Text-Only Version

Shifting Sensationalism: How Headlines and Content Differ in Sensationalism for Events in the U.S. Media

Deeya Bodas, Grace Sherman, Sharika Kottapalli, Alan Tao, Oju Chaudhary

1 Abstract

This research project performs an observational study on a set of articles published by varying American news entities over a 2 year period. Through gathering and analyzing relevant metrics such as the topic, politicalness, and level of sensationalism of these articles, we hope to gain further insight on how the media shifts over time and in response to events.

RQ0 (Comparing distributions):

How many articles of each topic exist over the timeline of our dataset?

RQ1 (Sensationality Gap):

How does the Sensationality Gap between article headlines and their corresponding content differ across each topic?

RQ2 (Time & Seasonality):

What are the trends in the Sensationality Gap between article headlines and corresponding content across time for each topic?

RQ3 (Events):

What is the eff ect of speci c events on the Sensationality Gap? How does this diff er for political

Text-Only Version

and non-political topics?

By distinguishing how media outlets adapt their framing in varying contexts, this research aims to contribute to a broader understanding of media bias, clickbait strategies, and the evolving role of sensationalism in shaping public discourse.

2 Introduction

As news stations have to compete for clicks in today's fast-paced media landscape, more and more emphasis is placed on headlines to capture public interest to improve engagement. These short, prominen t phrases often shape a reader's perception before they even access the article's full content. However, the extent to which headlines sensationalize information, especially when compared to the more balanced and informative article body, is still not well understood. In this study, we address this gap by introducing and measuring the ?Sensationality Gap? (SG): the difference in sensational tone between an article's headline and its full content. This gap not only re ects stylistic and journalistic choices but may also in uence how readers interpret information conveyed in articles. Ultimately, the language that articles use shapes how readers interpret current events.

Not all forms of news exhibit sensationalism in the same way. Different topics, such as politics, health, or entertainment, show different degrees of headline exaggeration. Our rst goal is to explore these topic-based differences in SG. In this study, we also aim to examine temporal patterns: are there predictable seasonal trends in how headlines diverge from article content? Does the SG increase during election seasons or end-of-year news cycles? Lastly, we try to further examine the nuances of time on SG with event-based analysis. Through this, we investigate how media sensationalism responds to signi cant political and non-political events.

To address our research questions, we collected 1.24 million online news articles published between January 2016 and January 2018, categorized into multiple topics. Using a set of 98 manually labeled articles as a training set, we used few-shot classi ers to measure sensationalism and topics in both headlines and article bodies, de ning the Sensationality Gap as the difference between headline and content sensationalism scores. We applied ARIMA models to analyze long-term and seasonal trends in SG across topics. Additionally, we used the Difference-in-Differences (DiD) approach to quantify the causal impact of speci c events (like presidential inaugurations, royal weddings, and BREXIT) on sensationality scores.

By identifying patterns in how sensationalism evolves across topics, time, and events, our study contributes to a broader understanding of media bias, audience manipulation, and journalistic ethics. In doing so, we offer a new way to examine the ways that headlines and articles shape public perception, and also how news outlets strategically use language to navigate the attention economy.

3 Background

Existing literature sugg ests that both political and entertainment-oriented news are prone to sensationalism, although direct comparative studies between political and non-political topics are limited.

1

3.1 De ning Sensationalism

First, we must de ne sensationalism: according to the study ?An Analysis of Sensationalism in News? [

1

], sen-

sationalism is coined as the tactic news organizations use to grab the reader's attention by ?provoking an emotional response in readers?.

3.2 Clickbait

Another study, ?Analyzing Sensationalism in News on Twitter (X): Clickbait Journalism by Legacy vs. Online-Native Outlets and the Consequences for User Engagement? [

Text-Only Version

2

] by Khawar and Boukes assesses sensationalism by analyzing ten key features commonly associated with clickbait. Similar to the previous study, these features include: hyperbolic words and phrases, listicles, forward referencing, slang, and informal punctuation. Khawar and Boukes, in their study, rst utilized manual content analysis by assigning binary variables to these ten sensationalist features (?is? vs. ?is not?). ?Analyzing Sensationalism? also assigns binary variables for topic categories (politics, government, celebrity news, etc.) to later categorize into broad political/entertainment groupings. While Khawar and Boukes used tweets to establish a correlation between sensational features in article headlines and user engagement on social media, our study would instead focus on how sensationalism changes after different events. In other words, while the article answers the question: ?why are sensational tactics used in news headlines?? our project answers the question: ?when are sensational tactics used the most??. While the paper examines sensationalism in political versus non-political headlines and its impact on public perception, this temporal factor is not discussed.

3.3 Sensationality Gap

While the perceptual eff ects of sensational words and lan-

Text-Only Version

guages in article headlines are well situated in today's literature, a gap remains in understanding how the sensationalization of headlines relative to article bodies varies in response to different types of news events. For example, the article ?Towards a pragma-linguistic framework for the study of sensationalism in news headlines? [

3

],

bridges textual and audience research to better understand how sensationalism operates in news discourse. The study also discusses how future research could expand on how sensationalist strategies vary across media outlets. In this way, while current studies may focus on general media sensationalism, they do not differentiate trends in sensationalism over time as well as how media outlets adjust their reporting strategies for differing events. In addition, while headlines may contain sensational words, the body of the article may remain unbiased: this nuance is not discussed in any of the studies cited. To address this gap, our research will compare sensationalist tendencies in headlines versus article bodies. This term that we de ne as ?Sensationality Gap? aims to help us understand how publications use the readers' attention once it is gained. By quantifying this gap, we can assess the extent to which media sources rely on emotionally charged headlines to at-

Text-Only Version

tract engagement while delivering more neutral or factual content in the article itself.

