A Mixed-Method Design Approach for Empirically Based Selection of Unbiased Data Annotators

Gautam Thakur **Janna Caspersen** **Drahomira Herrmannova**
Bryan Eaton **Jordan Burdette**

Oak Ridge National Laboratory
1 Bethel Valley Road
Oak Ridge, TN 37830
{thakurg,caspersenjr,herrmannovad,eatonbm,burdetteja}@ornl.gov

Abstract

Implicit bias embedded in the annotated data is by far the greatest impediment in the effectual use of supervised machine learning models in tasks involving race, ethics, and geopolitical polarization. For societal good and demonstrable positive impact on wider society, it is paramount to carefully select data annotators and rigorously validate the annotation process. Current approaches to selecting annotators are not sufficiently grounded in scientific principles and are limited at the policy-guidance level, thereby rendering them unusable for machine learning practitioners. This work proposes a new approach based on the mixed-methods design that is functional, adaptable, and simpler to implement in selecting unbiased annotators for any machine learning problem. By demonstrating it on a real-world geopolitical problem, we also identified and ranked key inane profile characteristics towards an empirically-based selection of unbiased data annotators.

1 Introduction

Human annotation is crucial for many supervised natural language processing problems. Because judgments of meaning can be subjective and vary depending on age, knowledge, intuition, etc. A number of previous works have studied the quality and reliability of manually generated annotations (Snow et al., 2008; Bhardwaj et al., 2010; Aker et al., 2012; Peldszus and Stede, 2013). What these works have in common is that they typically compare annotators from the perspective of agreement, i.e., the degree to which different annotators produce the same labels.

In this work, we study whether and to what extent inherent annotator beliefs and biases affect their answers in labeling tasks. We design a labeling task focused on inflammatory language surrounding Brexit and as part of this task we presented to participants several surveys developed to measure their attitudes towards issues such as race and religion, specifically the Modern Racism Scale (MRS) (McConahay, 1986) and the Moral Foundations Questionnaire (MFQ) (Davies et al., 2014), and surveys capturing their knowledge, demographics, and other relevant information. This labeling task was presented to hired Amazon Mechanical Turk workers (over 100 turkers were hired from 26 countries) in addition to a Subject Matter Expert (SME) who provided ground truth labels. We find that specific results on the Modern Racism Scale and the Moral Foundations Questionnaire are correlated with specific levels of agreement with the subject matter expert, while the same is not the case for knowledge of the topic. Furthermore, we show that the most accurate annotators share similar beliefs as measured by the MRS and MFQ questionnaires. We believe our findings can inform the selection of annotators for such difficult annotation tasks as inflammatory and hateful language detection. To the best of our knowledge, the proposed work is the first to study the relation between annotator belief and biases and the accuracy of their labels.

To that end, this work focuses on a much-needed approach to the systematic selection of unbiased annotators towards improving the outcomes and realism of machine learning decisions. This research investigates and attempts to answer these questions: (i) To what extent do domain expertise and/or psychological bias impact the quality of qualitative data annotations to develop labeled data for machine learning classification? (ii) Is there a significant difference between how individuals label training data? (iii) What are the commonalities between participants whose annotations best matched the ground-truthed dataset? iv) To what extent do those variations in data annotation impact the resulting automated machine learning classification?

Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021, pages 1930–1938

(v) How do we develop a profile that quantifies the implicit deviations within various machine learning models?

While benchmarks exist to validate the performance of ML in such scenarios, albeit the degree to which data annotators affect ML outcomes for such nuanced language patterns is unclear, nor certain is the presence of a systematic framework to discover implicit biases among the annotators that might have been responsible. For instance, in (Sap et al., 2019) authors investigated the presence of racial bias in automatic hate speech detection models, racial bias in an algorithm used to manage the health of populations was discovered in (Obermeyer et al., 2019; Caliskan et al., 2017). Besides, there exists several documented evidences of how machine learning-driven automation has resulted in catastrophic dangers to human security (Brown et al., 2006; West et al., 2019; Buolamwini and Gebru, 2018; Hamidi et al., 2018; Noble, 2018), infrastructure resiliency (Osoba and Welser, 2017; Vladeck, 2014), education (Pedro et al., 2019), and economy (Anderson, 2019; Furman and Seamans, 2019). Several approaches are proposed to counter them, including the development of guidelines and policies surrounding the ethical use of machine learning (Wiens et al., 2019), platforms and toolboxes for diagnosing biases (Brundage et al., 2020), recommended practices, and frameworks for evaluating the fairness and explainability (Hardt et al., 2016; Beutel et al., 2019; Pleiss et al., 2017; Gunning, 2017). In (Sheng et al., 2008) authors suggested when labeling is not perfect, selective acquisition of multiple labels is a strategy that data miners should consider. The purpose of this research is to model the extent to which professional background and mental biases affect analytic results from ML algorithms that are specifically applied to complex social media analysis.

