A Murder and Protests, the Capitol Riot, and the Chauvin Trial: Estimating Disparate News Media Stance

Sujan Dutta[1] , **Beibei Li**[2] , **Daniel S. Nagin**[2*] and **Ashiqur R. KhudaBukhsh**[1*]
[1]Rochester Institute of Technology
[2]Carnegie Mellon University
sd2516@rit.edu, beibeili@andrew.cmu.edu, dn03@andrew.cmu.edu, axkvse@rit.edu

Abstract

In this paper, we analyze the responses of three major US cable news networks to three seminal policing events in the US spanning a thirteen month period–the murder of George Floyd by police officer Derek Chauvin, the Capitol riot, Chauvin's conviction, and his sentencing. We cast the problem of aggregate stance mining as a natural language inference task and construct an active learning pipeline for robust textual entailment prediction. Via a substantial corpus of 34,710 news transcripts, our analyses reveal that the partisan divide in viewership of these three outlets reflects on the network's news coverage of these momentous events. In addition, we release a sentence-level, domain-specific text entailment data set on policing consisting of 2,276 annotated instances.

1 Introduction

The thirteen month period spanning the murder of George Floyd by police officer Derek Chauvin on May 25, 2020, the Capitol Riot on January 6, 2021, Chauvin's conviction for the Floyd murder on April 21, 2021, and his sentencing on June 25, 2021 were momentous events for policing in the United States in part due to the sustained media attention given to police practice, conduct and function. These four events, however, cast the police in diametrically different roles. The Floyd murder and Chauvin's conviction and sentencing focused attention on police use of excessive force, especially against Black Americans, whereas the Capitol riot cast police in the role of valiant warriors in the defense of the nation's political foundation.

In this paper, using advanced natural language processing (NLP) methods, we analyze the different responses of three major media outlets, Fox News, CNN and MSNBC to these seminal events. The differences that we document are consistent with the partisan divide in the viewership of the three outlets. According to a Pew sponsored survey[1], 93% of Republicans or those leaning Republican name Fox News as one of their preferred sources of political news whereas only 6% of Democrats or those leaning Democrat name Fox as a preferred source. For CNN and especially MSNBC, the percentages flip, CNN and MSNBC are preferred sources, respectively, for only 4% and 1% for Republicans/lean Republicans and for Democrats/lean Democrats they are, respectively, 79% and 95% preferred sources.

While the key focus in this paper is to address the specific task of estimating aggregate media stance on policing, our work has valuable potential for advancing AI for social good. In particular, we demonstrate how advanced NLP and machine learning technologies can help policy makers, content producers, and general public better understand the political disparity and partisan divide in the content creation of major media outlets. The mode of analysis that we demonstrate could just as well be applied to other contentious issues such as abortion, immigration, and gun control that divide contemporary America. Casting the task of estimating stance as a natural language inference (NLI) task [Bowman *et al.*, 2015], our study proposes a robust, scalable, and measurable indicator - media entailment ratio - for measuring the disparate stance of different major media outlets toward seminal societal events. It also provides a sustainable lens through which we can better understand, access, and evaluate the news content and public opinion. While focusing on three high-profile policing and justice events from May 2020 to June 2021, the proposed methods and analyses can also be generalized to other contexts for social good.

To summarize, our contributions are the following:

- ***Social***: We analyze the different responses of three major news outlets to momentous events for policing using sophisticated NLP methods. To our knowledge, our analysis is the first to show that across cable networks coverage of politically salient events responds quickly and dramatically to the partisan preferences of their viewership.
- ***Method***: Casting the problem of aggregate stance mining as a natural language inference task [Bowman *et al.*, 2015], we present an active learning pipeline drawing from existing sampling strategies [Sindhwani *et al.*, 2009; Scheffer *et al.*, 2001; Palakodety *et al.*, 2020] and a novel strategy, dubbed inconsistency sampling.
- ***Resource***: We release a sentence-level, domain-specific text entailment data set on policing consisting of 2,276 anno-

[*]Daniel S. Nagin and Ashiqur R. KhudaBukhsh are joint corresponding authors.

[1]https://www.pewresearch.org/fact-tank/2020/04/01/americans-main-sources-for-political-news-vary-by-party-and-age/

tated instances[2].

2 Data Set

We consider news transcripts from three major US cable news networks: Fox News, CNN, and MSNBC. We scrape all transcripts available on their official web sites uploaded between May 25 2019 and July 24 2021. We divide this time period of 26 months into four non-overlapping temporal slices:

1. *Pre-George Floyd* (denoted by $\mathcal{T}_{\textit{pre-Floyd}}$): baseline time period starting on May 25 2019 and ending on May 24 2020.
2. *George Floyd's murder, protest, and civil unrest* (denoted by $\mathcal{T}_{\textit{Floyd}}$): starts on May 25 2020, the day George Floyd was brutally murdered, and ends on Dec 31 2020.
3. *Capitol insurrection* (denoted by $\mathcal{T}_{\textit{capitol}}$): starts on January 1 2021 and ends on March 7 2021, a day before the trial of Derek Chauvin begun.
4. *Chauvin's conviction and sentencing* (denoted by $\mathcal{T}_{\textit{trial}}$): starts on March 8 2021, the day Chauvin's trial begun and ends on July 24 2021, one month after Chauvin's conviction.