3.4 GPT Models as Linguistic Classi cation

Lastly, although we may use previous methodologies of manual content analysis followed by algorithmic content analysis, GPT could also be used for linguistic classi ca- tion. The study ?GPT is an effective tool for multilingual psychological text analysis? [4], discusses how GPT models could be used to accurately detect psychological con- structs in text. Psychological constructs were de ned in the article as sentiment, discrete emotions, offensiveness, and moral functions and the model outcomes were com- pared to dictionary-based methods and machine-learning models. The results showed that GPT outperforms these traditional methods and offers a scalable, user-friendly al- ternative. This way, our study off ers a novel methodology to classify sensationalism in both article headlines and content body.

4 Data

To train and test our model, we used a collection of articles covering a 2-year span published from 01-01-2016 to 12- 31-2017. Two datasets were used:

Text-Only Version

?

All the News 2 Dataset:

https:

//components.one/datasets/

all- the- news- 2- news- articles- dataset/

?

AllSides Media Bias Chart:

https:

//www.allsides.com/media- bias/

media- bias- chart

To understand the media bias of each article, we used

the AllSides Media Bias Chart which categorized each

publication into one of the following categories, based

on their political slant: left, lean left, center, lean right,

right. The following gures depict the distribution of

articles based on the publications media bias [gure 1a],

the article volume of each publication [gure 1b], and the

media bias of each publication [gure 1c]

In the AllSides Media Bias Chart, not all of the publica-

tions covered in the All the News 2 Dataset were included.

However, the website contains media bias ratings for pub-

lications not mentioned in the chart, so those ratings were

(a) Article Volume by Media Bias

Text-Only Version

(b) Article Volume by Publica-

tion

(c) AllSides Media Bias Chart

Figure 1: Media Bias and Article Volume Representations

used for analysis of our ndings to answer RQ-3. Pro-

cessing and reading the data that we were looking to use

created quite a few challenges for research, and required

us to use methods for data processing that we had not

explored in class before. We created a Google Colab

Notebook to use a shared workspace for processing the

data [

1

]. We had challenges with working with the entire

dataset in-memory so we opted to cut down the data and

store it separately for the purpose of our research. We

also had to do some data-cleaning to reformat the dates

for the dataset, for which we wrote a custom script in the

notebook [

1

].

5 Methods

5.1 Labeling

In order to conduct our analysis on sensationalism of

articles in relation to topics, time, and events between

the 2-year period of January 2016 - Decemb er 2017, we

Text-Only Version

began by assigning the article titles and bodies a series

of labels. There were three different types of multi-label

classi cation tasks that we needed to conduct:

5.1.1 Establishing Topics

Classifying titles and articles into 1 of 10 categories

How: To ensure consistency and minimize ambiguity

in classi cation, we observed topic distributions across a

range of mainstream news sources (e.g., AP News, BBC,

NYT). After reviewing common editorial tags we curated

a set of 10 categories broad enough to cover diverse con-

tent while speci c enough to support accurate labeling.

?

U.S. Government/Military

?

World News

?

Economy/Business

?

Health/Lifestyle/Personal Finance

?

Science/Technology

?

Entertainment/Celebrity News

?

Sports

Text-Only Version

?

Human-Interest Story/Society

?

Crime/Law and Order

?

Other

An example of the manual topic classi cation process

is shown in gure 2a.

3

5.1.2 Classifying Politicality

Classifying titles and articles as political or non-political

How: Any article within the U.S. News/Military cate-

gory was automatically classi ed as political, while the

political-ness of articles within the rest of topics were

speci c to the text. An example of the manual politicality

classi cation process is shown in gure 2b.

5.1.3 Sensationalism Scoring

Classifying titles and articles as sensational or not

Criteria:

1.

Hyperbole: Extravagant language and superlatives

used to boost clicks and perceived news value.

2.

Forward Referencing: Phrases that create curiosity

Text-Only Version

by hinting at information only revealed after clicking,

(e.g., ?This is why. . . ?).

3.

Listicles: Headlines that present content as ranked

or numbered lists for quick, easy consumption (e.g.,

?17 Real-Life Secrets About. . . ?).

4.

Interrogative Structure: Headlines framed as ques-

tions to spark curiosity and prompt clicks (e.g., ?Will

it really matter??).

5.

Overuse of Capitalization: Excessive capitalization

used to add emphasis and grab attention, often sig-

naling clickbait.

6.

Entertainment/Celebrity News

7.

Informal Punctuation and Slang: Casual language

and expressive symbols (e.g., ?!!!?, slang) used to

grab attention and convey emotion.

How: We used a 6-point Likert scale based on six crite-

ria. These criteria were adapted from a similar study by

Khawar and Boukes that focused on scoring sensational-

ism in article bodies [

2

Text-Only Version

]. Since our evaluation includes

article titles and bodies, we selected only the criteria appli-

cable to both, ensuring consistency and relevance across

formats. The existence of a sensationalism criteria in a

title or a body grants the text one sensationalism point,

implying that a piece of text can have a sensationalism

score ranging from [0-6] after totaling the points. Later in

this study, we used the median score to de ne a threshold

of sensationalism. An example is shown in gure 2c.

In order to classify all articles on topic, politicalness,

and sensationalism, we divided this task into two parts:

Manual Labeling and Automated Labeling. Manual label-

ing was conducted by the 5 researchers on a total of 100

articles. These 100 labels were used to build the training

and evaluation datasets for the models that were used to

classify the rest of the articles automatically.

