No NLP task should be an island: multi-disciplinarity for diversity in news recommender systems

Reuver, M.; Fokkens, A.; Verberne, S.; Toivonen, H.; Boggia, M.

No NLP Task Should be an Island: Multi-disciplinarity for Diversity in News Recommender Systems

Myrthe Reuver[♡] **Antske Fokkens**[♡♣] **Suzan Verberne**[♢]
♡ CLTL, Dept. of Language, Literature & Communication, Vrije Universiteit Amsterdam
♣ Dept. of Mathematics and Computer Science, Eindhoven University of Technology
♢ Leiden Institute of Advanced Computer Science, Leiden University
{myrthe.reuver,antske.fokkens}@vu.nl,
s.verberne@liacs.leidenuniv.nl

Abstract

Natural Language Processing (NLP) is defined by specific, separate tasks, with each their own literature, benchmark datasets, and definitions. In this position paper, we argue that for a complex problem such as the threat to democracy by non-diverse news recommender systems, it is important to take into account a higher-order, normative goal and its implications. Experts in ethics, political science and media studies have suggested that news recommendation systems could be used to support a deliberative democracy. We reflect on the role of NLP in recommendation systems with this specific goal in mind and show that this theory of democracy helps to identify which NLP tasks and techniques can support this goal, and what work still needs to be done. This leads to recommendations for NLP researchers working on this specific problem as well as researchers working on other complex multidisciplinary problems.

1 Introduction

The field of Natural Language Processing (NLP) uses specific, self-defined definitions for separate tasks – each with their own leaderboards, benchmark datasets, and performance metrics. When dealing with complex, societal problems, it may however be better to take into account a broader view, starting from the actual needs to solve the overall societal problem. In particular, this paper addresses the complex issue of non-diverse news recommenders potentially threatening democracy (Helberger, 2019). We focus on a theory of democracy and its role in news recommendation, as described in Helberger (2019), and reflect on which NLP tasks may help address this issue. In doing so, we consider work by experts on the problem and domain, such as political scientists, recommender system experts, philosophers and media and communication experts.

News recommender systems play an increasingly important role in online news consumption (Karimi et al., 2018). Such systems recommend several news articles from a large pool of possible articles whenever the user wishes to read news. Recommender systems usually attempt to make the recommended articles increase the user's interaction and engagement. In a news recommender system, this typically means optimizing for the individual user's "clicks" or "reading time" (Zhou et al., 2010). These measures are considered a proxy for reader interest and engagement, but other metrics could also be used, including the time spent on a page or article ratings.

Recommender systems are tailored to individual user interests. For other types of recommender systems, e.g. entertainment systems (recommending music or movies), this is less of a problem. However, *news* recommendation is connected to society and democracy, because news plays an important role in keeping citizens informed on recent societal issues and debates (Helberger, 2019). Personalization to user interest in the news recommendation domain can lead to a situation where users are increasingly unaware of different ideas or perspectives on current issues. The dangers of such news 'filter bubbles' (Pariser, 2011) and online 'echo chambers' (Jamieson and Cappella, 2008) due to online (over)personalization have been pointed out before (Bozdag, 2013; Sunstein, 2018).

Political theory provides several models of democracy, which each also imply different roles for news recommendation. We follow the deliberative model of democracy, which states citizens of a functioning democracy need to get access to different ideas and viewpoints, and engage with these and with each other (Manin, 1987; Helberger, 2019) (a further explanation of this model is given in Section 2). A uniform news diet and personalization to only personal interests can, in theory if not

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 45–55
April 19, 2021

in practice, lead to a narrow view on current issues and a lack of deliberation in democracy. When considering this model, it becomes clear that news personalization on user interest alone is potentially harmful for democracy. The normative goal of a recommender system then becomes: supporting a deliberative democracy by showing a diverse set of views to users. NLP can play a role here, by automatically identifying viewpoints, arguments, or claims in news texts. Output of such trained models can help recommend articles that show a diverse set of views and arguments, and thus support a deliberative democracy.

The explicit goals and underlying values of democracy expressed in the model of deliberative democracy can help in defining what NLP tasks and analyses are relevant for tackling the potential harmful effects of news recommendation. This can increase the societal impact of relevant NLP tasks. We believe considering such theories and normative models can also help work on other complex concepts and societal problems where NLP plays a role. In this paper, we outline societal challenges and a theoretical model of the role of non-diverse news recommenders in democracy, as developed by experts such as political scientists and media experts. We then argue that argument mining, viewpoint detection, and related NLP tasks can make a valuable contribution to the effort in diversifying news recommendation and thereby supporting a deliberative democracy.

This position paper provides the following contributions to the discussion: We argue that taking normative and/or societal goals into account can provide insights in the usefulness of specific NLP tasks for complex societal problems. As such, we believe that approaching such problems from an interdisciplinary point of view can help define NLP tasks better and/or increase their impact. In particular, we outline the normative and societal goals for diversifying news recommendation systems and illustrate how these goals relate to various NLP tasks. This results in a discussion on how, on the one hand, news recommendation can make better use of NLP and, on the other hand, how the goal of diversifying news provides inspiration for improving existing tasks or developing new ones.

