Understanding Group Dynamics to Identify Sources of Bias

Ravie Lakshmanan
Columbia University
New York City, NY
rl2857@columbia.edu

ABSTRACT

As algorithms are increasingly used to make decisions that can have a significant impact on human lives and society at large in a variety of ways, concerns about fairness of algorithmic decision making systems have been raised. While most of the current research has been focussed towards fairness definitions and formulating fairness metrics, this paper aims to identify sources of bias by showing how humans' subconscious tendency to categorise information leads to polarisation, in turn altering judgement and leading to biased outcomes. The paper also attempts to view this problem from an in-group/out-group perspective, and how preconceived biases and affiliations to different groups can colour our perception about other people's beliefs and opinions.

KEYWORDS

Algorithmic Fairness; Algorithmic Discrimination; Fairness in Machine Learning; Fairness Metrics, Social Networks, Group Dynamics, Social Networks

ACM Reference format:

Ravie Lakshmanan. 2018. Understanding Group Dynamics to Identify Sources of Bias. 4 pages.

1 INTRODUCTION

In the last decade, advances in machine learning, coupled with the availability of vast amounts of data, has led to a monumental shift in the way software is used, increasingly enabling them to make autonomous decisions on behalf of humans. Today, decisions made by software determines who is shortlisted for a job [5], who gets a bank loan, what products to buy, what movies to watch, who to make friends with on online social networks, and if a person should be jailed or set free [6]. Unsurprisingly, the enormity and the potential impact of these decisions on human lives have raised concerns about fairness in the field of algorithmic decision making [6].

To mitigate such disparities in algorithmic decision-making scenarios, several techniques have been recently proposed. This includes devising mechanisms for non-discriminatory learning [7], testing software for bias by determining causal relationships between inputs and outputs [2], as well as examining the feasibility of making non-discriminatory decisions [8].

There have also been studies to understand what people perceive as fair by making them answer the question: "Is it fair to use a feature (F) in a given decision making scenario (S)?" [3].

This paper, however, approaches fairness from an alternative point of view, by taking into account humans' innate tendency to categorise information and their group dynamics to identify sources of bias and conflict and minimise them.

2 JUDGING BIAS

2.1 Categorisation as the Basis for Preconceived Biases

Categorisation has been one of the key strategies the human brain uses to process information [9]. For example, when we think of food types, we automatically consider an apple and a banana to be in the same category (fruit) even though they appear different, while we consider an apple and a red ball to be in two different categories despite their similarity in appearance. One of the principal ways categorisation happens is by maximising the importance of certain differences while minimising the relevance of other features. By taking this aspect into account, it has been found that when a certain set of objects are treated as belonging to one group, and a second set of objects as belonging to another, people tend to perceive those within a group as more similar than they are and those in different groups as less similar than they are [10,11].

This shows that merely placing objects in groups (i.e. categories) can distort humans' judgement of those objects. The effect of polarisation by gross simplification and

elimination of subtle nuances in favour of clear-cut distinctions can lead to inappropriate results, at the same time affecting people's view of others based on a variety of factors like race, gender, religion and nationality.

For example, when people tend not to know a person very well, more often than not, our minds subconsciously turn to his or her social category for answers, taking in incomplete data to complete the picture, at times leading to inaccurate results. This form perceptual bias in categorisation has been proven to be the result of implicit stereotyping [12], which detects how strongly a person associates traits with a particular social category.

The implicit association test (IAT) has found that people exhibit a strong or moderate bias toward associating men with science and women with the arts, while an analogous test has found that black people tend to have an unconscious pro-white bias.

Recent research into applying IAT to machine learning models [13] have found them to acquire the same stereotyped biases from textual data, thus stressing the need for addressing source of bias, both from a cultural and technological point of view.

