Calibrated and Diverse News Coverage

Anonymous Author(s)

ABSTRACT

In recent years, there has been a debate about whether automated news aggregators, like Google News, which recommend articles from various online sources, lead readers to content that reinforces their existing beliefs and restricts their exposure to a diversity of perspectives [15]. We design methods for selecting a set of articles that simultaneously cover all possible viewpoints, while also taking into account user preferences. First, we develop and publicly release a pipeline that, given a collection of news articles, automatically extracts the news stories, the topics, the entities in the articles, as well as the articles' stances towards the entities. Based on this pipeline, we explore the trade-off between different viewpoint-coverage strategies and identify the most preferable strategy for our requirements. Finally, we evaluate several algorithms that select a small number of articles, such that different viewpoints are respected, while taking into account the user preferences. Through extensive experiments and case studies, we show that even though news aggregators may cover all news stories using articles from diverse media outlets, this can still lead to biased stances towards the entities mentioned in the articles. We show how this issue can be mitigated while still covering all stories and while keeping the number of recommended articles small. We believe that our results yield important insights into the trade-off between covering all viewpoints and keeping a small set of selected news articles.

1 INTRODUCTION

The consumption of news in the digital age has undergone a significant transformation. With the abundance of information available through various online platforms, readers are often overwhelmed by the sheer volume of content, leading to limited attention span and selective exposure to news sources. This phenomenon has contributed to the formation of echo chambers, where individuals primarily consume news that align with their existing beliefs and opinions [11]. Consequently, there is a growing need for diversification in news consumption to promote a well-informed and balanced perspective on current events.

The importance of diversifying news consumption cannot be overstated, particularly in light of the declining readership of traditional news sources and the increasing political polarization in society. Efforts to encourage exposure to diverse viewpoints are crucial for fostering a more informed and engaged citizenry. However, achieving this goal is not without its challenges. Given the limited attention span of readers, *striking a balance between coverage of various topics and individual interests is essential.* For any news recommendation system to be effective and practical, it must take into account user preferences to ensure that recommended content is well received by users, especially when dealing with polarizing news topics.

WSDM'25, March 10–14, 2025, Hannover, Germany
2024. ACM ISBN 978-x-xxxx-xxxx-x/YY/MM
https://doi.org/10.1145/nnnnnnn.nnnnnnn

Previous attempts to address the problem of providing well-calibrated and diverse news coverage have fallen short in several aspects. For instance, many platforms focus on low-quality/click-bait content [25], or on highly-personalized topics [19], seeking to maximize user engagement while ignoring aspects of diversity [16]. On the other side, other platforms may account for diversity by selecting some contrarian content [15], but fail to consider the propensity of users to accept the content recommended to them, therefore not succeeding in keeping the users engaged [10, 32]. Such limitations hinder the practical application and effectiveness of existing systems in promoting diverse news consumption.

News aggregators, such as Google News, have been criticized for their role in fostering biased consumption [15]. These platforms aggregate articles based on theme or story, frequently drawing from a single news source with a particular ideological bias, and recommend content without considering alternative sources [27]. This approach has contributed to the creation of echo chambers, where users are not exposed to a variety of viewpoints. In response to such criticisms, these platforms adjusted their algorithms to include a more diverse set of stories. However, these changes are often too simplistic, focusing on *source diversity*—such as merely offering articles from popular left-leaning and right-leaning news sources (e.g., The New York Times, Fox News)—without fully considering *viewpoint diversity* and the user preferences towards those viewpoints. As a result, these adjustments may not necessarily enhance the user experience [9]. While research has explored potential biases in news aggregator recommendations, such as stance, sentiment, and entity representation [3, 7, 22], comprehensive best practices for providing diverse recommendations are still not well established.

Our contributions. We present a new approach for selecting a small set of news articles that provide a diverse coverage of news stories and balance different points of view, while thematically aligning with user interests. Our approach is based on a novel annotation pipeline, which extract meta-information from given news articles. It allows us to assess the bias of groups of news articles, and it enables us to give novel combinatorial algorithms which cover all viewpoints while still calibrating to user preferences.

Our main contributions are as follows:

A comprehensive annotation pipeline: We provide and publicly release an annotation pipeline, powered by state-of-the-art large language models (LLMs), used for extracting news stories, topics, and entities with stances from articles.

Automatic bias detection and coverage of all viewpoints: We introduce realistic and practical diversity measures that allow us to detect bias in the coverage of news stories, and stances extracted from news articles. We present novel combinatorial algorithms for covering all viewpoints present in a set of news articles, while still taking into account user preferences.

Extensive empirical evaluation: We extensively evaluate our algorithms and compare them against interpretable baselines. We find that even if news aggregators present articles for all stories from a diversity news sources (e.g., conservative, as well as liberal news outlets), this can still lead to bias regarding the entities (like politicians) mentioned in the articles. We show that our algorithms can mitigate this issue, presenting a less-biased coverage of the entities, while still covering all stories and while still satisfying the user preferences.

2 RELATED WORK

The digital age has ushered in new challenges for information consumption, particularly in the realm of news media. A growing concern among researchers and social scientists is the potential for digital systems, including automated news aggregators, to create echo chambers and filter bubbles. These phenomena can inadvertently limit users' exposure to diverse viewpoints, potentially endangering informed public discourse. The problem is compounded by the existence of bias in news reporting, users' limited attention span, and the human tendency towards confirmation bias, where individuals gravitate towards information that aligns with their existing beliefs.

Recognizing these challenges, researchers have been exploring ways to diversify users' news diets for over a decade. The goal is to expose readers to multiple viewpoints while respecting their time constraints and ensuring they derive maximum value from their news consumption. Abbar et al. [1] address the problem of identifying diverse recommendations for a specific news article a user is currently reading. Building on this foundation, Nikookar et al. [26] expand the scope to consider the diversification of recommendations across multiple browsing sessions, proposing methods to optimize for diversity both within and across these sessions. Reuver et al. [28] use distributional language models to place users and news articles in a multi-dimensional semantic space, where diversity could be operationalized as distance and variance. This approach opens up possibilities for modeling individual "latitudes of diversity" for different users, potentially allowing for personalized viewpoint diversity in support of healthy public debate.

Mulder et al. [24] emphasize the difference between source diversity (e.g., Fox news being right leaning, and New York Times being left leaning) and viewpoint diversity (e.g., content from New York Times that might be right leaning). They note that existing systems, such as Google News, primarily diversify based on news sources rather than viewpoints. Tintarev et al. [31] develop a new distance measure for diversity within a topic. Their approach enables diversity while maintaining topic relevance, introducing an adaptation to the existing maximal marginal relevance (MMR) technique. This allows for the composition of diverse recommendation lists with increasing information content as users progress through the list. While most methods in this space focus on content-based diversification, Advani et al. [2] propose methods for detecting and diversifying news presentation by reordering the ranking of the recommendations.

A critical aspect of implementing diverse news recommendation systems is understanding and modeling user acceptance. Loecherbach et al. [21] develop a system to study when and why users make choices regarding diverse news content. Their study underlines the importance of incorporating user acceptance into recommendation models. This aligns with the findings of Mulder et al. [24] and Tintarev et al. [31] further emphasizing that user acceptance of diversification must be addressed in tandem with algorithmic solutions to enable a complete and effective approaches.

In this context, the current research aims to contribute to this important field by exploring the trade-off between covering all viewpoints and maintaining a manageable set of selected news articles while considering user preferences. Our approach distinguishes itself through a broader notion of diversity, encompassing political leaning, stance towards specific entities, and perspectives on particular topics. Crucially, we operationalize this comprehensive view of diversity in a practical manner by integrating user interests into our model.