(a) Topic Classi cation Task

(b) Politicalness Classi cation Task

(c) Sensationalism Classi cation Task

Figure 2: Manual Classi cation Tasks

5.1.4 Labeling - Manual

For the rst cycle of our manual processing, we began

by assigning each person 20 articles to evaluate [

3

]. To

Text-Only Version

prevent bias, we ensured that no person was assigned a headline and body pairing, preventing the researcher from giving a corresponding headline and body the same scoring. In our second cycle of evaluation, we evaluated another person's set of 20 articles - Reviewer A labeled all of the articles of Reviewer B, Reviewer B labeled all of the articles of Reviewer C, and so on. No two reviewers saw the same headline or body in both cycles, ensuring that each article had exactly two sets of eyes on it. We conducted the manual labeling process on Google Sheets

[

4

]. After the manual labeling process, there were various types of classi cation disputes that we had to discuss and resolve as a group.

1.

Mismatch between politicalness of title and body:

If there was discrepancy between the politicalness of the title and body of an article, we resolved the dispute by discussing and choosing the ?better? label for the entire article.

2.

Misalignment of topic labeling b etween 2 reviewers:

If two reviewers categorized the same article into different topics, we discussed and selected the least ambiguous topic labeling.

3.

Misalignment of sensationalism scores between 2 reviewers: In our nal labeled dataset, we took the average of both reviewer's scores for each criteria. For example, if both reviewers found the headline to be hyperbolic the headline received a score of 1, and if no reviewers found it to be hyperbolic it got a 0. If the researchers disagreed the headline got a score of 0.5 for hyperbole.

4

5.1.5 Labeling - Automated

After manually labeling 100 articles, we transitioned to automated methods to process the remainder of the dataset. Selecting an appropriate multi-label classi er or natural language model involved balancing several key factors: Quantity: Our dataset includes approximately 1.2 million articles, and even when using a subset, the length of article bodies posed dif culty due to token limitations in many large language models (e.g., Google Gemini). Processing full articles also demanded substantial memory resources. Quality: Topic classi cation was more effectively handled by traditional multi-label classi ers, while generative models showed stronger performance in scoring sensationalism. The labels needed to be close

in accuracy to the human labels in order to be usable.

Limited Training Data: With only 100 manually labeled

examples, training a model from scratch was not feasi-

ble. One-shot or few-shot classi cation using pre-trained

models offered a more practical solution.

To conduct classi cation, we employed a Hugging-

Face model called SetFit: a ?prompt-free framework

for few-shot ne-tuning of Sentence Transformers? and

it ?achieves high accuracy with little labeled data? [

5

].

Given that the tasks for topic, politicalness, and sensation-

alism classi cation required us to de ne a set of possible

classes for each, SetFit provided us a method to incorpo-

rate our 100 manually labeled articles and automatically

label the rest of our dataset.

5.1.6 Topic and Politicalness Labeling Model

In order to establish a SetFit model for topic and political-

ness classi cation, we decided to rst take a subsample

of 62,000 articles out of the initial 1.2 million, as we

observed that both the training and inference stages of

our labeling process were considerably lengthy and often

exhausted our limited GPU resources. First, we created

strati ed training evaluation datasets, with an 80/20 split, from our 100 manually labelled articles, which included classi cations for the topic and whether or not an article was political. After con guring the columns in order to align with SetFit's requirements, we initialized the SetFit model with speci c parameters. We set the number of epochs to 5, to ensure the model repeatedly learns from each limited example, which is important given the small size of our training dataset. Additionally, we set the number of iterations to 15, which allows SetFit to generate diverse pairs, helping the model generalize political relevance beyond the few labeled examples. After training and evaluation was complete, we leveraged the trained model to predict the topic and politicalness of the remaining articles that weren't manually labeled. Finally, we aggregated the results into a new dataset with columns for the model's predictions, which allowed us to answer our research questions. We created 2 colab notebooks to perform classi cation of topic and politicalness. [

6

] [

7

].

5.1.7 Sensationalism Labeling Model

Labeling articles in terms of our sensationalism criteria was a more dif cult task due to the subjective nature of

Text-Only Version

the traits, the multi-label structure of the scoring system, and the limited size of our manually labeled dataset.

Attempt 1 (Google Gemini):

First, we tried Google's Gemini 2.0 Flash, a high-performing generative model that showed strong results when evaluating both headlines and article bodies. We re ned our prompt to include the six criteria for sensationalism scoring, followed by the headline and body as input. Gemini was able to provide reasonably accurate and nuanced assessments that aligned well with human reviewers. However, the main limitation was its token constraint and accessibility. Due to the average length of article bodies, many entries exceeded Gemini's token limit, requiring content truncation and reducing the consistency of inputs. Additionally, Gemini is a closed-source model hosted exclusively on Google Cloud, which restricted its scalability for large datasets like ours due to both cost considerations and cloud-based access limitations. Here is an example of Gemini's evaluation:

Headline:

"China stocks have best week in ve on trade talks optimism"

Model Response/Evaluation: Hyperbole:

"Best

Text-Only Version

week in ve"

quali es as a superlative highlighting the

positive outcome, which is meant to drive referral traf c.

(1 point).

Forward Referencing:

Not present. (0 points).

Listicle:

Not present. (0 points).

Interrogative Struc-

ture:

Not present. (0 points).

Overuse of Capitalization:

Not present. (0 points).

