"You are no Jack Kennedy":
On Media Selection of Highlights from Presidential Debates

Chenhao Tan
Department of Computer Science
University of Colorado Boulder
Boulder, CO, USA
chenhao@chenhaot.com

Hao Peng
Paul G. Allen School of CS&E
University of Washington
Seattle, WA, USA
hapeng@cs.washington.edu

Noah A. Smith
Paul G. Allen School of CS&E
University of Washington
Seattle, WA, USA
nasmith@cs.washington.edu

ABSTRACT

Political speeches and debates play an important role in shaping the images of politicians, and the public often relies on media outlets to select bits of political communication from a large pool of utterances. It is an important research question to understand what factors impact this selection process.

To quantitatively explore the selection process, we build a three-decade dataset of presidential debate transcripts and post-debate coverage. We first examine the effect of wording and propose a binary classification framework that controls for both the speaker and the debate situation. We find that crowdworkers can only achieve an accuracy of 60% in this task, indicating that media choices are not entirely obvious. Our classifiers outperform crowdworkers on average, mainly in primary debates. We also compare important factors from crowdworkers' free-form explanations with those from data-driven methods and find interesting differences. Few crowdworkers mentioned that "context matters", whereas our data show that well-quoted sentences are more distinct from the previous utterance by the same speaker than less-quoted sentences. Finally, we examine the aggregate effect of media preferences towards different wordings to understand the extent of fragmentation among media outlets. By analyzing a bipartite graph built from quoting behavior in our data, we observe a decreasing trend in bipartisan coverage.

CCS CONCEPTS

• **Applied computing → Law, social and behavioral sciences**;

KEYWORDS

media bias, presidential debates, quotations, wording, conversations

1 INTRODUCTION

Televised public debates have become a focal point of election campaigns [34]. A famous example is from the 1988 U.S. vice presidential debate. After Dan Quayle compared himself to John F. Kennedy,

> **SANDERS**: **Do I consider myself part of the casino capitalist process by which so few have so much and so many have so little by which Wall Street's greed and recklessness wrecked this economy?** [...]
> **CLINTON**: [...] I think what Senator Sanders is saying certainly makes sense in the terms of the inequality that we have. [...]
> **SANDERS**: [...] **So what we need to do is support small and medium-sized businesses, the backbone of our economy, but we have to make sure that every family in this country gets a fair shake** [...]

Figure 1: Between the two bold sentences from Bernie Sanders from neighboring turns in the 2016 Democratic primary debates, the first sentence was quoted 23 times in newspapers within a week after the debate, while the second one was not quoted at all in our data. Yet in our experiments, 3 out of 5 humans thought the second one was quoted more.

Lloyd Bentsen dismissively replied, "you are no Jack Kennedy". This moment received wide post-debate coverage, and even pervades later debates and popular parodies.[1] We refer to moments that are frequently quoted by media outlets as *highlights*. Media-selected highlights in post-debate coverage shape how the public interprets election debates, because these highlights may be the only debate content consumed by many voters [17, 25, 26].

However, most highlights that media select are not as exceptional as "you are no Jack Kennedy", and it remains unclear what factors determine media selection. Consider the example in Figure 1. It was not obvious to participants in our experiments which of the two passages from Sanders was a highlight. Even knowing that the first one was highlighted, we can propose multiple plausible explanations for this choice of the media. It could be the catchiness of "casino capitalist", or the parallel structure of "so few have so much" and "so many have so little". It may also relate to the conversational dynamics (e.g., Clinton's agreement about inequality) or non-textual factors such as the media's political leanings.

Some qualitative studies have investigated the effect of language-related factors in how the media select highlights [1, 11, 19, 24]. For instance, to explain the popularity of "you are no Jack Kennedy", Clayman [11] suggests three important factors: 1) *narrative relevance* (how well a moment fits in a news story); 2) *conspicuousness* (how much a moment stands out in a debate); 3) *extractability* (how

---

[1]https://en.wikipedia.org/wiki/Senator,_you're_no_Jack_Kennedy#Legacy.

self-contained a moment is). However, it is nontrivial to computationally characterize these qualitative factors, and their predictive power remains unknown. What is also missing in the existing literature is an understanding of how consumers of news coverage, i.e., the public, interpret media outlets' selection of highlights.

Moreover, media selection of highlights holds promise for understanding media bias and polarization. Existing studies have shown that non-textual factors such as the media's preferences and biases can affect the selection of highlights [2, 22, 32, 37]. In particular, Niculae et al. [37] demonstrate implicit structure in the media (e.g., international vs. domestic) by analyzing the patterns in quotes of President Barack Obama. Since televised presidential debates have been happening for decades, analysis of debate coverage can shed light on the evolution of media preferences over time.

To quantitatively investigate these questions, we collect American presidential debate transcripts, including both general debates (the debates between general election candidates after primary elections) and primary debates (the earlier, within-party debates in primary elections), and post-debate coverage in newspapers (details in §2). Our dataset spans more than three decades.

The present work: the effect of wording on media choices (§3). The first thrust of this paper investigates the effect of wording on media choices and examines whether the public understands these choices. To do that, we propose a binary classification framework, where a well-quoted sentence (highlight) is paired with a sentence that is not a highlight, controlling for the speaker and the debate situation. The task is to identify which one was quoted more. Using this classification framework, we investigate how well humans and machine-learned classifiers can predict media choices and what the distinguishing textual factors are.

We find that media choices in the selection of highlights are not entirely obvious to humans. As a proxy for the general public, we request Mechanical Turk workers to perform the classification task and explain what factors they use in making predictions. Although they are able to identify some textual signals and outperform random chance (50%), they only achieve an average accuracy of 60%.

Meanwhile, there seem to exist more signals in the wording that are not salient to untrained humans. With carefully-designed features that build on past qualitative studies [1, 11, 19, 24], our classifiers achieve an accuracy of 66%. This result indicates that textual factors can predict media choices to a greater extent than average human performance suggests. In fact, the main performance gap comes from primary debates. One possible explanation for the gap in human performance between general debates and primary debates is the amount of past exposure: primary debates receive less media coverage than general debates and humans may have weaker memories of primary debate highlights.

We also observe interesting similarities and differences when comparing distinguishing factors that humans mentioned with those identified by data-driven methods. For instance, negativity is considered important in both approaches. However, the two approaches view conversational context differently. Only 3% of human responses mention that context matters, while our models suggest that it is a significant factor: highlights tend to be more different from the speaker's previous utterance, and are more likely to be picked up in later utterances.

The present work: quoting patterns over time (§4). The second thrust of this paper examines the media's own preferences and biases in selecting highlights over time. Instead of viewing all media outlets as a uniform body, we take advantage of the longitudinal nature of our data and study whether the news media have become more fragmented over time. Using a bag-of-sentences approach, we construct a bipartite graph over media outlets and the sentences they quoted. Consistent with existing studies on polarization [2], we observe a decreasing trend in bipartisan coverage in general elections, where a clear two-party structure exists. When we investigate the similarity between media outlets without partisan assumptions, we find an increasing trend in the tightness of local clustering, but do not observe that media outlets are becoming less similar to each other over time on average.