We propose *Mixed-Method Design (MMD)* as a new approach towards selecting unbiased data annotators. Broadly, MMD is a method that focuses on collecting, analyzing, and mixing both quantitative and qualitative data in a single study or series of studies. Its central premise is that the use of quantitative and qualitative approaches, in combination, provides a better understanding of research problems than either approach alone (Creswell and Clark, 2017; Shorten and Smith, 2017). In this work, the qualitative work involves conducting surveys and questionnaires on modern racism, moral foundation, demographics among 100 participants who will also label the data. Using Subject Matter Expert (SME) as a baseline, we built a suite of supervised machine learning models from the labeled data and compared the performance of all participants against the SME. Finally, using statistical analysis, we identified key profile characteristics of data labelers that played an important role in how they labeled the data. Our main contributions are:

- Develop a generalized framework for selecting unbiased data annotators for problems surrounding inflammatory language, hate speech, and others, and requiring labeled data.
- Develop an open-source web-based platform[1] and deploy it to perform mixed-method design.
- Design and develop privacy-preserving and problem-agnostic qualitative surveys and questionnaire towards discovering implicit biases among potential data labelers.
- Identify a subset of key profile characteristics that plays a critical role identifying reliable annotators.

2 Mixed Method Design

The Mixed-method design workflow is shown in Figure 1 to selecting data labelers for complex tasks that involve risks of bias and ethics violation. The rest of the section discusses each step in more detail.

2.1 Labeling Task (Study)

This study focuses on Brexit 2016 referendum as a narrative to investigate implicit biases among the data labelers. Since Britain split from the European Union and changed its relationship to the bloc on trade, security and migration was both welcomed and denigrated and polarized the UK/EU. In the past, several opinion polls and surveys revealed a remarkable divide between generations and demographics within and outside of UK and EU countries.

For our study, we selected a set of 10,000 English tweets related to Brexit between January 2019 and October 2019 which was narrowed down to 2,000 tweets for ground truth annotation and to 250 tweets for annotation by Amazon Mechanical

[1] https://thirdeye.ornl.gov

Figure 1: Mixed-Method Design Workflow

Turk workers.The labeling task primarily focused on identifying whether a given tweet contains inflammatory language, i.e., language that is intended to arouse anger, create and encourage disorder, provoke violent feelings, or excite strong feelings for or against something or someone. Additionally, the participants were asked to answer the following questions about each tweet: 1) Does this text attempt to advocate for violence, hatred, discrimination, or a specific policy? 2) Does this text express a problem with a specific characteristic of an 'other'? 3) Does this text contain a propaganda device? Before starting the labeling assignment all participants were asked to take a short training that provided definitions and explanations for all four label categories.

2.2 Ground Truth Label Generation

First, a Brexit Subject Matter Expert (SME) was hired to generate ground truth labels for the collected tweets. In addition to Brexit knowledge, the SME has previous experiences examining online communication. The study also assessed SME's geo-political and demographic background. The hired SME spent a significant amount of time in Britain and was well versed with the geo-political polarization in the region. However, the SME was not a British citizen allowing an unbiased perspective. In addition to the SME, each tweet was labeled by a social scientist with expertise in inflammatory language and online propaganda. Obtaining labels from two experts was done to allow us to assess the general difficulty of the labeling task.

2.3 Participant Selection

Next, Amazon Mechanical Turk was used to hire over 100 data annotators from selected countries around the world. Figure 2 shows the distribution of participants. On average each participant took three weeks to complete the study and data labeling process. By analyzing time to label each tweet and the generated labels we detected several participants who were providing random labels or who always provided the same answer. These were removed from the study.