News network	#sentences
CNN	94, 911
FOX News	12, 026
MSNBC	12, 642

Table 1: Data set details. We consider transcripts of three major US cable news networks between May 25 2019 to July 24 2021. All sentences in our data set contains the token `police`.

Preprocessing. Next, we preprocess these transcripts and remove all text corresponding to video clips indicated by "(BEGIN VIDEO CLIP)" and "(END VIDEO CLIP)" marks and any non-utterance, such as time information (e.g., [12:15:05]) and special marks (e.g., "(COMMERCIAL BREAK)"). Overall, we obtain 34,710 transcripts (CNN: 28,163, Fox news: 3,617, and MSNBC: 2,930 transcripts). We also remove the speaker information (e.g., HANNITY, COOPER or MADDOWS) preceding any sentence since their presence may bias machine learning models' predictions triggering shortcut learning [Geirhos *et al.*, 2020].

We observe that compared to MSNBC and Fox, the news transcript archiving of CNN is substantially more detailed and comprehensive. The temporal distribution of the news transcripts is summarized in Figure 1. As shown in Figure 1, the relative fraction of uploaded transcripts (with respect to an individual news network) remained comparable throughout all four time slices.

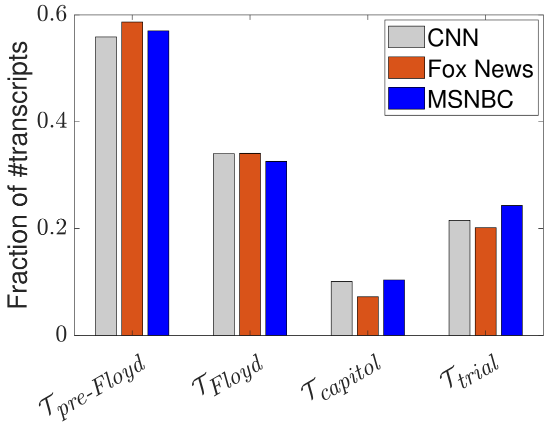

Figure 1: Relative proportion of transcripts (with respect to individual news networks) uploaded during each temporal slice.

Next, we construct three corpora, one from each news network transcripts, consisting of all sentences that contain the token `police` (denoted by $\mathcal{D}_{fox}$, $\mathcal{D}_{cnn}$, and $\mathcal{D}_{msnbc}$). We argue that this is a high recall approach to filter content discussing policing in the US without biasing the data set by filtering by topic or event related keywords. To support this, we note that MSNBC and Fox News have roughly equal number of sentences obtained through this filter. However, the number of sentences that mention Floyd is strikingly different (MSNBC: 713 sentences, Fox news: 198 sentences).

2.1 Separability Tests

When we manually inspect these transcripts, we see systematic differences. Here are three examples from $\mathcal{D}_{msnbc}$, $\mathcal{D}_{fox}$, and $\mathcal{D}_{cnn}$, respectively:

1. *Unarmed black person committing no crime shot and killed by police.*
2. *Number one, it's the media activists class that took a traffic – tragic incident and cast all police as corrupt which caused the Ferguson effect.*
3. *We've all seen the image of that capitol police officer who led that mob away from the Senate chamber.*

However, human biases on contentious social issues is well-documented [Breitfeller *et al.*, 2019]. Unlike human beings, machines are less likely to have labeling biases. We ask ourselves *how do we quantify this difference (if any) using automated methods?* Quantifying differences between large-scale text data sets is an emerging field and existing methods often require a vast amount of data [KhudaBukhsh *et al.*, 2021]. We instead conduct a simple experiment using classification accuracy as a proxy for text separability. Our instances are individual sentences with the token `police` from different networks, and the labels are the news networks. Our classifier takes a sentence with the token `police` from news transcripts as input, and outputs the predicted source news network. Our intuition is if the linguistic signals present in two networks' reporting on policing are distinguishable, the classifiers will be able to predict the source news network with high accuracy. We randomly sample equal number of examples from any two news networks and construct binary classifiers to predict the news network label. Note that, if the reporting has little or no separable linguistic signal, the classifier will not perform much better than chance (50% accuracy on an evenly balanced test set).

It may well be the case that Fox News typically uses long

[2] Data and additional resources are available at the project page: https://github.com/Suji04/IJCAI2022-Media-Stance.

