Crowdsourced Measure of News Articles Bias: Assessing Contributors' Reliability

Emmanuel Vincent[1], Maria Mestre[1]

[1] Factmata Ltd.,
114 Whitechapel High St,
London E1 7PT
{emmanuel.vincent, maria.mestre}@factmata.com

Abstract. We tackle the challenge of building a corpus of articles labelled for their political bias by relying on assessments provided by a crowd of contributors. The definition of 'bias' can be ambiguous to participants and both the targets of the ratings (articles) and the source of ratings (contributors) can be biased in some ways. In this paper, we explore techniques to mitigate this subjectivity and learn about the bias of both articles and contributors from the agreements and disagreements among their assessments. We report on the effectiveness of using a set of gold-standard articles to evaluate the reliability of contributors and discuss work in progress to evaluate the bias of contributors from their relative assessments of articles' bias.

1 Introduction

News providers are routinely accused of displaying political bias and this has become a pressing issue as the polarization is the population is increasing, notably in the US (Martin and Yurukoglu 2017). Social media platforms where users increasingly get their news have also been pointed as a source of increased polarization among the public and for favoring the rise of extremely biased information providers, whose inflammatory language is particularly prone to spreading on social media (Marwick and Lewis 2017).

Increased partisanship in news results in enhanced polarization in societies, which undermines democracy and is sometimes a factor in increasing ethnic violence (Minar and Naher 2018). For this reason, several governments have recently attempted to address this growing concern by developing legislation against "fake news". Advertisers are also increasingly interested in measures and detection of extreme bias in online content, as their brand values might be incompatible with funding hyper partisan or divisive content.

In this context, finding scalable ways to assess the bias of articles or information providers is a pressing challenge. This paper presents the first step of ongoing work in Factmata's effort to create a corpus of articles annotated for political bias relying on a crowdsourced approach and design of a system to identify the most reliable contributors.

2 Related Work

Several websites compile lists of news outlets characterized by their bias, one of the most prominent being Media Bias/Fact Check (MBFC)[1]. Like other similar initiatives, MBFC relies on the classification established by a few individuals and classifies news sources at the outlet level, based on analysis of a few articles published by the outlet. Approaches based on natural language processing (NLP) have been used to scale up bias detection, as Lazaridou and Krestel (2016), for example, who analyzed which politicians were being quoted by two major UK outlets, and showed this provided an indication of the outlets' political biases. Patankar and Bose (2016) have approached the challenge of determining bias at the individual news articles level using NPL tools that detect non-neutral sentence formulations based on Wikipedia non-NPOV corpus.

Further automation of bias detection based on Machine Learning approaches will need the creation of large datasets of labeled articles, and in this case crowdsourced solutions offer interesting scalability perspective. Budak, Goel and Rao (2016) performed a large-scale analysis of media bias in which contributors recruited on Mechanical Turk assessed the political bias of more than 10,000 articles from major media outlets covering US politics. However studies like this one did not investigate how to learn about the bias and reliability of contributors from their assessments. More insight can be learned in this respect from online rating systems, in relation to which research has been lead on trust and reputation to identify contributors' reliability and identify potentially biased or spam users (read Swamynathan, Almeroth and Zhao 2010 for an overview). The challenge is to develop a system that allows to learn both about the bias of the news articles that are being labeled and the bias and reliability of the contributors who provide the labels.

3 Data

3.1 Crowdsourcing bias assessments

We drew articles from a pilot study, representing a corpus of 1,000 articles on which ads had been displayed for the account of a customer; these thus form a sample of highly visited news articles from mainstream media as well as more partisan blog-like "news" sources. We used the Crowdflower platform[2] to present these articles to participants who were asked to read each article's webpage and answer the question: "Overall, how biased is this article?", providing one answer form the following five-point bias scale:

1. Unbiased
2. Fairly unbiased
3. Somewhat biased
4. Biased

[1] https://mediabiasfactcheck.com/

[2] https://crowdflower.com/

5. Extremely biased

To guide their assessments, we provided contributors with more details regarding how to classify articles in the form of a general definition of biased article as well as examples of articles with their expected classification (see Appendix 1 for details of the instructions). We chose a five-point scale to allow contributors to express their degree of certainty, leaving the central value on the scale (3) for when they are unsure about the article bias while the values 1 and 2 or 4 and 5 represent higher confidence that the article is respectively unbiased or biased to a more (1 and 5) or less (2 and 4) marked extent. Fifty participants contributed to the labeling and five to fifteen contributors assessed each article (see Appendix 2 for an example).

