Implementing Evaluation Metrics Based on Theories of Democracy in News Comment Recommendation (Hackathon Report)

Myrthe Reuver
CLTL
Dept. of Language, Literature & Communication
Vrije Universiteit Amsterdam
myrthe.reuver@vu.nl

Nicolas Mattis
Dept. of Communication Science
Faculty of Social Science
Vrije Universiteit Amsterdam
n.m.mattis@vu.nl

Abstract

Diversity in news recommendation is important for democratic debate. Current recommendation strategies, as well as evaluation metrics for recommender systems, do not explicitly focus on this aspect of news recommendation. In the 2021 Embeddia Hackathon, we implemented one novel, normative theory-based evaluation metric, "activation", and use it to compare two recommendation strategies of New York Times comments, one based on user likes and another on editor picks. We found that both comment recommendation strategies lead to recommendations consistently less activating than the available comments in the pool of data, but the editor's picks more so. This might indicate that New York Times editors' support a deliberative democratic model, in which less activation is deemed ideal for democratic debate.

1 Introduction

Recommender systems are a core component of many online environments. Such systems can be used to recommend movies or music to users where there is a large pool of potential recommendations. Their main task, as Karimi et al. (2018) put it, is "to filter incoming streams of information according to the users' preferences or to point them to additional items of interest in the context of a given object" (p. 1203). As such, they are usually designed in ways that maximise user satisfaction. Their performance is traditionally evaluated in terms of their "accuracy", which is often measured by proxies such as clicks, time spent on a page, or engagement. Simply put: the more attention a user pays to the content, the better the recommender system is deemed to be.

However, there is an increasing awareness in the recommender systems domain that "beyond-accuracy" metrics such as diversity or novelty are also important aspects of a meaningful recommender system evaluation (Raza and Ding, 2020; Kaminskas and Bridge, 2016). This is particularly true in contexts where the impact of recommendations extends beyond individual purchasing choices or movie selections, such as news recommendation. Given that exposure to diverse viewpoints is often regarded as beneficial for democratic societies (Helberger and Wojcieszak, 2018), scholars have recently highlighted the importance of exposure diversity in such systems (Helberger, 2019; Helberger et al., 2018). Not recommending diversity in news recommender systems could potentially lead to 'filter bubbles', where users only receive ideas and viewpoints they already know and/or agree with (Pariser, 2011).

Very recently, evaluation and optimization metrics by Vrijenhoek et al. (2021) have been specifically designed to align with potential goals of democratic news recommenders as suggested by Helberger (2019). As such, they move beyond the existing "beyond accuracy" evaluation metrics used in the recommender system field. These existing metrics range from "diversity", to "serendipity", "novelty", and "coverage" (Kaminskas and Bridge, 2016), but all of these implicitly aim at increasing user satisfaction rather than achieving normative goals.

In contrast, the metrics in Vrijenhoek et al. (2021) are explicitly linked to supporting democratic debate rather than user satisfaction. Specifically, these metrics are linked to models of democracy. One of these is the deliberative model of democracy, which states a functioning democracy consists of rational debate of viewpoints and ideas. Another model is the critical model, which contends a successful democracy has clashing and active debates of opposing viewpoints.

In this paper, we specifically focus on one of these metrics, "activation", and use it to evaluate

Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation, pages 134–139

two different recommendation strategies for New York Times user comments in response to news articles. In doing so, our goal is to explore the potential of, but also the challenges related to, such normative metrics, especially where it concerns Natural Language Processing (NLP) tools and strategies.

To better understand how different recommendation strategies in the NYT comment section perform in terms of this metric, we ask the following research question: *"How do different manners of recommending user comments on a news article affect the recommendation set's average activation scores?*

By comparing different comment recommendation strategies, we contribute to the ongoing discussion in three ways:

- We are the first, to our knowledge, to implement Vrijenhoek et al. (2021)'s evaluation metrics for democratic news recommenders on a dataset;
- We explicitly identify possibilities and problems related to NLP in the use of such metrics;
- We add to the literature on the deliberative value of user-comments as well as on editorial biases in comment selection.

