STATISTICAL METHODS FOR AUDITING THE QUALITY OF MANUAL CONTENT REVIEWS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large technology firms face the problem of moderating content on their online platforms for compliance with laws and policies. To accomplish this at the scale of billions of pieces of content per day, a combination of human and machine review are necessary to label content. Subjective judgement and bias are of concern to both human annotated content as well as to auditors who may be employed to evaluate the quality of such annotations in conformance with law and/or policy. To address this concern, this paper presents a novel application of statistical analysis methods to identify human error and these sources of audit risk.

1 BACKGROUND AND PROBLEM

Content moderation has been defined as 'governance mechanisms that structure participation in a community to facilitate cooperation and prevent abuse,' (Gorwa et al., 2020). Online platforms may use a combination of algorithmic and manual review methods to handle the volume of content they host (Binns et al., 2017). Audits of such methods may be required to ensure that they conform to regulatory, policy or other requirements (MacCarthy, 2020). During these audits, auditors may opine on the accuracy of such moderation decisions. The problem of inter-rater reliability during audit can result in an increased source of risk compared to consistency and neutrality requirements because human review is naturally subjective and it is difficult to maintain or measure consistency and equality among independent reviewers (Keller et al., 2020; Neal, 2022). There is also the risk of legal ramifications if reviews are not systematically equal (Ángel Díaz & Hecht-Felella, 2021). Additionally, independent external contractors are hired to assess the ground truth of content labels without bias from internal business processes. We want to quantify the internal reviewer's review quality and consistency compared to the ground truth. This paper evaluates quantitative methods to measure and to minimize audit risks as a result of human reviews in the form of statistical analysis.

2 METHODOLOGY: PROSPECTIVE ANALYSIS

Consider a scenario where there is a need to determine if certain products are being marketed towards the correct age demographics as some may require additional scrutiny during internal review. Review rubrics are developed based on policies, and products may then be routed to human reviewers to decide whether the product is suitable for a particular audience based on a majority vote. Our measure of reviewer consistency is the agreement rate, the agreement rate is the percentage of products for which there is complete reviewer consensus, assuming there are multiple reviewers reviewing the same content with multiple criteria questions.

By calculating the agreement rate, we can determine which questions reviewers have difficulty reaching a consensus on and may represent a source of bias. One of the approaches we can use when there are multiple rubric review questions is Fleiss Kappa (McHugh, 2012), which is a statistical measure for assessing the reliability of agreement between a fixed number of raters and analyzing the distribution of the reviewed result and the correlation of each rubric review question to the human review labeling result. This is applicable at both question level and sampled group level. For example, based on the Table 1, the increased level of Fleiss Kappa represents an increased level of agreement between reviewers on each question for the same sampled group, thus we are able to measure which questions are more prone to have inconsistent results between reviewers.

Table 1: Fleiss Kappa result for sampled review group

Sample Group	Rubric Question	Fleiss Kappa	Fleiss Kappa by Group
1	question 1	0.86	0.74
1	question 2	0.63	
2	question 1	0.77	0.65
2	question 2	0.58	

We wish to determine if there is a systematic relationship between the levels of each criterion or rubric question and the final reviewer classification. This can be done through a chi-square test. Based on our risk tolerance and industry acceptance standard it is suggested to test at the 5% alpha level, which if met would signify that there is 95% confidence to conclude that there is a relationship between a question's answer and the reviewer classification.

We also wish to determine if there is a systematic difference between review teams and the ground truth result for different products. In scenarios where there are only two review teams to compare across, this can be done through hypothesis testing with a t-test. In scenarios where there are more than two comparison groups, an ANOVA test is more appropriate to avoid multiple pairwise comparisons. This analysis allows us to determine if the difference between sampling error rates are systematically different between groups.

3 Methodology: Retrospective Analysis

Consider another scenario where reviewers sequentially instead of simultaneously conduct content review. This would result in one reviewer being able to see previous reviewers' review results. There is existing research to suggest human error exists in single blind vs double blind settings (A Tomkins, 2017) even in simultaneous review cases. We want to evaluate whether sequential review scenarios and their blinding effects would also affect review quality.

We can design an experiment among two groups of reviewers where one group can see the previous reviewers' result and the other one cannot by running an A/B test. The purpose of the test is to see if there is any significant difference between these two situations or if there are any interaction effects when considering the demographic/ historical review content, etc. of reviewers, which could introduce human error.

However, many audit scenarios might not allow for enough resources to conduct A/B testing as mentioned above. Consider a scenario where one either lacks the time to conduct A/B testing or a change in review structure has already been implemented. A method called Difference-in-Difference can use historical data to approximate the effects of review changes (Anders Fredriksson, 2019).

Consider two groups: pre and post review change. One of the groups has been subject to the review change, while the other has not. One can assume that these two groups would have behaved roughly identically were it not for the review change, and that any systematic differences can be attributed to the review change itself. By observing the difference in performance, we can quantify how big of a difference the review change had in these groups' behavior. Refer to Figure 1 in appendix, where the solid lines represent actual, historical trends, and the dotted line refers to a hypothetical performance, the counterfactual, from the treatment group where no changes to review structure had occurred. The difference between the actual vs hypothetical performance in the treatment group represents the treatment effect.