3.2 'Gold' dataset

To assess the reliability of contributors, we also asked two expert annotators (a journalist and a fact-checker) to estimate which bias ratings should be counted as acceptable for a quarter of all the articles in the dataset. For each article in this 'gold' dataset, the values provided by the two experts are merged. Two values are typically found to be acceptable for an article (most often 1 and 2, or 4 and 5), but sometimes three values are deemed acceptable and less often one value only: typically when both experts agree the article is either clearly extremely biased or not biased at all (e.g. because it covers a trivial and non-confrontational topic in the latter case). When experts disagree on the nature of the bias, providing a set of acceptable ratings as strictly greater than three for one and strictly lower than three for the other, the article is not considered in the gold dataset.

4 Analysis of results

4.1 Assessing contributors' reliability

As a first approach to guide us regarding the quality of data we collected, we performed a comparison of contributors' rating with the gold dataset ratings. Building on the "Beta reputation system" framework (Ismail and Josang 2002), we represent users' reliability in the form of a beta probability density function. The beta distribution $f(p|\alpha,\beta)$ can be expressed using the gamma function Γ as:

$$f(p|\alpha,\beta) = \Gamma(\alpha+\beta)/(\Gamma(\alpha).\Gamma(\beta))\,.p^{\alpha}(1-p)^{\beta-1}\,. \quad (1)$$

where p is the probability a contributor will provide an acceptable rating, and α and β are the number of 'correct' (respectively 'incorrect') answers as compared to the gold. To account for the fact that not all incorrect answers are as far from the gold, we further weight the incorrect answers as follows: an incorrect answer is weighted by a factor of 1, 2, 5 or 10 respectively if its shortest distance to an acceptable answer is 1, 2, 3 or 4 respectively. So β is incremented by 10 (resp. 2) for a contributor providing a rating of 1 (resp. 4) while the gold is 5 (resp. 2) for example. We use the expectation value of the beta distribution $R = \alpha/(\alpha + \beta)$ as a simple measure of the reliability of each contributor. See figure 1 for examples of reputation function obtained for (a) a user with few verified reviews, (b) a contributor of low reliability and (c) a user of high reliability.

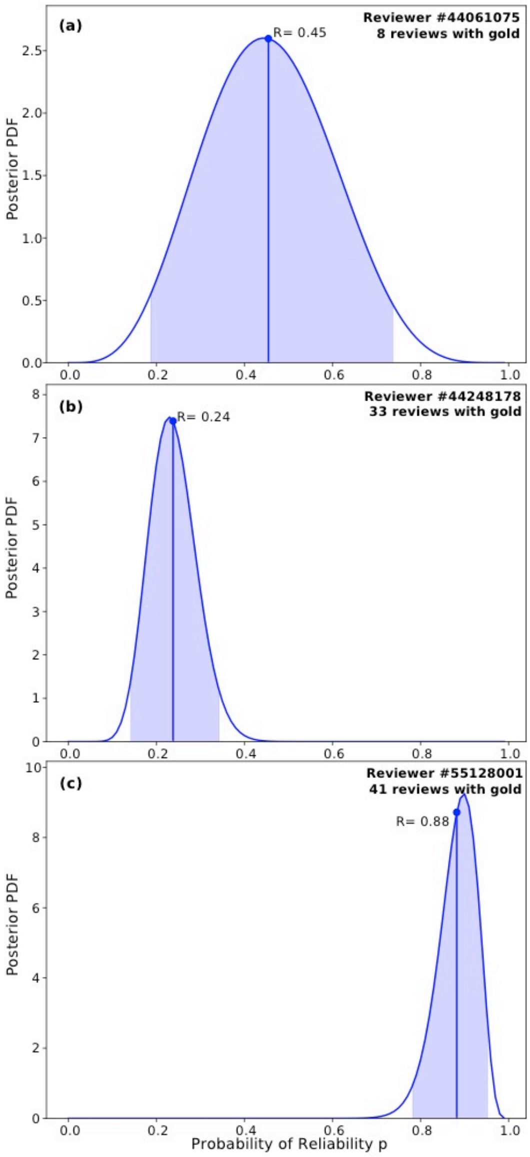

Fig. 1. Examples of reputation function obtained for (a) a user with few verified reviews for whom the uncertainty is still large, (b) a contributor of low reliability and (c) a user of high reliability. Shading shows the 95% probability interval.