Our goal was to "test-drive" one or more of the theory-driven evaluation metrics in Vrijenhoek et al. (2021), and see where we ran into conceptual or practical problems preventing us from answering a research question aimed at comparing different recommendation strategies on the basis of this metric.

2 Method

2.1 Dataset

Although not exactly the same as news articles in a news recommender system, user comments are particularly interesting in this context because of their deliberative implications. That is, they provide a public space where users can share, consume and engage with different ideas and viewpoints (Rowe, 2015). As such, they constitute an excellent context for the test of Vrijenhoek et al. (2021)'s activation metric.

The dataset (Kesarwani, 2018), one of the datasets linked to in the hackathon resources (Pollak et al., 2021), contains 9.450 articles with 2.176.364 comments and other related metadata from the New York Times. The articles were published from January 2017 to May 2017 and January 2018 to May 2018. The mean number of comments per article is 230, with an SD of 403.4.

The comment data set contains the text and timestamps of the individual comments, as well as unique identifiers for each comment and the article that it belongs to. In addition, for each comment it also contains the number of user likes (called "recommendations") as well as information on whether or not the comment was selected by the NYTimes editorial board. According to their website, "NYT Picks are a selection of comments that represent a range of views and are judged the most interesting or thoughtful. In some cases, NYT Picks may be selected to highlight comments from a particular region, or readers with first-hand knowledge of an issue." (Sta) In most cases, the editors select 1 comment per debate, but the spread is large, with the mean being 13 recommended comments per article (SD = 11).

2.2 Two recommendation strategies

We recommend the top 3, top 5, and top 10 comments for each news article in two ways:

- N most-liked by users
- N editorial recommendations (in order of appearance)

We also considered comparing these two recommendation strategies to maximizing intra-list diversity based on a representation with Google News word embeddings, but ran out of time to do so. This strategy is based on Lu et al. (2020), who use this strategy to implement the "editorial value" diversity.

We compare these strategies with the evaluation metric "activation" from Vrijenhoek et al. (2021). We then analyze what the different levels of Activation in different recommendation strategies say about the implicit support for the different democratic models outlined in Helberger (2019). A higher activation might indicate an implicit support of the *critical* model of democracy, where conflict needs to be emphasized in order to obtain a lively, healthy debate. A lower activation score might indicate an implicit support of the *deliberative* model of democracy, where rational and calm debate is deemed important for democratic debate.

2.3 Test and validation sets

In order to test our approaches, we used two samples of the dataset. Our validation set was February 2018. Our unseen test set was February 2017. We chose the same month so time-sensitive differences in comments or topics were avoided. February 2017 consisted of 1.115 articles, with M = 186 comments (SD = 298) per article. February 2018 had 885 articles, with M = 263 (SD = 466) comments per article.

3 Implementing the Metric

3.1 Exploring which metric to implement

Early in the hackathon, we found two of the five metrics in Vrijenhoek et al. (2021) require user data, such as previous watch or read history. The three metrics suitable to our research needs, and our data without such documentation, were "activation", "representation", and "alternative voices". However, the latter two presented too much of a challenge for the short time of a three-week, part-time hackathon.

"Representation" requires the identification of different viewpoints and perspectives in text. NLP has several manners of doing so: tasks such as claim detection, argument mining, and stance detection. For an overview of such NLP tasks and approaches useful for viewpoint diversity in news recommendation, see Reuver et al. (2021). These approaches take time to be done correctly, and we felt the short time available to us in this hackathon did not allow us to properly identify viewpoints in the comments.