3 ANNOTATION PIPELINE

In this section we introduce our pipeline for processing news articles, which we refer to as DNA, for Diverse News Annotation. The pipeline utilizes semantic annotations to enhance news datasets by incorporating state of the art best practices from news-story discovery, topic modeling, entity extraction, and stance detection.

3.1 Input and output to the pipeline

Input. We start with a set of articles $\mathcal{A} = \{a_1, \dots, a_n\}$, each of which is given as plain text. For each article a_i, we also consider the time $t(a_i)$ that it was published. This is essentially a minimal input—news articles and the time they are published—and therefore we expect that our pipeline is generalizable and will be useful for other researchers in the future.

Output. The DNA pipeline enhances the input with various facets such as topics, story identification, entities and bias labels.

Topic distributions. First, we consider a set of m topics, e.g., politics, sports, etc. For each article a_i, we return a vector $\mathbf{x}_i \in [0,1]^m$, which encodes its topic distribution over the m topics, i.e., $\sum_{j=1}^m \mathbf{x}_i(j) = 1$, where $\mathbf{x}_i(j)$ is the fraction of content in article a_i about topic j. We let $\mathbf{X} = \{\mathbf{x}_1, \dots, \mathbf{x}_n\}$ denote the set of vectors of all articles. This step corresponds to Part 1.1 of Figure 1.

News stories. We return a grouping of the articles into a set of *news stories* $\mathcal{S} = \{S_1, \dots, S_{|\mathcal{S}|}\}$, where $\mathcal{S}$ forms a partition of $\mathcal{A}$. Intuitively, each news story S_j consists of articles having a common theme. A story could be, e.g., a set of articles about the US basketball team at the Olympics. This step corresponds to Part 2.1 of Figure 1. Note that, unlike for topics where each article may belong to multiple topics, each article is assigned to exactly one story.

Entities. For each article a_i, we return the set of *entities* $E_i = \{e_1, e_2, ..., e_{|E_i|}\}$ mentioned in the article. Entities can be, for instance, names of politicians or countries. We write E to denote the set of all entities in $\mathcal{A}$.

Bias labels. We consider the *bias* of articles in different aspects. First, we consider the *political leaning of news sources*, e.g., whether they are left- or right-leaning. Second, we return the *stance toward the entities* mentioned in the articles, i.e., for each article we return whether its entities are discussed positively, negatively or in a neutral manner. This corresponds to Parts 2.2 and 2.3 of Figure 1.

3.2 Pipeline implementation

We present the methods used to compute the outputs described in the previous section. Due to space limitations, only an overview is

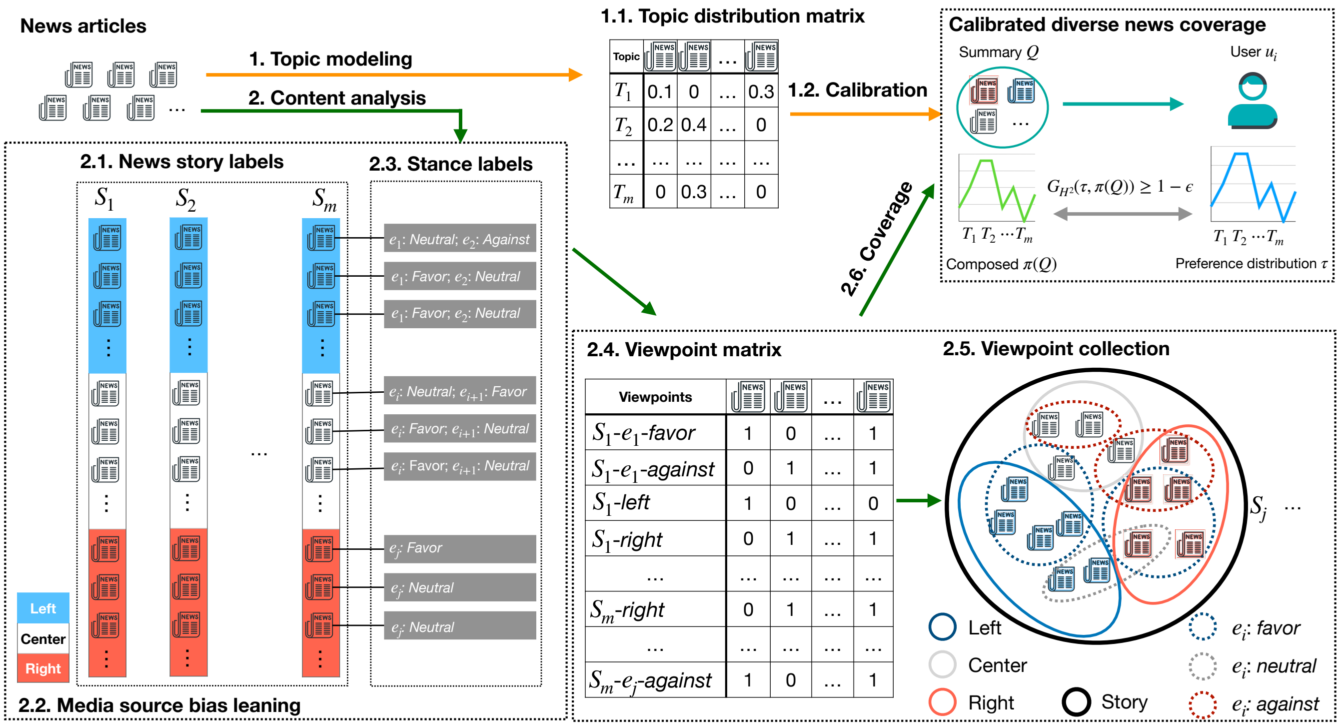


Figure 1: Overview of the methodology. The DNA pipeline analyzes news content using topic modeling, news story discovery, media bias label and stance detection module. The orange path shows the user preference model that measures the distance between selected articles Q with the user preferences. The green path shows the annotations obtained from the DNA pipeline. Article annotations includes story labels (2.1), entities and stances of article (2.3). A bias matrix is formed by combining annotation labels (2.4). Finally, articles form a collection of subsets of articles that capture all viewpoints (2.5).

provided; detailed explanations are available in the supplementary material. We have employed state-of-the-art methods to ensure robustness and accuracy. Additionally, our implementation is designed for flexibility, allowing each subroutine to be easily replaced as advancements in algorithms occur.

Topic modeling. We approach topic modeling as a soft-clustering problem. We use BERTopic [13] and best practices reported therein to obtain article embeddings, article clusters, topic representations, and topic distributions for the article–topic vectors $\mathbf{x}_i$.

News story discovery. For news story discovery we use hard-clustering methods, i.e., each news article is assigned to only one story. Here, we employ the method proposed by Yoon et al. [33] to compute the stories S_j. The method considers the set of articles $\mathcal{A}$ as a time-based stream and it clusters the articles based on their similarity, as computed in a sliding time window of a given size. This method achieves the best performance in our experiments, compared to other baselines [20, 23, 30], while it is robust and scalable.

Entity detection. We extract article entities using the BERT-based model NER[1] on a sentence level. We then postprocess the results to improve the data quality, where we take into account background information about politicians, political parties, and countries. This process gives us the set of entities E.

Bias detection. We obtain the biases for each article as follows. First, to obtain the *political leaning* of news sources, we use the data from Media Bias/Fact Check (MBFC),[2] which is a widely used resource for evaluating bias and reliability of news sources [29]. The media source bias rating falls into the following five categories: *{"left", "left-center", "center", "right-center", "right"}*, and we label each article with the leaning of its news source. Next, we consider the article's *stances* towards these entities, where we consider the labels *{"in-favor", "neutral-or-unclear", "against"}*. To obtain these labels, we use the zero-shot learning approach given by the FLAN-T5 model [8].