This paper is structured as follows: We first describe the problem that personalized news recommendation could pose for democracy, as well as the importance of an interdisciplinary approach to solving this problem in Section 2. Section 3 provides an overview of literature tackling diversity in news recommendation as a solution to this problem, and points out remaining gaps in these efforts, specifically connected to the idea of a deliberative democracy. Section 4 outlines several related NLP tasks and their connection to this overarching normative goal. In Section 5, we discuss what we think the NLP community should take away from this reflection, and in Section 6 we will conclude our paper.

2 Personalization in the News, Theories of Democracy, and Interdisciplinarity

The online news domain has increasingly moved towards personalization (Karimi et al., 2018). In the news domain, such personalization comes with specific issues and challenges. A combination of personalizaton and (political) news can lead to polarization, Filter Bubbles (Pariser, 2011), and Echo Chambers (Jamieson and Cappella, 2008). This trend to personalize leads to shared internet spaces becoming much more tailored to the individual user rather than being a shared, public space (Papacharissi, 2002). Such phenomena could negatively impact a citizen's rights to information and right to not be discriminated (Eskens et al., 2017; Wachter, 2020). Evidence for filter bubbles is under discussion (Borgesius et al., 2016; Bruns, 2019), but empirical work does indicate that especially fringe groups holding extreme political or ideological opinions may end up into such a conceptual bubble (Boutyline and Willer, 2017).

Helberger (2019) points out that a lack of diversity in news recommendation can also harm democracy. This clearly holds for the *deliberative* model of democracy. This model assumes that democracy functions on deliberation, and the exchange of points of view. A fundamental assumption in this model is that individuals need access to diverse and conflicting viewpoints and argumentation to participate in these discussions (Manin, 1987). News recommendations supporting a deliberative democracy should then play a role in providing access to these different viewpoints, ideas, and issues in the news (Helberger, 2019).

The threat to democracy of non-diverse news recommenders is a complex problem. It requires input from different academic disciplines, from media studies and computer science to political science and philosophy (Bernstein et al., 2020). Political

theory can provide a framework that helps define what is needed from more empirical and technical researchers to address this problem. In the next section, we will discuss recent work in diversity in news recommendation. We point out remaining gaps in these efforts, specifically connected to the idea of a deliberative democracy.

3 Diversity in News Recommendation

3.1 Recent Diversity Efforts

Previous work on diversity in news recommender systems has mainly focused on assessing the current state of diversity in news recommendation (Möller et al., 2018), or on assessing diversity especially at the end of a computational pipeline, in the form of (evaluation) metrics (Vrijenhoek et al., 2021; Kaminskas and Bridge, 2016), or on computational implementations of diversity (Lu et al., 2020). Less attention has been given to defining and identifying the viewpoints, entities, or perspectives that are being diversified, or to the underlying values and goals of diversification.

Within the recommender systems field, there are several ideas and concepts related to diversity, especially where it concerns evaluation or optimization metrics. Diversity, serendipity, and unexpectedness all are metrics used in the recommender systems literature that go beyond mere click accuracy (Kaminskas and Bridge, 2016). There are two gaps we see in many of these earlier metrics. Firstly, these metrics rarely focus on linguistic or conceptual features or representations of (aspects of) diversity in the news articles. Or, when they do, the NLP approaches are simplified (e.g. topic models in Draws et al. (2020b)) to centralize the recommendation algorithm and its optimization. Secondly, such "beyond user interest" optimization in recommender systems is usually not connected to normative goals and societal gains, but still geared towards user interest and the idea that users react positively to unexpected or previously unseen items. However, several fairly recent works (Lu et al., 2020; Vrijenhoek et al., 2021) have attempted to go beyond "click accuracy" for user interest and tackle the diversity in news recommendation problem while also explicitly considering normative values.

Lu et al. (2020) discuss how to implement "editorial values" in a news recommender for a Dutch online newspaper. Editorial values were defined as journalistic missions or ideals found important by the newspaper's editors and journalists. One of these values is diversity, but their case-study concerns implementing and optimizing for "dynamism" – a diversity-related metric the authors define as "how much a list changes between updates". The authors note the computational difficulty of measuring and optimizing for diversity, and propose a proxy. They define "intra-list diversity" as the inverse of the similarity of a recommendation set. This similarity is calculated over pre-defined news categories of the articles, such as 'sports' and 'finance', as well as over different authors. Viewpoints or perspectives are not mentioned. Lu et al. (2020)'s "editorial values" seem to correspond to the public values mentioned in Bernstein et al. (2020), and implicitly also relate to the democratic values described by Helberger (2019). Both mention diversity as a central important aspect, but Lu et al. (2020) still centralize the user's satisfaction, rather than public values or democracy.