Informal Punctuation:

Not

present. (0 points)

Attempt 2 (Ollama with Mistral model):

To address

these constraints, we turned to Ollama, a framework for

running open-source LLMs locally [

8

]. After testing sev-

eral models, we selected Mistral due to its quality of

performance on our task. Like Gemini, we structured our

input to include the criteria followed by the headline or

Text-Only Version

body content. Mistral produced reliable results and was

more exible than Gemini, with no token limit. However,

performance came at the cost of speed. Because the base

training prompt had to be passed in full for each sample,

evaluation time was high, making it impractical for label-

ing a dataset of our size. For reference, it took about 40

seconds to label one headline so it would take about an

hour every 100 headlines - we had 1.2 million headlines.

While promising in terms of quality, the lack of scalability

led us to consider alternative approaches.

Attempt 3 (DistilBERT):

DistilBERT was the rst

model we tried that successfully ran end-to-end within

Type of Labeling

Accuracy

Topics

0.70

Political vs Non-political

0.90

Sensationalism of Headlines

0.75

Sensationalism of Content

0.60

Text-Only Version

Table 1: Accuracy of SetFit on Labeling Tasks

Google Colab, using the NVIDIA T4 GPU to accelerate

training [

9

]. It off ered a good balance between speed and

performance for multi-label classi cation. However, due

to our small training set of 100 labeled articles, the model

began to over t to highly represented traits like hyperbole

and forward referencing. This imbalance led to skewed

predictions and limited sensitivity to less common traits.

While similar research has addressed this issue by apply-

ing F1-score?based class weighting, we chose to proceed

with SetFit before implementing such adjustments.

Final (SetFit):

Ultimately, we returned to SetFit, the

same model architecture used for our topic and politi-

calness classi cation tasks [

10

]. Because SetFit accepts

only labeled examples (not prompts) and produces binary

outputs, we adjusted our training setup accordingly. We

revised our training data by summing the sensationalism

points across the six criteria for each headline or article.

Any sample with a total score greater than or equal to

1.0 was labeled as ?sensational?, while those below were labeled ?not sensational.? This threshold was selected af-

ter testing a range of cutoffs to best match manual labels.

This binary framing allowed SetFit to learn effectively

from the limited examples, and the model demonstrated

good generalization when applied to unseen samples. We

created two colab notebooks [

11

] [

12

] to perform classi -

cation of sensationalism on article headlines and content.

The accuracies of SetFit, which was calculated by cal-

culating the predicted and actual classi cations in the

evaluation dataset, are shown in Table 1.

5.2 Evaluation Methods

5.2.1 RQ 0:

Once the model was validated, we used pandas to orga-

nize the data into topic groups and visualize the distribu-

tion over every month in the two-year period [gure 3].

Additionally, we created a Correlation Matrix of Article

Volume Between Topics to determine if there was a sta-

tistically signi cant correlation between a rise in article

volume of one topic, and a corresponding rise in article

volume of another [gure 4]. To assess whether the diff er-

ence in article volume for a given topic between Month A and Month B was statistically signi cant, we conducted a t-test. Due to time constraints, we did not perform this test for every topic and month combination. Instead, we manually selected a few topic-month pairs that showed a noticeable difference in article volume on the graph. For these pairs, we used a t-test to compare the article volume difference between Months A and B for Topic X against the differences observed in other months.

5.2.2 RQ 1:

In order to plot visualizations for RQ1 for manually labeled data, matching was required between article headlines and article bodies. As manual labels of sensationality were on a six point scale, articles that had a score equal to or above one were deemed ?sensational? (1), while articles with a score less than one were deemed ?not sensational?. When the 62k articles labeled by SetFit were used, the same binary classi cation of sensationality was used, and was converted into an integer representation. Sensationality Gap was calculated as the difference between the sensationality of the headline and the sensationality of the article body (can be equal to ?-1?, ?0?, or ?1?). To assess the data on a discrete timeline scale, articles were divided by the week and month they were published. Visualizations were created with matplotlib: a stacked bar graph for Average SGs were calculated by averaging (mean function) the SG of articles from each

Week or Month.

5.2.3 RQ 2:

In order to visualize trends in our articles, we used

ARIMA (autoregressive integrative moving average),

which is a statistical model that performs forecasting

on data from a time series format. We decided to use

pmdarima, which is a Python library meant for ARIMA

modelling, to help us conduct this analysis. First, we

reorganized our 62,000 articles to be sorted by date. Then,

we decided on 5 speci c scenarios for which we visualize

our current data and future predictions, using weeks at

the time increment.

?

The amount of ?sensational? headlines, grouped by

topic

?

The amount of ?sensational? article bodies, grouped

by topic

?

The amount of ?sensational? headlines, grouped by

politicalness

?

The amount of ?sensational? articles bodies, grouped

by politicalness

Text-Only Version

?

Amount of articles, grouped by three categories

(?Aligned? = Both headline and body are ?sensa-

tional?, ?Headline more sensational? = Headline is

6

?sensational? while body is not, ?Article more sen-

sational? = Body is ?sensational', while headline is

not)

To t an ARIMA model on the data, we needed to estab-

lish the optimal parameters, which are de ned by Duke

University's Statistical Forecasting page [

13

] as ?p - the

number of autoregressive terms?, ?d - the number of non-

seasonal differences needed for stationarity, and ?q - the

number of lagged forecast errors in the prediction equa-

tion?. Essentially, these parameters (p, d, q) allow us to

change how much we use past values to predict future

ones, how much we difference our data to allow stationar-

ity, and how much we use past errors in our forecasting.