"Alternative Voices" requires the identification of whether mentioned people are a member of a minority group. This metric is difficult to implement for several reasons. Conceptually, for comments it may be relevant to know whether the *commenter* has a marginalized background (rather than any mentioned named entities). However, we did not have such information in our dataset. Additionally, who is marginalized depends likely on context - which makes detection by one model difficult. There are also technical hurdles when considering this metric. It is relatively difficult to identify whether someone mentioned comes from a marginalized background based on only the text. This could possibly be solved with open data such as Wikipedia, but this allows only well-known named entities to be recognized. Furthermore, there is a bias in Wikipedia itself: especially women are less often mentioned. Another method would for instance utilize techniques such as large-scale language models to recognize names or terms related to certain marginalized groups. However, this in itself also has bias, and could lead to racist or otherwise unwelcome associations in the representation, as pointed out in Bender et al. (2021).

The "Activation" metric, in contrast, is related to the polarity in the text. Polarity detection is a common task in NLP, and one with extensive support in terms of tools and methods. For this project, we chose to specifically focus on Vrijenhoek et al. (2021)'s activation metric. The core idea behind this metric is to gauge to what extent certain content might spark action among the readers, and is related to emotion. Past research shows that both negative and positive emotions can affect the processing and effects of textual content (Brady et al., 2017; Ridout and Searles, 2011; Soroka and McAdams, 2015). As such, emotional content can produce various effects that may or may not contribute to healthy democracies. Indeed, activation is not universally appreciated in democratic theory. In the models of democracy, activation has different desired values, as outlined in Helberger (2019). For example, from a deliberative democratic perspective, it could be argued that neutral and impartial content facilitates reasoned reflection and deliberation. However, from a more critical democratic perspective one could also argue that emotional content is more valuable as it may generate additional interest and engagement.

3.2 Implementation

We implemented activation in the following manner, based on (Vrijenhoek et al., 2021)'s description of how it should be used. Each article has a certain set of comment recommendations, and also a set of all potential comments. For each comment, we calculate the "compound" polarity value. For both sets we take the mean of the absolute polarity value of each article, which we use as an approximation for Activation. We then remove the mean polarity from all possible articles from the mean of the recommendation set. This results in an output with a range [-1, 1]. According to Vrijenhoek et al. (2021), a negative value indicates the recommender shows less activating content than available in the pool of data, while a positive value means the recommendation system generally selects more activating content than generally in the data.

The use of "polarity" is related to that of "sentiment". We follow Vrijenhoek et al. (2021) and use the VADER dictionary-based approach (Hutto and Gilbert, 2014), since the "compound" value of polarity used in the operationalization of the activation metric seems to be based on this method. However, we are aware this is not the only approach of polarity analysis of text, and in fact may not have the most concept and empirical validity from the social science perspective (van Atteveldt et al., 2021), nor is considered the state of the art for sentiment analysis on user generated text in the computer science field (Zimbra et al., 2018). We discuss this in more detail in the Discussion section. As of now, we use no lemmatization or normalization on the text data. We will also discuss implications of this in the Discussion section. Our code for implementing the metrics, preprocessing the data, and eventually testing the metrics on the data can be viewed here: https://github.com/myrthereuver/Hackathon_MediaComments/blob/main/Hackathon_comments_script.ipynb

4 Results

Our results are visible in Table 1 and Table 2 below. Visible is that the editorial picks are considerably more negative, and thus are *less* Activated, than the recommendations based on user likes. However, both systems pick comments that are negative, and thus *lower* in activation than in the general pool of data.[1]

Recommendation	NYTimes Picks	Likes
Top 3	-0.083	-0.076
Top 5	-0.059	-0.053
Top 10	-0.041	-0.032
Mean all systems	-0.061	-0.053
all NYTimes Picks vs other comments	-0.039	X

Table 1: Results on the feb 2018 set. The left column shows the editorial picks, while the right column shows the recommendations based on user likes. Activation scores can range from [-1, 1], where a negative value denotes the recommender picks items less activating than in the general pool, while a positive value indicates the items are more activating.

