I've Seen That Before! Towards Understanding Hard News Exposure from Soft News Outlets

Jason Yan,[1] **Tong Lin,** [1] **Yanna Krupnikov,** [2] **Kerri Milita,** [3] **Sabina Tomkins** [1]

[1] School of Information, University of Michigan
[2] Department of Communication and Media, University of Michigan
[3] Department of Politics and Government, Illinois State University
jasonyan@umich.edu, tonglin@umich.edu, yanna@umich.edu, kmilita@ilstu.edu, stomkins@umich.edu

Abstract

A well-educated electorate is generally considered to be normatively good for democracy. Previous studies have shown that much political information comes from soft news sources, leading to concerns about the electorate's ability to participate effectively in democratic processes. However, limitations exist with prior works: the discussion around the role of soft news for broader political engagement does make critical assumptions. Specifically, researchers often implicitly assume distinctive contents as intrinsic delineators of hard and soft news categorization. However, this overlooks the possibility that even within soft news outlets, there exist pockets where political events are being covered as "hard" news, revealing a gap in the literature: what is the content of political news in soft news outlets, and how are they fundamentally different from that of hard news, if any? To address this, however, requires a systematic clarification of hard and soft news, which has proved to be complicated. Our work addresses this gap by leveraging a machine learning framework for identifying hard news within soft news outlets. Collectively, our on-going research makes two contributions to the literature: (1) we are the first to conduct a large-scale annotation task to identify hard news content in soft news outlets. In doing so, (2) we provide a systematic clarification of the information found in hard and soft news, agnostic of style and framing.

Introduction

A long-term problem in political science is whether people are informed enough to be active and effective participants in democracy. A well-educated electorate is generally considered to be normatively good for democracy (Prior 2003). Ostensibly, as modern technology makes information far more accessible than at any time in the past, the concern of a sufficiently well-informed voter base should not be an issue. However, with political discourse in the United States becoming increasingly polarized (Hare and Poole 2014; Heltzel and Laurin 2020) and past works showing that most people are not following news outlets (Prior 2003; Wojcieszak et al. 2024; Baum 2003), the task of informing the public seems to only increase in difficulty.

A good deal of exposure to political content comes from outlets that cannot be classified as "hard news" (Wojcieszak et al. 2024). Rather, people are getting political information from non-news and, importantly, "soft news" outlets. Soft news outlets are outlets that are more likely to present news in a sensationalist manner – often aiming to entertain, rather than inform (Grabe and Bucy 2022; Otto, Glogger, and Boukes 2017). Because many soft news outlets blur the line between reliable news and sensationalism/"infotainment", scholars argue that soft news becomes a gateway to the "believability of disinformation" (Grabe and Bucy 2022, p.582)

Yet the connection between exposure to soft news outlets and disinformation could be more nuanced. Soft news is a journalistic style – one that often attempts to make information more entertaining for the public (e.g Otto, Glogger, and Boukes 2017). The informal and sensational style of some soft news *could* leave people more open to disinformation (Grabe and Bucy 2022; Otto, Glogger, and Boukes 2017), but this argument depends on the *content* rather than the *style* of soft news journalism. Yet style and framing, while salient facets of journalism, are distinct dimensions that can be disentangled from the validity and coverage of information. In other words, solely accessing soft news outlets does not necessarily preclude individuals from acquiring broad and factual information. Rather - and even surprisingly - soft news may encourage the uptake of political engagement as a byproduct of it being an entertainment medium and leads to more representative voting behaviors (Baum 2002; Baum and Jamison 2006). Finally, there is some limited evidence suggesting soft news consumption does substantively increase factual knowledge attainment (Baum 2003). In concert, these works argue that soft news are beneficial for spreading political knowledge.

There do exist limitations with prior works: the discussion around the role of soft news for broader political engagement does make critical assumptions. Specifically, researchers often implicitly assume distinctive contents as being intrinsic delineators of hard and soft news categorization. However, this overlooks the possibility that even within soft news outlets, there exist pockets where political events are being covered as "hard" news, revealing a gap in the literature: what is the content of political news in soft news outlets, and how are they fundamentally different from that of hard news, if any? To address this, however, requires a systematic clarification of hard and soft news, which has proved to be difficult (see Reinemann et al. (Reinemann et al. 2012) for an overview).

