as an input the list of media storm candidates. We apply expert validation to ascertain which candidate corresponds to a genuine media storm. For each anomaly cluster, the expert reviewed newspaper articles from the associated dates and cross-referenced the time frame with historical events from the corresponding dates. Only anomaly clusters found to correspond to a genuine occurrence were provided descriptive labels by our expert and added to the set of media storms. More detailed information and guidelines regarding the expert validation can be found in Appendix A.

3.4 Experimental Setups

We utilized this two-step procedure in two distinct setups: In-Period and Out-Period implementations.

In-Period. In this setup, we focused on a target period between 1996-2006, aiming to expand a seed list and detect all other storms in same period. We started by applying the two-step procedure described in 3.3 to the seed list described in 2.2 and

[3]Smoothed by finding the 7-day rolling mean

[4]We assume that our initial storm list is but a portion of the real media storms in our target period. Therefore, we prioritize maximizing our identification of these real storms, before maximizing the sensitivity of the model.

the dispersion signals for the target years described in 3.1. The output list of validated storms from the first iteration was saved, and then used to initialize a second iteration of the procedure. The output of this iteration became the seed of the subsequent iteration. We continuously add the validated media storms to a list of finalized media storms over all iterations. We continued the iterations until reaching convergence, defined by identifying new media storms amounting to less than 1% of our current list of finalized media storms. We note that it can be necessary to curate the finalized list of media storms to consolidate duplicate storms. These were primarily due to small variations in the anomaly dates in each iteration that may still encapsulate a single media storm time frame.

Out-Period. In this setup we utilize the two-step procedure in 3.3, but begin the first step with input seed storm lists for one period, to uncover an output of occurrences in a second, unlabeled time period. Specifically, we compile data from an analyzed period together with additional, unlabeled data. As per Step 1, we use the already-labeled storms to run the random search and find the optimal anomaly detection instance. Then, we implement Step 2 on the media storm candidates for the new time period. In this way, we leverage information from a previous time frame to create a list of validated media storms for the unlabeled data.

These two experimental setups correspond with two common research scenarios. The In-Period deployment demonstrates the ability to leverage a handful of qualitatively-identified media storms to curate a comprehensive list encompassing a full target period. This challenge becomes especially pronounced when transitioning from qualitative, small-scale studies to more systematic, big-data-driven research. The Out-Period deployment demonstrates the ability to leverage an analyzed time period to detect media storms in a new time frame. This offers promise both for expanding datasets and for predictive prospects.

4 Results

Table 3 shows the results of the In-Period experimental setup. We performed four rounds of our procedure until reaching convergence – adding a single new media storm to our collection of 100 finalized storms. For each round, we count the number of storm candidates found by the anomaly detection model, the number of candidates validated as new storms, and the number of candidates found to not correspond with storms, as described in 3.3. Additionally, since in this setup we run multiple rounds on the same period, we note the completely newly-discovered media storms – instances that were not detected in previous rounds.

Iteration	**1**	**2**	**3**	**4**
Storm candidates	116	141	132	133
Storms validated	94	95	94	93
Not validated	22	46	38	40
New storms	71	18	4	1

Table 3: In-Period iterations

Table 4 displays, for each pair of signal types, the Pearson Correlation between the anomalies detected based solely on each of the signal types. An analysis of these correlations reveals that each signal contains exclusive information. Notably, the Plot signal shows the lowest correlations, perhaps due to the NEAT model being more discourse-grounded than vocabulary-based.

	Entities	**LLM**	**Plot**
Topics	0.69	0.64	0.46
Entities		0.72	0.47
LLM			0.53

Table 4: Anomaly-based Pearson Correlations

In our implementation of the Out-Period experiment, we ran a single round of the two-step procedure (described in Section 3.3) for each year between 2007 and 2016 in our data, utilizing the media storms found in the previous nine years as seeds for detection in the final year. For example, we utilized the media storms identified in the In-Period experiment in the years 1997-2006 as our input to find the media storms of 2007. Then, to analyze the year 2008, we utilized the storms from the years 1998-2007, and so forth.

Table 5 displays the results from our Out-Period experiments. There are slight fluctuations in the results of each round. For example, in 2007 and 2008 we identified only 10 candidates, while reaching peaks of 20 candidates in 2014 and 2016. Additionally, there is a slight variance in the number of candidates verified as media storms (second row) and the number of candidates not corresponding to genuine storms. The existence of slight fluctuations seems reasonable; we would expect slight

Year	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016
Storm candidates	10	10	15	15	16	15	19	20	15	20
Storms validated	6	9	12	11	12	14	14	16	13	13
Not validated	4	1	3	4	4	1	5	4	2	7

Table 5: Out-Period iterations

differences between periods when working with long-period temporal data.