2 AN OVERVIEW OF THE DATASET

Our dataset consists of two parts: debate transcripts and post-debate news coverage. We extract transcripts from general debates since 1960 and primary debates since the 2000 presidential election from the *American Presidency Project*.[2] Throughout this work, we define a *turn* as an uninterrupted utterance by a single speaker.

In order to collect post-debate news coverage, we use LexisNexis Academic[3] to search all newspapers within seven days after each debate. As LexisNexis indexes newspapers since 1980, we study media highlights of presidential debates from 1980 to 2016. To achieve high recall, we use debate type ("presidential", "vice", "democratic" and "republican") and the word "debate" as the query. Although newspapers are a subset of news coverage, they comprise a long-standing and often-studied segment of the media and are highly amenable to replication studies. We leave exploration of additional media sources to future work.

Inspired by existing studies [31, 37, 47, 52], we define "highlights" based on quotes in news articles that directly come from the debates. In this work, we differentiate quotes from quotations. We refer to any texts in news articles that are enclosed in quotation marks as *quotations*, and *quotes* are a subset of quotations that can be matched to a turn at a debate. We determine whether a quotation matches a turn in the presidential debate based on word overlap and fuzzy matching. The extraction process produces pairs of quotes and quoted sentences from the corresponding debate. We will present the formal definition of highlights in §3.1. Our dataset and supplementary material are available at https://chenhao.com/papers/debate-quotes.html.

Table 1 shows overall statistics of our dataset. We next discuss basic properties of our dataset. In particular, we observe an increasing trend of news media quoting presidential debate moments.

A diverse set of newspapers (Figure 2a). It is important to point out that there are more newspapers over time in our dataset, partly because more media outlets began to quote presidential debates and partly because LexisNexis gradually improves their collection of newspapers. Only four newspapers quoted general debates in

---

[2]http://www.presidency.ucsb.edu/debates.php. In fact, presidential debates are a relatively recent phenomenon, despite their prominence nowadays. After the first general debate between John F. Kennedy and Richard Nixon in 1960, there were no general debates until 1976. For more historical details, refer to http://www.cnn.com/2012/09/30/opinion/greene-debates/.

[3]http://www.lexisnexis.com/hottopics/lnacademic/.

(a) Cumulative fraction of quotes by media outlet.

(b) Fraction of debate sentences being (partially) quoted.

(c) Fraction of texts that are quotes in news articles.

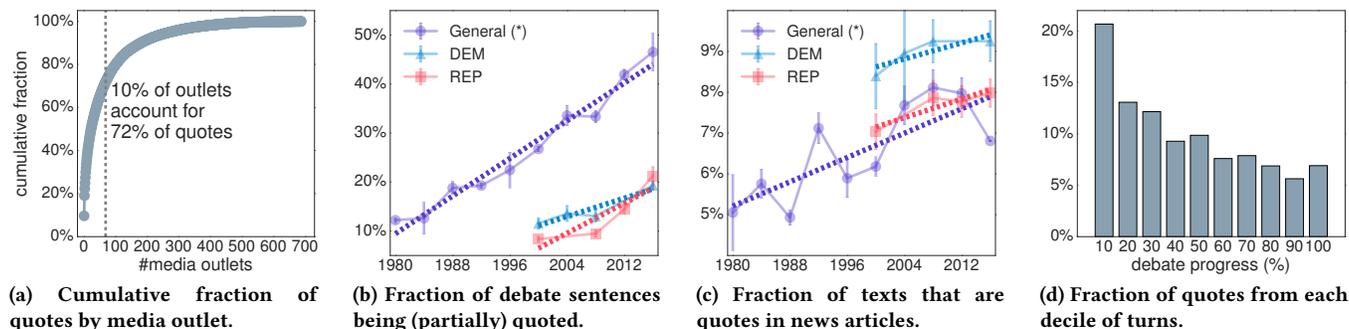**(d) Fraction of quotes from each decile of turns.**

Figure 2: In Figure 2a, media outlets are sorted by total number of quotes, and there is a heavy tail. Figure 2b shows that an increasing fraction of sentences in the debates are (partially) quoted by the media over time. Figure 2c indicates that an increasing fraction of texts in news articles are direct quotes from the debates. Figure 2d shows that more quotes are from the beginning of a debate. Throughout the paper, error bars represent standard error, dashed lines show the best linear fit and * in a legend indicates that the linear coefficient is statistically significantly different from 0 with $p < 0.05$.

debate type	#debates	avg. #sent's	avg. #tokens	avg. #quotes
general	26	1064.9	16278.0	944.2
vice	9	1018.2	15974.0	618.4
Democratic	38	1070.6	16028.3	330.7
Republican	59	1270.8	17781.1	369.1

Table 1: Dataset statistics. The last three columns show the average number of sentences, the average number of tokens, and the average number of quotes per debate, respectively.

1980; 334 newspapers did in 2016. There are around 700 unique newspapers in total and the top 10% of the newspapers in quoting debates account for 72% of all the quotes. The *New York Times* and *Washington Post* are consistently the top newspapers with the most quotes since 1980. We also have small newspapers (e.g., *Rhode Island Lawyers Weekly*) and international newspapers (e.g., *The Guardian*).
An increasing trend of quoting (Figure 2b and Figure 2c). Since there are more media outlets over time, it is expected that an increasing fraction of sentences in the debates are quoted in the news media. In comparison, general debates are quoted much more than primary debates. But it is unexpected that the fraction of texts that are direct quotes in news articles is also growing over time, as we observe. This suggests that directly quoting the candidates is an increasingly common way to cover debates.
More quotes are from the beginning of a debate (Figure 2d). Later turns in a debate are less likely to be quoted in the media. This decreasing likelihood is robust across different types of debates and echoes findings on movie quotes [13].

3 THE EFFECT OF WORDING ON MEDIA CHOICES

We study how textual factors associate with media selection of highlights from presidential debates for three reasons. First, media-selected highlights are the only debate content consumed by voters who do not watch the debate and hence rely on post-debate coverage. How well the public understands media selection of highlights is worth studying because the public uses this understanding to

interpret news coverage. Second, debating candidates cannot control their popularity or the news media's political preferences, but they can always choose the wording when they seek to deliver a message. Understanding the effects of wording can thus inform political communication. Third, it is valuable to know the extent to which we are able to predict media choices using *only* textual information—*although of course textual information alone may not fully predict media choices.*

To study the effect of wording on media choices, we propose an experimental framework that controls for the speaker and the debate situation and formulate a binary classification task (§3.1). We then study the public understanding of media choices by evaluating human performance on this task and analyzing free-form explanations in human surveys (§3.2). We further build on existing theories and develop quantitative features for data-driven classifiers (§3.3), and examine their prediction performance in §3.4.