In order to determine the (p, d, q) values for each scenario,

we used autoarima, which is provided by the pmdarima to

 nd the best parameters to make predictions. We created

a colab notebook [

Text-Only Version

14

] to perform ARIMA modelling.

5.2.4 RQ 3:

In order to explore differences in sensationalism across articles, we created a notebook [

15

] that performs propensity score matching and diff erence-in-diff erences analysis across a series of topics and days, inspired by RL-4 and RL-5. Our analysis focuses on understanding whether there is a statistically meaningful difference in the number of sensational headlines between political and non-political articles around a speci c event. To ensure a fair comparison, we used propensity matching based on key covariates such as publication, model-predicted topic, and date. Once matched, we calculated the number of sensational articles per day to observe how these counts varied in the days before and after the event. To help visualize these ndings, we used daily-level stacked bar plots to track changes in sensational content across both topics and political alignment. We focused our analysis on a two-week time window?seven days before and after the chosen event?in order to maximize context while minimizing the in uence of overlapping news events. This period allows us to clearly see whether a spike in sensational content is tied to the event itself or to broader trends. The visualizations are designed to support interpretation of the DiD results by showing patterns in sensationalism over time and across topics and polticiality.

6 Results

6.0.1 RQ 0:

Our graph [gure 3] displayed the volume of articles per month, per topic. To examine the correlation between article volumes in any topic groups A and B, we created a correlation matrix to visualize their relationship [gure 4].

Figure 3: Article Volume Per Month Per Topic

For each topic (U.S. Government/Military, World News, Economy/Business, Health/Lifestyle/Personal Finance, Science/Technology, Entertainment/Celebrity News, Sports, Human-Interest Story/Society, Crime/Law and Order, Other), the average number of articles remained relatively consistent across months. Occasional spikes in article volume were observed within certain categories. A notable example is the `Sports' category, which experienced a clear numerical increase in August 2016 (average article count = 6,666), compared to both July (4,578) and September (4,103). Although these differences were not statistically signi cant ($p = 0.14$ for August vs. July, and $p = 0.88$ for August vs. September), the volumetric increase is notable.

Over the two-year period, Economy/Business, U.S. Government/Military, and Entertainment/Celebrity News consistently received the highest average number of articles per month. Economy/Business reached particularly high article counts between months 15 to 22 (March to October 2017). Science/Technology, Human-Interest Story/Society, and Crime/Law and Order occupied the mid-range in article volume, and showed moderate and steady coverage. Health/Lifestyle/Personal Finance, Sports, Other, and World News consistently received the lowest average article counts per month.

The correlation matrix displays how article volumes across topics corellated over time. Strong positive correlations were shown between multiple pairs of topics. Notably, Entertainment/Celebrity News displayed strong correlations with Science/Technology (r = 0.87), Human-Interest Story/Society (r = 0.88), and World News (r = 0.79), which suggests that article volumes for these topics often increased or decreased together. This pattern may indicate that events related to these categories are more likely to occur together or be covered concurrently in the media.

U.S. Government/Military was also highly correlated with Entertainment/Celebrity News (r = 0.81) and World

News (r = 0.70). Contrastingly, Sports and Other topics showed low correlations with most categories (e.g., Sports and U.S. Government/Military, r = 0.21; Other and

7

Figure 4: Correlation Matrix for Article Volume accross Topics

Economy/Business, r = 0.08), which indicates that article volume in these topics uctuates independently of others.

6.0.2 RQ 1:

Collectively, the visualizations from RQ-1 provide evidence that sensationalism in news SG varies substantially by topic. Figure 5a (Average SG per Topic per Month) shows the average sensationality gap (SG) per topic across time, which reveals that certain topics, like Entertainment/Celebrity News, Human-Interest Story/Society, and Other, consistently contribute more to the overall SG. This indicates that headlines in these categories tend to be more exaggerated compared to their article content. In contrast, topics like Economy/Business, U.S. Government/Military, and Sports show lower or even negative SG values, which suggests more balanced or understated headline tones.

Figure 5b (Overall Average SG over Time) illustrates the overall average SG over time. While it aggregates across topics, uctuations in sensationalism levels likely

Text-Only Version

re ect changes in abundance of articles of a certain topic. For example, periods with lower SG may coincide with the dominance of less sensational topics such as politics or economics. This implies that topic distribution affects overall sensationalism levels.

Figure 5c (Overall Average SG per Topic) illustrates the average SG per topic across the full dataset timeline (two years). Here, the same sensational topics identi ed in Graph 1 stand out with the highest average SG values, while traditionally ?hard news? topics remain consistently low. This con rms that sensationalism is not evenly distributed but instead varies systematically depending on the news category.

(a) Average Sensationality Gap per Topic

(b) Average Sensationality Gap Over Time

(c) Average Sensationality Gap per Topic by Month

Figure 5: Sensationality Gap

6.0.3 RQ 2:

The following graphs display the observed time series data for our articles based on the ve scenarios, along with forecast predictions for the corresponding metrics for the next 10 weeks.

In general, varying speci c (p,d,q) values indicate how the time series is structured. For example, (n,0,0), or n-order autoregressive model, means that the data could be forecasted as a multiple of its n previous values. (0,1,0) indicates a ?random-walk? model. (0,1,1) indicates simple exponential smoothing. [14]

6.0.4 RQ 3:

The graphs that correspond to the six events can be found in the appendix.