2.2 Effect of In-groups and Out-groups

The immense popularity of social media and online social networks has fundamentally changed the way humans share information and form opinions. While disagreement and polarisation have existed in human societies for millennia, they are now increasingly taking place in the online world, with a huge impact on society. This is further exacerbated by the group behaviour exhibited by people on social media platforms, thanks in part due to the proliferation of online communities.

The question therefore is this: when people tend to categorise themselves as belonging to one group, as being connected to one another by a certain trait or ideology, how does that affect the way one treats those within the group and those on the outside?

The effect of "us" (in-group) versus "them" (out-group) is that people tend to think differently about members of groups they are part of, and those in groups they are not part of, regardless of whether the intention to discriminate between the groups is conscious or not. Studies have, for instance, found that it is not necessary for someone to share any traits with other group members to feel a kinship with an in-group, that people are willing to make financial sacrifices to establish a feeling of belonging to an in-group they aspire to be a part of [14], that they tend to like fellow in-group members more [15] and that common group membership can even outweigh negative personal traits [16].

Another study of in-group and out-group distinction found that people like to think of fellow in-group members as more variegated and complex than those in the out-group [17], while it has also been found to influence the way a person feels about himself or herself, the way he behaves and sometimes, even his performance, in addition to judging others [18].

With the simple act of knowing that one belongs to a group triggering his in-group affinity [19], an answer to this question is therefore central to identifying sources of bias, which can in turn be gleaned by studying patterns of interaction between communities [1], at the same time mitigate them designing meaningful interactions that keep disagreements between individuals low and expose each other to viewpoints that can decrease the overall polarisation [4].

3 PROPOSED APPROACH

The proposed approach is built upon two existing researches into group dynamics and conflict mitigation in online social networks. While the study conducted by Srijan Kumar, William L. Hamilton, Jure Leskovec and Dan Jurafsky focusses on studying inter-community interactions on Reddit across 36,000 communities, examining instances of conflict and devising techniques to mitigate them [1], a parallel study by Cameron Musco, Christopher Musco and Charalampos E. Tsourakakis looks at designing a social network structure such that it minimises disagreement and polarisation simultaneously [4].

The overall approach is explained as follows -

3.1 Gather Data for Analysis

The purpose of this stage is to acquire the necessary data for analysis. For example, in the context of Twitter, it could be tweets that contain a specific hashtag, such as "#northkorea", over a given time-period. In the case of Reddit, it could be the comments posted by users on each post, all of which is publicly accessible [22].

3.2 Identify conflict zones

In the next step, using Srijan Kumar et al. as a basis [1], we identify potential communities/topics of conflict. This is done by adapting their recurrent LSTM model whose input sequences are the concatenation of the author (u_i), source

community (c_s), target community (c_t) and word embeddings of the post [w_0,…,w_L] to predict negative mobilisation by taking the mean of the hidden states from each time step and feeding the resulting vector to a softmax layer:

$$[h_1, \ldots, h_{L+3}] = LSTM([u_i, c_s, c_t, w_1, \ldots, w_L])$$

$$y = \frac{1}{1 + exp\left(\Theta^T \left(\frac{1}{L+3}\right) \sum_{t=1}^{L+3} h_t\right)},$$

where y signifies the probability of the post leading to negative mobilisation, and h_t are the t LSTM hidden states.

3.3 Quantify Conflict

In the third stage, for each of the conflict zone identified, the conflict is quantified using the approach defined by Garimella et al. [21]. This method involves (i) building a conversation graph about a specific topic; (ii) partitioning the conversation graph to extract two partitions to identify potential sides of the controversy; and (iii) computing the value of controversy from characteristics of the graph to determine how controversial the topic is. The intuition behind adopting this approach is that the more controversial a topic is, the more scope there is for bias.

3.4 Identify Biases and Their Sources

Once controversial topics have been identified, the next step is to identify the biases themselves by classifying the tweets and comments using natural language processing techniques such as sentiment analysis [23]. In addition, the influence of the author (estimated via PageRank) on the social network is also taken into account to determine its impact.