3.1 Experimental Framework

To investigate how textual factors associate with media selection of highlights from presidential debates, we need to control for other confounding factors such as who the speaker is and what state the debate is in. Inspired by "natural experiments" and previous studies about the effect of wording on message sharing, memorability, and persuasion [13, 15, 54, 55], we propose a classification task that asks humans and machines to decide which sentence was quoted more in the news media between two "similar" sentences.
Binary classification framework. To formally define highlights, we use a sentence as the basic unit of analysis. In our natural experiment framework, we find a matching "negative" sentence for each media-selected highlight and evaluate whether humans or machines can tell the highlighted one from the not-highlighted one in a pair. Because debating candidates have varying popularity and the debate progresses with different levels of importance (see Figure 2d), we match each highlight with a not-highlighted sentence of similar length within three turns by the same speaker.[4]

We consider a sentence *highlighted* if it was among the most quoted $t\%$ sentences from the corresponding debate. We opted for

---
[4]There exist alternate ways to account for topic shift, e.g., [35].

category	% humans
circular (sound bite, newsworthy)	30.0
provocative, sensational	25.5
surprising, funny	17.0
issues, informative	16.0
controversial	15.0
memory, past exposure	12.0

Table 2: Top factors in human surveys and the percentage of humans that mentioned them.

this instead of absolute-count thresholding because the number of quotes increases over time (see Figure 2b). We experiment with $t = 1, 2, \ldots, 10$. The results are robust across choices of t, and we thus report only the results for $t = 10$ (except for overall accuracy).

Following the above definition, we extract ~14K pairs of sentences from all the debates. For a pair of sentences, we randomize the order and predict whether the first one was highlighted. A random guess gives an accuracy of 50%. We randomly select 80% of our data for training and hold out the other 20% for testing. To build machine learning classifiers in §3.4, we construct a vector representation of each pair by extracting features from each sentence and take the difference between them. We use logistic regression with ℓ_2-regularization. This approach is equivalent to a linear framework for the ranking task within a pair [28]. We grid search the best ℓ_2 coefficient based on five-fold cross-validated accuracy on the training set over $\{2^x\}$, where x ranges over 20 values evenly spaced between −8 and 1.

3.2 Human Interpretation of Media Choices

Using the above classification framework, we first examine the public understanding of how the news media select highlights. We recruit 200 U.S.-based Mechanical Turk workers as a sample of untrained humans (the public) to perform the prediction task on randomly sampled pairs from the held-out set. In addition, we ask our participants to explain the important factors that they use to make predictions in free-form responses.

Specifically, we request each participant to label 25 pairs and finish an exit survey to explain what factors they used to make predictions as well as their experience of watching debates and their political ideology. For a pair, we show the highlighted sentence, the not-highlighted sentence, and a few surrounding sentences[5] in the order they occurred in the debate and ask the participants to guess which one was quoted more in the news media. To make sure that they understand the task, we prepare three training pairs and require a comprehension quiz before they start. We also provide a bonus for each correct guess to incentivize participants to try their best. Further details of the human experiments are in the appendix. **Media choices are not obvious to the public.** The average human accuracy is 60% and Fleiss' κ between human labels is 0.2,

indicating slight agreement.[6] These observations suggest that media choices are not obvious to humans, at least based on textual content. One plausible explanation is that the textual information is insufficient to explain media choices: media choices are influenced by external factors such as statements made outside presidential debates and public opinion shifts. However, as we will show later, there seem to exist signals in the wording that are not salient to untrained humans. Another more pessimistic explanation is that humans have a limited understanding of how the news media select highlights.

Important factors in human surveys. To examine the important factors from the perspective of humans, we categorize free-form explanations in human surveys and present the top factors in Table 2. The most common factors cited are circular; i.e., 30% of the participants mentioned that they made decisions based on which one is newsworthy or which one makes a good sound bite. This suggests that it is nontrivial for humans to reason about media selection of highlights.

Among the next five most frequently mentioned categories, participants mentioned *sensational* (emotional, negative, shocking, etc.) and *surprising or funny*. Most of these factors are difficult to operationalize computationally. Interestingly, *memory or past exposure* was explicitly mentioned by 12% of the participants, indicating that humans may predict media choices even less accurately without unavoidable media exposure.

These top factors do not directly align with existing qualitative studies. For instance, Clayman [11], the most relevant work, points out three important factors: 1) *narrative relevance* (how well a moment fits in a news story); 2) *conspicuousness* (how much a moment stands out in a debate); 3) *extractability* (how self-contained a moment is). It is unclear how to map the factors that our participants mentioned to these three.

Notably, only 3% of the participants mentioned that context in which a sentence occurred matters, while an equal number of people brought up appealing to liberal voters as a criterion (none discussed the other direction). Extractability in Clayman [11], or "can be taken out of context" was viewed important by 4% of the participants. However, "potential to be twisted", a slightly different but more malicious interpretation, was explicitly mentioned by 6.5% of the participants. These observations indicate a negative attitude toward the news media, or at least some skepticism about their role in American politics.

3.3 Quantitative Features

Building on the above human intuitions and existing studies, we develop two sets of features: sentence-alone features, and conversation-flow features that attempt to capture conversational dynamics. In this section, we use training data to identify important features that distinguish highlights from non-highlights, and compare the signals from data-driven methods with the factors from humans' free-form explanations.

In addition to these two sets of features, we will employ bag-of-words features as a strong baseline in §3.4, i.e., unigrams and bigrams that occur at least 5 times in the training set.

---

[5]For both sentences in a pair, we include up to 3 sentences before it and 3 sentences after it to provide some context for the participants to understand the current state of the debate.

[6]Surprisingly, there is no clear relationship between an individual's prediction performance and their self-reported level of experience or political ideology.

Feature set	Related theories/intuitions and brief description	Significance	
Informative-ness	We use length as a proxy of informativeness. Longer sentences are more likely to be highlighted despite our control on length as discussed in §3.1. This echoes findings in Tan et al. [54, 55].	length	↑↑↑↑
Emotions	We consider positive and negative words in Pennebaker et al. [40]. Highlighted sentences use significantly more negative words, while there is no difference in positive words. This is consistent with negativity bias [44] and the negativity found in the news media [18].	posemo negemo	↑↑↑↑
Contrast	We use negations and negative conjunctions (e.g., *not, but, although*) to capture contrast. Our result echoes Atkinson [1], which demonstrates the importance of contrast.	negation negative conj.	↑↑↑↑ ↑↑↑↑
Personal pronouns	In general, highlighted sentences use more personal pronouns except first person plural and third person plural. One explanation for the contrast between *I* and *we* is that media outlets prefer statements about candidates themselves to unifying statements using *we*.	i, you, she, he they we	↑↑↑↑ ↓↓↓↓
Uncertainty/ subjectivity	Hedging is a common way to express uncertainty [30] and we use a dictionary from Tan and Lee [53]. In the debate context, hedges may also represent subjectivity.	hedges	↑↑↑↑
Strong emphasis	Superlatives represent the extreme form of an adjective or an adverb and can be used to put emphasis on a statement. Surprisingly, highlighted sentences do not use more superlatives.	superlatives	
Generality	We count indefinite articles to measure generality. Our findings are consistent with Danescu-Niculescu-Mizil et al. [13], Shahaf et al. [45], Tan et al. [54].	indef. articles	↑↑↑
Language model	To capture surprise or conspicuousness, we compute language model scores based on NYT texts and unigram part-of-speech (POS) tags in the WSJ portion of Penn Treebank. However, the only significant feature is that highlighted sentences are more similar to NYT texts in unigram usage. This finding is consistent with message sharing [54] but is different from memorable movie quotes [13].	unigram bi-, trigram POS {1, 2, 3}-gram	↑↑
Parallelism	Using parallel sentence structure is a rhetorical technique, e.g, the first sentence in Table 1 and "I've never wilted in my life, and I've never wavered in my life". We use average longest common sequences between sub-sentences to measure it [48].	parallelism	↑

Table 3: Testing results of sentence-alone features. Upward arrows indicate that highlighted sentences have larger scores in that feature, while downward arrows suggest the other way around (↑↑↑↑: $p < 0.0001$, ↑↑↑: $p < 0.001$, ↑↑: $p < 0.01$, ↑: $p < 0.05$, the same for downward arrows; p refers to the p-value after the Bonferroni correction).

