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

Myrthe Reuver & Nicolas Mattis
Free University of Amsterdam

Introduction

- News recommender systems = personalization to user interest or attention. Danger: potential filter bubbles, threatening democracy.

- **Models of democracy** specify healthy democratic debate. News recommender system could support such (Helberger, 2019). **New evaluation metrics** with normative meaning by Vrijenhoek et. al. (2021).

- Aim: “test-drive" one or more of these metrics to see whether feasible to implement.
  The question we "test-drove":
- *“How do different manners of recommending user comments on a news article affect the recommendation set’s average activation scores?"*

Method

Dataset: New York Times Comment dataset: 9.450 articles with 2.1+ million comments, from 2017 & 2018.
Comment Recommendation Methods, picking top 3, 5, and 10 comments based on:
- user votes (“recommendations")
- editorial picks (“NYT picks")

Test+validation months: February 2017, February 2018. Activation score of recommendations higher than of non-recommended comments?

Challenging to **implement evaluation metrics for news (comment) recommendation** that are not only **valid and feasible**, but also carry some **normative meaning**.

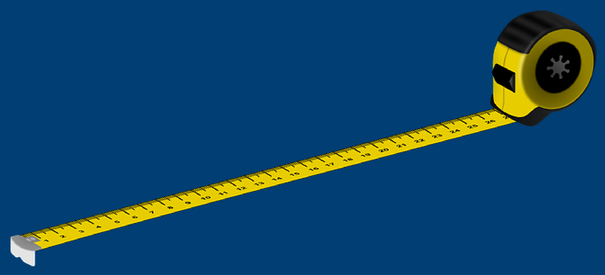

image 1: CLKer Free Vecor Images @ Pixabay, Simplified Pixabay License
image 2: Joseph Mucira @ Pixabay, Simplified Pixabay License

Implementation

Five metrics in Vrijenhoek et. al. (2021):

Calibration

$$Calibration_{(r,q)} = \sum_c r(c|u) log \frac{r(c|u)}{\tilde{q}(c|u)}$$

Fragmentation

$$RBO(Q_1, Q_2, s) = (1-s) \sum_{d=1} s^{d-1} \cdot A_d$$

Affect/Activation

$$Activation(p,q) = (|polarity(q)| - |polarity(p)|)/2$$

Representation

$$Representation_{(p,q)} = \sum_o p(o) log \frac{p(o)}{\tilde{q}(o|u)}$$

Alternative Voices

$$AlternativeVoices = \frac{q^+/p^+}{q^-/p^-}$$

- Calibration, Fragmentation –> less relevant (user data needed, less relevant to comments).
- Alternative Voices & Representation –> challenging to implement technically & conceptually.
- **"Activation"** –> relevant to comment debates and reasonably easy to implement
- But even activation has issues with validity and choices that influence outcome (which method etc.)

Results

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

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

Tables 1 and 2 above show *less activation* for recommended comments than possible comments. Different activation for user & editor recommendations.

Discussion

	Calibration (topic)	Calibration (style)	Fragmentation	Affect	Representation	Alternative Voices
Liberal	High	High	High	-	-	-
Participatory	Low	High	Low	Medium	Reflective	Medium
Deliberative	-	-	Low	Low	Equal	Medium
Critical	-	-	-	High	Inverse	High

Table 1: Overview of the different models and expected value ranges for each metric. Note that for the metrics reflecting distance of a distribution (Calibration and Representation), a "High" target value actually means that the resulting value should be close to zero.

Table from Vrijenhoek et. al. (2021): different models of democracy require different values of the metrics

- **Deliberative** requires **low affect**, **Critical** model requires high affect.
- **Methodological limitations and challenges**. More careful analysis needed to actually conclude something on comments’ activation.
- This study does **not** measure the democratic value of or intent behind recommendations.

References

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

Sanne Vrijenhoek, Mesut Kaya, Nadia Metoui, Judith Möller, Daan Odijk, and Natali Helberger. 2021. Recommenders with a mission: assessing diversity in news recommendations. In the Proceedings of the SIGIR Conference on Human Information Interaction and Retrieval (CHIIR), 173–183. https://doi.org/10.1145/3406522.3446019.