MSNBC	CNN	Fox News
And all the calls for significant and substantive police reform, which I support fully, will not heal what ails us because you have court systems and you have that structure for part of this larger criminal justice picture.	The murder trial of Derek Chauvin resume this morning following a day filled with emotional testimony and never-before-seen video of George Floyd's deadly encounter with police.	And as a result of her attacks on the police, we are seeing the murder rates skyrocket in St. Louis.
And today is another indication that we are still having lives lost senselessly at the hands of police officers.	Emotional and heartbreaking testimony on Tuesday from a number of witnesses in the trial of the police officer charged with killing George Floyd.	And meanwhile, you can loot and burn Police cars in New York, and youll probably just be just fine
With all that we're facing, I am so proud to bring the perspective of a black woman, a daughter of immigrants, the wife and mother of a husband and kids who sadly are more vulnerable to police violence because of their color. . .	Day eight now witness testimony in the trial of the Former Minneapolis Police Officer Derek Chauvin use-of-force and if Chauvin violated police practices and policy when he pinned George Floyd with his knee on his neck, a big focus of the testimony again this morning.	Aoc all-out crazy, who, by the way, has armed security officers with her at all times, like the rest of her Democrat socialist of America, her justice warriors who are sanctimonious hypocrites, would deny public safety to average people by defunding the police, by not wanting prisons, and by putting handcuffs on the police instead of the criminals.

Table 2: Illustrative examples highlighting disparate portrayals of policing shown in MSNBC (left column), CNN (middle column), and Fox (right column) news transcripts. These examples are high confidence predictions of automated methods trained to predict the network label given an input sentence with the token `police`.

sentences as opposed to short sentences used by MSNBC or CNN. To verify if any such quark present in the data set is influencing the classification accuracy through triggering shortcut learning [Geirhos *et al.*, 2020], as control, for each news network pair, we train another classifier on a training set of randomly selected sentences from news transcripts i.e. without the restriction that the sentences must contain the token `police`. The control allows us to compare how randomly selected sentences from two different news networks are distinguishable as opposed to sentences with the token `police`.

	CNN	*Fox News*	*MSNBC*
CNN	-	74.8 ± 0.4% 60.4 ± 2.0%	62.9 ± 1.2% 59.8 ± 0.9%
Fox News	74.8 ± 0.4% 60.4 ± 2.0%	-	73.5 ± 0.5% 62.2 ± 0.7%
MSNBC	62.9 ± 1.2% 59.8 ± 0.9%	73.5 ± 0.5% 62.2 ± 0.7%	-

Table 3: Separability results between different news networks. A cell $\langle i, j \rangle$ presents the accuracy on the binary classification task of predicting the news network given a sentence from a news transcript with label choices network i and j. Each cell, $\langle i, j \rangle$, summarizes the classification accuracy as a / b where where a (top) is the classification accuracy when models are trained and tested on sentences containing the token `police`; b (bottom) is the classification accuracy when models are trained and tested on any sentence from the news transcripts of network i and j. All models are trained on a data set of 11,000 sentences each from two relevant corpora (5,500 each) with an 80/20 train/test split. Each experiment is repeated five times with five different splits. We fine-tune `BERT` [Devlin *et al.*, 2019], a well-known pre-trained language model, for this classification task. Further details are presented in the project page.

Table 3 summarizes our separability results. We first compare the prediction accuracy on sentences with the token `police` in them. We notice that the classification tasks involving Fox news yielded substantially higher accuracy than the classification task involving the pair $\langle CNN, MSNBC \rangle$. Hence, CNN and MSNBC's portrayal of police could be linguistically more akin to each other than Fox news. We find that distinguishing between randomly chosen sentences is considerably more challenging than distinguishing between sentences related to police. The classification accuracy on randomly chosen sentences does not differ by much regardless of the choice of the network pair. Hence, the general nature of news coverage of Fox is less distinct than the network's coverage of police related events. In Table 2, we consider the classification tasks involving the network pairs $\langle CNN, Fox \rangle$ and $\langle Fox, MSNBC \rangle$ and list a few illustrative examples where the classifier correctly predicts the network label with high confidence.

Our classification results indicate that Fox's portrayal of policing is considerably different from CNN and MSNBC. In Section 5, we present critical insights into *how* these portrayals are different and provide methodological details and necessary technical background in Section 4.

3 Related Work

Beyond the rich social science research on differential policing and media portrayals [Chaney and Robertson, 2013; Lawrence, 2000; Brunson, 2007], previous attempts to quantify the differential treatment of police towards the white and black communities have employed computational linguistic methods on the transcripts of videos captured from police body cameras [Voigt *et al.*, 2017]. Voigt *et al.* [2017] established a linguistic framework to estimate the level of respect in the police officer's speech and found that officers speak with significantly less respect to black people than white. In contrast, our work focuses on differential media portrayals of police in prominent US cable news networks and how such portrayals reflect on news consumer responses.