Vrijenhoek et al. (2021) connect several democratic models to computational evaluative metrics of news recommender diversity. The paper discusses several metrics that could be used as optimization and evaluation functions for diversity for news recommender systems supporting a deliberative democracy, such as one to measure and optimize for the "representation" of different societal opinions and voices, and another to measure the "fragmentation": whether different users receive different news story chains. These evaluation metrics are, to our knowledge, the first to explicitly consider normative values and models of democracy in news recommender system design. However, this work does not discuss how to *represent* or *identify* different voices in news articles. The NLP-related components discussed are limited to annotating different named entities.

We argue that the inclusion of more fine-grained and state-of-the-art NLP methods allows more precise identification of different "voices" and viewpoints in support of diverse news recommender systems. The connection of these NLP tasks to diversifying news recommendation is as follows. We compare the building of diverse news recommenders in support of a deliberative democracy to building a tower, with the identification of the different voices or viewpoints as the base of that tower. When an approach can reliably and consistently identify different viewpoints or arguments, we can also diversify these viewpoints in recom-

mendations. A solid definition of viewpoints and reliable methods to detect them thus form the foundation of our diverse news recommendation tower, and builds it towards the goal of a functioning deliberative democracy.

3.2 Technical and Conceptual Challenges

The news is a specific domain for recommender systems, with much faster-changing content than for instance movie or e-commerce recommendation. This leads to a number of unique technical challenges.

Two specific technical and conceptual challenges to a (diverse) news recommendation have been addressed in previous work. The first is the cold start problem (Zhou et al., 2010), which occurs when a news recommender needs data on articles to decide whether to recommend the article to a (new) user. Recommendation, in news as well as in other domains, often uses the interaction data of similar users to recommend data to new users, such as in the method "collaborative filtering". Such data is missing on the large volumes of new articles added in the news domain every day, which makes such approaches less useful in this domain. This leads to other recommendation techniques being more common in the news recommendendation domain.

The second challenge specific to our problem is the continuous addition of new and many different topics, issues, and entities in public discussion and in the news. This makes detecting viewpoints with one automated, single model and one set of training data difficult. Previous work often explores one well-known publicly debated topic, such as abortion (Draws et al., 2020a) or misinformation related to COVID-19 (Hossain et al., 2020). However, in an ideal solution we would also be able to continuously identify all kinds of new debates and related views.

We believe that a combination of state-of-the-art NLP techniques such as neural language models can help address this problem without resorting to manual or unsupervised techniques. A possible interesting research direction is zero-shot or one-shot learning as in Allaway and McKeown (2020), where a model with the help of large(-scale) language models learns to identify new debates and viewpoints not seen at training time. In our case, this would mean identifying new debates and new viewpoints without explicit training on these when training for our task. We elaborate on potentially useful NLP tasks to focus on for our problem in the following section.

4 Relevant NLP Tasks

Within the NLP, text mining, and recommender systems literature, there are several (related) tasks that deal with identifying viewpoints, perspectives, and arguments in written language. We define a task in NLP as a clearly defined problem such as "stance detection", with each task having connected methods, benchmark datasets, leaderboards and literature. The literature is currently fragmented in different related tasks and also definitions of viewpoint, argument or claim, and perspective. Researchers also use different datasets and content-types (tweets and microblogs, internet discussions on websites like debate.org, or news texts).

In this section we discuss NLP tasks that are related to viewpoint and argumentation diversity as defined in relation to the normative goal of a healthy deliberative democracy. Recall that a deliberative model assumes that participants of a democracy need access to a variety of (conflicting) viewpoints and lines of argumentation. As such, we focus on NLP tasks that help identify what claims, stances, and argumentation are present in news articles, and how specific items in the news are presented or framed.

An important distinction that needs to be made is the one between *stance* and *sentiment*: a negative sentiment does not necessarily mean a negative stance or viewpoint on an issue, and vice versa. An example would be someone who supports the use of mouth masks as COVID-19 regulation (positive stance), and expresses negative sentiment towards the topic by criticizing the shortage of mouth masks available for caregivers. In this paper, we concern ourselves with stance *on* issues (being in favor of masks) rather than with sentiment expressed *about* such issues (being negative about their shortage).

The remainder of this section is structured as follows. We first describe work on recommender systems that explicitly refers to detecting viewpoints. We then address three relatively established NLP tasks: argumentation mining, stance detection and polarization, frames & propaganda. We then briefly address work that refers to 'perspectives'.

4.1 Viewpoint Detection and Diversity

The recommender systems literature specifically uses the term 'viewpoint' in relation to diversifying

recommendation. In these viewpoint-based papers, we notice a systems-focused tendency. Defining a viewpoint is less of a concern, nor is evaluating the viewpoint detection. Instead, researchers centralize viewpoint *presentation* to users, or how these respond to more diverse news, as in Lu et al. (2020) and Tintarev (2017). As a result, there is no standard definition of 'viewpoint' and the concept is operationalized differently by various authors.

Draws et al. (2020a) use topic models to extract and find viewpoints in news texts with an unsupervised method, with the explicit goal to diversify a news recommender. They explicitly connect different *sentiments* to different *viewpoints or perspectives*. For this study, they use clearly argumentative text on abortion from a debating website. The words 'viewpoint' and 'perspective' are used interchangeably in this study.