Year	# Storms	Duration Avg.	Duration STD
1996	9	8.33	5.96
1997	9	6.56	1.59
1998	14	9.14	4.59
1999	9	7.78	3.80
2000	11	9.73	7.40
2001	4	9.00	6.73
2002	10	12.60	7.82
2003	10	19.00	22.77
2004	11	13.00	10.14
2005	9	8.33	5.32
2006	5	7.80	3.35
Total	**101**	**10.38**	**9.54**

Table 6: Storms statistics – 1996 to 2006

Year	# Storms	Duration Avg.	Duration STD
2007	7	10.57	4.04
2008	9	9.22	4.94
2009	12	8.50	5.28
2010	11	7.73	5.66
2011	12	8.33	4.66
2012	14	8.64	5.33
2013	14	8.79	4.25
2014	16	8.56	6.36
2015	13	10.54	5.50
2016	12	9.42	4.64
Total	**120**	**8.96**	**5.06**

Table 7: Storms statistics – 2007 to 2016

The end result of these experiments is 101 storms for the first period (1996-2006), and 120 storms for the second period (2007-2016) for a total of 221 media storms found in our corpus. These lists included many significant events, such as Hurricane Katrina (2015), the Sandy Hook school shooting and ensuing gun control debate (2012), and the Snowden NSA revelations (2013). For a descriptive overview, see Appendix B.

In addition to these unanticipated events, many of the storms detected correspond to routine, planned events such as elections or sporting events. However, there were also intriguing cases such as a 2010 spike in discussion on issues of airline security and privacy. That storm does not correspond to any specific major event, perhaps arising due to the proximity to the Thanksgiving transit peak. This is an interesting example of a media storm – public discussion of important issues – that arises not from any specific event directly linked to the issue (We stress that this is merely a hypothesis that invites focused examination).

What is particularly interesting about these statistics is the relative consistency of the results between the two setups. Upon examination of the results in Tables 6 and 7, we see that there are no strongly discernible differences between the media storms found in each of the setups. During the years 1996 to 2006, the annual average number of storms was 9.18. This contrasts with the period from 2007 to 2016, which recorded an average of 12 storms annually. This difference was statistically significant, $t(18) = -2.422$, $p = 0.026$. However, it would appear such differences might be due to real-world trends over time. Specifically, we see that the first years of the second period (2007 and 2008) reveal fewer storms than some of the first setup's years. Meanwhile, an examination of the storm durations does not reveal statistical differences ($t(146.15) = 1.343$, $p = 0.181$). Such results support the utility of both setups, suggesting that both are detecting the same phenomena.

Finally, to understand the importance of the domain expertise, we examined the validation statistics between the two setups: The four rounds of the first setup found 522 media storm candidates - anomaly clusters flagged by the Prophet model. Of these, 28% did not correspond to a true media storm according to the expert. The yearly rounds of storm detection in the second setup yielded a total of 155 media storm candidates, of which 22% were not deemed as storms by the expert. These numbers seem to justify the role of human validation.

5 Automated Validation

The second step in our proposed procedure (Section 3.3) involves manual validation of media storm candidates by an expert. In order to estimate the possibility to automate this step, we performed an experiment designed to test the capability of a generative LLM to perform this task. We assembled all the media storms candidates produced in the first step in the procedure, during both experimental setups described in Section 3.4, resulting in a set of 320 unique candidates. For each candidate, we prompted the GPT-4 model (OpenAI, 2024) to decide whether or not it constitutes a media storm, providing it with a sample of 75 news articles from the relevant dates as well as their pairwise cosine-similarities (see Appendix C for full details). Table 8 shows the confusion matrix summarizing GPT's decisions vs. our expert validation.

Notably, the expert and GPT-4 were in agreement about 45% of the storm candidates. Among these, they agreed on the storm's label in 74% of the cases. However, GPT-4 failed to identify a large number of media storms found by the expert. These include some clear cases, such as the British Petroleum oil spill in the Gulf of Mexico (2010), the shooting of U.S. Representative Giffords in Arizona (2011), and the Ebola outbreak (2014). While these results justify the human-in-the-loop approach, they merit further exploration into the possibility of utilizing computational models in performing (or at least aiding in) the validation step.