Several recent lines of work have analyzed discussions around controversial issues on mainstream and social media and how political bias influences framing [Demszky *et al.*, 2019; Sap *et al.*, 2020; Roy and Goldwasser, 2020]. Demszky *et al.* [2019] provide an NLP framework to quantify and analyze political polarization in 4.4M tweets on 21 mass shooting events. A weakly supervised learning approach was proposed in Roy and Goldwasser [2020] to analyze the nuanced frames present in news articles on contentious topics (abortion, immigration, and gun control). In Mendelsohn *et al.* [2020], the authors analyzed the presence of dehumanizing language in discussions related to the LGBTQ community in the New York Times using computational linguistics methods. The authors applied supervised learning methods on a novel immigration-related tweet data set to detect the frames.

This study also analyzed how users' region and ideology correlate with the framing choices and how other users react to their posts. While these lines of research focus on a broad range of vital issues that include mass shooting [Demszky *et al.*, 2019], abortion [Roy and Goldwasser, 2020], immigration policy [Mendelsohn *et al.*, 2021] and dehumanization [Mendelsohn *et al.*, 2020], an extensive analysis of the media stances on the issue of police reform does not exist.

Unlike prominent supervised and unsupervised approaches to mine stances [Darwish *et al.*, 2020], in line with a recently proposed approach [Hossain, 2021], we treat the problem of stance mining as a natural language inference (NLI) task. NLI to mine stances has been deployed in another recent work which is similar to our work both in terms of method and research focus [Halterman *et al.*, 2021]. Halterman *et al.* [2021] explored the role of police in the Gujrat riot (2002, India) by analyzing $1,257$ news articles using BERT-based language models. Our work contrasts with Halterman *et al.* [2021] in three key ways. First, we analyze media portrayals of police in multiple seminal events over longitudinal corpora spanning thirteen months. Second, instead of focusing on a single news source, we consider multiple news outlets and conducts a contrastive analysis. Finally, we present a label-efficient active sampling method that combines multiple active learning strategies and a novel sampling technique that exploits logical consistency. The susceptibility of textual entailment systems to negation is well-documented in Kang *et al.* [2018]. Our work leverages this knowledge to set up an effective active learning pipeline detecting challenging premises that produce logically inconsistent entailment inferences.

4 Technical Background and Methods

4.1 Text Entailment

Our proposed approach casts the problem of mining stance as a natural language inference (NLI) task [Dagan *et al.*, 2005]. Given a premise $\mathcal{P}$ and a hypothesis $\mathcal{H}$, the NLI task involves predicting entailment, contradiction, or semantic irrelevance. Textual entailment is much more relaxed than pure logical entailment and can be viewed as a human reading $\mathcal{P}$ would infer most likely $\mathcal{H}$ is true. For example, the hypothesis *some men are playing a sport* is entailed by the premise *a soccer game with multiple males playing*[3].

Following Halterman *et al.* [2021], we cast the task of estimating the aggregate stance on policing as an NLI task. We consider two semantically equivalent hypotheses:

1. *Police protect us* (denoted by $\mathcal{H}_{protect}$)
2. *Police make us safe* (denoted by $\mathcal{H}_{safe}$)

and estimate the endorsement or support for police through *entailment ratio* described next.

Entailment ratio. Let $NLI(\mathcal{P},\mathcal{H})$ takes a premise $\mathcal{P}$ and a hypothesis $\mathcal{H}$ as inputs and outputs $o \in \{entailment, contradiction, neutral\}$. For a corpus $\mathcal{D}_i$ and a hypothesis $\mathcal{H}$, we compute the entailment ratio (denoted by $ent(\mathcal{D}_i, \mathcal{H})$) as the fraction of the individual sentences present in $\mathcal{D}_i$ that entails $\mathcal{H}$: $ent(\mathcal{D}_i, \mathcal{H}) = \frac{\Sigma_{\mathcal{P}\in\mathcal{D}_i}\mathbb{I}(NLI(\mathcal{P},\mathcal{H})=entailment)}{|\mathcal{D}_i|}$ where $\mathbb{I}$ is the indicator function. The larger the value of $ent(\mathcal{D}_i, \mathcal{H})$ is the greater is the support for $\mathcal{H}$ in the corpus.

While state-of-the-art, off-the-shelf NLI systems can perform well on a broad range of examples, relying on NLP systems to draw conclusions on a subject of high social impact needs careful consideration [Hovy and Spruit, 2016; Olteanu *et al.*, 2019]. In our work, we are interested in estimating the support for police in each of the corpora, a domain considerably different from the data on which our NLI system is trained on [Bowman *et al.*, 2015]. We thus finetune the existing NLI system on domain-specific examples. In order to make the learning process label efficient, we use active learning [Settles, 2009], a well-known supervised machine learning technique described next.

