UNDP Gender Social Media Monitoring – Pilot (July 2022)

Objective: Track hate speech occurrence and patterns on social media, specifically Twitter, with a focus on hate speech targeting women and girls across three pilot countries.

Data Source: Twitter

Time Frame: August 2021 – Onwards; Data is updated daily.

Country Coverage: Colombia, Philippines, Uganda

Languages: English, Spanish, Filipino

Methodology:

- Tweets from the pilot countries are identified based on geocoding tools
- Tweets are collected using keyword queries based on a list of terms developed by the UNDP BPPS Gender Team across the following categories: education, employment, politics, reproduction and gender-based violence
- Across the identified categories, a hate speech classifier (algorithm) is implemented to classify tweets between hate speech and non-hate speech
- Hate speech terms are identified using hate speech data sets from
    - https://hatespeechdata.com/
    - https://www.conductaenredes.org/
- Topic modelling is used to identify dominant themes across the categories
- A gender classifier (algorithm) is implemented to predict the gender of the Twitter user

Next Steps (August 2022):

Further refine the methodology by applying the following:

- Manual annotation of hate speech tweets to further improve classification
- Manual annotation of tweets to further improve topic sub-classification (further disaggregate topics with high number of tweets, e.g. politics, employment)
- Perform sentiment analysis to quantify sentiments and attitudes according to positive, negative or neutral
- Get feedback from UNDP Country Offices to enhance local insight