[1]Note that for the Picks, we took the most recent Top N editorially picked comments. The results may differ with a random Top of recommended comments, or another manner of selecting the Top editorial picks.

Recommendation	NYTimes Picks	Likes
Top 3	-0.067	-0.078
Top 5	-0.038	-0.052
Top 10	-0.021	-0.034
Mean all systems	-0.042	-0.055
all NYTimes Picks vs other comments	-0.013	X

Table 2: Results on the feb 2017 set. The left column shows the editorial picks, while the right column shows the recommendations based on user likes. Activation scores can range from [-1, 1], where a negative value denotes the recommender picks items less activating than in the general pool, while a positive value indicates the items are more activating.

5 Discussion

5.1 "Test-driving" theory-driven metrics

We implemented Vrijenhoek et al. (2021)'s activation metric, used to assess the relation of recommendations with democratic theory. We found that even the concrete metric as described in this work requires extensive NLP (pre-)processing choices that could significantly alter the outcome of evaluation. Not only selecting which sentiment tools, but also how to tokenize and lemmatize the texts could alter the polarity scores, as does text normalization for especially spelling mistakes in comments. For instance, whether or not to normalize the word "happines" (presumably meaning "happiness") could significantly alter the polarity score of texts, especially if spelling errors are frequent - as they could be in user-generated texts such as comments.

Additionally, selecting a sentiment tool for polarity scoring is not an easy task. As noted before, recent work in social science (van Atteveldt et al., 2021) has indicated NLP sentiment tools are not as reliable and valid as one would hope, and especially dictionary-based methods do not compare to human labelling. In the computer science field, such methods are also not considered the state of the art (Zimbra et al., 2018), performing well below more complex ensemble models of several machine learning methods.

Also, we found that some of the theory-based metrics are easier to generally apply to several datasets, contexts, and research questions than others. We already pointed out that some metrics require information on individual users, such as reading history, which is often not easily available

as open, shared data. Additionally, we found that implementing "Activation" generally makes sense to the comment recommendation context, while "Protected Voices" is more difficult to conceptually define, and the "Representation" metric requires more complex NLP analysis of viewpoints than available in standard tools or models.

Very important to note is that these theory-driven metrics are by no means "plug and play". Using these metrics does not translate 1:1 into a score that measures the democratic valu of content. In this context, it gives an indication if and to what extent a recommendation set lives up to democratic ideals set by different models, but drawing a meaningful line on whether content becomes valuable for a given model of democracy is difficult. These metrics also do not capture more complex concepts such as *intent* when designing recommender systems.

Moreover, these metrics are based on averages: they do not show possible spread of activation across comments as well as articles. We could assume that some articles, as well as some topics, simply attract more activating comments, while others attract a more nuanced and "deliberative" discussion. Future research may, next to implementing the other metrics, also research whether certain topics or categories of news articles and/or comments have significantly more or less activating comments when using these recommendation approaches.

5.2 Results implications for Democratic Debate in NYTimes Comments

We researched whether different recommendation strategies in the New York Times comments dataset lead to different Activation values for the recommendations as presented in Vrijenhoek et al. (2021), and in turn what this means for the democratic models related to these systems. We found editor selections are on average less activating than the most-liked comments. In 2018 this effect is clear, in the 2017 sample less so - even slightly opposite. This could mean several things from a media theory perspective. Perhaps, journalists implicitly select comments in accordance with deliberative ideals. Another explanation of these results is that more activating content is also more likely to be profane, which, as Muddiman and Stroud (2017) showed, makes their selection less likely. The idea behind the activation metric is that activating content increases engagement, maybe the fact that liked comments are more activating is due to that.

Either way, connecting our results to the idea of democratic recommendation, it appears that user selection favours a more critical notion of democracy whereas editor selection favours a comparably more deliberative notion. At the same time, our results also suggest that on the whole, both recommendation styles result in a selection of comments that is slightly less activating than the overall subset. This suggests that both recommendation strategies favour less activating content, which might indicate implicit support of a deliberative model of democracy, where rational and calm debate is preferred over activating and clashing content.