4.2 Active Learning to Strengthen the NLI Model

Active Learning is a powerful and well-established form of supervised machine learning technique [Settles, 2009]. It is characterized by the interaction between the learner, aka the classifier, and the teacher (oracle or labeler or annotator) during the learning process. *Pool-based active learning* is a popular variant; in this setting, the learner is initially trained on a small seed set of labeled examples and has access to a large collection of unlabeled samples. At each iteration, the learner employs a sampling strategy to select an unlabeled sample and requests the supervisor to label it (in agreement with the target concept). The dataset is augmented with the newly acquired label, and the classifier is retrained on the augmented dataset. The sequential label-requesting and re-training process continues until some halting condition is reached (e.g., annotation budget is expended or the classifier has reached some target performance). At this point, the algorithm outputs a classifier, and the objective for this classifier is to closely approximate the (unknown) target concept in the future. The key goal of active learning is to reach a strong performance at the cost of fewer labels through an active participation of the learner. Since training and inference on a very large pool of samples can be computationally prohibitive, lines of work have examined the trade-offs of batch active learning [Yang and Carbonell, 2013]. In this paper, we follow the pool-based batch active learning pipeline where instead of requesting one label at a time, the learner requests labels for a batch of unlabeled samples.

Our active learning pipeline draws inspiration from Palakodety *et al.* [2020] and uses multiple sampling strategies to construct an evenly balanced train set with diverse and challenging examples. Since our paper's main thrust is analyzing a research question of high social impact, in what follows, we present a succinct description and rationale behind our sampling strategies.

Random sampling. In order to capture a diverse set of examples, we include 900 randomly sampled instances giving equal weightage to each of the four temporal slices and three news networks. These 900 instances form the seed set of our active learning pipeline. In this phase, we obtain 160 entailments, 202 contradictions and 538 neutrals confirming that neutral is the majority class.

[3] This example is taken from Bowman *et al.* [2015].

Certainty sampling. Minority class certainty sampling is useful in rectifying high-confidence misclassifications involving short documents [Sindhwani *et al.*, 2009]. We conduct minority class certainty sampling and added 200 instances that the model predicts entailment or contradiction with highest confidence.

Inconsistency sampling. This novel task-specific sampling strategy is guided by the intuition that instances where the model's predictions are logically inconsistent could be potentially challenging. We first note that $\mathcal{H}_{safe}$ is a semantic variation of $\mathcal{H}_{protect}$ and $\forall p, NLI(p, \mathcal{H}_{safe}) = NLI(p, \mathcal{H}_{protect})$. We further note that the same premise cannot both entail or contradict the hypothesis and its negation. Let $\mathcal{H}'$ denote the negation of a hypothesis. $\forall p \; such \; that \; NLI(p, \mathcal{H}) \in \{entailment, contradiction\}, NLI(p, \mathcal{H}) \neq NLI(p, \mathcal{H}')$. We leverage these two observations to construct a novel active learning sampling strategy dubbed *inconsistency sampling*. We include 336 instances on which the model exhibits these two types of logical inconsistencies. To ensure that the data is balanced across different time slices and networks, while conducting sampling, we equally weight each network (CNN, Fox news, and MSNBC), each hypothesis ($\mathcal{H}_{protect}$ and $\mathcal{H}_{safe}$), and each of the four temporal slices ($\mathcal{T}_{pre\text{-}Floyd}$, $\mathcal{T}_{Floyd}$, $\mathcal{T}_{capitol}$, and $\mathcal{T}_{trial}$), each of the two types of logical inconsistencies and sampled from every combination of a network, hypothesis, logical inconsistency, and temporal slice.

Uncertainty sampling. Uncertainty sampling is one of the most well-known sampling strategies used in active learning [Settles, 2009]. Since we have multiple label categories in our prediction task, we follow the uncertainty sampling variant, margin sampling, proposed by Scheffer *et al.* [2001].

To summarize, our active learning pipeline consists of the following steps:

1. Construct an initial seed set of 900 instances ($\mathcal{D}_{seed}$: 158 entailment, 204 contradiction and 538 neutral instances) using random sampling.
2. Conduct minority class certainty sampling and add 200 samples ($\mathcal{D}_{certainty}$: 91 entailment, 86 contradiction and 23 neutral instances).
3. Conduct inconsistency sampling and add 336 samples ($\mathcal{D}_{inconsistency}$: 86 entailment, 123 contradiction and 127 neutral instances).
4. Finally, conduct uncertainty sampling (margin sampling) and add 540 samples ($\mathcal{D}_{uncertainty}$: 189 entailment, 199 contradiction and 152 neutral instances).

Overall, we obtain 1,976 instances (526 entailment, 610 contradiction, and 840 neutral instances). All annotations were conducted by two annotators in two phases with the two annotators independently annotating in the first phase followed by an adjudication step to resolve disagreements in the second phase. Before the adjudication process, Cohen's κ value was 0.82 indicating high inter-rater agreement.