Our work addresses this gap by contributing a machine learning framework for identifying hard news within soft news outlets. This framework will allows us to analyze the extent to which soft news serves as an instrument of knowledge transfer.

Method

We propose a large-scale systematic analysis of a major soft news outlet: People Magazine[1]. Our goal is to quantify the prevalence and gradient of hard news content among articles from People Magazine. To do so requires a ground-truth dataset which we can use as references to develop language models for downstream analysis. These models have seen success in prior works, such as detecting hyperpartisan news with various neural network architectures (e.g., BERT, ELMo, and Word2vec) (Naredla and Adedoyin 2022). To this end, we are conducting an annotation task consisting of annotators labeling each article with predetermined characteristics of interest. The data acquisition process will be discussed in the next section. The four characteristics are as follows: discussion of politicians, celebrities, political events, and implications. Politicians and celebrities refer to a discussion of these groups of individuals, which is a basic measure of relevance. Similarly, political events refer to the discussion of any events that are broadly associated with politics. Finally, and perhaps most crucially, is the measure of implication: does the article provide information that can be categorized as having implications? Measuring implications proves to be more complicated than the previous measures as it requires anchoring: implications for whom? An implication for politicians may not necessarily be an implication for ordinary people and would yield different labels. Thus, we propose two categories: implications for ordinary people or politicians/political parties. Table 1 summarizes our proposed dimensions as measures of prevalence and validity of hard news content.

Data Acquisition

We scraped articles from People Magazine using a keyword search dictionary. Articles that contain key words in their title were included in the scraping process. The words contained in this dictionary pertain to political topics.[2] The search dates ranged from 2017 to 2024. Currently, our corpus contains around 30,000 articles.

Exploratory Analysis

Exploratory analysis reveals crucial insights on news articles from People Magazine. A core aspect of hard news is the use of formal language. To understand the formality of news articles in People Magazine and potential temporal effects, we analyzed the formality of language across time. Here, we measure formality using a supervised learning classifier.[3] Figure 1 is a binned time-series plot of formality across time by months. We observe a sharp increase in formality, starting in 2020. This increased level of formality persisted but held constant until 2022, when it saw an instant dropoff to values below pre-pandemic levels.

From a cursory analysis, the increase in formality in 2020 is likely attributed to the COVID-19 pandemic, which caused nationwide lockdowns in the United States and much of the world. This might suggest that news articles, on average, became more formal in stylistic features. However, this may also be due to the pandemic shuttering much of societal functions, which greatly limits the types of news that could be published. If so, the increased level of formality from 2020 to mid-2022 may have consolidated from a higher proportion of actual "hard news" from People Magazine during that time horizon rather than news generally becoming more formal. Using BERTopic (Grootendorst 2022), we constructed a set of topics which we then assigned each article to. Figure 2 displays the average formality score of 30 randomly selected topics out of 364 constructed topics. We see that there exist variations in topic formality, which in conjunction with the time-series analysis, provides preliminary evidence for our hypothesis to be true.

Contributions

Collectively, our on-going research makes two contributions to the literature: (1) we are the first to conduct a large-scale annotation task to identify hard news content in soft news outlets. In doing so, (2) we provide a systematic clarification of the information found in hard and soft news, agnostic of style and framing.

References

Baum, M. A. 2002. Sex, lies, and war: How soft news brings foreign policy to the inattentive public. *American Political Science Review*, 96(1): 91–109.

Baum, M. A. 2003. Soft news and political knowledge: Evidence of absence or absence of evidence? *Political communication*, 20(2): 173–190.

Baum, M. A.; and Jamison, A. S. 2006. The Oprah effect: How soft news helps inattentive citizens vote consistently. *The Journal of Politics*, 68(4): 946–959.

Grabe, M. E.; and Bucy, E. 2022. Moral panics about the integrity of information in democratic systems:Comparing tabloid news to disinformation. *Journal of Broadcasting & Electronic Media*, 66(4): 565–591.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*.

Hare, C.; and Poole, K. T. 2014. The polarization of contemporary American politics. *Polity*, 46(3): 411–429.

Heltzel, G.; and Laurin, K. 2020. Polarization in America: Two possible futures. *Current opinion in behavioral sciences*, 34: 179–184.

[1] https://people.com/

[2] For example, Trump, Biden, Democrat, Republican, Insurrection, Abortion, Climate Change, etc.