A link to the code for our pipeline, which takes any set of news articles and enhances them with all the metadata can be found here.

4 NOTATION AND PROBLEM DEFINITION

In this section, we formally introduce the problem that we study in this paper. Our goal is to select a subset of news articles from a collection of news articles, in a way that it takes into account the user interests (calibration), while simultaneously ensuring diversity by coverage of different viewpoints.

Composed topic distributions and overlap. First we discuss how to measure calibration, that is, the alignment of a set of articles with the interests of a user. Consider an *ordered* set of articles $Q \subseteq \mathcal{A}$

[1]https://huggingface.co/dslim/bert-base-NER

[2]https://mediabiasfactcheck.com

and a weight vector $\mathbf{w} \in [0,1]^{|Q|}$ with $\|\mathbf{w}\|_1 = 1$. The vector $\mathbf{w}$ represent the user's attention. For instance, if $\mathbf{w}_i = \frac{1}{|Q|}$, for all i, then all articles receive the same attention, whereas if $\mathbf{w}_i > \mathbf{w}_{i+1}$, for all i, then older articles receive less attention, which can be seen as a reader's decaying attention with time. In the main text we only consider uniform weights and we discuss decaying attention in the supplementary.

The *composed topic distribution* of the articles Q is given by $\pi(Q) = \sum_{i=1}^{|Q|} \mathbf{w}_i \mathbf{x}_i$, where $\mathbf{w}_i$ is the weight of the article i, and recall that $\mathbf{x}_i$ represents the topic distribution of article i. Additionally, we let $Q^{\leq j}$ be the subset of articles in Q assigned up to position j, and we set $\pi(Q^{\leq j}) = \sum_{i=1}^{j} \mathbf{w}_i \mathbf{x}_i$ to the subdistribution composed by $Q^{\leq j}$. Note that $\|\pi(Q^{\leq j})\|_1 \leq 1$ for all $j \leq |Q|$.

To measure how well a composed topic (sub-)distribution π approximates a target distribution τ, we use the overlap measure

$$G_{H^2}(\tau, \pi) = \sum_{i=1}^{m} \sqrt{\tau_i \cdot \pi_i}, \tag{1}$$

which is the squared Hellinger distance and proposed by Kleinberg et al. [17, Sec. 6] for its desirable properties: (1) $G_{H^2}(\tau, \pi) \in [0,1]$ for all τ and π; (2) for fixed τ; $G_{H^2}(\tau, \pi)$ is uniquely maximized at $\tau = \pi$, and (3) $G_{H^2}(\tau, \pi)$ is a monotone and ordered-submodular function when adding articles to Q and when τ and $\mathbf{w}$ are fixed.

Viewpoint collections based on bias labels. We consider a *viewpoint collection* $C = \{C_1, \ldots, C_{|C|}\}$, where each C_i is a subset of articles and we allow viewpoint sets C_i and C_j to overlap. Intuitively, C_i is a set of articles with the same viewpoint. For instance, a collection C_i might contain all articles that have a negative stance towards a given entity, or all articles from left-leaning media outlets about a given story. See Part 2.5 in Figure 1.

We consider three concrete viewpoint collections:

- The *story-based* viewpoint collection consists of all sets $C_{i,j}$ for each story S_i and news outlet leaning $j \in$ *{"left", "left-center", "center", "right-center", "right" }*. This is inspired by platforms like Google News or Ground News[3] that aggregate and present information based on stories. However, they do not take into account stance bias within individual articles.
- The *entity-based* viewpoint collection consists of all sets $C_{e,j}$ containing all articles with entlty e and news outlet leaning $j \in$ *{"left", "left-center", "center", "right-center", "right" }*, as well as all sets $C_{e,k}$ containing all articles with entity $e \in E$ and stance $k \in$ *{"in-favor", "neutral-or-unclear", "against"}* towards e. Note that the *entity-based* viewpoint collection considers the viewpoints regarding entities, as well as the bias of media outlets, while disregarding specific story information.
- The *hybrid* viewpoint collection is the most exhaustive collection we consider. It consists of all sets from the *story-based* viewpoint collection, as well as all sets $C_{i,e,j}$ containing all articles from story S_i with stance j towards entity e.

It is useful to encode the viewpoint information from C in matrix form. Thus, we let $\mathbf{d}_i \in \{0,1\}^{|C|}$ denote the *viewpoint vector* of article a_i over $|C|$ viewpoints, where $\mathbf{d}_i(j) = 1$ indicates that article $a_i \in C_j$, i.e., a_i expresses the viewpoint given by C_j, and $\mathbf{d}_i(j) = 0$, otherwise. We let $\mathbf{D} \in \{0,1\}^{|C| \times n}$ denote the *viewpoint* matrix, which contains each $\mathbf{d}_i$ as its i-th column. This corresponds to Parts 2.4 and 2.5 of Figure 1.

Problem definition. At a high level, we want to select a subset of articles that covers diverse viewpoints. Formally, our goal is to select a minimum number of articles Q such that *at least one* article from each viewpoint subset $C_i \in C$ is contained in Q, indicating that all viewpoints are covered. Since the sets C_i are overlapping, we can expect that $|Q| \ll |C|$. Further, observe that finding such a set Q of minimum cardinality is equivalent to solving the *hitting-set problem*, which is known to be **NP**-hard.

Besides taking into account the diversity of the articles, we are also interested in the user preferences. Thus, we introduce a utility constraint using the overlap measure from Equation (1) to guide the article-selection process to satisfy the user preferences. To do this, we consider a target distribution τ which represents the user preferences over the set of m topics. Given a selected subset of articles Q and the weight vector $\mathbf{w}$, we define the *user utility* as $G_{H^2}(\tau, \pi(Q))$. Naturally, we want to maximize the user utility.

Now, we formally present the calibrated diverse news coverage problem. The task is to select a minimum subset of articles that intersects all subsets in the viewpoint collection, while covering topics that closely approximate the user preferences.

Problem 1 (Calibrated diverse news coverage)**.** *Given a set of articles $\mathcal{A}$, a viewpoint collection C, the topic distribution vectors in $\mathbf{X}$, a threshold parameter $\epsilon \in [0,1]$, and a target distribution τ, the goal is to find a subset $Q \subseteq \mathcal{A}$ to satisfy*

$$\begin{aligned} \min_{Q \subseteq \mathcal{A}} \quad & |Q| \\ \textit{such that} \quad & Q \cap C_i \neq \emptyset \qquad \textit{for all } \; C_i \in C, \; \textit{and} \\ & G_{H^2}(\tau, \pi(Q)) \geq 1 - \epsilon. \end{aligned} \tag{2}$$

The first constraint in Problem 1 is a *hitting-set constraint*, stating that all viewpoints should be covered by the selected articles. The second constraint is a *calibration constraint* given by the overlap measure, and a threshold parameter ϵ.

Since the overlap measure is nonnegative, for $\epsilon = 1$, Problem 1 generalizes the hitting-set problem. Thus, we obtain the following:

Lemma 1. *Problem 1 is* **NP***-hard.*

We also consider a budget-constrained version of Problem 1. This variant addresses the fact that the optimal size of Q might be larger than a user's attention. Thus, we introduce a budget parameter k to model the user's maximum attention, and we seek to maximize the coverage rate of subsets in the viewpoint collection when picking at most k articles. The formal definition is as follows.