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 the hackathon organizers for organizing the event, and for excellently supporting all teams working on challenges. All remaining errors are our own.

References

New York Times Statement on Comment Moderation. https://help.nytimes.com/hc/en-us/articles/115014792387-CommentsEach, last accessed on March 1, 2021.

Wouter van Atteveldt, Mariken ACG van der Velden, and Mark Boukes. 2021. The validity of sentiment analysis: Comparing manual annotation, crowd-coding, dictionary approaches, and machine learning algorithms. *Communication Methods and Measures*, pages 1–20.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency; Association for Computing Machinery: New York, NY, USA*.

William J Brady, Julian A Wills, John T Jost, Joshua A Tucker, and Jay J Van Bavel. 2017. Emotion shapes the diffusion of moralized content in social networks. *Proceedings of the National Academy of Sciences*, 114(28):7313–7318.

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

Natali Helberger, Kari Karppinen, and Lucia D'acunto. 2018. Exposure diversity as a design principle for recommender systems. *Information, Communication & Society*, 21(2):191–207.

Natali Helberger and Magdalena Wojcieszak. 2018. Exposure diversity. In Philip Michael Napoli, editor, *Mediated Communication*, volume 7, chapter 28, pages 535–560. Walter de Gruyter GmbH & Co KG.

Clayton Hutto and Eric Gilbert. 2014. Vader: A parsimonious rule-based model for sentiment analysis of social media text. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 8.

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.

Aashita Kesarwani. 2018. New York Times Dataset. https://www.kaggle.com/aashita/nyt-comments, last accessed on March 1, 2021.

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.

Ashley Muddiman and Natalie Jomini Stroud. 2017. News values, cognitive biases, and partisan incivility in comment sections. *Journal of communication*, 67(4):586–609.

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

Senja Pollak, Marko Robnik Šikonja, Matthew Purver, Michele Boggia, Ravi Shekhar, Marko Pranjić, Salla Salmela, Ivar Krustok, Tarmo Paju, Carl-Gustav Linden, Leo Leppänen, Elaine Zosa, Matej Ulčar, Linda Freienthal, Silver Traat, Luis Adrián Cabrera-Diego, Matej Martinc, Nada Lavrač, Blaž Škrlj, Martin Žnidaršič, Andraž Pelicon, Boshko Koloski, Vid Podpečan, Janez Kranjc, Shane Sheehan, Emanuela Boros, Jose Moreno, Antoine Doucet, and Hannu Toivonen. 2021. EMBEDDIA tools, datasets and challenges: Resources and hackathon contributions. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.

Shaina Raza and Chen Ding. 2020. A survey on news recommender system–dealing with timeliness, dynamic user interest and content quality, and effects of recommendation on news readers. *arXiv preprint arXiv:2009.04964*.

Myrthe Reuver, Antske Fokkens, and Suzan Verberne. 2021. No nlp task should be an island: Multidisciplinarity for diversity in news recommender systems. In *Proceedings of the EACL Hackashop on News Media Content Analysis and Automated Report Generation*. Association for Computational Linguistics.

Travis N Ridout and Kathleen Searles. 2011. It's my campaign i'll cry if i want to: How and when campaigns use emotional appeals. *Political Psychology*, 32(3):439–458.

Ian Rowe. 2015. Deliberation 2.0: Comparing the deliberative quality of online news user comments across platforms. *Journal of broadcasting & electronic media*, 59(4):539–555.

Stuart Soroka and Stephen McAdams. 2015. News, politics, and negativity. *Political Communication*, 32(1):1–22.

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*.

David Zimbra, Ahmed Abbasi, Daniel Zeng, and Hsinchun Chen. 2018. The state-of-the-art in twitter sentiment analysis: A review and benchmark evaluation. *ACM Transactions on Management Information Systems (TMIS)*, 9(2):1–29.