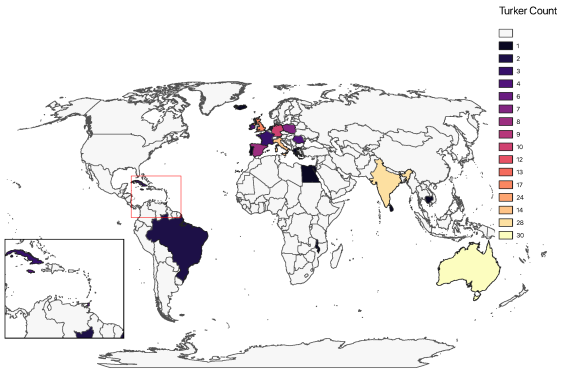

Figure 2: Data annotators were hired from all over the world from majority of the English speaking countries.

2.4 Survey

Both SME and the participants have to complete the following surveys before beginning the data labeling process. The surveys are designed to capture implicit biases.

2.4.1 Symbolic Racism 2000 Scale

Racist viewpoints are a key demographic of inflammatory hate speech. The SR2K is a widely used tool and is considered a reliable method for measuring an individual's racism (Sears and Henry, 2005). The questions within this scale measure contemporary attitudes (Henry and Sears, 2002). The SR2K was incorporated in this research because the participants are labeling data associated with racial ideologies and understanding how they feel about those ideologies is key to explaining their classification choices.

2.4.2 Moral Foundations Questionnaire

The Moral Foundations Questionnaire (MFQ) is considered to be a reliable measure of moral interests (Graham et al., 2011). The MFQ was developed based on the Moral Foundation Theory in 2008 and we used the 2011 version in this research. The MFQ was selected for use in this research because it allows us to assess participants' political leanings and their general moral preferences.

2.4.3 Knowledge Survey

The participants were given a 10 question knowledge survey that was designed to test participants' previous awareness of history associated with inflammatory speech and the Civil Rights Movement. This information is key to understanding participants' prior awareness of historic issues. The insights gained from this section will be used to compare participants' ability to identify inflammatory speech.

2.4.4 Demographics Survey

The demographic survey explored general information about the participants. The demographics collected included age, ethnicity, gender, education, marital status, employment, income, nationality, and political affiliation.

2.4.5 Social Media Background Survey

To better understand the participants' prior experience with social media the final section of the pre-survey had ten questions to address those experiences. The first four question of this section were meant to test their awareness of the most popular social media platforms. The following four questions were meant to generate a deeper understanding of the participants' personal interactions with social media platforms. The final two questions were open ended and focused on participants' profession/career, so that an understanding of their domain expertise may be gained without directly asking, as responses to a direct question on expertise may result in unreliable and exaggerated or understated responses.

2.5 Platform Development for Annotation

A interactive web-based application[2] is developed to help with the data labeling process of tweets related to Brexit. Besides, a training video with set of instructions are designed to guide users to label the data properly.

The remainder of the paper discusses the qualitative and quantitative analysis, as well as identification and ranking of characteristics indicative of a reliable data annotator.

3 Data Analysis

In this section we present the analysis of the collected data. We measure the agreement of each annotator with the SME and study the relation between the agreement and the annotator characteristics measured by the surveys. Table 1 shows the distribution of ground truth labels provided by the SME.

Table 1: Statistics of our ground truth dataset.

	True	False
Inflammatory	128	122
Problematization	49	201
Advocation	66	184
Propaganda	150	100

3.1 Agreement with Ground Truth

To measure the agreement of each annotator with the SME, we use Matthews Correlation Coefficient (MCC)MCC ranges from -1 (perfect disagreement) to 1 (perfect agreement) and can be calculated from confusion matrix using the following equation:

$$\mathrm{MCC} = \frac{\mathrm{TP} \times \mathrm{TN} - \mathrm{FP} \times \mathrm{FN}}{\sqrt{\mathrm{R}_+ \times \mathrm{R}_- \times \mathrm{P}_+ \times \mathrm{P}_-}}, \quad (1)$$

where R+ represents the sum of true positives and false negatives, R- represents the sum of false positives and true negatives, P+ represents the sum of true positives and false positives, and P- represents the sum of false negatives and true negatives. Figure 3 shows the range of MCC values for each label category. In the figure, each bar represents one annotator and the y-axis represents the MCC score. The orange bars represent labels generated by the social scientist with expertise in online inflammatory language.