[3] https://huggingface.co/s-nlp/roberta-base-formality-ranker. The model is a RoBERTa-based classifier that was fine-tuned on GYAFC (Rao and Tetreault 2018) and the Online Formality Corpus (Pavlick and Tetreault 2016).

Naredla, N. R.; and Adedoyin, F. F. 2022. Detection of hyperpartisan news articles using natural language processing technique. *International Journal of Information Management Data Insights*, 2(1): 100064.

Otto, L.; Glogger, I.; and Boukes, M. 2017. The Softening of Journalistic Political Communication: A Comprehensive Framework Model of Sensationalism, Soft News, Infotainment, and Tabloidization. *Communication Theory*, 27(2): 136–155.

Pavlick, E.; and Tetreault, J. 2016. An Empirical Analysis of Formality in Online Communication. *Transactions of the Association for Computational Linguistics*, 4: 61–74.

Prior, M. 2003. Any good news in soft news? The impact of soft news preference on political knowledge. *Political communication*, 20(2): 149–171.

Rao, S.; and Tetreault, J. 2018. Dear Sir or Madam, May I Introduce the GYAFC Dataset: Corpus, Benchmarks and Metrics for Formality Style Transfer. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 129–140. Association for Computational Linguistics.

Reinemann, C.; Stanyer, J.; Scherr, S.; and Legnante, G. 2012. Hard and soft news: A review of concepts, operationalizations and key findings. *Journalism*, 13(2): 221–239.

Wojcieszak, M.; Menchen-Trevino, E.; Clemm von Hohenberg, B.; de Leeuw, S.; Gonçalves, J.; Davidson, S.; and Gonçalves, A. 2024. Non-news websites expose people to more political content than news websites: Evidence from browsing data in three countries. *Political Communication*, 41(1): 129–151.

Dimensions	Measures	Questions
Politician	Mentions a politician(s)	Does this article mention politicians?
Celebrity	Mentions a celebrity(s)	Does this article mention entertainment or sports celebrity?
Political Event	Main focus is a political event(s)	Is this an article that mainly focuses on a political event(s) or politician(s)?
Implications	Discusses implications for ordinary people	Does this article explicitly discuss how a political figure, event, or trend affects ordinary people?
	Discusses implications for politician(s) or political party(s)	Does this article explicitly discuss how a political figure, event, or trend affects politician(s) or political party(s)?

Table 1: Dimensions that are measured in our survey to identify elements of hard news in articles from a soft news outlet.

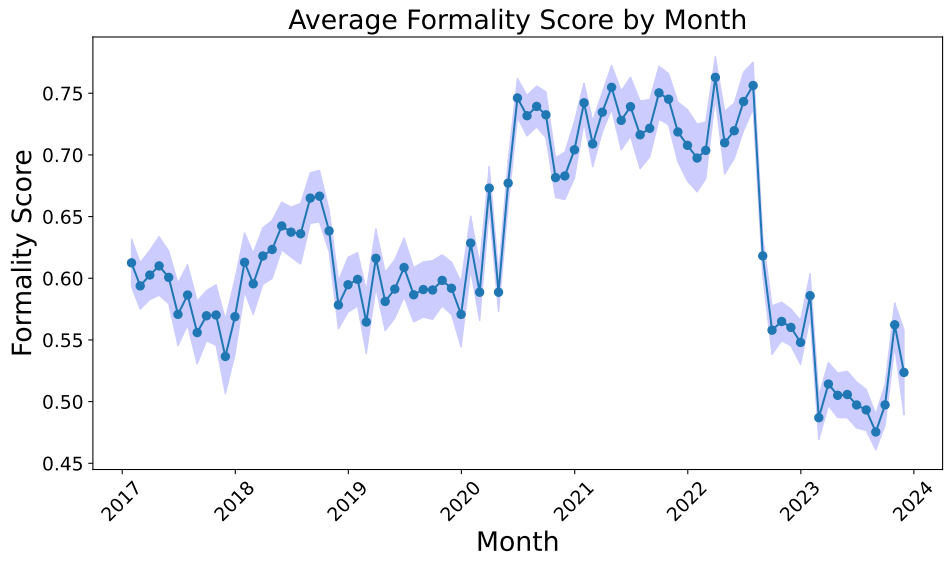

Figure 1: Binned time-series plot of the average formality score (y-axis) of news article for each month (x-axis). Alongside are the 95% confidence intervals of the average formality score by month. The formality scores were derived using a supervised learning classifier. The model is a RoBERTa-based classifier that was fine-tuned on GYAFC (Rao and Tetreault 2018) and the Online Formality Corpus (Pavlick and Tetreault 2016).