Inter-rater reliability. We calculated Krippendorff's alpha to measure the inter-rater agreement (Krippendorff 2011). When we include every worker, we obtain a value for alpha of 0.078, which can be interpreted as a very low agreement. However, inter-rater agreement is much higher when we perform the calculation only for contributors with a high reliability: the value of alpha is 0.40 (resp. 0.76) when we consider contributors with R greater than 0.5 (resp. 0.7).

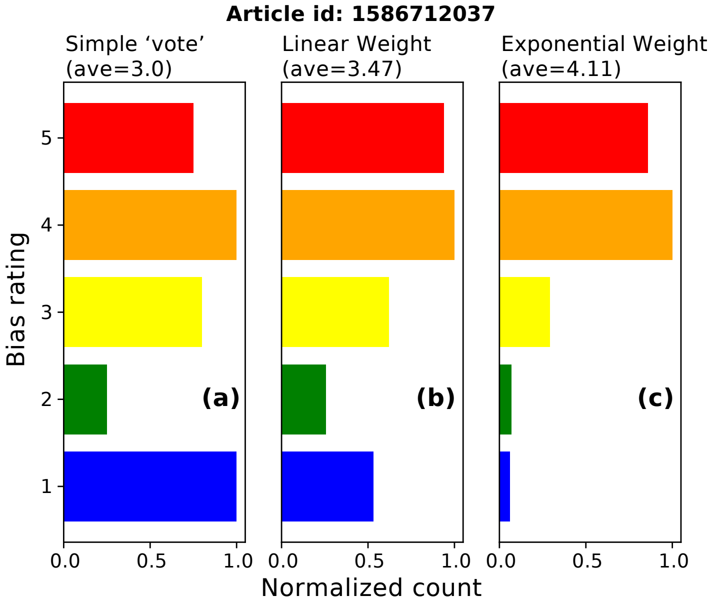

Fig. 2. Histogram displaying the bias ratings collected for an article titled "The invasion of Canada" (a) simple count of the number of users who provided each rating, (b) count weighted by users' reliability and (c) count exponentially weighted by users' reliability as explained in the text.

4.2 Assessing articles' bias based on contributors' ratings

Our goal is to determine the articles' bias and a degree of confidence in that classification based on signals provided by the crowd. A straightforward way to obtain an overall rating is to simply take each assessment as a 'vote' and average these to obtain a single value for the article.

However to try and get closer to an objective assessment of the article's bias, we tested the approach of weighting each rating by the reliability of the contributor. We tested a 'linear' weight for which a user's rating is weighted by its reliability R and a more aggressive 'exponential' weight for which a user's rating is weighted by $10^{4\times(R-1/2)}$ so that an absolutely reliable ($R = 1$) contributor's rating would weight a hundred times more than a contributor of reliability $R = 0.5$.

Figure 2 compares an article's ratings obtained with these different weightings applied. While the article's bias appears disputed from a simple vote perspective (Fig.

2a) with as many contributors judging the article as 'unbiased' (1) and 'biased' (4), it appears quite clearly biased when the exponential weight is applied (Fig. 2c). This reflects the fact that contributors who deem the article biased have higher reliability. In this case, the weighting is improving the clarity of the data collected since this reference-free, anecdote-based article in "Breaking Israel news" on "the invasion of Canada" by "hordes of illegal aliens from Syria, Haiti and anywhere else" can arguably be classified as biased.

5 Experiment: using this annotated dataset to improve the machine learning model

At Factmata we have a model to detect extreme political content online, which we provide as part of our commercial offering. One of the machine learning models was trained on a corpus of 35,236 articles scraped from domains that came from an open-source list of highly biased domains. This training dataset has noisy labels, so we decided to use the new labelled dataset described in this paper to estimate the performance of our algorithm, as well as understand how the performance would change if we added this dataset to the training data.

We first quantized the aggregated weighted scores, so that each article would fall into one of three categories: "very biased"", "unbiased" or "mixed/undecided". We only kept the first two categories, so we ended up with 280 biased instances (i.e. positives) and 260 unbiased instances (i.e. negatives). We split this dataset into training and test, splitting by domains. A domain tends to use similar language across all its pages, so by creating this test set, we are measuring how well a model generalizes to a new unseen domain. We ended up with the dataset described in the table below.