Sentence-alone features. We first examine features that do not rely on any contextual information in the debates and can be extracted from a sentence alone. We evaluate whether highlighted sentences are significantly different from not-highlighted sentences in each feature. Specifically, for each feature, we compute the feature values for both highlighted and not-highlighted sentences and conduct one-sided paired t-tests with the Bonferroni correction [4]. Table 3 presents intuitions and theories for each feature set, including related work. We present the full computational details in the appendix.

By comparing results in Table 3 with previously discussed factors from human surveys, we find that the top factors in human surveys also tend to be statistically significant signals from data-driven methods, such as length (informative), and negative emotions (sensational). But this is not always the case, e.g., positive emotions and strong emphasis were not statistically significant signals. Meanwhile, signals such as personal pronouns, hedges, and language model features arise from data-driven methods, but humans may not pay as much attention to them. A complete comparison between computational features and human factors would require operationalizing controversiality, sensationalism, humor, etc.; we leave this to future work.

Conversation-flow features. Although only a handful of participants in our human experiment mentioned that context matters, conversational dynamics in the debates may contribute to the selection of highlights [58]. We propose a novel set of conversation-flow features and indeed observe intriguing conversational dynamics around the highlights.

In order to capture the local context of a sentence (s), we compare the sentence with its neighboring turns. We use a window w and denote the content words in the next w turns by the same speaker as $\text{Words}_{\text{self}}^{\text{post}}(w)$, the content words in the previous w turns by the same speaker as $\text{Words}_{\text{self}}^{\text{prev}}(w)$. Similarly, we extract $\text{Words}_{\text{other}}^{\text{post}}(w)$ and $\text{Words}_{\text{other}}^{\text{prev}}(w)$ for other speakers. We compute Jaccard similarity between the sentence (Words_s) and its neighboring turns. For instance,

$$\text{Jaccard}_{\text{other}}^{\text{post}}(5) = \frac{|\text{Words}_s \cap \text{Words}_{\text{other}}^{\text{post}}(5)|}{|\text{Words}_s \cup \text{Words}_{\text{other}}^{\text{post}}(5)|}$$

measures the similarity between the sentence and the 5 turns by other speakers after the sentence.

- A kink exists in similarity to turns by the same speaker (Figure 3a). Highlighted and not-highlighted sentences present the same level of similarity to turns by the same speaker until the last turn before the sentence. An interesting kink shows up around the sentence: highlighted sentences are less similar to the turn immediately before but are more similar to turns after. We take this as a sign

(a) Similarity to one's own turns.

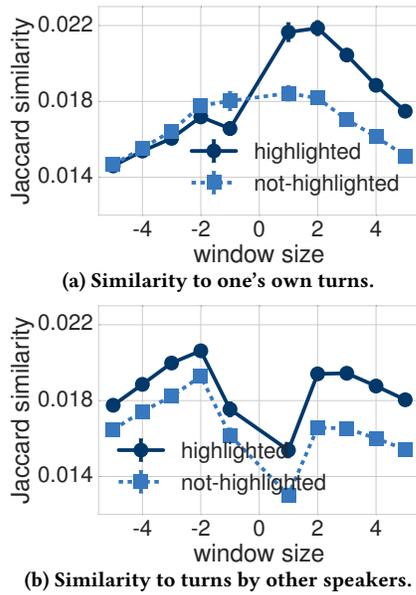(b) Similarity to turns by other speakers.

Figure 3: Figure 3a and Figure 3b present conversation-flow features that are based on Jaccard similarity between a sentence and its neighboring turns (negative windows for previous turns, positive windows for later turns, error bars are tiny). In Figure 3a, a kink exists around 0 in similarity to turns by the same speaker, while in Figure 3b, highlighted sentences are consistently more similar to turns by other speakers.

that "changepoints" in a monologue, where the speaker shifts in topic or style, are more likely to be quoted.

- Highlighted sentences are more similar to neighboring turns by other speakers (Figure 3b). Regarding the overall trend, both for highlighted and not-highlighted sentences, the similarity is smaller for the immediate neighboring turns (w is 1 or -1) than when more turns are included. This is because moderators often speak right before and right after candidates, and moderators speak distinctly from candidates, in terms of words (due to different communicative goals).

3.4 Prediction Performance

Finally, we study to what extent media choices can be predicted only from textual factors by examining classification performance on the held-out set. We also investigate the difference between general debates and primary debates.

Overall prediction accuracy (Figure 4a). Using only textual factors, our classifier achieves an accuracy of 66% on the held-out set for $t = 10$. The accuracy of both machines and humans increases for "easier" pairs, the ones in which highlighted sentences were quoted more frequently. This trend confirms that meaningful signals exist in the wording. The accuracy of machines ("all") is always above humans and the difference is statistically significant.

The bag-of-words (BOW) model already outperforms humans in this task. In comparison with the features that we propose in §3.3 ("all – BOW"), although "all – BOW" has far fewer features, it yields

a similar accuracy to BOW. "all – BOW" works relatively well when highlighted sentences in a pair were quoted more frequently and when there are fewer training instances, while BOW has an advantage when the highlighted sentence is closer to the 10% threshold (right side of Figure 4a). Combining all features (including BOW; "all") always leads to the best accuracy.

Note that the accuracies of machines and humans are not meant to be compared head-to-head, since machines rely on training data to identify the useful signals, while humans depend on their daily (potentially biased) media exposure. Instead, we view this accuracy gap as evidence suggesting that some signals in the wording are hard for humans to identify. This also points to the potential to inform the public with the help of machines.

Differences across debate types (Figure 4b). As primary debates have more candidates and receive less coverage than general debates, the news media may employ different criteria to select highlights. To explore the differences, we train classifiers on subsets of the training data from primary debates and test on different types of debate. Differences indeed exist in how wording affects media selection: the classifiers do not usually perform well when tested on other debate types, except from Republican primary debates to general debates. In fact, using all training instances does not improve over using the pairs only from the matching debate type, despite the latter's use of fewer training instances.[7]

Machines outperform humans only in primary debates. Human accuracy is much better in general debates than in primary debates. In fact, the advantage of our classifiers ("all") in Figure 4a mainly comes from primary debates. The reason may be that general debates receive more attention and more coverage, humans are thus more likely to remember what was selected as highlights; indeed, "memory, past exposure" was an important factor in the surveys. If this is the case, humans may have an even more limited understanding of media choices had there been no influence from previous exposure. We also observe that humans perform better in general debates after 2000 than in those before 2000.