3.5 Bias Mitigation and Fairness Score

In this final step, we take specific steps to resolve conflict by (i) making groups more heterogenous by moving past traditional in-groups based on gender, race or religion, (ii) giving two different groups a common goal that require them to cooperate, (iii) recommending potential new connections that minimise disagreement and polarisation [4], and (iv) boosting direct engagement between attackers and defenders, and determining the outcome in each case by quantifying conflict once again (refer to Section 3.3).

If over successive iterations, the controversy score comes down, it becomes telling that the conflict resolution strategy has worked, thus leading to informed opinions that are less biased and polarised.

4 FUTURE STEPS AND CONCLUSIONS

With software discrimination becoming a growing concern, it has become necessary to formulate different notions of fairness and understand which biases are desirable and unacceptable. This not only requires analysing software for any implementation bugs (which are often unintended), but also taking into account the developer's preconceived biases and checking training data for discrimination.

While there currently exists different methods to identify and measure discrimination in software, this work attempts to tackle the problem by identifying where such biases originate in online communities and offering remedial mechanisms to mitigate them.

Specifically, our personal knowledge of a specific member of a category can easily override category bias, but over time repeated contact with category members can act as an antidote to the negative traits society assigns to people in that category. In addition, giving the groups common goals that require them to cooperate can sharply reduce intergroup conflicts [24], and the more the people from different traditional in-groups based on race, class, gender or religion work together, the less they tend to discriminate against one another.

The future work therefore lies in putting this theory to test by implementing the approach explained in Section 3, and devise a baseline that can be used to eliminate prejudiced outcomes in intelligent systems as well as enable software developers who design such algorithms from perpetuating similar cultural stereotypes.

REFERENCES

[1] Srijan Kumar, William L. Hamilton, Jure Leskovec, and Dan Jurafsky. 2018. Community Interaction and Conflict on the Web. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France*. ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3178876.3186141

[2] Sainyam Galhotra, Yuriy Brun, and Alexandra Meliou. 2017. Fairness Testing: Testing Software for Discrimination. In *Proceedings of 2017 11th Joint Meeting of the European Software Engineering Conference and the ACM SIGSOFT Symposium on the Foundations of Software Engineering, Paderborn, Germany, September 4–8, 2017 (ESEC/FSE'17)*, 13 pages. https://doi.org/10.1145/3106237.3106277

[3] Nina Grgić-Hlača, Elissa M. Redmiles, Krishna P. Gummadi, and Adrian Weller. 2018. Human Perceptions of Fairness in Algorithmic Decision Mak- ing: A Case Study of Criminal Risk Prediction. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France*. ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3178876.3186138

[4] Cameron Musco, Christopher Musco, and Charalampos E. Tsourakakis. 2018. Minimizing Polarization and Disagreement in Social Networks. In *WWW 2018: The 2018 Web Conference, April 23–27, 2018, Lyon, France*. ACM, New York, NY, USA, 10 pages. https://doi.org/https://doi.org/10.1145/3178876. 3186103

[5] Ted Greenwald. How AI Is Transforming the Workplace. *The Wall Street Journal*, March 10, 2017. https://www.wsj.com/articles/how-ai-is-transforming-the-workplace-1489371060.

[6] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine Bias. *ProPublica*, May 23, 2016. https://www.propublica.org/article/machine-bias-risk-assessments-in-criminal-sentencing.

[7] Michael Feldman, Sorelle A. Friedler, John Moeller, Carlos Scheidegger, and Suresh Venkatasubramanian. 2015. Certifying and Removing Disparate Impact. In *KDD'15, August 10-13, 2015, Sydney, NSW, Australia*.

[8] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. 2017. Algorithmic Decision Making and the Cost of Fairness. In *Proceedings of KDD '17, August 13-17, 2017, Halifax, NS, Canada*, 10 pages. DOI: 10.1145/3097983.3098095.