Carlebach et al. (2020) also address what they call "diverse viewpoint identification". Here as well, we see a wide range of definitions and terms related to viewpoints and perspectives (e.g. 'claim', 'hypothesis', 'entailment'). The authors use state-of-the-art methods including large neural language models, but the study does not seem to consider carefully defining their task, term definitions, and the needs of the problem. As such, it is unclear what they detect exactly. This is mainly due to the detection itself not being the main focus of their paper.

With the more NLP-based tasks and definitions in the following sections, we explore how NLP tasks relate to this 'viewpoints' idea from the recommender systems community, and see what ideas and techniques these other tasks can add to diversity in news recommendation.

4.2 Argument Mining

Argument Mining is the automatic extraction and analysis of specific units of argumentative text. It usually involves user-generated texts, such as comments, tweets, or blogposts. Such content is often highly argumentative by design, with high sentiment scores. In some studies, arguments are related to stances, as in the Dagstuhl ArgQuality Corpus (Wachsmuth et al., 2017), where 320 arguments cover 16 (political or societal) topics, and are balanced for different stances on the same topic. These arguments are from websites specifically aimed at debating.

Stab and Gurevych (2017) identify the different sub-tasks in argumentation mining, and use essays as the argumented texts in question. For instance, one sub-task is separating argumentative from non-argumentative text units. Then, their pipeline involves classifying argument components into claims and premises, and finally it involves identifying argument relations. This first sub-task is also sometimes called *claim detection*, and is related to detecting stances and viewpoints when connecting claims to issues.

For a deliberative democracy, the work on distinguishing argumentative from non-argumentative text in argument mining is useful, since our goal requires the highlighting of deliberations and arguments, and not statements on facts. Identifying this distinction might enable us to identify viewpoints in news texts. The precise identification of claims and premises may also prove valuable, because supporting a deliberative democracy requires the detection of different deliberations and arguments in news texts.

4.3 Stance Detection

Stance detection is the computational task of detecting "whether the author of the text is in favor of, against, or neutral towards a proposition or target" (Mohammad et al., 2017, p. 1). This task usually involves social media texts and, once again, user-generated content. Commonly, these are shorts texts such as tweets. For instance, Mohammad et al. (2017) provide a frequently used Twitter dataset that strongly connects stances with sentiment and/or emotional scores of the text. Another common trend in stance detection is to use text explicitly written in the context of an (online) debate, such as the website debate.org and social media discussions.

A recent study on Dutch social media comments highlights the difficulties in annotating stances on vaccination (Bauwelinck and Lefever, 2020). The authors identify the need to annotate topics, but also topic aspects and whether units are expressing an argument or not. Getting to good inter-annotator agreement (IAA) is difficult, showing that these concepts related to debate and stance are not uniform to all annotators even after extensive training. The same is found by Morante et al. (2020): Annotating Dutch social media text as well as other debate text on the vaccination debate, they find obtaining a high IAA is no easy task.

Other work related to stance detection is more related to the news domain. The Fake News Classification Task (Hanselowski et al., 2018b) has a sub-task that concerns itself with predicting the stance of a news article towards the news headline. In their setup stances can be 'Unrelated', 'Discuss', 'Agree' or 'Disagree'. The Fake News Classification tasks also introduces *claim verification* as a sub-task. This task is also related to the claim detection task: in order to verify claims, one needs to detect them first.

Several papers specifically aim at stance detection in the news domain. Conforti et al. (2020) note that different types of news events, from wars to economic issues, might lead to stance classes that are not uniform across events. As a response, they decide to annotate stance on one specific type of news event: company acquisitions. The authors explicitly note here that textual entailment and sentiment analysis are different tasks from stance detection, but acknowledge that all these tasks are related. However, as stated before, in the news domain new topics or issues occur constantly. Data on only one type of news event is less representative of all texts in the news domain. Some recent work aims to address this through one-shot or zero-shot learning for detecting issues and viewpoints on issues (Allaway and McKeown, 2020). In such an approach, unseen topics or viewpoints would be detected even when they are very different from what is annotated or seen at training time.

Based on the above, there are three challenges involved in applying previous approaches on stance detection for diversifying news: First, most work on stance detection aims at short, high-sentiment user-generated texts with one specific stance. News articles are more complex. News texts might highlight a debate with several viewpoints of different people, with the emphasis on one rather than the other. Secondly, the authors of news articles generally do not express opinions explicitly, unlike authors of tweets or blogs. News articles can express viewpoints in more subtle ways, in the way a story is told or framed. Additionally, training data that does come from the news domain may not generalize well to new topics.

We conclude that stance detection is, in principle, a relevant task when aiming to ensure news recommendation supports a deliberative democracy, but the challenges generalizing to new topics and dealing with more subtle ways of expressing viewpoints must be addressed. One shot learning may provide means to deal with new topics in the every-changing news landscape. The focus on longer, less explicitly argumentative text is helpful for our goal, and exists in for instance the first subtasks of fake news detection (Hanselowski et al., 2018a) and other recent news-focused datasets and papers (Conforti et al., 2020; Allaway and McKeown, 2020).