We evaluate the performance of our trained models on an unseen evaluation set of randomly sampled 300 instances ($\mathcal{D}_{test}$: 69 entailment, 67 contradiction and 164 neutral instances) giving equal weightage to each of the four temporal slices and three news networks. We consider the base entailment model ($\mathcal{M}_{base}$), an NLI model published by Allen NLP[4], as our baseline. We in fact fine-tune this model using our data obtained through our active learning pipeline (the project page contains further details). As show in Table 4, (1) our model performs reliably on unseen data; and (2) our method considerably outperforms the baseline underscoring the need for training on domain-specific examples.

Data	Model	Macro F_1
-	$\mathcal{M}_{base}$	44.68
$\mathcal{D}_{seed}$	$\mathcal{M}_{seed}$	69.79 ± 1.64
$\mathcal{D}_{seed} \cup \mathcal{D}_{certainty}$	$\mathcal{M}_{certainty}$	72.39 ± 0.92
$\mathcal{D}_{seed} \cup \mathcal{D}_{certainty} \cup \mathcal{D}_{inconsistency}$	$\mathcal{M}_{inconsistency}$	74.15 ± 1.11
$\mathcal{D}_{seed} \cup \mathcal{D}_{certainty} \cup \mathcal{D}_{inconsistency} \cup \mathcal{D}_{uncertainty}$	$\mathcal{M}_{uncertainty}$	75.42 ± 0.65

Table 4: Performance comparison of baselines on our test set consisting 300 randomly sampled premise sentences from all three networks (100 from each network). $\mathcal{M}_{base}$ denotes a baseline NLI model published by Allen NLP. Subsequent models are fine-tuned on top of this. For all other models trained in this paper, performance is reported over five different training runs.

5 Results and Discussions

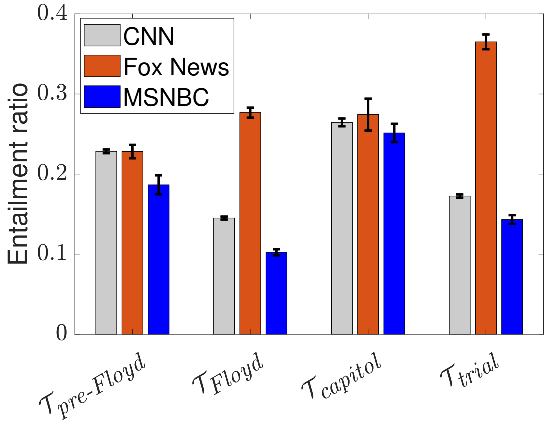

Figure 2: Temporal trend of news network support for $\mathcal{H}_{protect}$ in terms of entailment ratio. For a corpus $\mathcal{D}_i$ and a hypothesis $\mathcal{H}$, we computed the entailment ratio as the the overall fraction of the individual sentences present in $\mathcal{D}_i$ that entails $\mathcal{H}$.

Figure 2 summarizes the findings of our entailment analysis of the hypothesis $\mathcal{H}_{protect}$. Results for the hypothesis $\mathcal{H}_{safe}$ are very similar. Each bar measures the proportion of sentences classified as endorsing the hypothesis. In the year prior to George Floyd's murder the entailment ratios for the three networks were very similar at about 0.2. Thereafter, they dramatically diverged. For period from Floyd's murder to the end of 2020 which included peaceful protests against police brutality but also riots in many cities the entailment ratios of CNN and MSNBC declined by 36.5% and 45.2%, respectively, whereas that for Fox it increased by 21.3%. The large reductions in agreement to the hypothesis "police protect us" for the Democrat favored outlets CNN and MSNBC and the large increase in support for the premise for the Republican

[4] https://github.com/allenai/allennlp-models/blob/main/allennlp_models/modelcards/pair-classification-roberta-snli.json

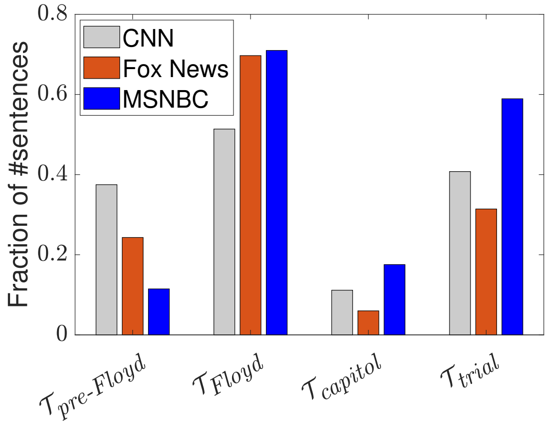

Figure 3: Relative proportion of sentences (with respect to individual news networks) containing the token `police` across four time periods of interest.

favor outlet Fox is consistent with news commentaries on partisan difference in reaction to the Floyd murder and ensuing events during the period. For example, a June 1, 2020 analysis appearing in the Washington Post[5] observed: “There are two things about the protests against police brutality ripping through the nation that Democratic and Republican lawmakers generally agree on:

1. George Floyd should not have died the way he did, after a white officer knelt on his neck.
2. Looting, vandalism and violence that have dominated some protests after dark are bad.