4 MEDIA PREFERENCES OVER TIME

Beyond the effect of wording, a media outlet's own preferences can potentially impact how highlights are selected. In fact, media polarization has attracted significant interest from both researchers and the public [2, 27]. We take advantage of the longitudinal nature of our dataset, and evaluate the extent of media fragmentation over time. Building on the intuition that outlets are similar if they quote the same sentences with similar sentiments, we employ two approaches to quantify the fragmentation level. We first consider the existing two-party structure in the U.S. and evaluate whether the media quote both parties "evenly". Second, we borrow concepts from the clustering literature and examine the overall similarity between media outlets beyond the partisan assumption.

4.1 Bipartisan Coverage

Because U.S. presidential elections typically involve two major parties, we first take advantage of this two-party structure and

---

[7]A more comparable setup is to subsample training instances to match the size in a particular debate type. Doing this, training only using the matching debate type (known as "in-domain") always outperforms using all debate types.

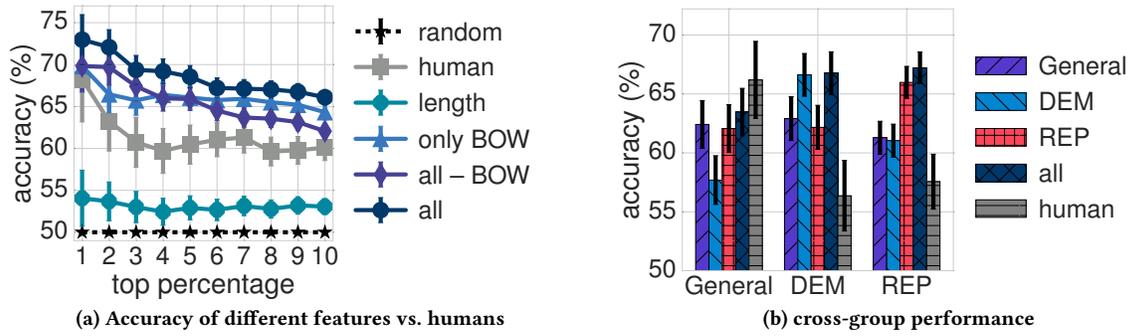(a) Accuracy of different features vs. humans

(b) cross-group performance

Figure 4: Classification accuracy. In Figure 4a, each point measures the accuracy of a feature set (indicated by the color) on the subset of held-out data where the quote count of the highlighted sentence in a pair is among the top $x\%$ quoted sentences in the debate. In other words, smaller x values correspond to "easier" pairs where the highlighted sentence is more prominent. y values for $x = 10$ give the accuracies on the full held-out data (all, 66.0% vs. human, 60.1%, $p = 0.0008$). Figure 4b illustrates the accuracy when we apply the classifier trained on a debate type to another debate type. Different colors represent the debate type of the training data, and the x-axis represents the debate type of the test data. Note that "human" reports the accuracy of human predictions on the corresponding debate type in the test data, and is not a function of the training set.
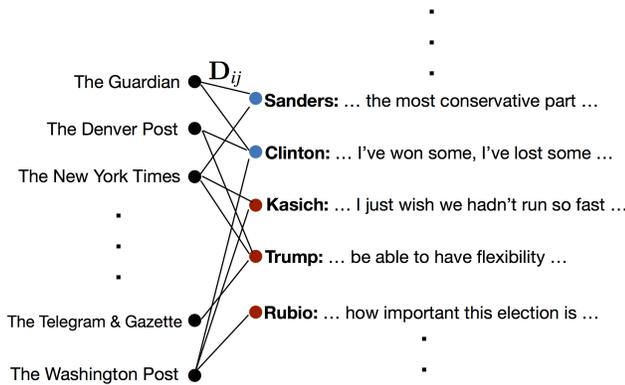

Figure 5: Bipartite graph between media outlets and highlights from presidential debates. The edges are sampled from presidential debates in 2016.

evaluate fragmentation by how much "Democratic-leaning" media outlets quote Republican candidates and vice versa.

Bipartite graph representation. A natural representation of quoting patterns is a bipartite graph between outlets and sentences, where an edge between media outlet i and candidate sentence j indicates that i quotes j (e.g., Figure 5). This graph can be represented using a media-sentence matrix $\mathbf{D} \in \mathbb{R}^{M \times S}$, where each row represents a media outlet (M is the number of outlets) and each column represents a sentence from a candidate (S is the number of sentences). To obtain $\mathbf{D}_{ij}$, we use three methods to account for both the frequency and the sentiment of a media outlet quoting a candidate utterance: *a. count* ($\mathbf{D}_{ij}$ is the number of times that sentence j was quoted in outlet i); *b. positive context* ($\mathbf{D}_{ij}$ is the number of positive words in the 30 words around each quote of sentence j in outlet i); *c. negative context* (similar to positive context but counting negative words). We normalize each row so that the ℓ_2-norm is 1 and remove media outlets that used fewer than 10 quotes in an election.

Using min-cut to identify bipartisan coverage. We focus on news coverage of general debates, because there has always been one presidential candidate and one vice-presidential candidate from the Democratic party and the Republican party during the past three decades.[8] We thus build matrices for the bipartite graphs based on general debates in each presidential election.

To the extent that outlets "lean" one way or the other, we expect them to quote one party or the other more (or to positively quote one party more, or to negatively quote one party more). In the extreme case, a subset of media outlets might only quote the Republican candidates and the rest only quote the Democratic candidates, in other words, there exists no bipartisan coverage. These intuitions align with the idea of using min-cut to identify bipartisan coverage. If we apply the min-cut algorithm to separate sentences from Democratic candidates and sentences from Republican candidates in the bipartite graph [49], then the extreme case where no bipartisan coverage exists leads to a min-cut of 0. Conversely, the more costly the min-cut, the more entangled the two sides are.

Decreasing bipartisan coverage (Figure 6a). We thus compute the fraction of weights in the min-cut and smaller values in this statistic indicate that media outlets quote largely from one of the two sides and little from the other.[9] Figure 6a shows that cross-cutting coverage in the min-cut is declining over time, under all three definitions. It is worth noting that the fraction of weights in the min-cut is not small (about 40%, upper bounded by 50%) despite the declining trend, which suggests that media outlets tend to at least cover both sides. The fact that there exists less cross-cutting coverage in sentiment (both positive and negative) than in counts indicates that although media outlets quote both sides, the sentiment differs.[10]

---

[8]We ignore all independent candidates in this analysis.

[9]Note that media outlets that preferentially quote Democratic candidates may not support the Democratic party, because a quote can be presented in a negative light. We thus also use sentiment information in the context of quotes to populate $\mathbf{D}$.

[10]In terms of the partition resulted from the min-cut, in most years, the majority of the media outlets are in the partition associated with Republican candidate sentences, at least consistent with a recent report in 2016 [39]. This is especially true for big media outlets such as the *New York Times* and *Washington Post*. A notable exception is that the *Washington Post* is in the Democratic partition for positive contexts in 2008.

(a) The fraction of weights in the min-cut.

(b) Average pairwise similarity across all media outlets.