4.4 Polarization, Frames, and Propaganda

Some work already explicitly takes into account the more complex political dimension of news texts when defining an NLP task. This work is often interdisciplinary in nature, with NLP researchers working with political scientists or media scholars. The idea of (political) perspectives is prominent in these papers, though researchers in this subfield use different definitions and names for similar tasks.

'Frames', 'propaganda', and 'polarization' are loaded terms, with less nuance than terms such as 'stance' and 'argument'. Terms like 'polarization' are (ironically) more polarizing due to their political connotations. An explicitly political aspect in the task definition can be useful for our societal problem – as stated, the deliberative democracy goal is also inherently connected to political debates. However, it can also lead to a confusion of terminology or the use of (accidentally) loaded terminology, for instance terms that are controversial in related disciplines such as communication science or media studies.

An example is a recent shared task on Propaganda techniques (Da San Martino et al., 2019). It distinguishes 18 classes of what the authors call 'rhetorical strategies' that are not synonymous with, but related to, propaganda. These include 'whataboutism', 'bandwagon', and 'appeal to fear and prejudice', as well as 'Hitler-comparisons'. These terms are, incidentally, also known as cognitive biases (the bandwagon effect) or framing (appeal to fear) and argumentation flaws (Hitler-comparisons, on the internet known as Godwin's Law). Such confusion of terminology, especially in a politically sensitive context, makes it less straightforward to see how this task can be used for viewpoint diversification in support of a deliberative democracy.

Sometimes, the task of identifying different viewpoints on an issue or event in the news is translated to 'political bias'. In such work, the

viewpoints are related to a certain ideology or political party (Roy and Goldwasser, 2020) or 'media frames'. However, we would argue that a viewpoint in the public debate does not have to be a political standpoint related to a specific political ideology. Limiting ourselves only to detecting debates and viewpoints explicitly related to political parties would also limit the view on public debate and deliberative democracy, and thus would not support our normative goal to its full extent.

Other NLP work that addresses the political nature of news texts and perspectives is Fokkens et al. (2018). In this work, stereotypes on Muslims are detected with a self-defined method known as 'micro-portrait extraction'. This paper is an example of work where other disciplines (communication and media experts) are heavily involved in task definition and execution, aiding clear and careful definitions and aiding to the problem and the societal complex issue (stereotypes in the news) at hand.

'Fake news' related tasks are also connected to the political content of news. The Fake News Classification Task (Hanselowski et al., 2018b) has the explicit goal to identify fake news. It consists of several sub-tasks related to argument mining and stance detection. The debate on (fake) news has recently shifted away from the simple label 'fake news', since it is not only the simple distinction between fake and true that is interesting. This again shows the importance of multi-disciplinary work: computational tasks are often aimed at a simple classification such as 'true' versus 'false', while social scientists and media experts call for different labels not directly related to the truth of an entire article or claim, such as 'false news', 'misleading news', 'junk news' (Burger et al., 2019), or 'clickbait'. All these are terms for a media diet with lower quality (or with less 'editorial values' to use the term from Lu et al. (2020)).

It can be useful for a deliberative democracy-supporting diverse news recommender when tasks already incorporate the political dimension of news texts. However, it can also be harmful when the political or social science definitions are not clear and uniform, or when the political dimension actually narrows what a deliberative democracy is by only considering explicitly political viewpoints, or only views tied to political parties or ideologies.

4.5 Perspectives

In NLP, definitions of 'perspective' range from 'a relation between the source of a statement (i.e. the author or another entity introduced in the text) and a target in that statement (i.e. an entity, event, or (micro-)proposition)' (Van Son et al., 2016) to stances to specific (political) claims in text (Roy and Goldwasser, 2020). These definitions are similar to those seen in the Stance Detection literature. Sometimes, it is unclear what the difference is between a stance and a perspective.

Common debate content used for analysis and task definition of perspectives is political elections (Van Son et al., 2016), vaccination (Morante et al., 2020), and also societally debated topics like abortion. Perspectives are especially useful for our goal, since they assume different groups in society are seeing one issue from different angles. This allows us to identify an active debate in society, which explicitly supports a deliberative democracy.

5 Discussion

In the previous section, we have outlined a number of relevant NLP tasks, and made their possible contribution to the support of a deliberative democracy through diverse news recommendation explicit. In the following section, we discuss the implications and considerations following from these separate tasks for diversity in news recommendations, and provide some advice for NLP researchers.

5.1 Evaluation

There has been a general push in NLP evaluation to go "beyond accuracy" (Ribeiro et al., 2020) and in recommender systems to go "beyond click accuracy" (Lu et al., 2020; Zhou et al., 2010) in evaluation and optimization. We believe that going beyond these evaluations might also mean looking at normative, societal goals and values, and the implications for the task and its effect on these goals and values. A possible advantage of a higher-level evaluation with a normative goal is that it allows the measurement of real-world impact. One explicit problem however is how to evaluate whether support of a deliberative democracy has been achieved.