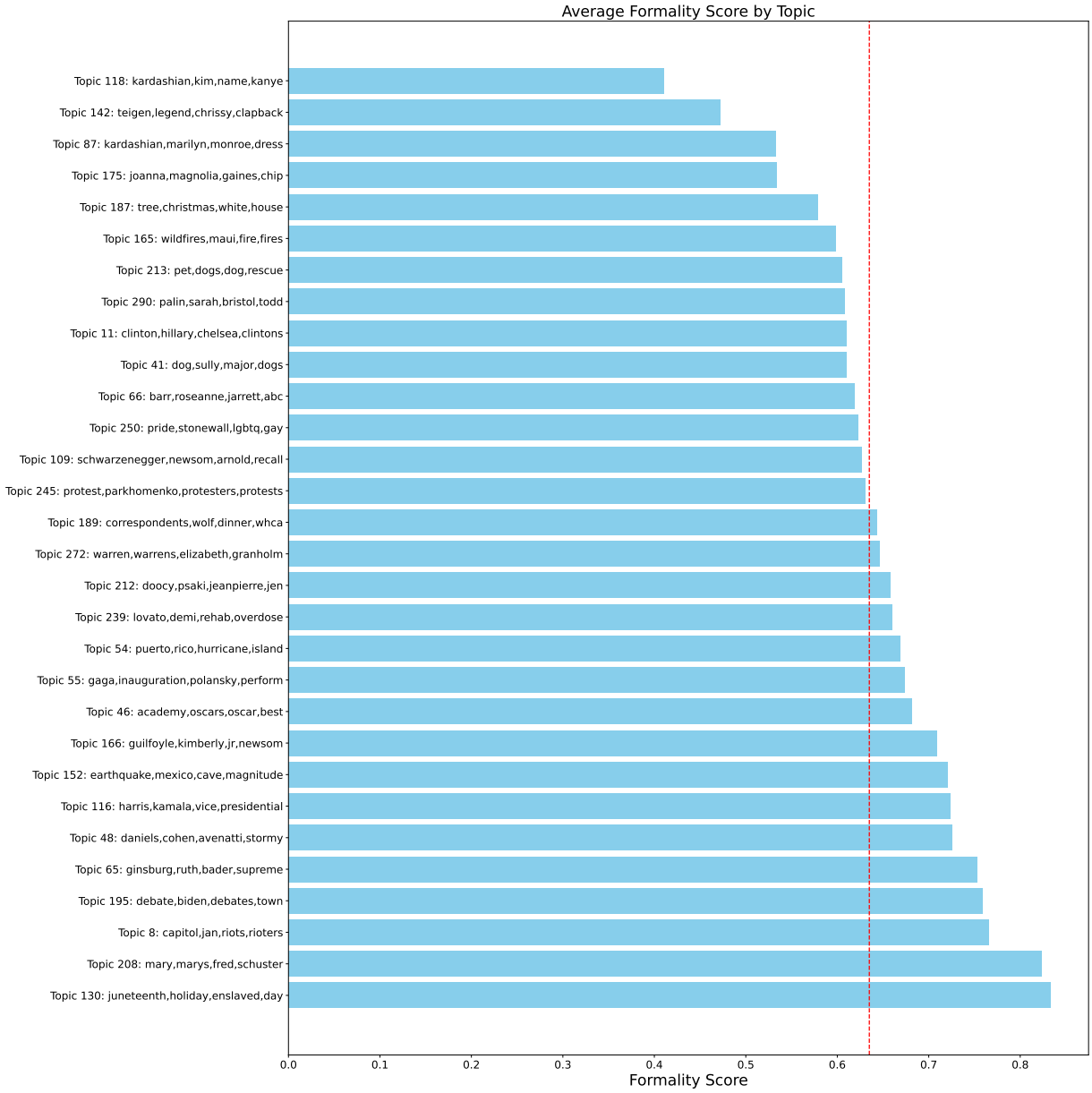

Figure 2: Average formality scores (x-axis) of the topics (y-axis) constructed by BERTopic (Grootendorst 2022). The red vertical bar indicates average formality across all articles in the corpus. For presentation clarity, we show 30 randomly selected topics out of the 364 constructed topics.