Dataset	**Number of positives**	**Number of negatives**
Original training	8971	26265
Original + manual training	9133	26439
Manual test dataset	86	118

We ran an experiment, where we trained the model on the aggregated training dataset, as well as the original. We then measured the performance improvement on the manually labeled test set. The results are in the table below:

Performance metrics on manual test set	**Precision**	**Recall**	**F1-score**	**ROC-AUC**
Original training	0.74	0.78	0.76	0.52

Original + manual training	0.71	0.88	0.78	0.65

As we can see, the largest improvement was seen in the recall of the new model, likely because the manually labeled dataset has captured types of political bias that do not occur in the open-source dataset. Even though we only increased the training data by less than 1%, the ROC-AUC improved by 25% and the recall by 13%. This is a promising result showing that a small addition of manually labeled data can make a significant improvement in the predictive power of a model trained on noisy labels.

6 Conclusion, and future work

In this paper, we have presented work in progress to create a corpus of news articles labeled for political bias and development of a method to identify reliable contributors. As a first step, we compute a reliability score for each contributor by comparing their assessment to a set of experts-created acceptable assessments on a subset of the articles. Using a probabilistic framework allows us to estimate the confidence we can have in users' reliability scores. Weighting users' contributions by their reliability score increases the clarity of the data and allows us to identify the articles that have been confidently classified by the consensus of high reliability users to train our machine learning algorithms. This notably allows us to note that high reliability contributors disagree on the bias rating for about a third of the articles, which we use to train our machine learning model to recognize uncategorizable articles in addition to biased and unbiased.

This research is very preliminary. An important next step will be to learn about potential contributors' bias from the pattern of their article ratings: for instance a contributor might be systematically providing more "left-leaning" or "right-leaning" ratings than others, which could be taken into account as an additional way to generate objective classifications. This would turn a low quality input into useful data. Another avenue of research will be to mitigate possible bias in the gold dataset. This can be achieved by broadening the set of experts providing acceptable classification and/or by also calculating a reliability score for experts, who would start with a high prior reliability but have their reliability decrease if their ratings diverge from a classification by other users when a consensus emerges.

References

1. Budak, C., Goel, S. and Rao, J.M., 2016. Fair and balanced? quantifying media bias through crowdsourced content analysis. *Public Opinion Quarterly*, *80*(S1), pp.250-271.
2. Ismail, R. and Josang, A., 2002. The beta reputation system. *BLED 2002 proceedings*, p.41.
3. Krippendorff, K., 2011. Computing Krippendorff's alpha-reliability.
4. Lazaridou, K. and Krestel, R., 2016. Identifying Political Bias in News Articles. *Bulletin of the IEEE TCDL, 12*.

5. Martin, G.J. and Yurukoglu, A., 2017. Bias in cable news: Persuasion and polarization. *American Economic Review*, 107(9), pp.2565-99.
6. Marwick, A. and Lewis, R., 2017. Media manipulation and disinformation online. New York: Data & Society Research Institute.
7. Minar, M.R. and Naher, J., 2018. Violence originated from Facebook: A case study in Bangladesh. arXiv preprint arXiv:1804.11241.
8. Patankar, A.A. and Bose, J., 2016, June. Bias Based Navigation for News Articles and Media. In *International Conference on Applications of Natural Language to Information Systems* (pp. 465-470). Springer, Cham.
9. Swamynathan, G., Almeroth, K.C. and Zhao, B.Y., 2010. The design of a reliable reputation system. *Electronic Commerce Research*, *10*(3-4), pp.239-270.

Appendix

1. Article bias assessment instructions provided to contributors

Definition

Biased articles provide an unbalanced point of view in describing events; they are either strongly opposed to or strongly in favor of a person, a party, a country… Very often the bias is about politics (e.g. the article is strongly biased in favor of Republicans or Democrats), but it can be about other entities (e.g. anti-science bias, pro-Brexit bias, bias against a country, a religion…).

A biased article supports a particular position, political view, person or organization with overly suggestive support or opposition with disregard for accuracy, often omitting valid information that would run counter to its narrative.

Often, extremely biased articles attempt to inflame emotion using loaded language and offensive words to target and belittle the people, institutions, or political affiliations it dislikes.