Problem 2 (Calibrated diverse news coverage with a budget)**.** *Given a set of articles $\mathcal{A}$, a viewpoint collection C, the topic distribution vectors in $\mathbf{X}$, a threshold parameter $\epsilon \in [0,1]$, a target distribution τ, and a budget parameter k, the goal is to find a subset $Q \subseteq \mathcal{A}$ to satisfy*

$$\begin{aligned} \max_{Q \subseteq \mathcal{A}} \quad & \sum_{C_i \in C} \mathbb{I}(Q \cap C_i \neq \emptyset) \\ \textit{such that} \quad & G_{H^2}(\tau, \pi(Q)) \geq 1 - \epsilon \;\; \textit{and} \;\; |Q| \leq k, \end{aligned} \tag{3}$$

where $\mathbb{I}(\cdot)$ is the binary indicator function with $\mathbb{I}(Q \cap C_i \neq \emptyset) = 1$ if $Q \cap C_i \neq \emptyset$ and 0 otherwise.

[3] https://ground.news/

Note that while both Problems 1 and 2 guarantee that the user's preferences are satisfied, Problem 2 takes the ordering of the selected articles in Q into account: it prioritizes articles that cover many unseen viewpoints due to the limited budget.

5 PROPOSED METHOD

We now describe our main algorithm, which combines a *greedy multi-objective optimization process* with *local-search postprocessing*. Our main challenge is that articles that are useful for satisfying the hitting-set constraint may not be helpful to improve the overlap measure from the calibration constraint, and vice versa. We consider two different strategies to trade-off between these competing goals.

Let us first define the gain obtained by adding an article a_i to a set of articles $Q^{\leq j}$. We let $H(Q^{\leq j}, C) = \sum_{C_i \in C} \mathbb{I}(Q^{\leq j} \cap C_i \neq \emptyset)$ denote the number of subsets C_i intersecting (hitting) the first j items in Q.

The *coverage gain* of adding element a_i to Q is

$$F_H(a_i, Q^{\leq j}) = H(Q^{\leq j} \cup a_i, C) - H(Q^{\leq j}, C).$$

Observe that $F_H(a_i, Q^{\leq j}) \in [0, |C|]$, for all a_i.

Similarly, we define the *overlap measure gain* as

$$F_{G_{H^2}}(a_i, Q^{\leq j}) = G_{H^2}(\tau, \pi(Q^{\leq j} \cup a_i)) - G_{H^2}(\tau, \pi(Q^{\leq j})).$$

Observe that $F_{G_{H^2}} \in [0, 1]$.

To solve Problems 1 and 2, we use a greedy algorithm, which at each iteration, picks the article with the highest gain. However, the orders of elements sorted by overlap measure gain and coverage gain may be inconsistent. For instance, the article with the largest overlap measure gain may have the lowest coverage gain. In addition, the two gains are not on the same scale. To deal with these issues, we introduce two subroutines to pick elements that simultaneously increase both gains.

Subroutine 1: Index minimization (SR-1). In our first subroutine, we compute $F_H(a_i, Q^{\leq j})$ and $F_{G_{H^2}}(a_i, Q^{\leq j})$ for all a_i and we sort them by descending gains. Based on this ordering, we define the ordered article lists L_H and $L_{G_{H^2}}$. We let $I(a_i, L_H)$ denote the index of a_i in list L_H and $I(a_i, L_{G_{H^2}})$ the index of a_i in list $L_{G_{H^2}}$. Now the *index score* of article a_i is given by $I(a_i) = I(a_i, L_H) + I(a_i, L_{G_{H^2}})$. Note that low index scores correspond to elements that have high coverage gain and high overlap gain. Our subroutine returns the article a_i with the smallest index score $I(a_i)$.

Note that subroutine SR-1 considers the coverage gain and overlap gain independently, and chooses the article ranked the highest in the two lists. This strategy circumvents the issue that $F_H(a_i, Q^{\leq j})$ and $F_{G_{H^2}}(a_i, Q^{\leq j})$ have different scales. However, as we only consider the ordering, it is still possible that the gains for the two measures are unevenly distributed and the method does not allow us to prioritize one of the two aspects.

Subroutine 2: β trade-off (SR-2). Our second subroutine alleviates the problem mentioned above and allows us to trade-off between coverage and overlap gain.

More concretely, we consider the *normalized coverage rate gain*

$$\bar{F}_H(a_i, Q^{\leq j}) = \frac{F_H(a_i, Q^{\leq j})}{F_H^*(Q^{\leq j})},$$

where $F_H^*(Q^{\leq j}) = \max_\ell F_H(a_\ell, Q^{\leq j})$ is the maximum coverage gain achieved by any article. Note that $F_H^*(a_i, Q^{\leq j})$ could be 0 (if all subset in C are hit) and in this case we set $\bar{F}_H(a_i, Q^{\leq j}) = 0$. Similarly, we consider the *normalized overlap measure gain*

$$\bar{F}_{G_{H^2}}(a_i, Q^{\leq j}) = \frac{F_{G_{H^2}}(a_i, Q^{\leq j})}{\max_\ell F_{G_{H^2}}(a_\ell, Q^{\leq j})}.$$

Due to the normalization, both scores are in the interval $[0, 1]$.

To trade-off between the normalized gains, we introduce a weight parameter $\beta \in [0, 1]$, which allows us to prioritize either calibration or coverage. We consider the following score:

$$F(a_i, Q^{\leq j}) = \beta \cdot \bar{F}_{G_{H^2}}(a_i, Q^{\leq j}) + (1 - \beta) \cdot \bar{F}_H(a_i, Q^{\leq j}).$$

Now the subroutine chooses the article a_i with the largest $F(a_i, Q^{\leq j})$.

Greedy initial solution. To obtain an initial solution of k articles, we start by initializing an empty set of articles $Q \leftarrow \emptyset$. Then we greedily add articles to Q based on Subroutine 1 or Subroutine 2 until Q contains k articles.

Local search. Optionally, we also consider a post-processing step using local search. The goal is to remove articles from Q while still ensuring that we hit all viewpoint sets C_j and that the calibration constraint is still met. The local search proceeds in iterations and stops when one of the constraints in Problem 1 is violated. In each iteration, the local search checks the neighboring solutions of Q, which can be obtained by deleting a single article from Q. We replace Q with the neighboring solution that decreases the objective function the least. Note that when the size of Q changes, the corresponding weight vector $\mathbf{w}$ (that we use to obtain $\pi(Q)$) should also change; here, we equally redistribute the weight value of the last position to all remaining positions.

5.1 Baselines

For our empirical evaluation we consider the following baselines.

BL1: Hitting set. As mentioned before, without the calibration constraint, Problem 1 is equivalent to the hitting-set problem. Therefore, BL1 is a greedy hitting-set algorithm which we obtain by running our algorithm with Subroutine 2 and $\beta = 0$. This baseline allows us to understand how "coverage without calibration" may differ from user preferences.

BL2: Calibration. BL2 aims at finding a small set of articles that are close to the user's target distribution, while ignoring any coverage constraints. Similar to BL1, we obtain an algorithm by applying our algorithm from above with Subroutine 2 with $\beta = 1$; this corresponds to replacing the objective with the calibration constraint and removing the threshold parameter ϵ. This baseline allows us to understand how "calibration without coverage" may lead to the missed coverage of important topics.