		Expert	
		Storm	Not Storm
GPT	Storm	19%	10%
	Not Storm	45%	26%

Table 8: Expert-GPT confusion matrix

This analysis further offers a unique opportunity to explore possible false-negatives by the expert (media storms they had missed). A total of 32 candidates were validated as media storms by the GPT-4 model but not by the expert. After reviewing these, five were determined to qualify as media storms by our expert: one new event, the Khobar Tower Bombing (1996), and four cases of additional peaks in coverage surrounding media storms previously validated as such by the expert.

6 Conclusion & Future Work

In this paper, we offer several contributions. First, we present a human-in-the-loop method to detect media storms in a large corpus of news texts. We describe a two-step iterative procedure, combining unsupervised anomaly detection and expert validation, to identify these rare events within a larger dataset. Significantly, whereas previous studies build upon 'arbitrary' statistical thresholds, we utilize an unsupervised anomaly detection algorithm to allow the media dynamics to reveal themselves in the data. Our expert input comes into play in validating these patterns, confirming they correspond to the theoretical concept. Consequently, we are able to uncover additional, more nuanced media storms than in previous studies. By incorporating expert validation, we can set the granularity or type of the storms which we seek to identify; researchers can express their research agenda to decide what types of media storms they are interested in detecting. Additionally, we performed a comparison between the expert and GPT-4, demonstrating that while not fully capable of replacing a human expert, there is some potential in utilizing a generative LLM during the validation process.

Second, our method offers a procedure that can be applied in various research scenarios, over diverse and large corpora, while leveraging expert knowledge for validation. Within the realm of this paper, we included three English-language newspapers for a specific time-frame. However, the method could plausibly be applied on any news corpora in any language, provided the necessary techniques could be utilized (e.g., entity-detection, sentence transformers). Additionally, researchers might be able to use this approach on non-mainstream media sources as well, including identifying periods of textual convergence in social media platforms and digital news.

Third, through the two experimental setups, we collected a comprehensive list of media storms. This time frame we chose to focus on is of particular significance for media scholars. Between 1996 and 2016, the media landscape underwent dramatic transformations, with the rise of 24-hour news cycles, the interactivity of social media and the fragmentation of the attention landscape (Chadwick, 2017; Edy and Meirick, 2018). These validated storms provide opportunities to examine intriguing theoretical questions, including how the volatility of the media landscape has evolved, changes in the events triggering storms, and perhaps developing predictive capabilities regarding storm outbursts and durations. Thus we use the results of this study to provide a dataset consisting of media storms with their start and end dates, which will be made publicly available to researchers together with the dispersion signals extracted from the corpus.

7 Limitations

We note two main limitations of this project. First, the procedure described here assumes that our media storms are all mutually exclusive. We locate time frames of anomalous coverage and associate each period with a single, discrete media storm. In reality, a single time frame might contain more than one major news story, or the anomaly might actually be identified as one story declines and the other begins. Such findings correspond to issues that arose during the expert validation stage: some anomalous clusters contained a few potential storm stories. Only upon close examination of the time series' peaks and the articles that were published in correspondence with them, could we decide on a single story for the storm. Additionally, some of the periods actually did include two separate media storm stories, one following the other (See comments in Appendix A). In this project, we limited ourselves to choosing a single media storm per each period. In future work, however, we could integrate a clustering method to further distinguish and track stories within the media storms.

A second limitation is that our method does not include systematic steps to prevent the existence of false negatives - media storms undetected by the anomaly detection. Since we do not have a gold-standard to initiate our storm detection, there remains a possibility that our procedure may have failed to detect instances within our corpus. In general, our approach relies on high-quality seeds to initiate the search for additional media storms. We assume that these instances fully represent the phenomenon, and that, therefore, all media storms should be similar enough in characteristic to them. In this way, multiple iterations of anomaly detection should uncover all true media storms. However, we note that this is not a complete solution to the issue of false negatives. In future work, we would examine potential solutions, such as randomly sampling the non-storm time periods to examine for storms, utilizing computational models to produce "competing" validations (as in the preliminary experiment described in Section 5, or perhaps generating additional textual signals which might reveal more storm instances.

References

Dimo Angelov. 2020. Top2vec: Distributed representations of topics. *Preprint*, arXiv:2008.09470.

J. Bergstra and Y. Bengio. 2012. Random search for hyper-parameter optimization. *Journal of Machine Learning Research*, 13(Feb):281–305.