Recent work by Vrijenhoek et al. (2021) has identified evaluation metrics to evaluate whether a recommender system supports specific models of democracy, one of which is the deliberative model. They propose a number of evaluation metrics for recommender system diversity that are explicitly

connected to different models of democracy. These metrics could be used to evaluate different aspects of diversity related to a (deliberative) democracy. The aspects discussed are the representation of different groups in the news, whether alternative voices from minority groups are represented in the recommendations, whether the recommendations activate users to take action, and the degree of fragmentation between different users.

However, Vrijenhoek et al. (2021) does not address the evaluation of the NLP tasks involved. Where specific, clearly defined NLP tasks can generally be evaluated through hand-labelled evaluation sets, such sets do not provide the necessary insights to determine their role in supporting a deliberative democracy. In the end, we need to find a way to connect accuracy of NLP technologies to the overall increased diversity of news offers. Ideally, we would then also measure the ultimate impact on the users of a diverse recommender system diversifying viewpoints or stances with an NLP method. Such an evaluation is highly complex and clearly requires expertise from various fields (including technology, user studies and methods for investigating social behavior). It could for instance involve longitudinal studies on user knowledge of issues and viewpoints.

5.2 No NLP Task is An Island

We argue that NLP tasks have a clear role in the development of diverse recommender systems. Especially recent developments in the field, such as the use of pre-trained language models and neural models, could be used to obtain a reliable and useful representations of issues in the news, as well as viewpoints and perspectives on these issues. Such approaches are possibly more fine-grained and can be more reliable than the now commonly used unsupervised methods such as topic models.

Benchmarking with separate datasets, definitions, and shared tasks and challenges has brought our field far, and much progress has been achieved in this manner. However, we feel complex societal issues should be aimed at achieving a societal goal rather than evaluated on task-specific benchmarking dataset. When considering issues such as diversity in news recommendation and its effects on democracy and public debate, we are at the limit of what separate NLP tasks could bring us. We should dare to look past the limits of separate tasks, and attempt to oversee the over-arching normative goals and tasks related to such problems, especially when working on real-world impact.

As discussed in Section 4, the NLP field has many related tasks that seem to be relevant to the problem of news recommender diversity and especially the support of a deliberative democracy. However, we note that NLP tends to use their own definitions, and not consider other fields or even sub-fields, when designing these tasks. This means the field covers a wide array of different implementations and definitions related to perspectives and viewpoints in the news. We therefore urge NLP researchers to not only consider and evaluate their systems on their own definitions and tasks, but also consider the wider societal and normative goals their task connects to, and what other related tasks could be used to achieve the same or similar goals.

5.3 NLP and Other Disciplines

NLP, especially NLP working on societal real-world problems, should involve other fields, and expertise in other fields. This is especially true when working on complex problems like viewpoint diversity in news recommendation. This recommendation has also been made at the Dagstuhl perspectives workshop "Diversity, fairness, and data-driven personalization in (news) recommender systems" (Bernstein et al., 2020), but we would like to emphasize it more specifically for the NLP field.

One example where a lack of interdisciplinary seems to sometimes to lead to issues for our problem is in the Polarization, Frames, and Propaganda set of NLP tasks outlined in Section 4.4. Definitions used of 'frame', 'propaganda', and 'polarization' are sometimes seemingly made without consulting relevant experts, or without considering earlier theoretical work defining these terms. This leads to definitions that are easy to computationally measure with existing NLP techniques, such as classification. However, these definitions do not necessarily do justice to the complex problem the model or task is aimed at. Such work also does not consult earlier theoretical and empirical considerations of these terms and definitions.

We argue for the inclusion of experts from the social sciences and humanities in every step of the process – designing the tasks and definitions, evaluation of task success and usefulness, and tying the result to broader implications. For diversity in news recommenders, this means discussing and engaging with experts on political theory and philosophy,

ethics of technology, and media studies and communication science (Bernstein et al., 2020).

5.4 Ethical and Normative Considerations

When our goal is to foster a healthy democratic debate, we should consider whether we should highlight or recommend content with fringe opinions that might be dangerous to individuals or the debate itself, e.g. the anti-vaxxing argument in the vaccination debate, conspiracy theories on the state of democracy, or inherently violent arguments. The deliberative model of democracy values rational and calm debate, not emotional or affective language. While this is a question of whether to *recommend* such views, not whether to *detect* them, we find it important to stress such considerations here. In a complex problem with a high-level normative goal, it is important to make such considerations explicit, as these also influence whether we are actually fostering a healthy deliberative debate. This means a simple computational solution, e.g. *maximize diversity of viewpoints and debates*, might not always be the best manner to reach the normative goal (e.g. *foster a healthy deliberative democracy*).