But on social media, in statements and in interviews, there’s a clear partisan difference between which of these two points members of Congress choose to emphasize, with Republicans choosing to focus on the second while Democrats the first.”

The period January 1, 2021 to March 7, 2021 spans the Capitol Riot, in which police were both the principal victims and heroes, up to the conviction of Derek Chauvin. During this period the entailment ratio for Fox is unchanged whereas those for CNN and especially MSNBC increase dramatically by 82.4% and 145.9% respectively. The change differences for $\mathcal{T}_{capitol}$ are consistent with the partisan differences which persist to this day in the interpretation of the events of January 6 with the Republican National Committee now characterizing the events of January 6 as an exercise of “legitimate political discourse.” The fourth period from March 8, 2021 to July 24, 2021 covers the conviction and sentencing of Derek Chauvin. This period was associated with a 33.1% increase in the entailment ratio for Fox and 34.8% and 43.1% decreases for CNN and MSNBC, respectively, again not surprising given the political leanings of each network’s viewers.

Figure 3 reports for each network by period the fraction of that network’s total sentences with the police token. The results reinforce the conclusions of the entailment analysis. In the pre-Floyd period the police token appears proportionally far more often on Fox than CNN and MSNBC. In the second period the proportion of total mentions by network increases dramatically for both MSNBC and Fox even as their entailment ratios move in oppose directions. During $\mathcal{T}_{capitol}$ mentions of the police token drop substantially for all three networks (this is also the shortest time slice in our analysis) but on CNN and MSNBC the entailment moves up from their low levels during $\mathcal{T}_{Floyd}$. Finally, like with period $\mathcal{T}_{Floyd}$, mentions of the police token rise substantially for all three networks during the Chauvin trial period even as the entailment ratio declines for MSNBC and increases for Fox.

Much has been written about the partisan divide in cable news coverage. To our knowledge our analysis is the first to show that across cable networks coverage of politically salient events responds quickly and dramatically to the partisan preferences of their viewership. This finding, we believe, is disturbing because it speaks to the siloed nature of political coverage in contemporary America. A necessary condition for reaching compromise in a democratic society is having an understanding of how individuals with opposing views perceive the world even as one disagrees that world view. That network reporting to their already partisan siloed viewership responds so quickly to comport with their viewership’s world view is not an encouraging sign that partisan divides can be bridged anytime soon.

Ethical Statement

In this paper, we consider public-domain data available on the official web sites of three major US cable news networks. We thus do not foresee major ethical concern over choice of data source. In our work, we fine-tune large-scale pre-trained language models (`BERT` [Devlin *et al.*, 2019] and `RoBERTa` [Liu *et al.*, 2019]) for training purpose. While our entailment results are obtained after fine-tuning the large language model, recent works have indicated that these models have a wide range of biases that reflect the texts on which they were originally trained, and which may percolate to downstream tasks [Bender *et al.*, 2021]. Finally, in this paper, we present a new approach to analyze news media content that can help us identify differences in media portrayals of major policing event in the US. While we do not discuss how to react to these discoveries, our analyses can inform the society better about these differences using our quantifiable approach.

Acknowledgments

The authors like to thank Sharon Lynch for her valuable input to the paper.

References

[Bender *et al.*, 2021] Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. On the Dangers of Stochastic Parrots: Can Language Models Be Too Big? In *ACM FaccT*, pages 610–623, 2021.

[Bowman *et al.*, 2015] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In *EMNLP*, 2015.

[5] https://www.washingtonpost.com/politics/2020/06/01/difference-between-democratic-republican-reactions-protests-elevate-george-floyd-or-antifa/

[Breitfeller *et al.*, 2019] Luke Breitfeller, Emily Ahn, David Jurgens, and Yulia Tsvetkov. Finding microaggressions in the wild: A case for locating elusive phenomena in social media posts. In *EMNLP-IJCNLP*, 2019.

[Brunson, 2007] Rod K. Brunson. "Police don't like black people": African-American young men's accumulated police experiences. *Criminology & public policy*, 6(1):71–101, 2007.

[Chaney and Robertson, 2013] Cassandra Chaney and Ray V Robertson. Racism and police brutality in America. *Journal of African American Studies*, 17(4):480–505, 2013.

[Dagan *et al.*, 2005] Ido Dagan, Oren Glickman, and Bernardo Magnini. The pascal recognising textual entailment challenge. In *Machine Learning Challenges Workshop*, pages 177–190. Springer, 2005.

[Darwish *et al.*, 2020] Kareem Darwish, Peter Stefanov, Michaël Aupetit, and Preslav Nakov. Unsupervised user stance detection on twitter. In *International AAAI Conference on Web and Social Media (ICWSM)*, volume 14, pages 141–152, 2020.