4 Conclusion

The statistical analyses mentioned in the above scenarios help auditors to identify high risk groups, question design choices, and review methods that are more likely to result in human review error. This paves the way for remediation suggestions and targeted training designed to reduce human error and sources of experimental risk (Gino & Coffman, 2021). Through these methods, we are able to use statistical analysis and experimental design to quantitatively pinpoint sources of human error in audit procedures.

URM STATEMENT

The authors acknowledge that at least one key author of this work meets the URM criteria of ICLR 2023 Tiny Papers Track.

REFERENCES

WD Heavlin A Tomkins, M Zhang. Reviewer bias in single- versus double-blind peer review. *Proceedings of the National Academy of Sciences of the United States of America*, 114, 2017.

Gustavo Magalhães de Oliveira Anders Fredriksson. Impact evaluation using difference-in-differences. *RAUSP Management Journal*, 54(4):519–532, 2019.

Reuben Binns, Michael Veale, Max Van Kleek, and Nigel Shadbolt. Like trainer, like bot? inheritance of bias in algorithmic content moderation. In *Social Informatics: 9th International Conference, SocInfo 2017, Oxford, UK, September 13-15, 2017, Proceedings, Part II 9*, pp. 405–415. Springer, 2017.

Francesca Gino and Katherine Coffman. Unconscious bias training that works. *Harvard Business Review*, 2021.

Robert Gorwa, Reuben Binns, and Christian Katzenbach. Algorithmic content moderation: Technical and political challenges in the automation of platform governance. *Big Data & Society*, 7(1): 2053951719897945, 2020.

Daphne Keller, Paddy Leerssen, et al. Facts and where to find them: Empirical research on internet platforms and content moderation. *Social media and democracy: The state of the field and prospects for reform*, 220:224, 2020.

Mark MacCarthy. Transparency requirements for digital social media platforms: Recommendations for policy makers and industry. *Transatlantic Working Group*, 2020.

Mary L. McHugh. Interrater reliability: the kappa statistic. *Biochemia Medica*, 22(3):276–282, 2012.

et al Neal, T. M. S. A general model of cognitive bias in human judgment and systematic review specific to forensic mental health. *Law and Human Behavior*, 46(2):99–120, 2022.

By Ángel Díaz and Laura Hecht-Felella. Double standards in social media content moderation. *Brennan Center for Justice*, 2021.

A APPENDIX

1) Fleiss Kappa: Statistical measure for assessing the reliability of agreement between a fixed number of raters. Kappa is defined as:

$$\kappa = \frac{\bar{p}-\bar{p_e}}{1-\bar{p_e}}$$

Where $\bar{p}$ is the average reviewers' agreement and $\bar{p_e}$ is the sum of squares of the proportion of assignment to the index category.

2) Chi Square Equation: It is a statistical hypothesis testing that helps to identify if two categorical variables have significant relationship or not.

$$\tilde{\chi}^2 = \frac{1}{d}\sum_{k=1}^{n}\frac{(O_k - E_k)^2}{E_k}$$

3) T-test: A statistical test to determine if there is a significant differences between two groups.

Hypotheses:

H_o: Null hypothesis.
H_a: Alternative hypothesis.

Test Statistic:
Case 1: σ^2 is <u>known</u>
$z = \frac{\overline{x}-\mu_o}{\frac{\sigma}{\sqrt{n}}} \sim N(0,1)$
Case 2: σ^2 is <u>unknown</u>
$t = \frac{\overline{x}-\mu_o}{\frac{s}{\sqrt{n}}} \sim t_{n-1}$

4) ANOVA testing: A statistical test to determine if there are significant differences between two or more groups.

Hypotheses:

$H_o : \mu_1 = \mu_2 = ... = \mu_k$
$H_a : \mu_i \neq \mu_j$ for $i \neq j$

Test Statistic:

$F_{obs} = \frac{\text{MSTr}}{\text{MSE}} \sim F_{v_1,v_2}$

$v1 = df(\text{SSTr}) = k - 1$
$v2 = df(\text{SSE}) = n - k$

Reject H_o if $F_{obs} \geq F_{\alpha,v_1,v_2}$

5) Figure 1 Difference in Difference approach

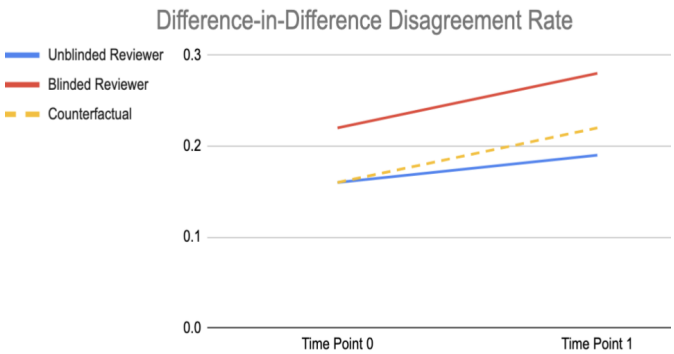

Figure 1: Difference in Difference approach

6) Further Analysis for Prospective Analysis:

It would be also useful to extrapolate sampling error rate to population result by calculating the confidence interval of the error rate in each population group using binomial distribution. This helps the audit to quantify the impact of the risk of human reviewer bias and identify high risk groups of products.

Lastly, in order to help identify to what extent what factors would affect reviewer bias, we can use both a regression model as well as classification model to find which metrics such as product metadata and rubric criterion have the most influence on the classification result based on the beta coefficients. In general, the classification model should give similar metrics as the regression model regarding the cause of bias.