BL3: Maximal marginal relevance. BL3 adapts the diversity-promoting retrieval method *maximum marginal relevance (MMR)* [5]. It is an iterative procedure that ranks elements based on a relevance score and a diversity measure. Bourgeois et al. [5] studied selection bias in news coverage and proposed a method to select a subset of diverse news sources based on MMR. We adapt their method such that it ranks the articles iteratively based on the score function

$$MMR(a_i) := \beta \cdot relevance(a_i, \tau) - (1 - \beta) \cdot \max_{a_j \in Q^{\leq k}} (sim(a_i, a_j)), \quad (4)$$

Table 1: Summary of our datasets. #Sources refers to the number of news sources among the articles.

Dataset	#Articles	#Stories	#Collections	#Sources	#Entities
COVID-Jan	4 933	88	216	197	40
NELA-Jan	9 094	97	1 055	110	105
WCEP-Small	11 591	225	1 005	389	146

where β is a parameter that controls the strength of the diversification and $Q^{\leq k}$ is the set of elements already selected. Here, we define the relevance function based on the overlap with the target distribution τ and the diversification based on the viewpoints. Concretely, we set $relevance(a_i) = F_{G_{H^2}}(a_i, Q^{\leq k})$, and $sim(a_i, a_j)$ is the cosine similarity of bias vector $\mathbf{d}_i$ and $\mathbf{d}_j$. Intuitively, the relevance measure aims to select the article with the largest overlap gain, while the diversification measure aims to select the article most different from the selected set with respect to the viewpoints.

BL4: Nonnegative orthogonal matching pursuit (NOMP). Lappas et al. [18] used the NOMP algorithm [6] for selecting a representative subset of elements. Formally, given a target vector $\tau_{\mathbf{D}}$ and an integer k, NOMP aims to find a binary vector $\mathbf{z}_Q \in \{0,1\}^n$, which contains at most $|Q|$ 1-entries, and such that $\|\tau - \mathbf{D}\mathbf{z}_Q\|_2^2$ is minimized. We adopt NOMP as our fourth baseline to find subsets of articles to approximate the target distributions based on $\mathbf{X}$ and $\mathbf{D}$. In particular, we use the average vector of $\mathbf{D}$ as the target vector $\tau_{\mathbf{D}}$ and apply NOMP to compute corresponding $\mathbf{z}_Q$ with budget k. We pick the articles corresponding to non-zero entries in $\mathbf{z}_Q$ to compose the selected articles.

6 EXPERIMENTAL EVALUATION

We empirically evaluate our method and compare with the baselines. For a detailed evaluation of the parameter settings, please refer to the supplementary material. Our implementation is available online [4].

6.1 Datasets and experimental setup

Datasets. We use three real-world news datasets. COVID is a publicly available dataset from AYLIEN,[4] which contains English news articles from 440 global sources between November 2019 and July 2020. All articles are related to COVID-19. NELA [14] contains 1 778 361 articles from 361 outlets between January 2022 and December 2022. WCEP [12] is a benchmark news dataset collected from the *wikipedia current event portal* and the *common crawl archive*. It provides ground truth story labels for each article.

For each dataset, we consider the articles of one month and apply our DNA pipeline for topic modeling, news story discovery, entity detection, and bias and stance annotations. Table 1 presents details on the annotated datasets. COVID-Jan and NELA-Jan are generated using news articles from January 2020 and 2022, respectively. WCEP is generated using news stories that contain at least 50 articles.

Experimental environment. We conduct our experiments on a Linux server with 2 AMD Epyc 7742 CPUs, 1 TB of RAM and 1 NVIDIA DGX-A100 GPU. Our code is written in Python v3.11.7.

Target distributions. We generate the user target distributions τ synthetically, as follows: we pick a number of seed topics uniformly at random from the m topics. We then compute the top 10 most similar topics to each seed topic by their keyword representation in BERTopic [13]. We obtain τ by assigning exponentially decaying weights to the seed topic and its similar topics.

Algorithms. We write ALG-SRi to denote the algorithm from Section 5, with Subroutine $i \in \{1,2\}$, and ALG-SR1-LS when combined with local-search postprocessing. The baselines are denoted by BLi, for $i \in \{1,2,3,4\}$. We solve Problem 2 and refer to the parameter k as *budget*, i.e., the number of articles selected by the algorithm.

Evaluation criteria. We evaluate our algorithms using the coverage objective function $\sum_{C_i \in C} \mathbb{I}(Q \cap C_i \neq \emptyset)$ from Problem 2 and the overlap measure G_{H^2} from Equation (1).

Additionally, we also consider the *entity balance score*, which measures the balance of news articles with respect to stances towards the entities. Formally, given a set of articles Q, we consider the balance vector $\mathbf{b} \in [0,1]^{3|E|}$, which contains three entries for each entity $e \in E$, corresponding to the fraction of articles in Q that have stance positive/neutral/negative towards e. Ideally, the value of these fractions for each entity is $\frac{1}{3}$. Thus, we estimate balance by computing the overlap with a scaled uniform distribution. In particular, we define the *entity balance score* as $B(Q) = G_{H^2}\left(\frac{1}{|E|}\mathbf{b}, \frac{1}{3|E|}\mathbf{1}\right)$, where $\mathbf{1}$ is the all-ones vector and the normalization ensures that both vectors are distributions.

6.2 Experiments

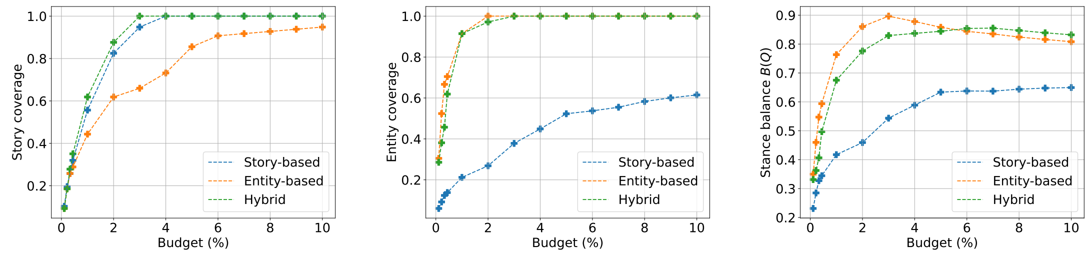
Our evaluation aims to answer the following questions:
Q1: What is the interaction between different coverage objectives? In particular, if we seek to cover story-based viewpoints, how well are entity-based viewpoints covered?
Q2: What is the best-performing method and best settings to achieve diverse viewpoints coverage while considering user preferences?
Q3: Can we achieve an effective trade-off between coverage and user preferences?

Q1: Interaction of different viewpoint coverage objectives. We use ALG-SR1 for covering the *story-based*, *entity-based*, and *hybrid* viewpoint collections. We present the results in Figure 2. In Figures 2(a) and 2(b), we observe that optimizing coverage on the *story-based* and *entity-based* collections yields very good performance. Covering the *hybrid* collection obtains a very good tradeoff, as it only requires few more articles to cover the entities and it even slightly outperforms the story coverage.

Interestingly, while the entity-based coverage works relatively well on the story-coverage (Fig. 2(a)), the story-based coverage performs rather poorly on the entity coverage (Fig. 2(b)). This indicates that while only considering the coverage of all stories and media bias in each story it may still lead to a biased set of articles with repsect to entity stances.

Next, in Table 2 we report the performance of ALG-SR1-LS for covering the three viewpoint collections on all datasets. We ensure that the viewpoint collections are perfectly covered and that the overlap measure score is at least $G_{H^2} \geq 0.9$. We report the minimum budget, stance balance score $B(Q)$, as well as the coverage of the *entity-based* and *story-based* collections. We observe that covering the *entity-based* collection requires the smallest budget to cover all viewpoints, and it achieves the highest balance score in all datasets.