In all four plots the y-axis range is -0.2 to 0.9 to allow easier comparison of results between the different categories. An interesting observation is that some label categories seem to have a higher disagreement with the ground truth. For example, this is the case for advocates violence category in the bottom left corner which tends to have lower MCC values than the other three categories. A possible explanation is that different people perceive this specific narrative category differently, while the other categories tend to be perceived more similarly.

3.2 Analysis of Survey Results

We used box plots to compare each individual survey with GT agreement. Specifically, annotators

[2] https://thirdeye.ornl.gov

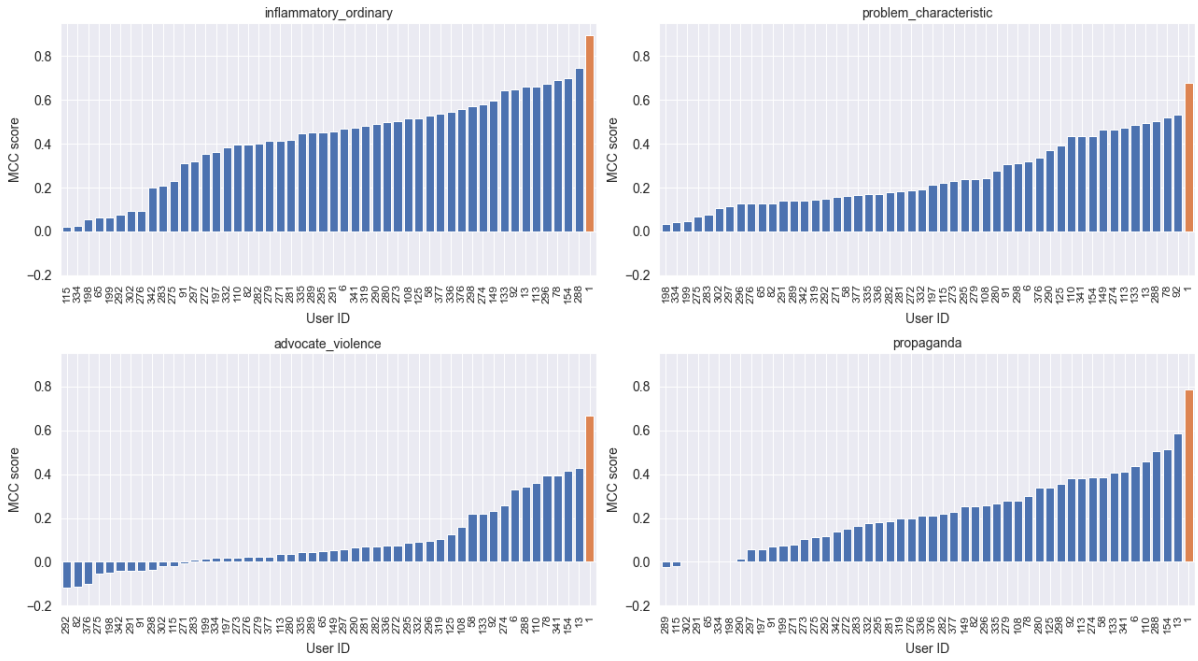

Figure 3: Annotator agreement with the ground truth (SME labels) for each of the four label categories. Each blue bar represents one annotator, while the orange bars represent annotations generated by a social scientist with expertise in online inflammatory language.

were divided into two groups according to their survey scores: group with a score lower than the median score and individuals with higher or equal to the median score. This was done for each survey separately. We plotted statistics of agreement with the GT for each of these two groups using boxplots. We also compare the groups using two-sample t-test with significance threshold of 5%. The results are snown in Figures 5, 7, and 9.