7 Discussion

7.1 Summary of Findings

7.1.1 RQ0: Comparing distributions

The observed spikes in article volume likely re ect major events that temporarily increased media attention. For example, in the case of the Sports category, the August 2016 spike aligns with the Rio de Janeiro Olympic Games (dated August 5?21, 2016), which suggests that the event may have resulted in increased media interest, despite a lack of statistically signi cant change. This illustrates how discrete, localized events can produce short-term shifts in article volume in otherwise steady topical trends.

Across this two-year span, Economy/Business, U.S. Government/Military, and Entertainment/Celebrity News consistently received the highest average number of articles per month. Economy/Business consistently publishes th e largest amount of articles, particularly between months 15 to 22 (March to October 2017), which suggests a period of heightened economic focus or change, possibly in relation to policy shifts, corporate news cycles, or political transitions. U.S. Government/Military coverage also remains popular and steady throughout the timeline. This can re ect continuous political developments and international relations during this period. Periods of increased coverage likely re ect large changes in the political sphere, such as the 2016 Presidential Election Results or the inaguration of Donald Trump.

The mid-level categories include Science/Technology, Human-Interest Story/Society, and Crime/Law and Order and show consistency in article volume, with relatively moderate counts. This steadiness may infer a baseline level of audience engagement or media commitment/obligation to reporting these types of events. The lower-volume topics such as Health/Lifestyle/Personal

(a) Weekly Count of Sensational Headlines, By Topic

(b) Weekly Count of Sensational Content, By Topic

(c) Weekly Count of Sensational Headlines, By Politicalness

(d) Weekly Count of Sensational Content, By Politicalness

(e) Weekly Count of Articles, By Comparison Between Headline and Content

Figure 6: ARIMA Forecasting

Text-Only Version

Finance, Sports, Other, and World News contribute less substantially to overall monthly article counts, which can imply a lower level of high-impact or notable events, and a lower level of audience engagement.

7.1.2 RQ1: Sensationality Gap

The results of this study demonstrate that sensationalism, as measured by the sensationality gap (SG) between headlines and article bodies, varies signi cantly across news topics. This gap is an indicator that headlines are usually exaggerated. A larger gap implies a stronger disconnect between how a story is presented and what actually entails, which is a sign of clickbait.

Across all time periods analyzed, topics like Entertainment/Celebrity News, Human-Interest Stories, and Society consistently show the highest average SG. These genres often rely on emotional appeal or narrative hooks to attract attention, which likely incentivizes more sensational headlines. On the other hand, topics like Economy/Business, U.S. Government/Military, and Sports tend to exhibit much smaller or even negative SGs, suggesting that their headlines more closely match article content and tone. This may be due to journalistic norms emphasizing objectivity and clarity in reporting hard news.

These ndings matter because headlines serve as the rst point of contact between audiences and information. When headlines consistently misrepresent content, they can distort public understanding and contribute to misin- formation

7.1.3 RQ2: Time & Seasonality

Overall, auto arima was able to generate varying sets of (p,d,q) parameters for different topics, levels of political- ness, and sensationalism comparison between headline and content, across the 5 different scenarios. However, the forecasted predictions did not display any signi cant trends or patterns, which could indicate that there isn't aren't a clear set of trends that we could use to determine metrics like how many sensational articles will appear weekly by topic or how many articles have more sensa- tional headlines than content. This could be due to several causes, such as the variability in our sampled dataset or in the manual and automated scores.

7.1.4 RQ2: Time & Seasonality

Overall, auto arima was able to generate varying sets of (p,d,q) parameters for different topics, levels of political- ness, and sensationalism comparison between headline and content, across the 5 different scenarios. However, the forecasted predictions did not display any signi cant trends or patterns, which could indicate that there isn't aren't a clear set of trends that we could use to determine metrics like how many sensational articles will appear weekly by topic or how many articles have more sensational headlines than content. This could be due to several causes, such as the variability in our sampled dataset or in the manual and automated scores.

7.1.5 RQ3: Events

We used our DiD model to assess two things: (1) Sensationality of article Headlines/Bodies by Politicality and (2) Sensationality of article Headlines/Bodies by Topic.

In general, we saw that sensationality decreased after a certain event occurred. While there were cases in which this did not occur (Olympics and Pulse Nightclub shooting), these results could re ect a shift in journalistic tone during or after major news events. In moments like these, journalists and media outlets may adopt more somber or fact-based reporting as a response to public scrutiny, risk to reputations, or the need to provide reliable information. This would be consistent with prior ndings suggesting that during crises or national tragedies, media coverage often becomes more restrained and aligned with public service objectives. However, we also observed exceptions to this trend. In the case of high-pro le events such as the Olympics and the Pulse Nightclub shooting, sensationality either increased or remained stable. These deviations could be explained by the nature of the events themselves.

For example, the Olympics, due to its topic of entertainment and sports, may easily lend itself to more sensational verbiage.

7.2 Implications

The ndings from this study can help researchers and media analysts understand how political events in uence the tone and sensationalism of news coverage, particularly when comparing political versus non-political reporting. By identifying spikes in sensational headlines tied to speci c topics or publications, future research can explore patterns of media overreaction or narrative ampli cation. This insight could support the development of tools or guidelines aimed at promoting more balanced reporting, ultimately contributing to a more informed and less panic-driven public discourse on the news.

7.3 Ethical Considerations

To maintain ethical integrity throughout our study, we implemented a two-reviewer blind review process. This approach helped reduce individual bias and encouraged consistency in labeling, particularly when identifying sensational traits in headlines and articles. By ensuring researchers had not previously read the headlines or article bodies, we aimed to prevent any prior knowledge from in uencing their scoring. Additionally, as mentioned before, no reviewer was assigned a headline and body pairing, preventing them from unintentionally giving a corresponding headline and body the same scoring.