[4] https://aylien.com/



(a) Coverage of stories (b) Coverage of entity stances (c) Stance balance

Figure 2: Performance of ALG-SR1 for covering *story-based*, *entity-based*, and *hybrid* viewpoint collections on NELA-Jan. The y-axis represents the coverage of the *story-based* collection in (a), the coverage of the *entity-based* collection in (b), and the stance balance score $B(Q)$ in (c). The x-axis represents the budget k on the selected number of articles, as a percentage of the total number of articles in the collection.

Table 2: Performance of ALG-SR1-LS for covering different viewpoint collections on all datasets. "Coverage E" and "Coverage $\mathcal{S}$" denote the coverage of the *entity-based* and *story-based* collections, respectively.

Dataset	Collection	k/n (%)	$B(Q)$	Coverage E(%)	Coverage $\mathcal{S}$(%)
COVID-Jan	story-based	7.0	0.46	62	100
	entity-based	2.1	0.69	100	47
	hybrid	9.0	0.58	100	100
NELA-Jan	story-based	5.3	0.63	52	100
	entity-based	4.2	0.86	100	79
	hybrid	7.0	0.85	100	100
WCEP-Small	story-based	11.1	0.72	82	100
	entity-based	5.8	0.82	100	71
	hybrid	12.5	0.78	100	100

Furthermore, it also covers a large proportion (47%–79%) of stories without encoding the story coverage in the objective.

By contrast, covering the *story-based* collection requires a larger budget, up to twice as much, to achieve the desired goal. This can be attributed to the fact that each article belongs to one story, but it may contain multiple entities. However, we observe that it obtains the smallest entity balance score.

Covering the *hybrid* collection achieves perfect coverage for both *entity-based* and *story-based*, while only using few more articles than covering the *story-based* collection. It obtains balance scores close those obtained when covering the *entity-based* collection.

In summary, covering the *story-based* collection risks obtaining biased content by ignoring the stances towards entities. Covering all stories needs potentially large budget. We find that covering the *entity-based* collection is the best practice for calibrated and diverse news coverage if we want to return few articles, as it achieves balanced full coverage of stances and news media bias with a moderate budget and covers a large proportion of news stories (while still satisfying the user preferences). When all stories must be covered, covering the *hybrid* collection provides more balanced results than covering the *story-based*, while using only few more articles.

Q2: Comparison of methods and varying budget. Next, we compare ALG-SR1 and ALG-SR2 against the baselines for varying the budget parameter k. We consider covering the *entity-based* collection on different datasets. The results are reported in Figure 3.

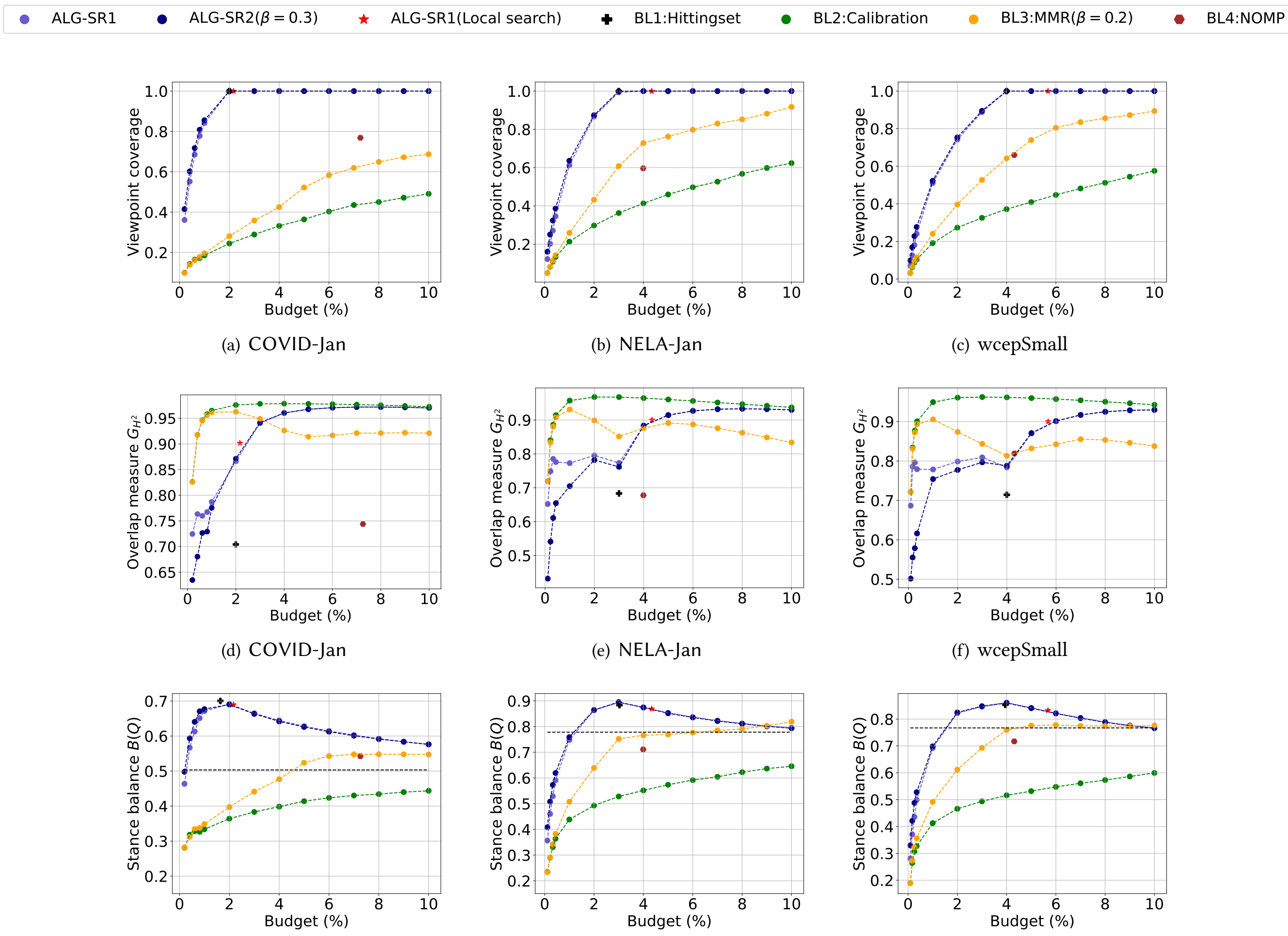
First, we report the result of the coverage rate in Figures 3(a)–(c). Note that full coverage of the viewpoint collection guarantees to cover all possible stances and media biases on each entity, but does not guarantee to cover all stories. As expected, BL1 achieves full coverage of the collection with the smallest budget (2%–4%), while BL2 achieves the lowest coverage. ALG-SR1 and ALG-SR2 perform similarly and achieve full coverage with the same budget as BL1 in all experiments, and they outperform BL2, BL3 and BL4 constantly by a large margin. We observe that initially the coverage increases sharply with k, but then flattens off.

Next, we report the result for the overlap measure G_{H^2} in Figures 3(d)–(f). A higher overlap score indicates that the selected articles are well-calibrated to the user's target distribution τ. As expected, BL2 achieves the highest overlap score (around 0.97 for $k \geq 2\%n$), while BL1 achieves the lowest (around 0.7). ALG-SR1 and ALG-SR2 perform similarly for $k \geq 2\%n$ and show an overall upward trend as the budget increases. For small k, ALG-SR1 and ALG-SR2 behave differently, with ALG-SR1 achieving $G_{H^2} \geq 0.75$ early on.