Amber E. Boydstun, Anne Hardy, and Stefaan Walgrave. 2014. Two faces of media attention: Media storm versus non-storm coverage. *Political Communication*, 31(4):509–531.

Andrew Chadwick. 2017. *The Hybrid Media System: Politics and Power*. Oxford University press.

Jill A Edy and Patrick C Meirick. 2018. The Fragmenting Public Agenda: Capacity, Diversity, and Volatility in Responses to the "Most Important Problem" Question. *Public Opinion Quarterly*, 82(4):661–685.

Mike Gruszczynski. 2020. How media storms and topic diversity influence agenda fragmentation. *International Journal of Communication*, 14(0).

Matthew Honnibal and Ines Montani. 2017. spaCy 2: Natural language understanding with Bloom embeddings, convolutional neural networks and incremental parsing. To appear.

Karin Kukkonen. 2014. Plot. In Peter Hühn et al., editors, *The Living Handbook of Narratology*. Hamburg University, Hamburg. Accessed: 12 Feb 2019.

Effi Levi, Guy Mor, Tamir Sheafer, and Shaul Shenhav. 2022. Detecting narrative elements in informational text. In *Findings of the Association for Computational Linguistics: NAACL 2022*, pages 1755–1765, Seattle, United States. Association for Computational Linguistics.

Benjamin Litterer, David Jurgens, and Dallas Card. 2023. When it rains, it pours: Modeling media storms and the news ecosystem. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 6346–6361, Singapore. Association for Computational Linguistics.

Josephine Lukito, Prathusha K Sarma, Jordan Foley, and Aman Abhishek. 2019. Using time series and natural language processing to identify viral moments in the 2016 U.S. presidential debate. In *Proceedings of the Third Workshop on Natural Language Processing and Computational Social Science*, pages 54–64, Minneapolis, Minnesota. Association for Computational Linguistics.

Nishanth Nakshatri, Siyi Liu, Sihao Chen, Dan Roth, Dan Goldwasser, and Daniel Hopkins. 2023. Using LLM for improving key event discovery: Temporal-guided news stream clustering with event summaries. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 4162–4173, Singapore. Association for Computational Linguistics.

Tom Nicholls and Jonathan Bright. 2019. Understanding news story chains using information retrieval and network clustering techniques. *Communication Methods and Measures*, 13(1):43–59.

OpenAI. 2024. Gpt-4 technical report. *Preprint*, arXiv:2303.08774.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

S.R. Shenhav. 2015. *Analyzing Social Narratives*. Routledge, New York.

Sean J Taylor and Benjamin Letham. 2018. Forecasting at scale. *The American Statistician*, 72(1):37–45.

Damian Trilling and Marieke van Hoof. 2020. Between article and topic: News events as level of analysis and their computational identification. *Digital Journalism*, 8(10):1317–1337.

Wouter van Atteveldt, Nel Ruigrok, Kasper Welbers, and Carina Jacobi. 2018. 2. news waves in a changing media landscape 1950-2014. In Peter Vasterman, editor, *From Media Hype to Twitter Storm: News Explosions and Their Impact on Issues, Crises and Public Opinion*, pages 61–82. Amsterdam University Press, Amsterdam.

Peter L.M. Vasterman. 2005. Media-hype: Self-reinforcing news waves, journalistic standards and the construction of social problems. *European Journal of Communication*, 20(4):508–530.

Stefaan Walgrave, Amber E. Boydstun, Rens Vliegenthart, and Anne Hardy. 2017. The nonlinear effect of information on political attention: Media storms and u.s. congressional hearings. *Political Communication*, 34(4):548–570.

Kasper Welbers, Wouter Van Atteveldt, Jason Bajjalieh, Doron Shalmon, Prashant V Joshi, Scott Althaus, and Marc Jungblut. 2021. Linking event archives to news: a computational method for analyzing the gatekeeping process. *Communication Methods and Measures*, 16(1):59–78.

Charlotte Wien and Christian Elmelund-Præstekær. 2009. An anatomy of media hypes: Developing a model for the dynamics and structure of intense media coverage of single issues. *European Journal of Communication*, 24(2):183–201.

Gadi Wolfsfeld and Tamir Sheafer. 2006. Competing actors and the construction of political news: The contest over waves in israel. *Political Communication*, 23(3):333–354.