Rules and Tips

Rate the article on the “bias scale” following these instructions:

- Provide a **rating of 1** if the article is **not biased at all**; the article might discuss cooking, movies, lifestyle… or talk about politics in a neutral and factual way.
- Provide a **rating of 2** if the article is **fairly unbiased**; the article might talk about contentious topics, like politics, but remains fairly neutral.
- Provide a **rating of 3** if the article is **somewhat biased** or if it is impossible to determine its bias, or the article is ambivalent (i.e. biased both for and against the same entity).
- Provide a **rating of 4** if the article is **clearly biased**; it overtly favors or denigrates a side, typically an opinion piece with little fairness.
- Provide a **rating of 5** if the article is **extremely biased / hyper partisan**; it overtly favors a side in emphatic terms and/or belittles the other ‘side’, with disregard for accuracy, and attempts to incite an action or emotion in the reader.

Please **do not include your own personal political opinion** on the subject of the article or the website itself. If you agree with the bias of the article, you still should tag is as biased. Try and remove any sense of your personal political beliefs, and critically examine the language and the way the article has been written.

Please do not pay attention to other information on the webpage (page layout, other articles, advertising etc.). **Only the content of the article is relevant** here: text, hyperlinks in it, photos and videos within the text of the article. Also, do not look at the title of the website, its name, or how it looks - just examine the article in front of you and its text.

Do not answer randomly, we will reject submissions if there is evidence that a worker is providing spam responses. Do not skip the rating, providing an overall bias is required.

Examples

- Example of sentences from an hyper-partisan article with many mentions about Donald Trump, clear opposition towards him and loaded language in bold (such an article should be rated as 5):

"*This is how a trickle-down of* ***vileness*** *acquires a fire hose. But the big story doesn't stop with Trump's globe-wide gift to the* ***worst devils*** *of human nature. The big story is that Trump, or his trusted Ministers of Internet Intake, inhabits a* ***bottom-barrel*** *world in which Fox News and Infowars and Gateway Pundit and—sure—Britain First loom large. They're picking this stuff up, combining through it, repurposing it all the time*"

- Example of another hyper-partisan article, with loaded anti-Clinton language in bold, and a call to action at the end for others to join and support the ideology:

"It's a neat little magic trick. It is also **incredibly unethical** and most likely illegal… but then again, that never stopped **the Clinton machine** before. Please press "Share on Facebook" if you think these **dirty tricks** need to be exposed!"

- Example of a biased article (should be rated as 4 on the 1-5 scale). Here, there is less loaded language, but clearly the article is one sided towards Trump:

"*President Trump's stock market rally is historical! No President has seen more all time highs (63) in their first year in office than President Trump. President Trump set the record earlier this year for the most all time closing stock market highs during his first year in office. Currently the Dow has set 80 closing highs since last year's election and 63 since President Trump's inauguration. (As a comparison, President Obama had no stock market highs his entire first term.)*"

- Example of an article talking about a trivial topic. Even though the article speaks positively about money orders and Rite Aid, this shouldn't be marked as biased (should be rated as 1):

"*For people who want to pay bills, purchase goods, or simply want to send guaranteed funds without the risk associated with exchanging cash, money orders are a trusted method of payment. Rite Aid money orders are convenient because of the low fees, numerous locations, and long hours.*"

2. Sample annotation data

pageurl	worker_id	article_bias
url1	44278209	3.0
url1	43718845	4.0
url1	38202325	4.0
url1	37881503	4.0
url1	44164300	4.0
url1	55128002	4.0
url1	55128001	3.0
url1	55128003	2.0
url2	31613324	3.0
url2	44128742	2.0
url2	39793872	5.0
url2	38202325	5.0
url2	44303394	5.0
url2	37881503	4.0
url2	55128002	4.0
url2	55128003	4.0
url2	55128004	5.0
url3	31613324	4.0
url3	44128742	5.0
url3	16718271	1.0
url3	43951421	1.0
url3	44303394	3.0
url3	38202325	4.0
url3	37881503	2.0
url3	55128002	1.0
url3	55128001	1.0
url3	55128003	1.0
url3	55128004	1.0

- url1: http://www.stlamerican.com/news/local_news/privilege-at-the-protest-white-allies-demonstrate-without-incident-outside/article_543f4ba2-9f5f-11e7-95d0-c3a75bed0e90.html
- url2: http://www.theamericanconservative.com/buchanan/trump-embraces-the-culture-war/
- url3: http://www.thegatewaypundit.com/2017/10/breaking-active-shooter-reported-usc-campus-lockdown-videos/