Case Studies

Bias and Fairness in ML: KDD 2020 Tutorial

CASE STUDY A – Disaster Relief

Description: There has recently been a severe storm and many areas have been badly damaged, and people who have lost homes, jobs, and other property. The National Emergency Relief Agency is exploring the use of social media data to assess the damage and determine where to intervene with what types of resources at what level.

Goal: Accurately assess damage and send appropriate relief resources
Data: twitter posts, facebook posts geocoded with lat-long within disaster area and keywords/hashtags related to the storm
Analysis: intensity and type of damage by neighborhood
Actions: assessment and allocate relief effort (type and amount)

Breakout Session 1 — Sources of Bias:

What are some potential sources of bias in the underlying data?

[The bullet points below are filled out answers from the participants in the KDD 2020 Tutorial]

Sample bias:
Missing people who don't use social media

- Old people don't use social media.
- Lack of access to internet for some places affected / may not be able to tweet/social media
- Data only from FB / twitter users

- Data missing geocoding (posts about the disaster without knowing where from)

How active people of impacted area are on social platforms
People who are posting might be biased to people of a specific ethnicity
Retweets reposts

People more likely to post about damage/negative outcomes than positive

How complete is the dataset that FB and Twitter are able to share? There are some firehose restrictions, and some privacy settings from users, that would limit from seeing all posts.

What is the social and demographic distribution in the area - different groups might have different propensity to report/use platform.

Reliability:

Some posted contents may not be true.

Geo-information can be wrong (with high noise)

How might biases be introduced in the data science pipeline? (Think about ETL, record linkage, feature engineering, labels, modeling, and model selection)

[The bullet points below are filled out answers from the participants in the KDD 2020 Tutorial]

Ground Truth: Could we use past data to evaluate? How about the official report?

How is the damage quantified? Who is quantifying the damage?
Assessing damage due to storm vs baseline or post-storm
Decision to weight likes/comments/retweets and the introduction of folks from non-affected areas doing these activities

How do we treat outliers? Is the modeling favoring means?

How encoded : male or female may be exclusionary

Record linkage - deduplication Possible overlapping users using facebook and twitter

What are the risks to fairness in downstream applications and deployment of the model described?

[The bullet points below are filled out answers from the participants in the KDD 2020 Tutorial]

Who would we over-allocate resources to based on the biases?
“Rich gets richer.”

Who would we underallocate resources to?

By age or types of housing? Not sure on the housing part

Can game system by just complaining more

Do we have a comparison method to check against?
Facebook and Twitter likely optimize for Click-Through-Rate and Engagement - might skew even further.

Maybe places that are more densely populated get prioritized b/c more social media users

Breakout Session 2 — Bias Metrics:

How would you describe a **false positive in this problem** to a policymaker or business owner? What's the potential harm/cost of one?

How would you describe a **false negative** to a policymaker or business owner? What's the potential harm/cost of one?

What confusion matrix metric (e.g., FPR, FNR, TPR, FDR, etc.) would you choose to focus on in terms of equity for this case? Think of the fairness tree here.