Hate speech is in the eye of the beholder Exploring bias on hate perception

PI: Dr Antonela Tommasel - ISISTAN, CONICET-UNICEN, Argentina - antonela.tommasel@isistan.unicen.edu.ar

Collaborating Researchers (in alphabetical order):

- Dr Daniela Godoy - ISISTAN, CONICET-UNICEN, Argentina - daniela.godoy@isistan.unicen.edu.ar
- Ms Aiqi Jiang - Queen Mary University of London, UK - aiqi.jiang@yahoo.com
- Dr Arkaitz Zubiaga - Queen Mary University of London, UK - a.zubiaga@qmul.ac.uk

An important goal for hate speech detection techniques is to ensure that they are not unduly biased towards or against particular norms of offence. Training data is usually obtained by manually annotating a set of texts. Thereby, the reliability of human annotations is essential. Meanwhile, the ability to let big data "speak for itself" has been questioned as its representativeness, spatiotemporal extent and uneven demographic information can make it subjective. We hypothesize that *demographics substantially affect hate speech perception*. In this context, the research question guiding this project is:

How do latent norms and biases caused by demographics derive in biased datasets, which affects the performance of hate speech detection systems?

AREA OF FOCUS

The project will focus on whether demographics influence the perceiver's judgements of the level of offensiveness (or hatefulness) and harm to the target. Particularly, we chose to use gender, age, geographical residence and ethnicity as the demographic variables under analysis They are easily understood and accessible demographic attributes that allow to select large populations that could provide labelling of hate speech instances that we hypothesised that would differ in their definitions.

METHODOLOGY AND PLANNED ACTIVITIES

The study is going to be divided into three phases.

- Phase 1: Exploring bias in existing datasets. This phase aims at studying whether bias can be detected on existing datasets. To this end, datasets with known demographic taggers will be used (such as the one from the Wikipedia Detox Project) for training hate speech detection models. Then, the evaluation of each training model will be carried out against all datasets. Considering the possibility that the different datasets might comprise different types of hate speech, which could differ amongst the diverse datasets, only binary classification will be used. This would allow to determine whether models trained based on taggers of specific demographics tend to misclassify instances tagged by different demographic groups in a greater proportion than those tagged by the same demographic group; or whether there is a certain bias to a class of a particular dataset. This could hint the possibility of negative impacts against the same populations the systems are designed to protect. For example, if the model has a bias to consider speech from a minority group as hate speech, victims might be unfairly penalised, while at the same time, abuse against them will be neglected.

  Even though restricting the language of datasets also restricts the diversity of demographic characteristics, only English datasets will be considered. Additionally, only datasets belonging to the same social media site will be analysed, to reduce the possible bias introduced by the characteristics of the different sites.

- Phase 2: Assessing diverse utterances from diverse demographic backgrounds. To generate a large variety of different textual instances, we will design an online survey to ask participants to assess the hatefulness of a selection of the statements extracted from datasets belonging to different categories (e.g. racist, homophobia, sexism, misogyny, gender, religion), and whether they personally agree with them. This will allow to assess the agreement between the selected participants and the people that originally tagged the selected utterances. Additionally, as collected datasets often rely on a set of predefined set of keywords for creating datasets, which might not be representative of unstructured hate speech utterances in social media, we will also collect posts from diverse media profiles expressing hate in more subtle ways that using hate specific keywords. We will not provide any "undecided" option to encourage participants to take a side. We will select at least 100k utterances. As it has been shown that annotating hate speech with a numeric or binary scale is a challenging task that is latter associated to low inter-agreement between participants, the assessment will be conducted based on a comparative approach in which each participant has to select the most and least hateful statement in a set. Statements are guaranteed to appear with uniform frequency in the different sets, and each set is expected to be evaluated by participants from every demographic group.

  The selection of participants will consider different demographic characteristics, including gender, age, religion, ethnicity and geographic location. We expect to get in the order of 750 participants belonging to each category, who will be asked to tag at least 100 utterances. It is worth noting that as the variables are not mutually exclusive, the effect of the combination of variables could also be studied. To that end, we aim at getting in the order of 100 participants for each

binary combination of variables. Participants will be recruited online by sharing the survey on different communication channels. Additionally, participants will be hired in Amazon Mechanical Turk. Participants will be informed of the goal of the data collection and be assured that all data will be anonymous. Once the dataset is collected, the evaluation of the first phase will be replicated.

- Phase 3: Assessing contextualised hate. Most available datasets consider individual posts, neglecting that in social media, posts are inherently related, and that there is additional information available regarding, for example, reactions to the posts, the user that shared the post and his/her social relations. In this context, considering existing datasets, we will crawl the social context of the individual posts. Particularly, we will retrieve the number of likes (or the equivalent in the social media site under analysis), the number of times if was re-shared, and statistics of the user that published it, in terms of number of friends, number of shared posts, age of the account and frequency of posting. This information can summarise the popularity of the post under analysis. The goal of this phase is to analyse the effect of perceived popularity in the assessment of hatefulness in combination with the demographics of participants. Following the same methodology as before, we will ask selected participants to assess the hatefulness of social posts. Participants will be divided into two groups: one will assess the posts without any context, and the other will have access to the popularity indicators of the post to perform the assessment. The same set of posts will be shown to both groups, and each group will include participants belonging to every demographic group, and combinations of demographic variables. Once the dataset is collected, the evaluation of the first phase will be replicated.

The following Table presents the work schedule for the project, detailing the activities to be carried out organised in bimesters. The publication of results in conferences and/or journals will be made in parallel to the described activities.

Phase 1: Exploring bias in existing datasets

Collecting datasets with available demographic information from taggers.	X					
Cross-evaluating the existing datasets.		X				

Phase 2: Assessing diverse utterances from diverse demographic backgrounds

Selecting the utterances to be tagged.		X				
Obtaining the crowd-sourced annotations on the selected dataset.			X			
Cross-evaluating the collected dataset.				X		

Phase 3: Assessing contextualised hate

Collecting the contextual data of social hate speech posts belonging to existing datasets.				X		
Obtaining the crowd-sourced annotations on the selected dataset.					X	
Cross-evaluating the collected dataset.						X

OUTCOME AND CONTRIBUTIONS

The primary expected outcomes of this project are:

- A detailed study on bias in hate speech detection due to the demographic characteristics of dataset taggers.
- A study of the effect of popularity perception on the assessment of hatefulness according to the demographic characteristics of taggers.

As a result of the study, two datasets will be created and made publicly available on diverse platforms under a CC-by license, always in agreement with the TOS of social media sites. All shared data will be anonymised.

- A new hate speech dataset with comprehensive demographic information of its taggers. The dataset will include utterances with both explicit and implicit examples of hate speech. The available demographic information of taggers will include gender, age, geographic residence and expressions of ethnicity.
- A new hate speech social media dataset including both popularity indicators of posts and comprehensive demographic information of its taggers.

PREVIOUS WORK

This project is founded on previous research that focused on the analysis of social media of both participating groups. Particularly, they have explored the dynamics of social networks in terms of their users, and how they behave and express on social media. Dr Tommasel and Dr Godoy have also started to study the diffusion of unwanted or malicious content by tackling the detection of aggressive content. On the other hand, Dr Zubiaga took a leading role on the collection and annotation of rumour datasets for the EC-funded project PHEME, which were used to study how to assess and mitigate online harm.