In our project, we used an AI model to label our data, and while the accuracy was reasonable, it could have been improved with more time. We worked to remain as unbiased as possible given the sensitive nature of the data, but our personal belief systems may have in uenced how we classi ed content as ?political? or ?nonpolitical,? how we assigned topics, and how we rated the sensationality of headlines and article bodies. This potential bias was likely ampli ed by the use of machine learning models trained on our relatively small and potentially biased dataset. Fortunately, because our study was purely observational and did not involve human participants, we did not need to consider any psychological effects on subjects.

7.4 Limitations and Future Directions

One major limitation of our project was the computational and nancial constraints. Because we were working with full-length news articles - many of which exceeded typical to ken limits - we required models capable of handling large input sizes. However, many models that support longer token windows or produce more robust outputs come with usage fees or require premium access, which

Text-Only Version

we could not afford. As a result, we were restricted to free-tier models with limited memory capacity, which in uenced both the quality and scope of our labeling. Additionally, we lacked access to high-capacity GPUs, which prevented us from scaling our processing pipeline. Instead of labeling the full dataset of 1.2 million articles, we had to rely on a random subsample of 62,000, which may not fully re ect the distribution of topics or sensationalism patterns in the larger corpus. This necessary trade-off likely introduced sampling bias and reduced the representativeness of our nal model.

Another limitation was the small size of the labeled training dataset due to limited available manpower for manual annotation. We were only able to label 100 articles, which constrained the depth and variability of our training set. With an 80:20 split between training and evaluation, the amount of usable data for model learning was further reduced. This likely hindered the model's ability to capture nuanced patterns in sensationalism and limited its generalizability across a broader range of headlines and article types.

A third limitation lies in the inherent subjectivity of our labeling tasks, including sensationalism, topic classi cation, and political assignment. While we implemented a blind review process to reduce individual bias, traits

Text-Only Version

like hyperbole, forward referencing, and even political- ness are fundamentally interpretive. What our research group deems sensational or political may not align with another's judgment, especially in politically or culturally sensitive contexts. Similarly, assigning topics to articles can involve subtle distinctions and assumptions that re- ect personal perspectives or prior knowledge. These subjective judgments inevitably introduced variance into our training data, which may have been further ampli- ed when used to train machine learning models. As a result, our model's outputs could re ect and reinforce these initial biases rather than providing a purely objective classi cation.

In future iterations of this project, we would like to extend our analysis to the full dataset of 1.2 million ar- ticles - without having to truncate article bodies - which would offer a more comprehensive and representative un- derstanding of sensationalism trends across topics and time. With greater computational resources and more re ned models, we could better assess the sensationality of text. Additionally, expanding the labeled training data - both in size and diversity - would enhance model accuracy and reduce subjectivity, enabling us to capture a wider range of linguistic and contextual nuance.

Another promising direction is to explore how media framing strategies shift across different sociopolitical contexts. By distinguishing how media outlets adapt their tone, content, and stylistic choices based on the topic, event, or audience, this research could con tribute to a broader understanding of media bias, clickbait strategies, and the evolving role of sensationalism in shaping public discourse. These insights may also inform tools that promote media literacy and transparency in digital news consumption. Our current research questions off er a valuable starting point for this future exploration. By investigating the distribution of article topics over time (RQ0), we have begun to map the landscape of media coverage across a large corpus. By connecting these ndings to measures such as the Sensationality Gap (RQ1), its temporal dynamics (RQ2), and the in uence of speci c events (RQ3), we can begin to trace how media tone shifts in response to real-world developments - offering a more dynamic and context-aware model of sensationalism. This line of research holds the potential to not only deepen academic understanding but also support public tools that help readers critically navigate the modern information ecosystem.

8 Contribution Statement

8.1 Deeya:

I worked on my share of labelling the data and performed

the initial setup of the dataset and initial cleaning to lter
it into one we used to run our model through. During
RP-2 I also worked on writing part of the introduction

and preliminary analysis, and all of the study design, and putting our entire project into Latex format to turn in. I worked mainly on helping Oju troubleshoot with Distil-BURT as well as spending time on training the models used for Set t Topic and Piolitically labelling tasks. For the nal report I made the RQ 3 notebook and wrote parts of the Data section, ethical concerns, methods for RQ-3, and wrote the overall outline for the pap er as well. I also suggested and implemented a series of managerial tasks throughout the course of the project to keep the group on track. These tasks include regular check-ins, scheduling group meetings, and deadline-setting (while they were a small part of my contributions I believe they helped us successfully complete this project!)

8.2 Grace:

I worked mainly on RQ-0, where I created the data visu-alization for article volume across topics and calculated the correlation between the AV for each topic. I also con-tributed to RQ-1, where I helped determine the average Sensationality Gap for each topic. I also created most of the visualizations throughout the project and led the manual labeling and manual checking tasks. In this paper, I wrote all the sections for RP0, contributed to the introduction, and input everything into Overleaf along with Alan.

8.3 Sharika:

I worked to sort our preprocessed data to create a set of 100 articles to manually label. I also extracted information from the nal manually labeled spreadsheet les to sort out discrepancies in article headline/body topic and politicality mismatches. I also worked on RQ-1 to prepare visualizations of SG over topics and time for both manually labeled and SetFit-labeled data. In this paper, I worked on the introduction, related works, and methods/results sections corresponding to RQ-1.