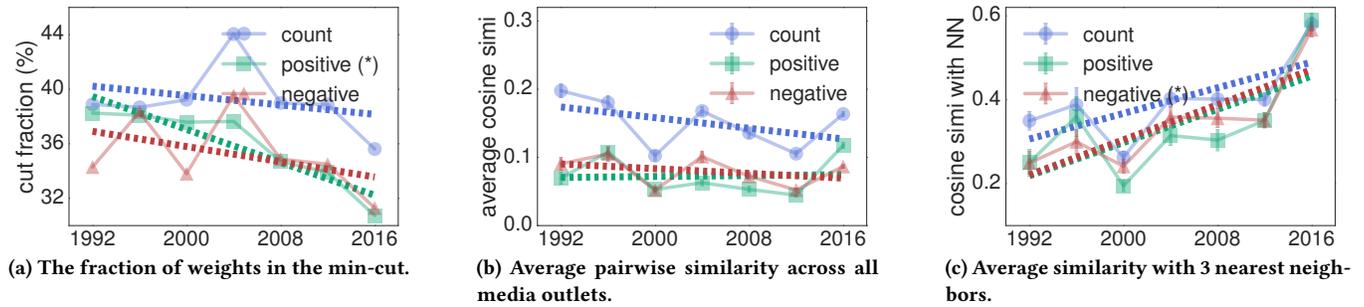(c) Average similarity with 3 nearest neighbors.

Figure 6: Figure 6a estimates whether the outlets quote evenly across two parties and shows a declining trend over time. Figure 6b gives the global average pairwise similarity and there are no clear trends, while Figure 6c shows that "local similarity" (average similarity with top nearest neighbors) increases over time. There are not enough media outlets with at least 10 quotes to compute meaningful results in the 80s, so we exclude those years. In all figures, different colors represent different ways to represent media outlets with their quoting patterns.

4.2 Beyond Partisan Assumptions

An alternative way to estimate fragmentation is to study how well media outlets cluster together without partisan assumptions. We build matrices based on both primary debates and general debates in each presidential election. We use each row D_i to represent media outlet i and investigate the quality of clustering between media outlets.

Clustering purity is typically evaluated at two levels: in a pure clustering, inter-cluster distances are large and intra-cluster distances are small [43]. When we use the Silhouette score (a measure of purity) to identify the optimal number of clusters in a K-means clustering of media representations in D, the optimal number is close to the total number of media outlets, M, which suggests that there are many isolated small clusters.

We thus examine fragmentation at the above two levels by considering each media outlet as a singleton: whether media outlets have become less similar to each other overall (analogous to inter-cluster distances); and whether media outlets have become more similar to their nearest neighbors (analogous to intra-cluster distances). We use cosine similarity to measure the similarity between a pair of media outlets.

No clear trends in "inter-cluster" similarity (Figure 6b). To evaluate whether the news media become less similar to each other, we calculate the global mean of all pairwise similarities. A decreasing global mean would indicate fragmentation at the global level, but we do not observe consistent trends or any statistically significant correlation with time (Figure 6b). The similarity in positive context and negative context is always smaller than the similarity in usage frequency. This again suggests that, although different media outlets may quote the same sentences, they present different opinions around the quotes.

Increasing "intra-cluster" similarity (Figure 6c). To capture "local" similarity that is analogous to intra-cluster distances, we propose a statistic that measures the average similarity between a media outlet and its K nearest neighbors. We refer to this as *local similarity*. A growing local similarity suggests increasing tightness at the local level. We observe consistent increasing trends across three definitions, and this observation is robust with choices of K. This observation is related to the fact that there are increasingly

many media outlets over time and it is thus more likely for a media outlet to have a close nearest neighbor. However, this hypothesis is insufficient to explain our observations, because it also suggests that "inter-cluster" similarity should increase, which does not hold.
Discussion. Our observations are derived from a three-decade dataset and are consistent with past work on polarization and partisan selective exposure [2, 51], but the reasons behind the decreasing bipartisan coverage and the increasing local similarity require further investigation.

Our results are certainly limited by the relatively short history of presidential debates. It is also important to note that our study does not take into account the influence among media outlets themselves [6, 20, 29]. For instance, Golan [20] shows a correlation between the morning *New York Times* and three evening television news programs. Further studies regarding the diffusion in media selection of highlights can shed more light on our observations.

5 RELATED WORK

We have discussed the most relevant studies throughout the paper. Here we discuss three additional strands of related work.
The effect of post-debate coverage on public opinion. Studies have shown that media choices about coverage can have serious consequences [8, 17, 23, 25, 26, 38, 56]. For instance, Fridkin et al. [17] show that in the 2004 U.S. election, citizens who only read the news coverage rated Kerry more negatively compared to those who watched the debate firsthand, because media outlets highlighted the moment of Kerry outing Cheney's lesbian daughter, although this moment did not catch much attention from the live audience. Boydstun et al. [5] develop a mobile app to collect real-time feedback for presidential debates. Patterson [38] discuss the media's critical tendency and its partisan consequences in the U.S.
Influences between the media and politicians. Although our work focuses on media selection of highlights, politicians often behave based on their beliefs about media preferences, which suggests complex dynamics between the media and politicians [3, 9, 12, 21, 42]. For instance, Blumler and Kavanagh [3] discuss politicians' increasing adaptation to different news values and formats in the presence of media abundance. Also relevant is research on the influence of politicians on the media, including agenda-setting, rhetorical positioning, and framing [10, 16, 33, 46, 57].

Power dynamics in debates and other types of coverage. Studies have shown that language use and topic control in debates can reflect influence between candidates and indicate power dynamics [35, 36, 41]. More recently, social media have also become an important channel to monitor public opinion on debates in real time [7, 14] and potentially change news media coverage.

6   CONCLUSION

In this paper, we conduct the first systematic study on media selection of highlights from presidential debates, using a three-decade dataset. We introduce a computational framework that controls for the speaker and the debate situation to study the effect of textual factors. First, we find that media choices are not obvious to Mechanical Turk workers, suggesting that the public may have a limited understanding of how the news media choose highlights in news coverage. Second, although machines and humans achieve similar accuracy in general debates, machines significantly outperform humans in predicting media-chosen highlights in primary debates. Our findings indicate that there exist signals in the textual information that untrained humans do not find salient. In particular, highlights are locally distinct from the speaker's previous turn, but are later echoed more by both the speaker and other participants. We further demonstrate a declining trend of bipartisan coverage using macro quoting patterns and analyze the quality of clustering between media outlets without partisan assumptions.

The news media play an important role in connecting the public and politicians. Our work indicates that the public may not understand what factors matter in media choices. Quantitative studies in this domain can complement qualitative theories to improve the understanding of political and media communication for both scholars and the public.

A   APPENDIX

A.1   Amazon Mechanical Turk Labeling Details

Detailed task and filtering criteria. For each threshold $t \in \{2, 4, 6, 8, 10\}$, we randomly sampled 200 pairs from the corresponding heldout set, in total 1,000 pairs. In our Amazon Mechanical Turk experiments, we required participants to be from the United States and have done at least 10 HITs with at least 97% acceptance rate. We requested each participant to label 25 pairs in an assignment (5 for each threshold) and each participant can only finish one assignment. We paid $1.00 ($0.04 a pair) for an assignment and gave $0.02 bonus for each correct guess to incentivize participants to try their best. We removed participants who did not answer survey questions in good faith and spent too short time on the task. To compensate for such filtering, we added assignments on Amazon Mechanical Turk until we had 200 valid assignments. In the end, we gathered 5,000 labels in total (5 labels per pair).