A Guidelines for Expert Validation

At the conclusion of the first step in the procedure described in Section 3.3, the expert received a set of storm candidates (i.e. anomaly period) encompassing news articles' start and end dates. For each such anomaly period, the expert was given two tasks: (1) decide if a media storm is occurring, and (2) if a storm has been identified, decide on a descriptive label of the dominant news story or group of stories.

In order to address these tasks the expert performed the following steps:

Review the news article titles. For this purpose, the expert was aided by t-distributed stochastic neighbor embedding (t-SNE) visualization of the spread of all the articles published during this period. The visualization embedded the articles in the latent semantic space based on the *all-mpnet- 335base-v2* sentence-embedding model as described in 3.1. The t-SNE visualization allows for improved efficiency in browsing news coverage, helping to identify clusters of similar articles and understand if there is a dominant story or group of stories among them. The expert reviewed the titles of news articles and, if necessary, further explored the articles in context.

Examination of historical context of storm candidates. For this purpose, the expert used lists of key events (such as `www.infoplease.com/current-events`) and other sources, such as Google and Wikipedia. We note that key historical events helped identify many media storms; however, in some cases, media storms evolved from increased attention to specific issues or policy domains, rather than historical events.

B Media Storms

Between the years 1996 and 2016, we found 221 media storms utilizing our method. These storms include several categories of news stories. First, 43 of the instances were relating to U.S. elections and election campaigns - including the elections themselves, debates, party primaries, and coverage of the campaign trail.

Another relatively prevalent category are unanticipated violent events. These include the Versace murder (1997), the Columbine School Shooting (1999), the September 11th terror attacks (by far the most prominent storm as attested to by the convergence levels), the shooting of U.S. Representative Giffords in Arizona (2011), and the riots killing of police officers in Dallas (2016). Overall, there were 30 such media storms.

42 of the media storms were considered foreign news, in that they occurred outside of the U.S. These include wars in the Balkans (1997-1999), violent outbreaks in the Middle East (e.g., 2002, 2012, 2013), disasters (e.g., the 2010 earthquake

in Haiti, the 2005 tsunami in the Indian Ocean and the Fukushima nuclear accident in 2011), and significant deaths (e.g., Princess Diana in 1997 and Pope John Paul II in 2005).

Another category of interest was media storms that included intense coverage of stories that did not correspond to a specific event, but rather related to policy-driven matters. For example, there have been several periods of intense attention on the U.S. involvement in Iraq that would encompass multiple stories - daily insurgent attacks, visits by U.S. government officials, interviews with local leaders - occurring long after specific events such as the original invasion or the start of the "Surge" troop increase. These were cases where we could discern intense discussion of an issue for a period, without linking the media storm to a specific trigger. Another interesting and surprising example of such a storm occurred in 2010, when the media coverage reveals high levels of attention to issues of air travel, airport security and debates about passenger privacy. While we could not find any clear trigger event behind such coverage, the proximity of the discussion to the Thanksgiving holiday rush hints at what might be a heightened public attention to such issues. Perhaps an online discussion on a social media platform might have even initiated such a media discussion.

Table 9 summarizes the 10 longest media storms found in our dataset.

Title	Year	Duration
2003 invasion of Iraq	2003	80
2004 Presidential Election	2004	41
Iraq War coverage	2003	30
US Ebola outbreak	2014	30
2000 Presidential Election	2000	29
Trent Lott Scandal	2002	26
Operation Defensive Shield	2002	23
AIG Bonuses	2009	23
1996 Olympics	1996	22
2010 Midterm Elections	2010	22

Table 9: 10 longest media storms

C GPT-4 Prompts

For each media storm candidate (anomaly instance), we provided the following prompt to the model via the OpenAI API:

"*A media storm is a dramatic increase in media attention to a specific issue or story for a short period of time. In such a case, we expect most news articles for a given period to discuss a single story or issue. I have a corpus of news articles published between [START DATE] and [END DATE]. For this period, please use the article titles and the dates to first decide if a media storm is occurring. If a media storm is occurring, respond with 'YES' and provide a label to describe the story behind the media storm. If a media storm is not occurring, respond with 'NO'. Please respond concisely in the format: 'YES: [LABEL]' or 'NO'.*"

This prompt included the dates of the anomaly period, the titles of a random sample of news articles published during that period, and a matrix containing the pairwise cosine distances between the sample articles' embeddings. This information was provided to match the details provided to the human coder in the validation stage.

We randomly sampled articles for each period due to the large number of documents for each anomalous interval. We experimented with several sample sizes, finding that sampling 75 articles to provide with the prompt yielded the best results.