[Demszky *et al.*, 2019] Dorottya Demszky, Nikhil Garg, Rob Voigt, James Zou, Jesse Shapiro, Matthew Gentzkow, and Dan Jurafsky. Analyzing polarization in social media: Method and application to tweets on 21 mass shootings. In *NAACL:HLT*, pages 2970–3005, 2019.

[Devlin *et al.*, 2019] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *NAACL:HLT*, pages 4171–4186, 2019.

[Geirhos *et al.*, 2020] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard S. Zemel, Wieland Brendel, Matthias Bethge, and Felix A. Wichmann. Shortcut learning in deep neural networks. *Nat. Mach. Intell.*, 2(11):665–673, 2020.

[Halterman *et al.*, 2021] Andrew Halterman, Katherine A. Keith, Sheikh Muhammad Sarwar, and Brendan O'Connor. Corpus-level evaluation for event QA: the indiapoliceevents corpus covering the 2002 gujarat violence. In *ACL/IJCNLP 2021*, Findings of ACL, pages 4240–4253, 2021.

[Hossain, 2021] Tamanna T. Hossain. *COVIDLies: Detecting COVID-19 misinformation on social media*. PhD thesis, UC Irvine, 2021.

[Hovy and Spruit, 2016] Dirk Hovy and Shannon L. Spruit. The social impact of natural language processing. In *ACL*, pages 591–598, 2016.

[Kang *et al.*, 2018] Dongyeop Kang, Tushar Khot, Ashish Sabharwal, and Eduard Hovy. AdvEntuRe: Adversarial training for textual entailment with knowledge-guided examples. In *ACL*, 2018.

[KhudaBukhsh *et al.*, 2021] Ashiqur R. KhudaBukhsh, Rupak Sarkar, Mark S. Kamlet, and Tom M. Mitchell. We don't speak the same language: Interpreting polarization through machine translation. In *AAAI 2021*, pages 14893–14901, 2021.

[Lawrence, 2000] Regina G. Lawrence. *The politics of force: Media and the construction of police brutality*. Univ of California Press, 2000.

[Liu *et al.*, 2019] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, M. Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *ArXiv*, abs/1907.11692, 2019.

[Mendelsohn *et al.*, 2020] Julia Mendelsohn, Yulia Tsvetkov, and Dan Jurafsky. A framework for the computational linguistic analysis of dehumanization. *Frontiers in artificial intelligence*, 3:55, 2020.

[Mendelsohn *et al.*, 2021] Julia Mendelsohn, Ceren Budak, and David Jurgens. Modeling framing in immigration discourse on social media. In *NAACL:HLT*, pages 2219–2263, June 2021.

[Olteanu *et al.*, 2019] Alexandra Olteanu, Carlos Castillo, Fernando Diaz, and Emre Kıcıman. Social data: Biases, methodological pitfalls, and ethical boundaries. *Frontiers in Big Data*, 2:13, 2019.

[Palakodety *et al.*, 2020] Shriphani Palakodety, Ashiqur R. KhudaBukhsh, and Jaime G. Carbonell. Voice for the Voiceless: Active Sampling to Detect Comments Supporting the Rohingyas. In *AAAI 2020*, pages 454–462, 2020.

[Roy and Goldwasser, 2020] Shamik Roy and Dan Goldwasser. Weakly supervised learning of nuanced frames for analyzing polarization in news media. In *EMNLP*, pages 7698–7716, 2020.

[Sap *et al.*, 2020] Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. Social bias frames: Reasoning about social and power implications of language. In *ACL 2020*, pages 5477–5490, 2020.

[Scheffer *et al.*, 2001] Tobias Scheffer, Christian Decomain, and Stefan Wrobel. Active hidden markov models for information extraction. In *International Symposium on Intelligent Data Analysis*, pages 309–318. Springer, 2001.

[Settles, 2009] Burr Settles. *Active learning literature survey*. University of Wisconsin-Madison Department of Computer Sciences, 2009.

[Sindhwani *et al.*, 2009] Vikas Sindhwani, Prem Melville, and Richard D Lawrence. Uncertainty sampling and transductive experimental design for active dual supervision. In *ICML*, pages 953–960, 2009.

[Voigt *et al.*, 2017] Rob Voigt, Nicholas P Camp, Vinodkumar Prabhakaran, William L Hamilton, Rebecca C Hetey, Camilla M Griffiths, David Jurgens, Dan Jurafsky, and Jennifer L Eberhardt. Language from police body camera footage shows racial disparities in officer respect. *Proceedings of the National Academy of Sciences*, 114(25):6521–6526, 2017.

[Yang and Carbonell, 2013] Liu Yang and Jaime Carbonell. Buy-in-bulk active learning. In *Advances in neural information processing systems*, pages 2229–2237, 2013.