Survey questions. The following questions were asked after participants finished labeling:

- Some people follow presidential debates most of the time, while others aren't that interested. How often would you say you have watched presidential debates from 1980 to 2016? (Never, 1-5 times, 5-10 times, more than 10 times)
- Generally speaking, when it comes to political parties in the U.S., how would you describe yourself? (Democrat, Independent close to Democrat, Independent (close to neither), Independent close to Republican, Republican, Other)
- Explain what factors influence your decision on which sentence was quoted more, simple comments such as several adjectives or nouns (e.g., issues, surprise) can help our research. (free-form response)
- If you have any comment about our task, please give us your feedback. (free-form response)

There are no clear trends in human performances regarding levels of experience or ideology. Most participants gave very positive feedback about our task, e.g., "Very fun to read some of the old transcripts and think about those conversations–thank you!".

A.2   Sentence-alone Feature Definitions and Testing Results

Detailed definitions. The following list is aligned with features in Table 2.

- Length is measured by the number of words in a sentence.
- The lexicons of positive words and negative words come from the corresponding category in LIWC [40].
- Negations and negative conjunctions. Negations include "n't", "not", "no", "cannot". Negative conjunctions include "although", "atho", "but", "nor", "whereas", "while", "though", "however", "otherwise", "tho" and "unless".'
- Personal pronouns. The list of definitions come from LIWC [40].
- Hedges. We use a list of regular expressions from Tan and Lee [53].
- Superlatives. We get a list of candidate lexicons by matching words with the regular expression "[a-zA-Z]+est", and then manual filter false positives. We also include "most", "least" and "worst".
- Indefinite articles. "a" and "an".
- Language model features. We use each sentence's likelihood w.r.t. both lexical level and part-of-speech level language models trained on newswire data. Specifically, we train 1, 2, and 3-gram lexical level language models on the NYT corpus[11], and we use the WSJ portion of Penn Treebank[12] to train POS-level models. Our implementation is based on SRILM [50].
- Parallelism. We measure parallelism using the average length of the longest common sequences between sub-sentences following Song et al. [48]. The basic unit in the longest common sequence is a word. Sub-sentences are split by ";" or ",".

Feature testing procedure and results. For each feature, we use a one-sided paired t-test to test whether, on our training pairs, our scoring function for that feature is larger in the highlighted sentences than in the not-highlighted sentences. Given that we did 20 tests in total, there is a risk of obtaining false positives due to multiple testing. To account for this, we only report significant results after the Bonferroni correction [4], i.e., we multiply each p-value by 20 and see whether it is smaller than (for example) 0.05.

---

[11]LDC number: LDC2008T19. https://catalog.ldc.upenn.edu/ldc2008t19
[12]LDC number: LDC99T42. https://catalog.ldc.upenn.edu/ldc99t42

REFERENCES

[1] Max Atkinson. 1984. *Our Masters' Voices: The Language and Body Language of Politics*. Psychology Press.

[2] Matthew A. Baum and Tim Groeling. 2008. New Media and the Polarization of American Political Discourse. *Political Communication* 25, 4 (2008), 345–365.

[3] Jay G. Blumler and Dennis Kavanagh. 1999. The Third Age of Political Communication: Influences and Features. *Political communication* 16, 3 (1999), 209–230.

[4] Carlo E. Bonferroni. 1936. *Teoria statistica delle classi e calcolo delle probabilita*. Libreria internazionale Seeber.

[5] Amber E. Boydstun, Rebecca A. Glazier, Matthew T. Pietryka, and Philip Resnik. 2014. Real-time Reactions to a 2012 Presidential Debate: A Method for Understanding Which Messages Matter. *Public Opinion Quarterly* 78, S1 (2014), 330–343.

[6] Thomas P. Boyle. 2001. Intermedia Agenda Setting in the 1996 Presidential Election. *Journalism & Mass Communication Quarterly* 78, 1 (2001), 26–44.

[7] Marcel Broersma and Todd Graham. 2012. Social Media as Beat: Tweets as a News Source During the 2010 British and Dutch elections. *Journalism Practice* 6, 3 (2012), 403–419.

[8] Jennifer Brubaker and Gary Hanson. 2009. The Effect of Fox News and CNN's Postdebate Commentator Analysis on Viewers' Perceptions of Presidential Candidate Performance. *Southern Communication Journal* 74, 4 (2009), 339–351.

[9] Karen Callaghan and Frauke Schnell. 2010. Assessing the Democratic Debate: How the News Media Frame Elite Policy Discourse. *Political Communication* 18, 2 (2010), 183–213.

[10] Dennis Chong and James N. Druckman. 2007. A Theory of Framing and Opinion Formation in Competitive Elite Environments. *Journal of Communication* 57, 1 (2007), 99–118.

[11] Steven E. Clayman. 1995. Defining Moments, Presidential Debates, and the Dynamics of Quotability. *Journal of Communication* 45, 3 (1995), 118–147.

[12] Steven E. Clayman and John Heritage. 2002. *The news interview: Journalists and public figures on the air.* Cambridge University Press.

[13] Cristian Danescu-Niculescu-Mizil, Justin Cheng, Jon Kleinberg, and Lillian Lee. 2012. You Had Me at Hello: How Phrasing Affects Memorability. In *Proceedings of ACL*.

[14] Nicholas A. Diakopoulos and David A. Shamma. 2010. Characterizing Debate Performance via Aggregated Twitter Sentiment. In *Proceedings of CHI*.

[15] John Dinardo. 2010. Natural Experiments and Quasi-Natural Experiments. In *Microeconometrics*. Palgrave Macmillan UK, 139–153.

[16] Robert M. Entman. 1993. Framing: Toward Clarification of a Fractured Paradigm. *Journal of Communication* 43, 4 (1993), 51–58.

[17] Kim L. Fridkin, Patrick J. Kenney, Sarah Allen Gershon, and Gina Serignese Woodall. 2008. Spinning Debates: The Impact of the News Media's Coverage of the Final 2004 Presidential Debate. *The International Journal of Press/Politics* 13, 1 (2008), 29–51.

[18] John G. Geer. 2012. The News Media and the Rise of Negativity in Presidential Campaigns. *PS: Political Science & Politics* 45, 03 (2012), 422–427.

[19] Elisabeth Gidengil and Joanna Everitt. 2003. Talking Tough: Gender and Reported Speech in Campaign News Coverage. *Political communication* 20, 3 (2003), 209–232.

[20] Guy Golan. 2007. Inter-media Agenda Setting and Global News Coverage. *Journalism Studies* 7, 2 (2007), 323–333.

[21] Noah Grand. 2015. *The Emergence of Newsworthiness: Inclusion, Exclusion and Inequality in Political News and Online Media*. Ph.D. Dissertation.

[22] Tim Groseclose and Jeffrey Milyo. 2005. A Measure of Media Bias. *The Quarterly Journal of Economics* (2005), 1191–1237.

[23] Kimberly Gross, Ethan Porter, and Thomas Wood. 2017. Presidential Debates in the Age of Partisan Media: A Field Experiment. (2017). Available at SSRN.