8.4 Alan:

I worked on my share of manually labelling the articles. Also, I did model exploration for labelling articles, and I conducted the few shot classi cation analysis with SetFit for topic, politicalness, and sensationalism. I prepared the dataset of 62k articles with their respective labelling, and this dataset was used to answer RQ 1 through 3. I also created the topic classi cations for the 1.2 million articles, which was used to answer RQ 0. Also, I worked on RQ 2 by performing ARIMA forecasting after the

Text-Only Version

labels for all articles were complete. In this paper, I contributed to the abstract, and worked on the methods, results, and discussion sections, in particular the parts relating to SetFit and ARIMA. Also, I helped convert our writing to the proper CHI format with Grace.

8.5 Oju:

I worked manually labelling my section of the articles, model exploration for sensationality labelling, writing for methods, limitations, and future direction sections. Also, I worked on creating the graphs for RQ-3 that measured the different metrics before and after the six events that we selected.

References

[1]

Bodas, Deeya. (2025). Data Processing.

https://colab.research.google.com/drive/1YvOs8F6dm7WHXVqH

h

iqur f M yK w

4

mqP

?

aut huser

=

2

Text-Only Version

scrol l T o

=

I C nQ

n

H H f xt r

:

[2]

Khawar, S., Boukes, M. (2024). Analyzing Sensation-

alism in News on Twitter (X): Clickbait Journalism

by Legacy vs. Online-Native Outlets and the Conse-

quences for User Engagement. Digital Journalism, 1?21.

https://doi.org/10.1080/21670811.2024.2394764.

[3]

Kottapalli, Sharika. (2025) Manual Labelling

Setup. https://colab.research.google.com/drive/1ZiDc-

SE7PANcteIuby8oQKbbzlHjZ-39?usp=sharing.

[4]

Bodas, Deeya., Chaudhary, Oju., Kotta-

palli, Sharika., Sherman, Grace., Tao,

Alan. (2025). Sensationality Ratings Example.

https://docs.google.com/spreadsheets/d/195CfwoQdi8lRYiHbsn-

tbnOpgICpxlZZ5CdycwKMH7w/edit?usp=sharing.

[5]

Hugging Face. (n.d.). SetFit Documen-

tation. Retrieved May 6, 2025, from

Text-Only Version

https://huggingface.co/docs/set t/en/index.

[6]

Tao, Alan., Bodas, Deeya.

(2025). SetFit Labelling (Topics)

https://drive.google.com/ le/d/13p7W

j

QS J yI t G

p

4

S yab

K

E ml N T

G

hr

=

view

?

us p

=

shar ing

:

[7]

Tao, Alan., Bodas, Deeya., Sherman,

Grace. (2025). SetFit Labelling (Political-

ness) https://drive.google.com/ le/d/1Rgt4mB2-

a1D8wXUb7bL6-8ifTWq51JTY/view?usp=sharing.

[8]

Chaudhary, Oju. (2025) Ollama La-

belling. https://drive.google.com/ le/d/1-

YfFiUvUaX07kASYnMzwiqIyUGuoSNX9/view?usp=sharing.

[9]

Chaudhary, Oju. (2025) DistilBERT Labelling.

https://colab.research.google.com/drive/1qM48eUlXSQGc9fqiTvjysI2v0HGGPwfj?usp=sharing.

[10]

Tao, Alan., Bodas, Deeya. (2025). SetFit Labelling.

https://colab.research.google.com/drive/1XWM32vQYzx8A8h7h-

KSgz67v51pP

8

4

O

?

us p

=

shar ing

:

[11]

Tao, Alan (2025). SetFit La-

belling (Sensationalism of Headlines)

Text-Only Version

https://drive.google.com/ le/d/1YSQQHE7HeICauLHTdClT4C45rw3oqnY6/view?usp=sharing.

[12]

Tao, Alan (2025). SetFit La-

belling (Sensationalism of Content)

https://drive.google.com/ le/d/1G3kBZ5oSrLXTYnlDIGx

r

U ngmgBM nz j s

=

view

?

us p

=

shar ing

:

[13]

Duke University Fuqua School of Business. (n.d). Intro-

duction to ARIMA: non-seasonal models. Retrieved May

6, 2025, from https://people.duke.edu/ rnau/411arim.htm.

[14]

Tao, Alan (2025). ARIMA Modelling

https://colab.research.google.com/drive/1dodKu2250-

hr1chzKPq48TBI6vaCfoZr?usp=sharing.

[15]

Bodas, Deeya. (2025). RQ-3

https://colab.research.google.com/drive/19mXv-

Text-Only Version

3fJL29LJe-eu5d5RkbrEipcl4b

?

us p

=

shar ing

:

9 Appendix

(a) Headlines by Politicality

(b) Bodies by Politicality

(c) Headlines by Topic

(d) Bodies by Topic

Figure 7: Articles regarding the Pulse Nightclub Shooting

(a) Headlines by Politicality

(b) Bodies by Politicality

(c) Headlines by Topic

(d) Bodies by Topic

Figure 8: Articles regarding the Trump Inauguration

(a) Headlines by Politicality

(b) Bodies by Politicality

(c) Headlines by Topic

(d) Bodies by Topic

Figure 9: Articles regarding Harambe

(a) Headlines by Politicality

(b) Bodies by Politicality

(c) Headlines by Topic

(d) Bodies by Topic

Figure 10: Articles regarding The Olympics

Figure 11: Enter Caption

(a) Headlines by Politicality

(b) Bodies by Politicality

(c) Headlines by Topic

(d) Bodies by Topic

Figure 12: Articles regarding The Royal Engagement