Figure 4 shows SR2K results for all annotators, and Figure 5 shows differences in agreement with the subject matter expert (SME) for individuals with low and high SR2K score. The figures are color-coded to match, i.e., the light blue bars in Figure 4 indicate which annotators belong to the light blue colored bars in Figure 5. Specifically, annotators were divided into two groups according to their SR2K score: group with a score lower than the median score (19) and individuals with higher or equal to the median score. There are 24 annotators in the former group and 25 annotators in the latter group. The figure shows there are significant differences between groups with high and low SR2K score in terms of agreement with the SME (GT), particularly when identifying tweets containing a problem characteristic and tweets advocating violence. Individuals with a higher SR2K tend to agree with the SME, on average, much less than individuals with a lower score when identifying such tweets. We compared the groups using two-sample t-test and in all cases but the first (inflammatory/ordinary) the differences in terms of MCC between the two groups are statistically significant (p-value was lower than 5%).

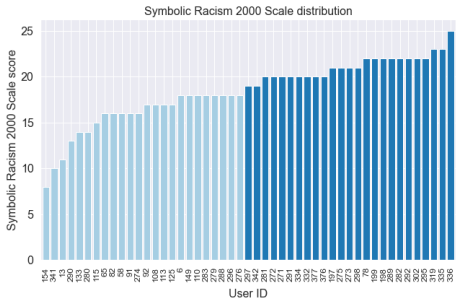

Figure 4: SR2K results for all annotators. Each bar in the figure represents one annotator. The bars are color-coded to indicate which annotators scored lower than median SR2K score, and which score higher or equal to median score.

Figure 6 shows MFQ results for all annotators. Similarly as in the case of the SR2K score, we compared the MFQ score (also called "progressivism" score) with the annotators' MCC score (Figure 7). The results show that individuals with higher progressivism score tend to, on average, agree with the SME more than individuals with lower progres-

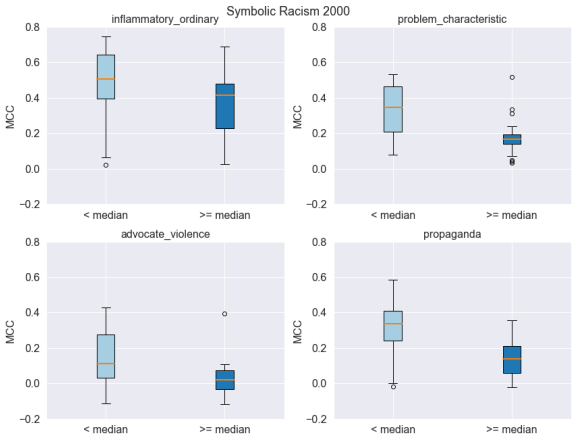

Figure 5: Comparison of the Symbolic Racism 2000 Scale results and agreement with the GT. The lighter colored bar represents the group with the lower than median SR2K score and the darker blue represents the group with higher or equal to median score.

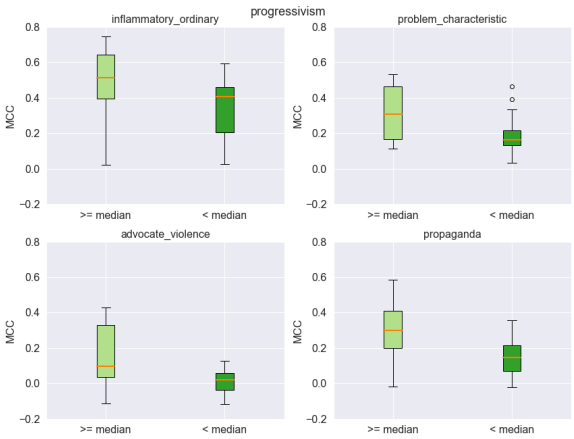

Figure 7: Comparison of the Moral Foundations Questionnaire results and agreement with the GT. The lighter colored bar represents the group with the higher or equal to median MFQ (progressivism) score and the darker bar represents the group with lower than median score.

sivism score. We compared the two groups using independent two sample t-test. The differences in MCC between the two groups statistically significant for all four narrative categories (p-value lower than 5% in all cases).

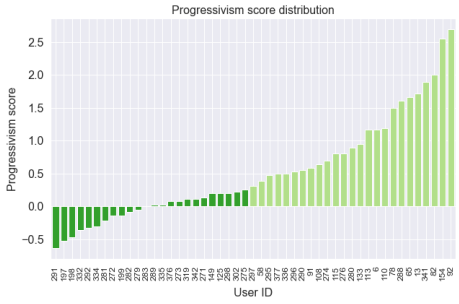

Figure 6: MFQ results for all annotators. Each bar in the figure represents one annotator. The bars are color-coded to indicate which annotators scored lower than median MFQ score, and which score higher or equal to median score.