Next, in Figures 3(g)–(i) we report the entity bias score of the selected articles Q. Notice that although the coverage constraint guarantees to cover all possible stances towards entities, it is possible that the distribution of stances is skewed, e.g., for some entities there might be more articles in favors of them than against. We find that ALG-SR1 and ALG-SR2 have less bias than the bias of the entire set of articles (given by the dashed line) after covering about 50% of the collection. We also observe that as k increases, their bias converges to the average bias (as expected). BL1 and BL4 also achieve good results here, whereas BL2 and BL3 perform worse than our algorithms.

In summary, the results for BL1 and BL2 show a large gap between the goal of unbiased coverage and calibration: well-calibrated articles may lead to biased content, while unbiased articles may fail to satisfy users. ALG-SR1 and ALG-SR2 can effectively balance the goals of providing unbiased coverage and calibration. With the local-search improvement, a budget of just 2%-5% is sufficient to create an unbiased news summary tailored to user preferences ($G_{H^2} \geq 0.9$).

Q3: Exploring the trade off between coverage and calibration. In Figure 4 we run ALG-SR2 with different choices of β. Recall that large values of β emphasize user preferences, whereas small values of β emphasize coverage. As expected, the algorithm's behavior aligns well with the choice of β. Inspecting the two figures for the dataset of interest allows the data scientist to select a value for the



(a) COVID-Jan (b) NELA-Jan (c) wcepSmall

(d) COVID-Jan (e) NELA-Jan (f) wcepSmall

(g) COVID-Jan (h) NELA-Jan (i) wcepSmall

Figure 3: Performance of all algorithms and baselines on different datasets, optimizing for covering the *entity-based* collection. The y-axis in (a)–(c) shows the coverage rate of viewpoint collection, in (d)–(f) the overlap measure G_{H^2}, and in (g)–(i) the stance balance $B(Q)$, where we also report the entity balance score for the entire set as a dashed line. We vary the budget $k \in [0, 0.1]$.

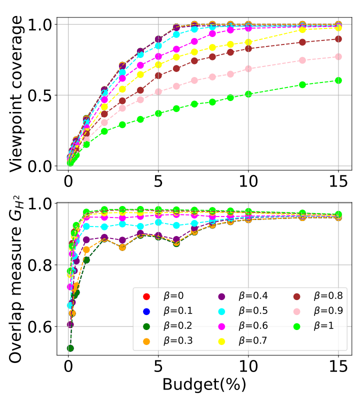


Figure 4: Performance of ALG-SR2 with varying parameter $\beta \in [0, 1]$ on NELA-Jan dataset. The y-axis on the first plot shows the coverage rate of the *hybrid* collection, and the y-axis in the second plot shows the overlap measure score G_{H^2}.

budget k that provides a meaningful trade off between coverage and calibration.

7 CONCLUSIONS

In this paper, we studied the problem of calibrated and diverse news coverage. We built and released an annotation pipeline that takes as input plain-text news articles and automatically identifies news stories, topic distributions, entities, as well as stances and media biases. Using this pipeline we showed that focusing solely on covering all news stories can lead to bias regarding the entity stances. To address this issue, we developed novel algorithms that simultaneously cover all news stories and respect the entity stances, incurring less bias. We believe that our annotation pipeline will be useful to researchers in different fields and that our findings are highly relevant for evaluating and mitigating the bias of news aggregators and recommender systems.

ETHICAL CONSIDERATIONS

In developing algorithms aimed at diversifying news consumption, we acknowledge the inherent duality of technological applications: the same tools designed to broaden perspectives can, paradoxically, be repurposed to narrow them. This duality is not merely theoretical but a realistic risk, evidenced by the manipulation of social media algorithms to foster echo chambers rather than dismantle them. Despite these risks, our work underscores the critical importance of addressing echo chambers, thereby exerting pressure on platforms to continue refining their technologies to support a healthier information ecosystem.

Our operational definition of diversity, focused primarily on political viewpoint diversity, admittedly adopts a narrow scope. This approach is intentional, designed to simplify the complex landscape of news perspectives into a more manageable framework, specifically within the U.S. context characterized by the prevalent left–right divide. This choice facilitates the development and implementation of our methodologies but may not capture the full spectrum of diversity that exists within and across societies. We recognize the limitations of this approach and encourage future research to explore and implement alternative definitions that capture other dimensions of diversity. Moreover, our U.S.-centric approach aims to constrain the problem space to enhance the feasibility of our solutions. However, this geographical and cultural limitation may not directly translate to or be effective in other regions with different political and media landscapes. Acknowledging these limitations is crucial as it frames the applicability and scalability of our findings.

Ultimately, our research seeks to contribute to a broader discourse on the necessity of diverse news consumption. By demonstrating the feasibility of coding and implementing political viewpoint diversity, we hope to inspire the adoption and rigorous real-world testing of these techniques. Such endeavors will be vital for evaluating the efficacy of our methods and understanding their broader implications. Through this work, we aspire to foster an information environment where users are empowered with the choice to access a multiplicity of perspectives, thereby enriching their understanding and engagement with the world.

REFERENCES

[1] Sofiane Abbar, Sihem Amer-Yahia, Piotr Indyk, and Sepideh Mahabadi. Real-time recommendation of diverse related articles. In Daniel Schwabe, Virgílio A. F. Almeida, Hartmut Glaser, Ricardo Baeza-Yates, and Sue B. Moon, editors, *22nd International World Wide Web Conference, WWW '13, Rio de Janeiro, Brazil, May 13-17, 2013*, pages 1–12. International World Wide Web Conferences Steering Committee / ACM, 2013. doi: 10.1145/2488388.2488390. URL https://doi.org/10.1145/2488388.2488390.

[2] Rishi Advani, Paolo Papotti, and Abolfazl Asudeh. Maximizing neutrality in news ordering. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 11–24, 2023.

[3] Mehwish Alam, Andreea Iana, Alexander Grote, Katharina Ludwig, Philipp Müller, and Heiko Paulheim. Towards Analyzing the Bias of News Recommender Systems Using Sentiment and Stance Detection. In *Companion Proceedings of the Web Conference 2022*, pages 448–457, Virtual Event, Lyon France, April 2022. ACM. ISBN 978-1-4503-9130-6. doi: 10.1145/3487553.3524674. URL https://dl.acm.org/doi/10.1145/3487553.3524674.

[4] Anonymous authors. Code and dataset. https://anonymous.4open.science/r/DNA-pipeline-7868.

[5] Dylan Bourgeois, Jérémie Rappaz, and Karl Aberer. Selection bias in news coverage: Learning it, fighting it. In Pierre-Antoine Champin, Fabien Gandon, Mounia Lalmas, and Panagiotis G. Ipeirotis, editors, *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 535–543. ACM, 2018. doi: 10.1145/3184558.3188724. URL https://doi.org/10.1145/3184558.3188724.

[6] Alfred M Bruckstein, Michael Elad, and Michael Zibulevsky. On the uniqueness of nonnegative sparse solutions to underdetermined systems of equations. *IEEE Transactions on Information Theory*, 54(11):4813–4820, 2008.

[7] Sihao Chen, William Bruno, and Dan Roth. Towards corpus-scale discovery of selection biases in news coverage: Comparing what sources say about entities as a start. *CoRR*, abs/2304.03414, 2023. doi: 10.48550/ARXIV.2304.03414. URL https://doi.org/10.48550/arXiv.2304.03414.

[8] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models, 2022. URL https://arxiv.org/abs/2210.11416.

[9] Ryan Evans, Daniel Jackson, and Jaron Murphy. Google news and machine gatekeepers: Algorithmic personalisation and news diversity in online news search. *Digital Journalism*, 11(9):1682–1700, 2023.