Such more nuanced and complex issues come to light when we consider public values such as diversity and the normative goal of a deliberative democracy. They are less explicit when only considering the NLP task as a separate task, which only needs to be evaluated by its performance on a benchmark dataset. However, questions such as these are especially important when considering that NLP and its technology is contributing to the solution of a societal problem. The attention to an over-arching normative goal helps NLP researchers to consider their responsibility and the implications of their work when it is used in real-world settings. This has been argued before by researchers in the NLP community (Fokkens et al., 2014; Bender et al., 2021), and we think it is a positive development when NLP researchers consider the wider ethical and normative considerations of their tasks and goals.

6 Conclusion

In this paper, we have provided an overview of several separate NLP tasks related to news recommender system diversity, especially considering the normative goal of a deliberative democracy. An explicit incorporation of such over-arching normative goals is currently missing in these tasks, while this is conceptually very useful and societally relevant. As such, taking this end goal into account can help improve social relevance of NLP and support NLP researchers in defining specific goals and next steps in their research.

Research on recommendation systems could benefit from more specific work that operationalizes the theoretical concepts in democratic theory. Such operationalizations should start with the groundwork laid by NLP tasks such as stance detection, argumentation mining and tasks aiming at detecting frames, propaganda and polarization. However, current NLP tasks do not address problems related to viewpoint diversity in news recommendation in its full complexity yet. NLP should take the complexities of news and the news recommendation domain into account. News texts often contain more than one stance or argument, and they tend to have more implicitly expressed viewpoints than other texts. Moreover, news comes with the challenge that new topics constantly appear and training data on detecting viewpoints in some issues may not generalize well to new data on other topics or issues.

This leads us to the following two concrete steps for future work, specifically in NLP: (1) researchers should further advance methods that aim to identify more subtle ways in which viewpoints occur in real-world news text; (2) methods should address the issue of constant changes in data, with one possible solution being one-shot learning. Last but not least, in order to find out how these tasks can truly be used to improve a deliberative democracy, we face the challenge of evaluating beyond assigning correct labels to pieces of text. This brings us back to the main message of this paper: Answering this question goes beyond the expertise of NLP researchers. In order to maximize the impact of our technologies for addressing this complex problem, we need expertise from other disciplines.

Acknowledgments

This research is funded through Open Competition Digitalization Humanities and Social Science grant nr 406.D1.19.073 awarded by the Netherlands Organization of Scientific Research (NWO). We would like to thank our interdisciplinary team members, and the anonymous reviewers whose comments helped improve the paper. All opinions and remaining errors are our own.

References

Emily Allaway and Kathleen McKeown. 2020. Zero-shot stance detection: A dataset and model using generalized topic representations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 8913–8931.

Nina Bauwelinck and Els Lefever. 2020. Annotating topics, stance, argumentativeness and claims in dutch social media comments: A pilot study. In *Proceedings of the 7th Workshop on Argument Mining*, pages 8–18.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big. *Proceedings of FAccT*.

Abraham Bernstein, Claes De Vreese, Natali Helberger, Wolfgang Schulz, and Katharina A Zweig. 2020. Diversity, fairness, and data-driven personalization in (news) recommender system (dagstuhl perspectives workshop 19482).

Frederik J Zuiderveen Borgesius, Damian Trilling, Judith Moller, Balázs Bodó, Claes H De Vreese, and Natali Helberger. 2016. Should we worry about filter bubbles? *Internet Policy Review*, 5(1).

Andrei Boutyline and Robb Willer. 2017. The social structure of political echo chambers: Variation in ideological homophily in online networks. *Political psychology*, 38(3):551–569.

Engin Bozdag. 2013. Bias in algorithmic filtering and personalization. *Ethics and information technology*, 15(3):209–227.

Axel Bruns. 2019. *Are filter bubbles real?* John Wiley & Sons.

Peter Burger, Soeradj Kanhai, Alexander Pleijter, and Suzan Verberne. 2019. The reach of commercially motivated junk news on facebook. *PloS one*, 14(8):e0220446.

Mark Carlebach, Ria Cheruvu, Brandon Walker, Cesar Ilharco Magalhaes, and Sylvain Jaume. 2020. News aggregation with diverse viewpoint identification using neural embeddings and semantic understanding models. In *Proceedings of the 7th Workshop on Argument Mining*, pages 59–66.

Costanza Conforti, Jakob Berndt, Mohammad Taher Pilehvar, Chryssi Giannitsarou, Flavio Toxvaerd, and Nigel Collier. 2020. Stander: An expert-annotated dataset for news stance detection and evidence retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: Findings*, pages 4086–4101.

Giovanni Da San Martino, Alberto Barrón-Cedeño, and Preslav Nakov. 2019. Findings of the NLP4IF-2019 shared task on fine-grained propaganda detection. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 162–170, Hong Kong. Association for Computational Linguistics.

Tim Draws, Jody Liu, and Nava Tintarev. 2020a. Helping users discover perspectives: Enhancing opinion mining with joint topic models. In *Proceedings of SENTIRE'20*.

Tim Draws, Nava Tintarev, Ujwal Gadiraju, Alessandro Bozzon, and Benjamin Timmermans. 2020b. Assessing viewpoint diversity in search results using ranking fairness metrics. In *Informal Proceedings of the Bias and Fairness in AI Workshop at ECML-PKDD (BIAS 2020)*.