Figure 8 shows Knowledge Test results for all annotators. Finally, we compared the knowledge test scores with the annotators' agreement with GT (Figure 9). In this case, we can see that there are no differences between the two groups of annotators, which was also confirmed by t-test (p-values for all four narrative categories range from 77-98%).

4 Discussion

The qualitative surveys captured a number of profile characteristics of both SME and other data annotators. As our final question, we aimed to rank

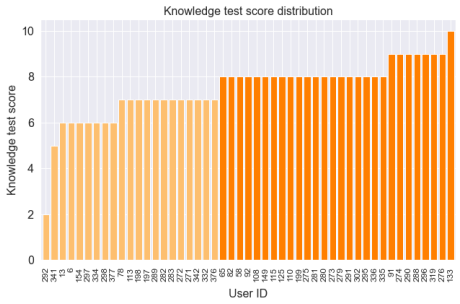

Figure 8: Knowledge Test results for all annotators. Each bar in the figure represents one annotator. The bars are color-coded to indicate which annotators scored lower than median Knowledge Test score, and which score higher or equal to median score.

the collected characteristics according to their importance towards indicating higher or lower agreement with the ground truth (SME). To do this, we calculated the following three statistics to capture the relation between each profile characteristic and agreement with the ground truth: mutual information score statistic, recursive feature elimination, and univariate linear regression test. All profile characteristics were ranked using each of these three statistics. A final rank for each profile characteristic was produced as a sum of all three individual ranks.

In Figure 10, we show the ranked characteristic from the least (Residence) to the most (purity_sanctity – a component of the MFQ) important. Thus, the results show the location of residence is the least important factor in indicating

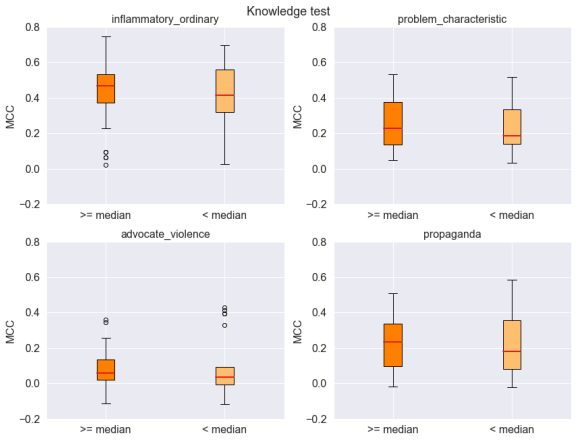

Figure 9: Comparison of the Knowledge Test results and agreement with the GT. The lighter colored bar represents the group with the lower than median Knowledge Test score and the darker bar represents the group with higher or equal to median score.

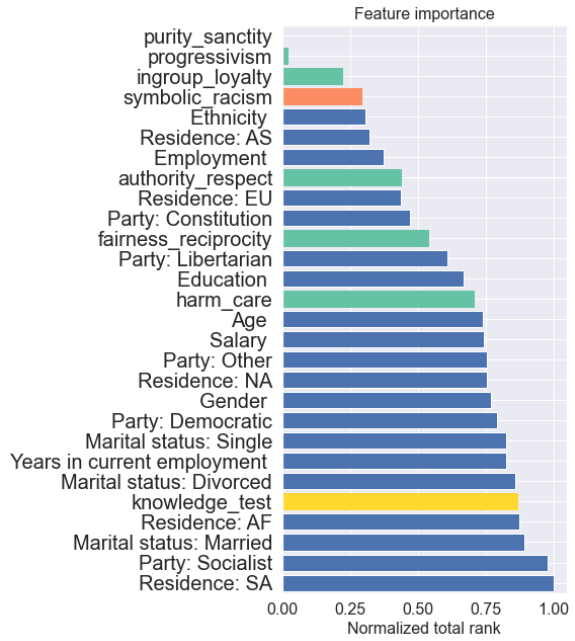

Figure 10: Normalized total rank of various attributes. Blue color bars are demographics while other from survey questionnaire.

possible agreement with the ground truth, while socio-demographic attributes such as purity, progressiveness, etc., are much more important for identifying more reliable annotators.