[9] David J. Freedman, Maximilian Riesenhuber, Tomaso Poggio, and Earl K. Miller. Categorical Representation of Visual Stimuli in the Primate Prefrontal Cortex. In *Science* 12 Jan 2001: Vol. 291, Issue 5502, pp. 312-316 DOI: 10.1126/science.291.5502.312

[10] Henry Tajfel, and A. L. Wilkes. Classification and Quantitative Judgement. 1963. In British Journal of Psychology, 54, pp. 101-114

[11] Olivier Corneille, Olivier Klein, Sophie Lambert, and Charles M. Judd. 2002. On the Role of Familiarity with Units of Measurement in Categorical Accentuation: Tajfel and Wilkes (1963) Revisited and Replicated. In *Psychological Science*, Vol. 13, No. 4 (Jul., 2002), pp. 380-383

[12] Anthony G. Greenwald, Debbie E. McGhee, and Jordan L. K. Schwartz. 1998. Measuring Individual Differences in Implicit Cognition: The Implicit Association Test. In *Journal of Personality and Social Psychology*, Vol. 74, No. 6, 1464-1480.

[13] Aylin Caliskan1, Joanna J. Bryson, and Arvind Narayanan. 2017. Semantics derived automatically from language corpora contain human-like biases. In *Science*, 14 Apr 2017: Vol. 356, Issue 6334, pp. 183-186 DOI: 10.1126/science.aal4230

[14] Blake E. Ashforth, and Fred Mael. 1989. Social Identity Theory and the Organization. In *Academy of Management Review*, Vol. 14, No. 1, pp. 20-39. https://doi.org/10.5465/amr.1989.4278999

[15] Markus Brauer. 2001. Intergroup Perception in the Social Context: The Effects of Social Status and Group Membership on Perceived Out-group Homogeneity and Ethnocentrism. In *Journal of Experimental Social Psychology*, 37, 15–31. doi:10.1006/jesp.2000.1432

[16] Kenneth L. Dion. 1973. Cohesiveness as a Determinant of Ingroup-Outgroup Bias. In Journal of Personality and Social Psychology, Vol. 28, No. 2, 163-171.

[17] Patricia W. Linville, Gregory W. Fischer, and Peter Salovey. 1989. Perceived distributions of the characteristics of in-group and out-group members: Empirical evidence and a computer simulation. In *Journal of Personality and Social Psychology*, 57(2), 165-188. http://dx.doi.org/10.1037/0022-3514.57.2.165

[18] Margaret Shih, Todd L. Pittinsky, and Nalini Ambady. 1999. Stereotype Susceptibility: Identity Salience and Shifts in Quantitative Performance. In *Psychological Science*, Vol. 10, No. 1 (Jan., 1999), pp. 80-83.

[19] Marilynn B. Brewer, and Madelyn Silver. 1978. Ingroup Bias as a Function of Task Characteristics. In *European Journal of Psychology*, Vol. 8, 393-400. https://doi.org/10.1002/ejsp.2420080312.

[20] Robert Kurzban, John Tooby, and Leda Cosmides. 2001. Can Race be Erased? Coalitional Computation and social Categorization. In *PNAS*, 98 (26) 15387-15392. https://doi.org/10.1073/pnas.251541498

[21] K. Garimella, G. De Francisci Morales, A. Gionis, and M. Mathioudakis. Quantifying Controversy in Social Media. In *Proceedings of the 9th ACM International Conference on Web Search and Data Mining (WSDM)*. ACM, 2016.

[22] Reddit data dump. http://files.pushshift.io/reddit/.

[23] Y. Mejova, A. X. Zhang, N. Diakopoulos, and C. Castillo. Controversy and sentiment in online news. *Symposium on Computation + Journalism*, 2014.

[24] Muzafer Sherif, O. J. Harvey, B. Jack White, William R. Hood, and Carolyn W. Sherif. 1961. Intergroup Conflict and Cooperation: The Robbers Cave Experiment. Retrieved from *Classics in the History of Psychology*.