[10] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Reducing controversy by connecting opposing views. In *Proceedings of the tenth ACM international conference on web search and data mining*, pages 81–90, 2017.

[11] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Political discourse on social media: Echo chambers, gatekeepers, and the price of bipartisanship. In *Proceedings of the 2018 world wide web conference*, pages 913–922, 2018.

[12] Demian Gholipour Ghalandari, Chris Hokamp, Nghia The Pham, John Glover, and Georgiana Ifrim. A large-scale multi-document summarization dataset from the wikipedia current events portal. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 1302–1308. Association for Computational Linguistics, 2020. doi: 10.18653/V1/2020.ACL-MAIN.120. URL https://doi.org/10.18653/v1/2020.acl-main.120.

[13] Maarten Grootendorst. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[14] Maurício Gruppi, Benjamin D. Horne, and Sibel Adali. NELA-GT-2021: A large multi-labelled news dataset for the study of misinformation in news articles. *CoRR*, abs/2203.05659, 2022. doi: 10.48550/ARXIV.2203.05659. URL https://doi.org/10.48550/arXiv.2203.05659.

[15] Mario Haim, Andreas Graefe, and Hans-Bernd Brosius. Burst of the filter bubble? effects of personalization on the diversity of google news. *Digital journalism*, 6 (3):330–343, 2018.

[16] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. The challenge of understanding what users want: Inconsistent preferences and engagement optimization. *Management science*, 2023.

[17] Jon M. Kleinberg, Emily Ryu, and Éva Tardos. Calibrated recommendations for users with decaying attention. *CoRR*, abs/2302.03239, 2023. doi: 10.48550/arXiv.2302.03239. URL https://doi.org/10.48550/arXiv.2302.03239.

[18] Theodoros Lappas, Mark Crovella, and Evimaria Terzi. Selecting a characteristic set of reviews. In Qiang Yang, Deepak Agarwal, and Jian Pei, editors, *The 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '12, Beijing, China, August 12-16, 2012*, pages 832–840. ACM, 2012. doi: 10.1145/2339530.2339663. URL https://doi.org/10.1145/2339530.2339663.

[19] Huyen Le, Raven Maragh, Brian Ekdale, Andrew High, Timothy Havens, and Zubair Shafiq. Measuring political personalization of google news search. In *The World Wide Web Conference*, pages 2957–2963, 2019.

[20] Bang Liu, Di Niu, Kunfeng Lai, Linglong Kong, and Yu Xu. Growing story forest online from massive breaking news. In *Proceedings of the 2017 ACM on Conference on Information and Knowledge Management, CIKM 2017, Singapore, November 06 - 10, 2017*, pages 777–785. ACM, 2017. doi: 10.1145/3132847.3132852. URL https://doi.org/10.1145/3132847.3132852.

[21] Felicia Loecherbach, Kasper Welbers, Judith Moeller, Damian Trilling, and Wouter Van Atteveldt. Is this a click towards diversity? Explaining when and why news users make diverse choices. In *13th ACM Web Science Conference 2021*, pages 282–290, Virtual Event United Kingdom, June 2021. ACM. ISBN 978-1-4503-8330-1. doi: 10.1145/3447535.3462506. URL https://dl.acm.org/doi/10.1145/3447535.3462506.

[22] Laura Mascarell, Tatyana Ruzsics, Christian Schneebeli, Philippe Schlattner, Luca Campanella, Severin Klingler, and Cristina Kadar. Stance Detection in German News Articles. In *Proceedings of the Fourth Workshop on Fact Extraction and VERification (FEVER)*, pages 66–77, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.fever-1.8. URL https://aclanthology.org/2021.fever-1.8.

[23] Sebastião Miranda, Arturs Znotins, Shay B. Cohen, and Guntis Barzdins. Multilingual clustering of streaming news. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 -*

November 4, 2018, pages 4535–4544. Association for Computational Linguistics, 2018. URL https://aclanthology.org/D18-1483/.

[24] Mats Mulder, Oana Inel, Jasper Oosterman, and Nava Tintarev. Operationalizing Framing to Support Multiperspective Recommendations of Opinion Pieces. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 478–488, Virtual Event Canada, March 2021. ACM. ISBN 978-1-4503-8309-7. doi: 10.1145/3442188.3445911. URL https://dl.acm.org/doi/10.1145/3442188.3445911.

[25] Kevin Munger. All the news that's fit to click: The economics of clickbait media. *Political Communication*, 37(3):376–397, 2020.

[26] Sepideh Nikookar, Mohammadreza Esfandiari, Ria Mae Borromeo, Paras Sakharkar, Sihem Amer-Yahia, and Senjuti Basu Roy. Diversifying recommendations on sequences of sets. *The VLDB Journal*, 32(2):283–304, 2023.

[27] Cornelius Puschmann. Beyond the bubble: Assessing the diversity of political search results. *Digital Journalism*, 7(6):824–843, 2019.

[28] Myrthe Reuver, Nicolas Mattis, Marijn Sax, Suzan Verberne, Nava Tintarev, Natali Helberger, Judith Moeller, Sanne Vrijenhoek, Antske Fokkens, and Wouter Van Atteveldt. Are we human, or are we users? The role of natural language processing in human-centric news recommenders that nudge users to diverse content. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, pages 47–59, Online, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.nlp4posimpact-1.6. URL https://aclanthology.org/2021.nlp4posimpact-1.6.

[29] Ronald E Robertson, Jon Green, Damian J Ruck, Katherine Ognyanova, Christo Wilson, and David Lazer. Users choose to engage with more partisan news than they are exposed to on google search. *Nature*, 618(7964):342–348, 2023.

[30] Todor Staykovski, Alberto Barrón-Cedeño, Giovanni Da San Martino, and Preslav Nakov. Dense vs. sparse representations for news stream clustering. In Alípio Mário Jorge, Ricardo Campos, Adam Jatowt, and Sumit Bhatia, editors, *Proceedings of Text2Story - 2nd Workshop on Narrative Extraction From Texts, co-located with the 41st European Conference on Information Retrieval, Text2Story@ECIR 2019, Cologne, Germany, April 14th, 2019*, volume 2342 of *CEUR Workshop Proceedings*, pages 47–52. CEUR-WS.org, 2019. URL https://ceur-ws.org/Vol-2342/paper6.pdf.

[31] Nava Tintarev, Emily Sullivan, Dror Guldin, Sihang Qiu, and Daan Odjik. Same, Same, but Different: Algorithmic Diversification of Viewpoints in News. In *Adjunct Publication of the 26th Conference on User Modeling, Adaptation and Personalization*, pages 7–13, Singapore Singapore, July 2018. ACM. ISBN 978-1-4503-5784-5. doi: 10.1145/3213586.3226203. URL https://dl.acm.org/doi/10.1145/3213586.3226203.

[32] De-Nian Yang, Hui-Ju Hung, Wang-Chien Lee, and Wei Chen. Maximizing acceptance probability for active friending in online social networks. In *Proceedings of the 19th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 713–721, 2013.

[33] Susik Yoon, Dongha Lee, Yunyi Zhang, and Jiawei Han. Unsupervised story discovery from continuous news streams via scalable thematic embedding. In Hsin-Hsi Chen, Wei-Jou (Edward) Duh, Hen-Hsen Huang, Makoto P. Kato, Josiane Mothe, and Barbara Poblete, editors, *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2023, Taipei, Taiwan, July 23-27, 2023*, pages 802–811. ACM, 2023. doi: 10.1145/3539618.3591782. URL https://doi.org/10.1145/3539618.3591782.