In Figure 11, we show top five characteristics and their average values that substantially influence the evidentiary based selection of unbiased annotators.

Specifically, this research has strongly indicated that data reviewers morals, prejudices, and prior

Figure 11: Key attributes and their measured average values.

knowledge of the narrative in question significantly impact the quality of labeled data and consequently, the performance of ML models. ML projects that rely on labeled text data to understand narratives must qualitatively assess their data reviewers worldviews if they are to make definitive statements about their results.

For the automated detection of complex narratives, it is important that models built should be free from any implicit biases of any kind. We hope this work contributes to the broad research presently taking place across the field of machine learning to analyze and understand massive amounts of electronic communications (i.e. social media posts, news, blogs, etc.) in complex scenarios involving issues related to ethics, race, gender, and biases surrounding them.

5 Conclusion

The discipline of natural language processing and machine learning has tremendously improved in the last decade. However, it still suffers from biases surrounding complex narratives related to education, health, climate, gender, race, and ethics; especially, it unfairly penalizes certain segments of the population, e.g. women and minorities. Societal, cultural, and demographic phenomena play a pivotal role in how the population conceptualizes policy decisions and complex events. Thus, it is critical for societal good that narratives should be carefully crafted for maximum impact. It is our collective responsibility that any automation (classification, prediction, etc.) surrounding these narratives must be free of any preconceived notions or predilections of any kind. One way to achieve this is by producing high-quality input labeled data curated by annotators aware of such biases. Inspired by this, we have proposed a new framework based on mixed-method design to improve the odds of selecting annotators, who can curate unbiased and high-quality labeled data. In doing so, we identified and ranked personal and professional traits critical to selecting a diverse pool of data annotators, so the resulting labeled data and the models built using

those data best matched the ground-truth. In the future, we would like to extend our study to cater to multi-lingual narratives and expand beyond existing issues of culture, region, and geopolitical dynamics.

Acknowledgments

This manuscript has been authored by UT-Battelle, LLC under Contract No. DE-AC05-00OR22725 with the U.S. Department of Energy. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes. The Department of Energy will provide public access to these results of federally sponsored research in accordance with the DOE Public Access Plan (http://energy.gov/downloads/doe-public-access-plan).

The authors are grateful to RJ Moquito of National Geospatial Intelligence Agency for their support and guidance. The authors also extend their thanks to Budhendra "Budhu" Bhaduri, Amy Rose, Marie Urban, Supriya Chinthavali for allocating necessary resources to complete the research work.

References

Ahmet Aker, Mahmoud El-Haj, M. Dyaa Albakour, and Udo Kruschwitz. 2012. Assessing Crowdsourcing Quality through Objective Tasks. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 1456–1461, Istanbul, Turkey. European Language Resources Association (ELRA).

Patrick Anderson. 2019. Damages caused by ai errors and omissions: Management complicity, malware, and misuse of data (part i of a special report). *Malware, and Misuse of Data (Part I of a Special Report)(June 21, 2019)*.

Alex Beutel, Jilin Chen, Tulsee Doshi, Hai Qian, Li Wei, Yi Wu, Lukasz Heldt, Zhe Zhao, Lichan Hong, Ed H. Chi, and Cristos Goodrow. 2019. Fairness in recommendation ranking through pairwise comparisons. In *KDD*.

Vikas Bhardwaj, Rebecca Passonneau, Ansaf Salleb-Aouissi, and Nancy Ide. 2010. Anveshan: A Framework for Analysis of Multiple Annotators' Labeling Behavior. In *Proceedings of the Fourth Linguistic Annotation Workshop*, pages 47–55, Uppsala, Sweden. Association for Computational Linguistics.

Gerald Brown, Matthew Carlyle, Javier Salmerón, and Kevin Wood. 2006. Defending critical infrastructure. *Interfaces*, 36(6):530–544.

Miles Brundage, Shahar Avin, Jasmine Wang, Haydn Belfield, Gretchen Krueger, Gillian Hadfield, Heidy Khlaaf, Jingying Yang, Helen Toner, Ruth Fong, et al. 2020. Toward trustworthy ai development: mechanisms for supporting verifiable claims. *arXiv preprint arXiv:2004.07213*.