[24] Daniel C. Hallin. 1992. Sound Bite News: Television Coverage of Elections, 1968–1988. *Journal of communication* 42, 2 (1992), 5–24.

[25] D. Sunshine Hillygus and Simon Jackman. 2003. Voter Decision Making in Election 2000: Campaign Effects, Partisan Activation, and the Clinton Legacy. *American Journal of Political Science* 47, 4 (2003), 583–596.

[26] Hyunseo Hwang, Melissa R. Gotlieb, Seungahn Nah, and Douglas M. McLeod. 2007. Applying a Cognitive-processing Model to Presidential Debate Effects: Postdebate News Analysis and Primed Reflection. *Journal of Communication* 57, 1 (2007), 40–59.

[27] Shanto Iyengar and Kyu S. Hahn. 2009. Red media, Blue media: Evidence of Ideological Selectivity in Media Use. *Journal of Communication* 59, 1 (2009), 19–39.

[28] Thorsten Joachims. 2002. Optimizing Search Engines Using Clickthrough Data. In *Proceedings of KDD*.

[29] Kyungmo Kim and George A. Barnett. 1996. The Determinants of International News Flow. *Communication Research* 23, 3 (1996), 323–352.

[30] George Lakoff. 1975. Hedges: A Study in Meaning Criteria and the Logic of Fuzzy Concepts. *Journal of Philosophical Logic* 2, 4 (1975), 458–508.

[31] Jure Leskovec, Lars Backstrom, and Jon Kleinberg. 2009. Meme-tracking and the Dynamics of the News Cycle. In *Proceedings of KDD*.

[32] Yu-Ru Lin, James P. Bagrow, and David Lazer. 2011. More Voices Than Ever? Quantifying Media Bias in Networks. In *Proceedings of ICWSM*.

[33] Maxwell E. McCombs and Donald L. Shaw. 1972. The Agenda-setting Function of Mass Media. *Public Opinion Quarterly* 36, 2 (1972), 176–187.

[34] Mitchell S. McKinney and Diana B. Carlin. 2004. Political Campaign Debates. In *Handbook of political communication research*. 203–234.

[35] Viet-An Nguyen, Jordan Boyd-Graber, and Philip Resnik. 2012. SITS: A Hierarchical Nonparametric Model Using Speaker Identity for Topic Segmentation in Multiparty Conversations. In *Proceedings of ACL*.

[36] Viet-An Nguyen, Jordan Boyd-Graber, Philip Resnik, Deborah A. Cai, Jennifer E. Midberry, and Yuanxin Wang. 2014. Modeling Topic Control to Detect Influence in Conversations Using Nonparametric Topic Models. *Machine Learning* 95, 3 (2014), 381–421.

[37] Vlad Niculae, Caroline Suen, Justine Zhang, Cristian Danescu-Niculescu-Mizil, and Jure Leskovec. 2015. QUOTUS: The Structure of Political Media Coverage as Revealed by Quoting Patterns. In *Proceedings of WWW*.

[38] Thomas E. Patterson. 2016. News Coverage of the 2016 General Election: How the Press Failed the Voters. (2016).

[39] Thomas E. Patterson. 2016. News Coverage of the 2016 Presidential Primaries: Horce Race Reporting Has Consequences. (2016).

[40] James W. Pennebaker, Martha E. Francis, and Roger J. Booth. 2007. *Linguistic Inquiry and Word Count: LIWC 2007*. Technical Report.

[41] Vinodkumar Prabhakaran, Ashima Arora, and Owen Rambow. 2014. Staying on Topic: An Indicator of Power in Political Debates. In *Proceedings of EMNLP*.

[42] Andrea Prat and David Strömberg. 2013. The Political Economy of Mass Media. In *Advances in Economics and Econometrics: Tenth World Congress*. Cambridge University Press.

[43] Peter J. Rousseeuw. 1987. Silhouettes: A Graphical Aid to the Interpretation and Validation of Cluster Analysis. *Journal of Computational and Applied Mathematics* 20 (1987), 53–65.

[44] Paul Rozin and Edward B. Royzman. 2001. Negativity Bias, Negativity Dominance, and Contagion. *Personality and Social Psychology Review* 5, 4 (2001), 296–320.

[45] Dafna Shahaf, Eric Horvitz, and Robert Mankoff. 2015. Inside Jokes: Identifying Humorous Cartoon Captions. In *Proceedings of KDD*.

[46] Yanchuan Sim, Brice D.L. Acree, Justin H. Gross, and Noah A. Smith. 2013. Measuring Ideological Proportions in Political Speeches. In *Proceedings of EMNLP*.

[47] Matthew P. Simmons, Lada A. Adamic, and Eytan Adar. 2011. Memes Online: Extracted, Subtracted, Injected, and Recollected. In *Proceedings of ICWSM*.

[48] Wei Song, Tong Liu, Ruiji Fu, Lizhen Liu, Hanshi Wang, and Ting Liu. 2016. Learning to Identify Sentence Parallelism in Student Essays. In *Proceedings of COLING*.

[49] Mechthild Stoer and Frank Wagner. 1997. A Simple Min-cut Algorithm. *Journal of the ACM* 44, 4 (1997), 585–591.

[50] Andreas Stolcke, Jing Zheng, Wen Wang, and Victor Abrash. 2011. SRILM at Sixteen: Update and Outlook. In *Proceedings of IEEE Automatic Speech Recognition and Understanding Workshop*.

[51] Natalie J. Stroud. 2010. Polarization and Partisan Selective Exposure. *Journal of Communication* 60, 3 (2010), 556–576.

[52] Chenhao Tan, Adrien Friggeri, and Lada A. Adamic. 2016. Lost in Propagation? Unfolding News Cycles from the Source. In *Proceedings of ICWSM*.

[53] Chenhao Tan and Lillian Lee. 2016. Talk it up or Play it down? (Un)expected Correlations between (De-)emphasis and Recurrence of Discussion Points in Consequential U.S. Economic Policy Meetings. (2016). Presented in Text as Data.

[54] Chenhao Tan, Lillian Lee, and Bo Pang. 2014. The Effect of Wording on Message Propagation: Topic- and Author-controlled Natural Experiments on Twitter. In *Proceedings of ACL*.

[55] Chenhao Tan, Vlad Niculae, Cristian Danescu-Niculescu-Mizil, and Lillian Lee. 2016. Winning Arguments: Interaction Dynamics and Persuasion Strategies in Good-faith Online Discussions. In *Proceedings of WWW*.

[56] Yariv Tsfati. 2003. Debating the Debate: The Impact of Exposure to Debate News Coverage and Its Interaction with Exposure to the Actual Debate. *The International Journal of Press/Politics* 8, 3 (2003), 70–86.

[57] Chris Wells, Dhavan V. Shah, Jon C. Pevehouse, JungHwan Yang, Ayellet Pelled, Frederick Boehm, Josephine Lukito, Shreenita Ghosh, and Jessica L Schmidt. 2016. How Trump Drove Coverage to the Nomination: Hybrid Media Campaigning. *Political Communication* (2016).

[58] Justine Zhang, Ravi Kumar, Sujith Ravi, and Cristian Danescu-Niculescu-Mizil. 2016. Conversational Flow in Oxford-style Debates. In *Proceedings of NAACL (short papers)*.