Sarah Eskens, Natali Helberger, and Judith Moeller. 2017. Challenged by news personalisation: five perspectives on the right to receive information. *Journal of Media Law*, 9(2):259–284.

Antske Fokkens, Serge ter Braake, Niels Ockeloen, Piek Vossen, Susan Legêne, and Guus Schreiber. 2014. Biographynet: Methodological issues when nlp supports historical research. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 3728–3735.

Antske Fokkens, Nel Ruigrok, Camiel Beukeboom, Sarah Gagestein, and Wouter van Atteveldt. 2018. Studying muslim stereotyping through microportrait extraction. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*.

Andreas Hanselowski, PVS Avinesh, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M Meyer, and Iryna Gurevych. 2018a. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874.

Andreas Hanselowski, Avinesh PVS, Benjamin Schiller, Felix Caspelherr, Debanjan Chaudhuri, Christian M. Meyer, and Iryna Gurevych. 2018b. A retrospective analysis of the fake news challenge stance-detection task. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1859–1874, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Natali Helberger. 2019. On the democratic role of news recommenders. *Digital Journalism*, 7(8):993–1012.

Tamanna Hossain, Robert L Logan IV, Arjuna Ugarte, Yoshitomo Matsubara, Sean Young, and Sameer Singh. 2020. Covidlies: Detecting covid-19 misinformation on social media. In *Proceedings of the 1st Workshop on NLP for COVID-19 (Part 2) at EMNLP 2020*.

Kathleen Hall Jamieson and Joseph N Cappella. 2008. *Echo chamber: Rush Limbaugh and the conservative media establishment*. Oxford University Press.

Marius Kaminskas and Derek Bridge. 2016. Diversity, serendipity, novelty, and coverage: a survey and empirical analysis of beyond-accuracy objectives in recommender systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 7(1):1–42.

Mozhgan Karimi, Dietmar Jannach, and Michael Jugovac. 2018. News recommender systems–survey and roads ahead. *Information Processing & Management*, 54(6):1203–1227.

Feng Lu, Anca Dumitrache, and David Graus. 2020. Beyond optimizing for clicks: Incorporating editorial values in news recommendation. In *Proceedings of the 28th ACM Conference on User Modeling, Adaptation and Personalization*, pages 145–153.

Bernard Manin. 1987. On legitimacy and political deliberation. *Political theory*, 15(3):338–368.

Saif M Mohammad, Parinaz Sobhani, and Svetlana Kiritchenko. 2017. Stance and sentiment in tweets. *ACM Transactions on Internet Technology (TOIT)*, 17(3):1–23.

Judith Möller, Damian Trilling, Natali Helberger, and Bram van Es. 2018. Do not blame it on the algorithm: an empirical assessment of multiple recommender systems and their impact on content diversity. *Information, Communication & Society*, 21(7):959–977.

Roser Morante, Chantal Van Son, Isa Maks, and Piek Vossen. 2020. Annotating perspectives on vaccination. In *Proceedings of The 12th Language Resources and Evaluation Conference*, pages 4964–4973.

Zizi Papacharissi. 2002. The virtual sphere: The internet as a public sphere. *New media & society*, 4(1):9–27.

Eli Pariser. 2011. *The filter bubble: What the Internet is hiding from you*. Penguin UK.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Shamik Roy and Dan Goldwasser. 2020. Weakly supervised learning of nuanced frames for analyzing polarization in news media. *EMNLP Findings*.

Christian Stab and Iryna Gurevych. 2017. Parsing argumentation structures in persuasive essays. *Computational Linguistics*, 43(3):619–659.

Cass R Sunstein. 2018. *# Republic: Divided democracy in the age of social media*. Princeton University Press.

Nava Tintarev. 2017. Presenting diversity aware recommendations: Making challenging news acceptable. In *The FATREC Workshop on Responsible Recommendation*.

Chantal Van Son, Tommaso Caselli, Antske Fokkens, Isa Maks, Roser Morante, Lora Aroyo, and Piek Vossen. 2016. GRaSP: A multilayered annotation scheme for perspectives. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 1177–1184.

Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a mission: assessing diversity in news recommendations. In *SIGIR Conference on Human Information Interaction and Retrieval (CHIIR) Proceedings*.

Henning Wachsmuth, Nona Naderi, Yufang Hou, Yonatan Bilu, Vinodkumar Prabhakaran, Tim Alberdingk Thijm, Graeme Hirst, and Benno Stein. 2017. Computational argumentation quality assessment in natural language. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 176–187.

Sandra Wachter. 2020. Affinity profiling and discrimination by association in online behavioural advertising. *Berkeley Technology Law Journal*, 35(2).

Tao Zhou, Zoltán Kuscsik, Jian-Guo Liu, Matúš Medo, Joseph Rushton Wakeling, and Yi-Cheng Zhang. 2010. Solving the apparent diversity-accuracy dilemma of recommender systems. *Proceedings of the National Academy of Sciences*, 107(10):4511–4515.