Joy Buolamwini and Timnit Gebru. 2018. Gender shades: Intersectional accuracy disparities in commercial gender classification. In *Conference on fairness, accountability and transparency*, pages 77–91. PMLR.

Aylin Caliskan, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. *Science*, 356(6334):183–186.

John W Creswell and Vicki L Plano Clark. 2017. *Designing and conducting mixed methods research*. Sage publications.

Caitlin L Davies, Chris G Sibley, and James H Liu. 2014. Confirmatory factor analysis of the moral foundations questionnaire. *Social Psychology*.

Jason Furman and Robert Seamans. 2019. Ai and the economy. *Innovation policy and the economy*, 19(1):161–191.

Jesse Graham, Brian A Nosek, Jonathan Haidt, Ravi Iyer, Spassena Koleva, and Peter H Ditto. 2011. Mapping the moral domain. *Journal of personality and social psychology*, 101(2):366–385.

David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web*, 2(2).

Foad Hamidi, Morgan Klaus Scheuerman, and Stacy M. Branham. 2018. Gender recognition or gender reductionism? the social implications of embedded gender recognition systems. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 113, New York, NY, USA. Association for Computing Machinery.

Moritz Hardt, Eric Price, and Nathan Srebro. 2016. Equality of opportunity in supervised learning. *arXiv*.

P. J. Henry and David O. Sears. 2002. The symbolic racism 2000 scale. *Political Psychology*, 23(2):253–283.

John B McConahay. 1986. Modern racism, ambivalence, and the modern racism scale.

Safiya Umoja Noble. 2018. *Algorithms of Oppression: How Search Engines Reinforce Racism*. NYU Press.

Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. 2019. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453.

Osonde A Osoba and William Welser. 2017. *The risks of artificial intelligence to security and the future of work*. RAND.

Francesc Pedro, Miguel Subosa, Axel Rivas, and Paula Valverde. 2019. Artificial intelligence in education: Challenges and opportunities for sustainable development.

Andreas Peldszus and Manfred Stede. 2013. Ranking the annotators: An agreement study on argumentation structure. In *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*, pages 196–204, Sofia, Bulgaria. Association for Computational Linguistics.

Geoff Pleiss, Manish Raghavan, Felix Wu, Jon Kleinberg, and Kilian Q Weinberger. 2017. On fairness and calibration. In *Advances in Neural Information Processing Systems*, volume 30, pages 5680–5689. Curran Associates, Inc.

Maarten Sap, Dallas Card, Saadia Gabriel, Yejin Choi, and Noah A. Smith. 2019. The risk of racial bias in hate speech detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1668–1678, Florence, Italy. Association for Computational Linguistics.

David O. Sears and P.J. Henry. 2005. Over thirty years later: A contemporary look at symbolic racism. volume 37 of *Advances in Experimental Social Psychology*, pages 95 – 150. Academic Press.

Victor S. Sheng, Foster Provost, and Panagiotis G. Ipeirotis. 2008. Get another label? improving data quality and data mining using multiple, noisy labelers. In *Proceedings of the 14th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, KDD '08, page 614622, New York, NY, USA. Association for Computing Machinery.

Allison Shorten and Joanna Smith. 2017. Mixed methods research: Expanding the evidence base. *Evidence-Based Nursing*, 20(3):74–75.

Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng. 2008. Cheap and Fast – But is it Good? Evaluating Non-Expert Annotations for Natural Language Tasks. In *Proceedings of the 2008 Conference on Empirical Methods in Natural Language Processing*, pages 254–263, Honolulu, Hawaii. Association for Computational Linguistics.

David C Vladeck. 2014. Machines without principals: liability rules and artificial intelligence. *Wash. L. Rev.*, 89:117.

Sarah Myers West, Meredith Whittaker, and Kate Crawford. 2019. Discriminating systems. *AI Now*.

Jenna Wiens, Suchi Saria, Mark Sendak, Marzyeh Ghassemi, Vincent X Liu, Finale Doshi-Velez, Kenneth Jung, Katherine Heller, David Kale, Mohammed Saeed, et al. 2019. Do no harm: a roadmap for responsible machine learning for health care. *Nature medicine*, 25(9):1337–1340.